2024.02.08 IC DeAI Technical Working Group Call Summary

Short

The group's discussion on Zero-Knowledge Proofs (ZKPs) and optimistic AI/ML on the Internet Computer (IC) highlighted the technology's potential for privacy and scalability in decentralized AI systems. ZKPs, particularly beneficial for privacy-preserving applications, allow for verifying data or computations without revealing underlying information. The IC's architecture is well-suited for ZKPs, offering advantages over platforms like Ethereum by facilitating on-chain storage and verification of proofs. The conversation also explored the challenges of deploying AI and ML models on the IC, such as the need for specialized GPU support and the limitations of WebAssembly for large AI models. Practical applications of ZKPs for AI on the IC were discussed, including privacy-preserving AI models and verifiable computation. The group acknowledged the technical challenges in implementing ZKPs but noted recent advancements that have improved efficiency. The discussion transitioned to vector databases on the IC, focusing on development efforts and the potential for scalable architectures to support AI applications. Enhancements in GPU support and increased instruction limits were identified as crucial for the performance of vector DBs on the IC. Participants expressed interest in sharing developments and collaborating on vector DB implementations.

Long

Part 1: Zero-Knowledge and Optimistic AI/ML

The group discussed the potential of ZKPs in ensuring privacy and security in decentralized AI systems. ZKPs allow for the verification of data or computations without revealing the underlying data, making them valuable for privacy-preserving applications.

Zero-Knowledge Proofs (ZKPs) and the Internet Computer: The discussion emphasizes the advantages of using ZKPs on the IC, highlighting its ability to verify computations without revealing the data or the computation itself. This characteristic makes ZKPs particularly suitable for ensuring privacy and security in decentralized applications. The IC's capacity to store and verify proofs on-chain was noted as a significant advantage over other platforms like Ethereum, which often require additional layers or specific proof systems (e.g., zk-SNARKs with trusted setups) to accommodate ZKPs. The IC's

capacity for handling large data storage and computation aligns well with the demands of ZKP applications.

Optimization for AI and Machine Learning: The conversation covered the challenges and strategies for deploying AI and ML models on the IC, focusing on the limitations posed by current technology and the potential of ZKPs to overcome these. The group discussed the need for specialized GPU support and the limitations of WebAssembly (Wasm) in handling large AI models. Zero-knowledge proofs were proposed as a solution to offload computation off-chain while still ensuring the integrity and correctness of the computations on-chain.

Practical Applications and Use Cases: Participants shared insights into practical applications of ZKPs for AI on the IC, including privacy-preserving AI models (where the model or its parameters are not revealed), verifiable computation (ensuring that an off-chain computation, such as AI model inferences, was performed correctly without revealing its specifics), and scalability (using ZKPs to reduce the on-chain footprint of complex computations). The discussion also touched on the potential of ZKPs to enable new types of applications that require a high degree of privacy and security, e.g. decentralized identity verification (utilizing ZKPs for identity verification processes without compromising user privacy).

Participants discussed the potential for zero-knowledge machine learning (zkML) to protect private data, models, and weights while still enabling verification of computations. This approach is seen as particularly beneficial for applications requiring data privacy, such as biometric authentication, where a model could run locally on a user's device without exposing sensitive data. The discussion also touched on the practicality of ZKPs for real-world applications, noting recent advancements that have significantly reduced overheads and made ZKPs more feasible for a range of use cases.

The conversation delved into various applications of ZKPs, including identity verification without data exposure, proof of asset ownership, and secure, tamper-proof oracles for financial data. The potential for ZKPs to enable private and efficient AI computations on the IC, even in scenarios where traditional methods would be impractical due to privacy concerns or computational limitations, was a key focus.

Participants expressed interest in exploring innovative applications of ZKPs within the IC ecosystem, such as decentralized secret management, AI model verification, and cross-chain integrations. The discussion underscored the potential of ZKPs to address fundamental challenges in decentralized AI, such as ensuring privacy, security, and trust,

while also acknowledging the need for ongoing research and development to fully realize this potential.

Participants also discussed the importance of efficient proving schemes and the challenge of making ZKP verifiers and provers both fast and portable. Recent developments in the field, such as improved proving schemes and the concept of ZKVMs (Zero-Knowledge Virtual Machines), were highlighted as promising avenues for making ZKPs more accessible and practical for a wider range of applications.

Challenges and Limitations: The technical challenges associated with implementing ZKPs, such as the computational overhead and the complexity of developing efficient proof systems, were acknowledged. The group also discussed the current state of research and development in the field, indicating ongoing efforts to make ZKPs more practical and accessible for developers.

Advancements in ZKPs: The discourse transitions into the significant improvements in ZKPs, particularly highlighting Halo 2 and STARKs developed in 2019. These advancements have made proofs quicker to generate, albeit with large proving sizes. The contrast between older and newer ZKP technologies is discussed, with a focus on the efficiency and speed of modern implementations.

Parallelization and Speed Improvements: A major point of discussion is the need for systems that allow fast proof generation and parallelization. Traditional centralized provers limit scalability, but newer schemes offer the possibility of distributing the proving process across multiple machines, drastically reducing computation time.

Specialized Circuits vs. General Purpose: The group discusses the benefits of specialized circuits for specific tasks over general-purpose solutions, particularly in the context of AI models. The conversation touches on the possibility of integrating specialized circuits as opcodes within a more general framework, allowing users to opt-in as needed.

Optimistic Machine Learning: The concept of optimistic machine learning is introduced, focusing on the idea of assuming correctness in computations and only verifying in case of disputes to avoid the overhead of verifying every computation upfront. Instead, the system allows for the possibility of fraud proofs, where computations can be retrospectively verified if there's a dispute or suspicion of incorrectness. This approach can significantly reduce the computational overhead by relying on the social deterrent of fraud proofs and shifting from proving correctness to proving fraud only when

necessary. While less secure than ZKPs, optimistic approaches offer a trade-off between speed and security, potentially making AI/ML applications more scalable.

Challenges and Future Directions: The group acknowledges the challenges in applying optimistic approaches to machine learning and AI, questioning how consumers can verify AI outputs without extensive background knowledge. The potential need for rerunning computations to verify outputs is discussed as a possible limitation of optimistic ML.

The conversation then transitioned into technical aspects of implementing ZKP and optimistic ML on the IC, including challenges and potential solutions for practical applications. The group discussed the importance of fast proving times, the potential for parallelization, and the challenges of implementing optimistic approaches in a secure and efficient manner.

Part 2: Vector Databases on the Internet Computer

The conversation shifts towards the application and development of vector databases for AI on the Internet Computer, highlighting ongoing projects and the desire for more efficient and scalable solutions.

Initial Efforts: Participants are actively developing and planning to port vector databases to the IC, with an emphasis on leveraging the unique capabilities of the platform, such as its distributed nature and the potential for utilizing specialized GPU support for enhanced computation and processing speeds.

Basic to Advanced Features: The development efforts range from creating basic vector DB implementations to more advanced systems that could potentially offer a wide range of features comparable to existing vector databases like Qdrant. The goal is to support a variety of AI and machine learning applications by providing efficient and scalable vector storage and retrieval mechanisms on the IC.

Porting Efforts: There is interest in porting existing open-source vector DBs to the IC, with Qdrant being mentioned as a specific example. The conversation highlighted the potential benefits of porting such databases, including leveraging their established features and community support. The ease of porting Rust-based implementations to the IC's WebAssembly environment was also discussed as a factor in selecting potential candidates for porting.

Scalable Architectures: The need for scalable architectures to support vector DBs on the IC was emphasized, with suggestions for utilizing multiple canisters and exploring distributed storage solutions to handle larger data volumes. This ties into broader discussions about scalable architectures for AI and machine learning applications on the IC.

GPU Support and Instruction Limits: Enhancements in GPU support and increased instruction limits were identified as key areas that could significantly benefit the development and performance of vector DBs on the IC. Such improvements would enable more complex computations and data processing tasks, making the IC a more attractive platform for deploying sophisticated AI applications.

Sharing and Collaboration: Participants expressed interest in sharing their developments and collaborating with others in the community. There's a clear desire to explore various implementations and learn from each other's experiences in tackling the challenges of developing vector DBs on the IC.

Part 3: Future call topics

In the concluding segment of the call, participants discussed forward-looking topics and logistical aspects to enhance future meetings. The conversation centered around the interest in delving deeper into vector databases (DBs) in subsequent discussions, with an emphasis on exploring specific projects related to this technology. There was also a proposal to focus on educational materials and data collection practices in upcoming sessions. The idea was to brainstorm and gather resources that projects could provide, share best practices for data collection, and discuss approaches to standardize and share data related to AI applications running on the Internet Computer (IC).

Additionally, the participants suggested hosting Twitter Spaces to reach a broader audience and potentially attract more participants to Discord meetings. This approach aims to complement the technical discussions with more general and educational content, thereby increasing visibility and engagement within the IC ecosystem.

The call wrapped up with affirmations of the session's value and the importance of staying updated on developments like Optimistic Machine Learning (OPML) and Zero-Knowledge Machine Learning (ZKML). The participants appreciated the insights shared about vector DBs and expressed eagerness for the educational focus in upcoming sessions. The discussion underscored the group's commitment to advancing

AI projects on the IC, fostering a collaborative learning environment, and exploring new ways to showcase their work to a larger audience.