# Cryptocurrencies, Blockchains, and Smart Contracts; Hardware for Deep Learning

**EXPERT-CURATED GUIDES TO THE BEST OF CS RESEARCH**

*Research for Practice combines the resources of the ACM Digital Library, the largest collection of computer science research in the world, with the expertise of the ACM membership. In every RfP column two experts share a short curated selection of papers on a concentrated, practically oriented topic.*

Our fourth installment of Research for Practice covers two of the hottest topics in computer science research and practice: cryptocurrencies and deep learning.

First, Arvind Narayanan and Andrew Miller, co-authors of the increasingly popular open-access Princeton Bitcoin textbook, provide an overview of ongoing research in cryptocurrencies. This is a topic with a long history in the academic literature that has recently come to prominence with the rise of Bitcoin, blockchains, and similar implementations of advanced, decentralized protocols. These developments—and colorful exploits such as the DAO vulnerability in June 2016—have captured the public imagination and the eye of the popular press. In the meantime, academics have been busy, delivering new results in maintaining anonymity, ensuring usability, detecting errors, and reasoning about decentralized markets, all through the lens of these modern cryptocurrency systems. It is a pleasure having two academic experts deliver the latest

updates from the burgeoning body of academic research on this subject.

Second, Song Han provides an overview of hardware trends related to another long-studied academic problem that has recently seen an explosion in popularity: deep learning. Fueled by large amounts of training data and inexpensive parallel and scale-out compute, deep-learning-model architectures have seen a massive resurgence of interest based on their excellent performance on traditionally difficult tasks such as image recognition. These deep networks are compute-intensive to train and evaluate, and many of the best minds in computer systems (e.g., the team that developed MapReduce) and AI are working to improve them. As a result, Song has provided a fantastic overview of recent advances devoted to using hardware and hardware-aware techniques to compress networks, improve their performance, and reduce their often large amounts of energy consumption.

As always, our goal in this column is to allow our readers to become experts in the latest topics in computer science research in a weekend afternoon's worth of reading. To facilitate this process, as always, we have provided open access to the ACM Digital Library for the relevant citations from these selections so you can read the research results in full. Please enjoy! —*Peter Bailis*

# CRYPTOCURRENCIES, BLOCKCHAINS, AND SMART CONTRACTS

BY ARVIND NARAYANAN AND ANDREW MILLER

Research into cryptocurrencies has a decades-long pedigree in academia, but decentralized cryptocurrencies (starting with Bitcoin in 2009) have taken the world by storm. Aside from being a payment mechanism "native to the Internet," the underlying blockchain technology is touted as a way to store and transact everything from property records to certificates for art and jewelry. Much of this innovation happens in the broader hobbyist and entrepreneurial communities (with increasing interest from established industry players); Bitcoin itself came from outside academia. Researchers, however, have embraced cryptocurrencies with gusto and have contributed important insights.

Here we have selected three prominent areas of inquiry from this young field. Our selections of research papers within each area focus on relevance to practitioners and avoid areas such as scalability that are of interest primarily to cryptocurrency designers. Overall, the research not only exposes important limitations and pitfalls of the technology, but also suggests ways to overcome them.

## Anonymity, privacy, and confidentiality

*Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., Savage, S. 2013. A fistful of Bitcoins: characterizing payments among men with no names. In Internet Measurement Conference: 127-140; https://www.*

*usenix.org/system/files/login/articles/03_meiklejohn-online.pdf.*

Bitcoin exists in a state of tension between anonymity (in the sense that real identities are not required to use the system) and traceability (in that all transactions are recorded on the blockchain, which is a public, immutable, and global ledger). In practice, the privacy of vanilla Bitcoin comes from obscurity: users may create as many addresses as they like and shuffle their coins around, even creating a new address for each transaction. But this paper demonstrates that "address clustering" can be very effective, applying a combination of heuristics to link together all the pseudo-identities controlled by an individual or entity.

Anonymity in cryptocurrencies is a matter of not just personal privacy, but also confidentiality for enterprises. Given advanced transaction graph analysis techniques, without precautions, the blockchain could easily reveal cash flow and other financial details.

*Sasson, E. B., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., Virza, M. 2014. Zerocash: decentralized anonymous payments from Bitcoin. IEEE Symposium on Security and Privacy; http://zerocash-project.org/media/pdf/zerocash-extended-20140518.pdf.*

There are many different proposals for improving the privacy of cryptocurrencies. These range from Bitcoin-compatible methods of "mixing" (or "joining") coins with

each other, to designs for entirely new cryptocurrency protocols that build in privacy from the beginning. Perhaps the most radical proposal is Zerocash, an alternative cryptocurrency design that uses cutting-edge cryptography to hide all information from the blockchain except for the *existence* of transactions; each transaction is accompanied by a cryptographic, publicly verifiable proof of its own validity. Roughly, the proof ensures that the amount being spent is no more than the amount available to spend from that address. The paper is long and intricate, and the underlying mathematical assumptions are fairly new by cryptographic standards. But this fact itself is food for thought: to what extent does the security of a cryptocurrency depend on the ability to comprehend its workings?

### Endpoint security

Turning to security, the Achilles' heel of cryptocurrencies has been the security of endpoints, or the devices that store the private keys that control one's coins. The cryptocurrency ecosystem has been plagued by thefts and losses resulting from lost devices, corrupted hard drives, malware, and targeted intrusions. Unlike fiat currencies, cryptocurrency theft is instantaneous, irreversible, and typically anonymous.

*Eskandari, S., Barrera, D., Stobert, E., Clark, J. 2015. A first look at the usability of Bitcoin key management. Workshop on Usable Security; http://users.encs.concordia.ca/~clark/papers/2015_usec.pdf.*

This paper studies six different ways to store and protect one's keys, and evaluates them on ten different criteria encompassing security, usability, and deployability. No solution fares strictly better than the rest. Users may benefit considerably from outsourcing the custody of their keys to hosted wallets, which sets up a tension with Bitcoin's decentralized ethos. Turning to Bitcoin clients and tools, the authors find problems with the metaphors and abstractions that they use. This is a ripe area for research and deployment, and innovation in usable key management will have benefits far beyond the world of cryptocurrencies.

### Smart contracts

One of the hottest areas within cryptocurrencies, so-called smart contracts are agreements between two or more parties that can be automatically enforced without the need for an intermediary. For example, a vending machine can be seen as a smart contract that enforces the rule that an item will be dispensed if and only if suitable coins are deposited. Today's leading smart-contract platform is called Ethereum, whose blockchain stores long-lived programs, called contracts, and their associated state, which includes both data and currency. These programs are immutable just as data on the blockchain is, and users may interact with them with the guarantee that the program will execute exactly as specified. For example, a smart contract may promise a reward to anyone who writes two integers into the blockchain whose product is RSA-2048—a self-enforcing factorization bounty!

*Luu, L., Chu, D-H., Olickel, H., Saxena, P., Hobor, A. 2016. Making smart contracts smarter. In ACM SIGSAC Conference on Computer and Communications Security: 254-269; http://queue.acm.org/rfp/vol14iss6.html.*

Unfortunately, expressive programming languages are hard to reason about. An ambitious smart contract called The DAO suffered a theft of an estimated $50 million thanks to a litany of security problems. (Ultimately, this theft was reversed by a networkwide "hard-fork" upgrade.) The authors study four classes of security vulnerabilities in Ethereum smart contracts, and build a tool to detect them based on a formalization of Ethereum's operational semantics. They find that thousands of contracts on the blockchain are potentially vulnerable to these bugs.

*Clark, J., Bonneau, J., Felten, E. W., Kroll, J. A., Miller, A. and Narayanan, A., 2014. On decentralizing prediction markets and order books. Workshop on the Economics of Information Security, State College, Pennsylvania; http://www.econinfosec.org/archive/weis2014/papers/Clark-WEIS2014.pdf.*

If smart-contract technology can overcome these hiccups, it could enable decentralized commerce—that is, various sorts of markets without intermediaries controlling them. This paper studies how one type of market—namely, a prediction market—could be decentralized. Prediction markets allow market participants to trade shares in future events (such as "Will UK initiate withdrawal from the EU in the next year?") and turn a profit from accurate

predictions. In this context the authors grapple with various solutions to a prominent limitation of smart contracts: they can access only data that is on the blockchain, but most interesting data lives outside it. The paper also studies decentralized order books, another ingredient of decentralized markets.

### Overcoming the pitfalls

Cryptocurrencies implement many important ideas: digital payments with no central authority, immutable global ledgers, and long-running programs that have a form of agency and wield money. These ideas are novel, yet based on sound principles. Entrepreneurs, activists, and researchers have envisioned many powerful applications of this technology, but predictions of a swift revolution have so far proved unfounded. Instead, the community has begun the long, hard work of integrating the technology into Internet infrastructure and existing institutions. As we have seen, there are pitfalls for the unwary in using and applying cryptocurrencies: privacy, security, and interfacing with the real world. These will be fertile areas of research and development in the years to come.

## HARDWARE FOR DEEP LEARNING

BY SONG HAN

DNNs (deep neural networks) have evolved to a state-of-the-art technique for machine-learning tasks ranging from computer vision to speech recognition to natural language processing. Deep-learning algorithms, however, are both

computationally and memory intensive, making them power-hungry to deploy on embedded systems. Running deep-learning algorithms in realtime at subwatt power consumption would be ideal in embedded devices, but general-purpose hardware is not providing satisfying energy efficiency to deploy such a DNN. The three papers presented here suggest ways to solve this problem with specialized hardware.

### The compressed model

*Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., Dally, W. J. 2016. EIE: efficient inference engine on compressed deep neural network. International Symposium on Computer Architecture; https://arxiv.org/pdf/1602.01528v2.pdf.*

This work is a combination of algorithm optimization and hardware specialization. EIE (efficient inference engine) starts with a deep-learning-model compression algorithm that first prunes neural networks by 9-13 times without hurting accuracy, which leads to both computation saving and memory saving; next, using pruning plus weight sharing and Huffman coding, EIE further compresses the network 35-49 times, again without hurting accuracy. On top of the compression algorithm, EIE is a hardware accelerator that works directly on the compressed model and solves the problem of irregular computation patterns (sparsity and indirection) brought about by the compression algorithm. EIE efficiently parallelizes the compressed model onto multiple processing elements and proposes an efficient way of partitioning and load balancing both the storage and the

computation. This achieves a speedup of 189/13 times and an energy efficiency improvement of 24,000/3,400 times over a modern CPU/GPU.

### Optimized dataflow

*Chen, Y.-H., Emer, J., Sze, V. 2016. Eyeriss: a spatial architecture for energy-efficient dataflow for convolutional neural networks. International Symposium on Computer Architecture; https://www.researchgate.net/publication/301891800_Eyeriss_A_Spatial_Architecture_for_Energy-Efficient_Dataflow_for_Convolutional_Neural_Networks.*

Deep-learning algorithms are memory intensive, and accessing memory consumes energy more than two orders of magnitude more than ALU (arithmetic logic unit) operations. Thus, it's critical to develop dataflow that can reduce memory reference. Eyeriss presents a novel dataflow called RS (row-stationary) that minimizes data-movement energy consumption on a spatial architecture. This is realized by exploiting local data reuse of filter weights and feature map pixels (i.e., activations) in the high-dimensional convolutions, and by minimizing data movement of partial sum accumulations. Unlike dataflows used in existing designs, which reduce only certain types of data movement, the proposed RS dataflow can adapt to different CNN (convolutional neural network) shape configurations and reduce all types of data movement through maximum use of PE (processing engine) local storage, direct inter-PE communication, and spatial parallelism.

### Small-footprint accelerator

*Chen, T., Wang, J., Du, Z., Wu, C., Sun, N., Chen, Y., Temam, O. 2014. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. International Conference on Architectural Support for Programming Languages and Operating Systems; http://pages.saclay.inria.fr/olivier.temam/files/eval/CDSWWCT14.pdf.*

Recent state-of-the-art CNNs and DNNs are characterized by their large sizes. With layers of thousands of neurons and millions of synapses, they place a special emphasis on interactions with memory. DianNao is an accelerator for large-scale CNNs and DNNs, with a special emphasis on the impact of memory on accelerator design, performance, and energy. It takes advantage of dedicated storage, which is key for achieving good performance and power. By carefully exploiting the locality properties of neural network models, and by introducing storage structures custom designed to take advantage of these properties, DianNao shows that it is possible to design a machine-learning accelerator capable of high performance in a very small footprint. It is possible to achieve a speedup of 117.87 times and an energy reduction of 21.08 times over a 128-bit 2-GHz SIMD (single instruction, multiple data) core with a normal cache hierarchy.

### Looking forward

Specialized hardware will be a key solution to make deep-learning algorithms faster and more energy efficient. Reducing memory footprint is the most critical issue. The papers presented here demonstrate three ways

to solve this problem: (1) optimize both algorithm and hardware and accelerate the compressed model; (2) use an optimized dataflow to schedule the data movements; (3) design dedicated memory buffers for the weights, input activations, and output activations. We can look forward to seeing more artificial intelligence applications benefit from such hardware optimizations, putting AI everywhere, in every device in our lives.

Peter Bailis *is an assistant professor of computer science at Stanford University. His research in the Future Data Systems group (futuredata.stanford.edu) focuses on the design and implementation of next-generation data-intensive systems. He received a Ph.D. from UC Berkeley in 2015 and an A.B. from Harvard in 2011, both in computer science.*

Arvind Narayanan *is an assistant professor of computer science at Princeton. He leads a research team investigating the security, anonymity, and stability of cryptocurrencies as well as novel applications of blockchains. He co-created a Massive Open Online Course as well as a textbook on Bitcoin and cryptocurrency technologies. Narayanan also leads the Princeton Web Transparency and Accountability Project, to uncover how companies collect and use our personal information. His doctoral research showed the fundamental limits of de-identification, for which he received the Privacy Enhancing Technologies Award.*

Andrew Miller *is an Assistant Professor in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, and received his Ph.D. from the University of*

*Maryland. He has studied cryptocurrencies since 2011, and has authored scholarly papers on a wide range of original research, including new proof-of-work puzzle constructions, programming languages for block chain data structures, and peer-to-peer network measurement and simulation techniques. He is an Associate Director of the Initiative for Cryptocurrencies and Contracts (IC3) at Cornell and an advisor to the Zcash project.*

Song Han *is a fifth-year Ph.D. student at Stanford University. His research focuses on energy-efficient deep learning, at the intersection between machine learning and computer architecture. He proposed deep compression that can compress state-of-the art CNNs (convolutional neural networks) by 10-49 times. He designed EIE (efficient inference engine), a hardware architecture that does inference directly on the compressed sparse model. His work received the Best Paper Award from the International Conference on Learning Representations in 2016.*