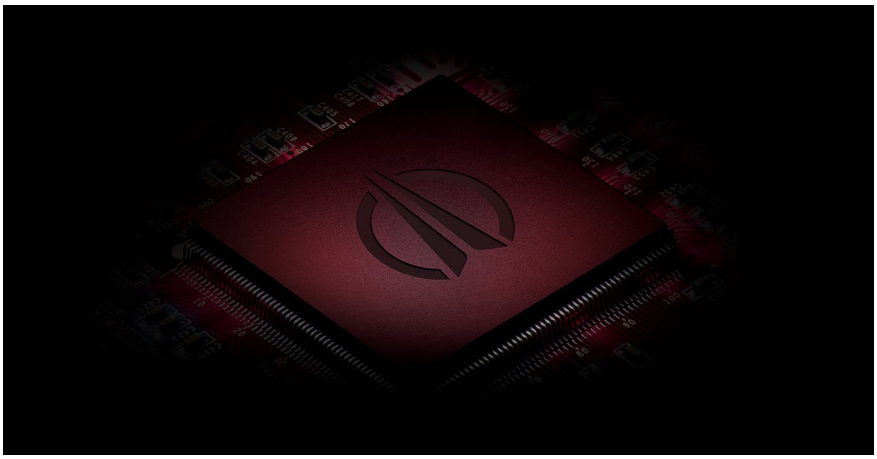
Choosing ASICs for Sia

David Vorick Follow
Jun 22, 2017 · 19 min read



We recently announced that we would be manufacturing and selling ASICs for Sia, an announcement that received a lot more heat and controversy than I was expecting. People primarily seemed to be concerned with mining centralization, and what happens if a small number of groups end up controlling all the hashrate—groups that may well include the developers, as the developers are the ones making the chips.

We feel strongly however that ASICs are the right move, because as the rest of this post explains, GPU based mining is a false panacea that ultimately leaves a cryptocurrency far more vulnerable to attack. The relative decentralization gained from having the active hashrate controlled by a larger number of parties is far outweighed by the fact that the currency ends up being far more vulnerable to 51% attacks by centralized parties. Not only this, but you decouple the mining from the incentives—in Bitcoin, miners lose big when the price drops. In the GPU mined altcoin world, the price dropping means that miners just hop to a more profitable coin.

This post dives deep into the mechanics of Proof of Work, and outlines why GPU based mining and ASIC resistance really is a bad choice, especially for smaller cryptocurrencies.

A Primer on Proof of Work

We've seen a lot of discouraging things play out in Bitcoin. At points, mining pools have controlled more than 51% of the hashrate, and today something like 80% of all Bitcoin mining chips are produced by a single company, a company that has not shied away from using their monopoly to make political moves. Ultimately you only need about 5 mining pools to get 51% of the hashrate in Bitcoin, and 10 to hit 75%.

The story is actually a bit worse in Ethereum—3 pools control more than 60% of the hashrate, and 6 pools will get you over 85%. I have tried to get information about how much of this hashrate is everyday users and how much is massive datacenters. Not surprisingly, the massive datacenters are not eager to advertise themselves, and it's difficult to get a good feel for the distribution. We know however though that there are very large Ethereum mining farms, and that these farms are able to use economies of scale to get significantly better cost efficiencies and energy efficiencies than what you can get with your GPU at home. Make no mistake, the centralization pressures that drove Bitcoin to where it is today are active in the Ethereum ecosystem as well—GPU mining is not a safe haven.

If you are a GPU mined PoW coin that is not Ethereum, the situation is a lot worse. But before we get too much further with that, let's step back and talk about the function and mechanics of Proof of Work. In Bitcoin (and blockchains generally), what we really care about is

transaction finality. We want to know when we receive money that it's our money, and that nobody can take that money away from us in the future. Proof of Work provides a powerful stamp that says 'this history cannot be changed without doing a lot of work'. When we get money in Bitcoin, we know that the only way we can lose that money is if an alternate history appears that doesn't include our payment, and that alternate history has more work than the history we see.

We also know that alternate histories are really, really expensive. Proof of Work provides an ingenious tether to resources in the physical world —we know that a block requires doing a ton of computation, and we know from the laws of physics that this type of computation is inherently very energy expensive. (There are types of computation that don't require much energy, or at least they require very little. PoW is not among them.) We know when we see a Bitcoin block, it took tens of thousands of dollars in electricity to produce, even using the most sophisticated hardware in the world. If someone is also making an alternate history, they are required by the laws of physics to also be spending at least that much money on their alternate history.

In terms of guarantees from the laws of physics, this is more or less where it stops. A Bitcoin block costs tens of thousands of dollars to produce, and a one block double-spend necessarily costs tens of thousands of dollars. If people are waiting 6 blocks, then an attacker trying to execute a 6 block double-spend is going to need to spend over one hundred thousand dollars executing their attack.

When you think about the sheer volume of transactions on the Bitcoin network, this doesn't paint a great picture, especially when you consider that an attacker could double-spend multiple people/services/exchanges all at once. If we stop here, it more or less becomes unsafe for the aggregate transaction volume of a single block to exceed the block reward. In practical terms though, an attacker is unlikely to be able to simultaneously double-spend every single participant in a block, especially if there are hundreds or thousands of participants, and many of them have trust with each other (meaning, you trust your friend not to double spend you when they send money, or you otherwise have some recourse if the transaction does get double-spent).

Luckily for Bitcoin, we are able to go a step farther than just counting the amount of money in electricity that would need to be spent to execute a double spend. We know that the requirement for building an alternate history is more than just a requirement of spending money on electricity. You need hardware that can convert that electricity into an alternate history. And even better, we know that the original history is going to be extended by all the non-attacking miners on the network, so you need access to more hardware than the rest of the network combined—a 51% attack.

Caveat: What matters is not only the volume of hardware, but the speed and efficiency of that hardware as well. The speed at which your hardware can convert electricity into Proof of Work is called hashrate, and for any double spend attempt to succeed, you need more hashrate than the rest of the network combined. And the efficiency of your hardware determines exactly how much money on electricity you'll have to spend to produce that alternate history. If your hardware is half as efficient (most mining hardware can see significant speed boosts if they are willing to sacrifice efficiency), then a one hundred thousand dollar attack turns into a two hundred thousand dollar attack.

In Bitcoin, if you are using hardware other than Bitcoin-specific ASICs to attack the network, your efficiency is going to drop by a factor of a thousand or more. The hundred thousand dollar attack becomes a hundred million dollar attack. For this reason, we don't typically worry about things like supercomputers—an entire supercomputer mining Bitcoin is overpowered by a handful ASICs, and the energy costs to produce a full alternate history are strictly prohibitive. If you are going to attack Bitcoin, you need Bitcoin ASICs, end of story.

We also know that ASICs are very expensive, and they are exclusively useful for Bitcoin mining. So if you have an ASIC, and you aren't using it for mining today, you've pretty much wasted your money. It's therefore unlikely that there is significant hashrate in the world that isn't actively mining on the Bitcoin network—any meaningful amount of such hashrate is a LOT of wasted dollars.

It is for this reason that, when talking about Bitcoin security, we usually ignore the raw energy costs of creating a block and instead focus on the active hashrate, and worry about what it would take to get 51% of the hashrate to rally together and attempt an attack against the network.

It's largely unrealistic (even when you consider governments) that there is meaningful hashrate out there which is not actively mining on Bitcoin.

Bitcoin has one more layer of defense, and that's through the incentive model. If you are going to try to build an alternate history, you are going to need access to billions of dollars of specialized hardware. This hardware can only make money by mining Bitcoin, which means the value of the hardware is inherently tied to the price of Bitcoin. If you attack the Bitcoin network and the attack is noticed, it is likely to shake confidence and drop the value of the coin. The value of your billions of dollars of hardware is going to drop right alongside it. So this hundred thousand dollar attack actually has secondary costs that are far, far more substantial. And I can guarantee you that any large "reorg" (when one version of blockchain history is replaced by another) is going to be noticed by the market, especially if that reorg involves large successful double spends.

As a quick recap:

1. Proof of Work provides a cryptographic assurance that a certain amount of money needs to be spent to create an alternate history. Therefore, any attack (double-spend or otherwise) using alternate histories must minimally have a payoff that is larger than the cost to create that alternate history (even if you already have 100% of the hashrate).

2. The original history is being continuously extended by the network. A successful alternate history requires having more hashrate than the rest of the network combined. So the barrier for creating alternate histories is higher than just being able to afford the electricity, you also need access to billions of dollars of hardware.

3. Specialized hardware is both faster and more energy efficient than standard hardware, by so many orders of magnitude that at least in Bitcoin and other ASIC mined coins, using ASICs is the only practical means to build an alternate history.

4. ASICs are very expensive, and inherently useful exclusively for mining that particular coin. So if the ASIC is not mining, it's just wasted money, and it's a LOT of wasted money. This makes it

unlikely that a meaningful number of ASICs exist today which are not actively mining, simply because its so expensive for them to exist and then do nothing. For ASIC mined coins only, this allows us to worry less about how expensive it is to build an alternate history and instead focus more on who owns the existing hashrate—that hashrate likely represents all practical hashrate that can be applied to attack the network.

5. For ASIC mined coins only, there is a very large amount of hardware that is useful exclusively for mining a particular coin. If the value of that coin falls, the value of the hardware necessarily falls with it. Performing attacks that could shake market confidence, scare away users, or otherwise affect the price of the coin ends up being a lot more expensive for the owner of the hashrate than just the money spent on electricity—they also have to consider the losses they incurred when the value of their hardware dropped.

Application to GPU-based Coins

The disadvantages of GPUs start most obviously at the fourth point—with ASIC mined coins, it's typically safe to assume that all of the visible hashrate represents more or less all of the global hashrate that

could exist on a coin. In a GPU or CPU based coin, this simply isn't true. And if you are one of the smaller GPU coins (like Sia is today), you actually have high visibility into hashrate that exists which is not on your network.

To put it in concrete terms, the Ethereum network has an estimated **2 million** GPUs mining on it. The two largest pools in Ethereum have about 500,000 GPUs each, and the next largest has about 250,000 GPUs on it (all rough estimates). In contrast, the Sia network only has an estimated 200,000 GPUs mining on it. That means there are 3 Ethereum pools that are powerful enough to 51% attack the Sia network today. Instead of being in a situation like Bitcoin where there are 5 pools that combined could control 51% of the hashrate, in Sia there are 3 pools that we know of which are each large enough individually to perform a full 51% attack on our network. It's a substantially worse situation!

And this is only the hashrate we know about. There are machine learning data centers with tens of thousands, and potentially millions of GPUs. Because things like this tend to be secret (Google would not want the competition to know how much they spend on machine learning computation), we have no idea if there are other parties or datacenters in the world completely unrelated to cryptocurrency that are also capable of launching 51% attacks, even against Ethereum.

We also don't enjoy any security related to the value of the hardware. If the price of Sia falls, all the GPUs actively mining Sia can be pointed to another coin. And if nothing else, they can be sold on Ebay, or even be used to participate in research networks. Attacking Sia with GPUs in a way that causes the price of the coin to fall does not destroy the value of the hardware required to perform the attack. The cost of the attack is therefore reduced to merely the cost of the electricity, which means the payoff of the attack needs only to match the electricity cost, and the attack itself can be spread across multiple participants.

To put it in concrete terms again, the cost of mining a block on Sia in terms of electricity is somewhere between $500 and $2000. If an Ethereum miner decides to attack Sia, they will also lose out on whatever profit they were making from mining Ethereum (margins tend to be low, but we'll be generous), so we'll call it $5000 to mine a block. 6 blocks would therefore cost about $30,000.

Between ShapeShift, Poloniex, Bittrex, OTC trading, etc, it's very likely that an attacker would be able to convert far more than $30,000 into another cryptocurrency in the space of a single block. Any escrow is likely going to release funds after 6 blocks (if not sooner), which means a single double-spend would allow the miner to successfully steal back the full $30,000 that they lost in creating the double-spend. And because they received payment in other cryptocurrency, the victims have no recourse.

This is an attack that is openly available to any of the large Ethereum miners, and it can be executed against any of the smaller (where smaller means 1/4th the size of Ethereum—so at this point in time… all of them) GPU mined coins. Most PoW altcoins today are sitting ducks to double-spends, and ASICs are the solution.

> *Most PoW altcoins today are sitting ducks to double-spends, and ASICs are the solution.*

When people think about mining centralization, they typically think only inside the scope of the single coin. Who are the miners of that coin, and how bad is the centralization there? Those are good questions to ask in Bitcoin, but you only get the luxury of focusing on those questions in Bitcoin because ASICs firmly protect the coin against external hashrate. GPU based coins don't enjoy these protections whatsoever. When considering the security of a GPU coin, you have to look at everyone who is able to execute a 51% attack, which includes miners and mining pools on other coins, and also includes datacenters and farms that have large volumes of GPUs for other reasons (e.g. machine learning).

Doubts about ASIC Resistance

The earlier parts of this post explain why ASIC resistance is not desirable. But even if it were desirable, I have doubts about how effective ASIC resistance really can be.

It boils down to a pretty simple fundamental argument. If you have a chip that is useful for general stuff (video gaming, computing, etc.), or anything that's not strictly relevant to mining, then it's going to have circuits and design decisions that it made to cater to these general uses. And you can always make a cheaper/faster/simpler chip just by cutting out the pieces that allow it to be useful for more general purposes.

The cost of producing chips is really high. Now that we're producing chips ourselves, I'm well aware of the barriers. I don't know the exact prices, but my guess is that even if you have the specification for a GPU, the cost of actually producing a batch of them is going to be between tens of millions and hundreds of millions of dollars. And if the ASIC resistant algorithm is effective, you might only be able to save 15% on your chip costs and electricity costs. So ASICs aren't going to appear until the block rewards are enough that a 15% advantage is going to cover the tens of millions to hundreds of millions that you had to spend producing the chip.

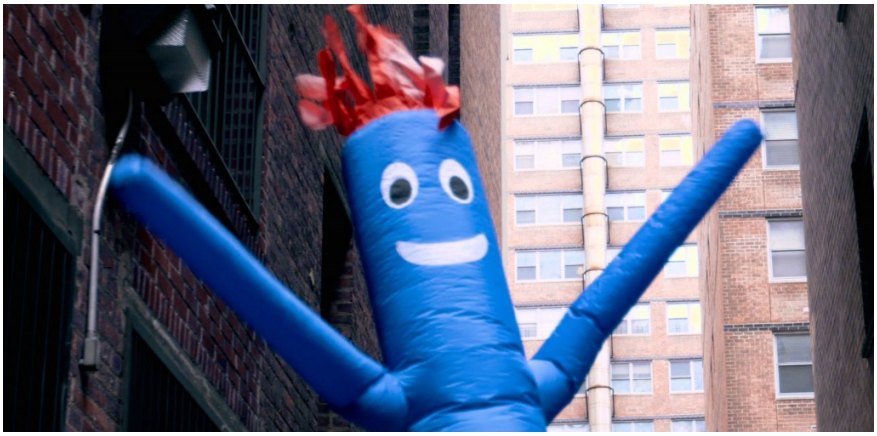
But as soon as one party is able to overcome that barrier, it's pretty much game over for everyone else. They get to enjoy the advantages of the 15% efficiency boost, and anyone else who wants to compete is

going to need to front tens of millions of dollars, and it's going to take 6 to 9 months (at minimum) for them to get their chips. And that means your only options at that point are centralization or a PoW shift.

In fairness, we haven't seen any ASICs yet for the recent ASIC resistance algorithms, despite high block rewards. The threat of losing your entire investment because the coin hardforked is a very serious threat, potentially enough to keep ASICs at bay.

Maybe ASIC resistance is possible in practice. I'm skeptical, but so far it has worked. Even so, it's not desirable, because you leave yourself open to all the GPU-coin problems discussed earlier.

On the Impotence of 51% Attacks



A lot of people seem to have the impression that you can do anything you want if you have 51% of the hashrate. 51% attacks are a lot less powerful than I think most people realize, and it's one of the greatest strengths of Bitcoin. Miners are beholden to the consensus rules. If miners create an illegal block, it doesn't matter how much hashrate they have or how much they extend the illegal chain—full nodes will just ignore them. This means that miners are unable to change consensus rules like the coin inflation or block size. Miners are unable to steal money that was never sent to them, and they can't force full nodes off of the network.

This fact, combined with all the other incentives that more or less force Bitcoin miners to keep the market + price happy, means that Bitcoin enjoys a huge amount of security even in the face of the miner

centralization that plagues it today. It is of course a far better situation if there is nobody who controls even 1% of the hashrate, but the situation today is not a dystopia. The miners have relatively little power, despite their hashrate centralization.

Still, miners with 51% hashrate are able to double-spend at will; they are able to censor transactions at will; and if they want, they can even mine exclusively empty blocks, effectively killing off the currency. But all of these actions have market-based consequences, and ultimately nullify billions of dollars of mining hardware. At least for ASIC coins, the incentives protect the network against these types of manipulation (though not entirely—a government for example may decide that censoring certain transactions is worth the resulting decline in value).

That said, we would rather not rely on incentives where we do not need to, and we will work to keep the hashrate for Sia as decentralized as possible. A network where the largest miner has 1% of the hashrate is far better than a network where 5 miners make up 51% of the hashrate, and we will do everything we can to steer Sia towards the path of greatest decentralization.

The Economics of Preventing Hashrate Monopolies

One of the controversial things that we have said is that we will not sell enough chips to any single party for them to own more than 20% of the hashrate. This has seemed alarmingly high for many, especially juxtaposed against the ideal of having no miner above 1% total hashrate.

Unfortunately, we can't control what people do. Chip manufacturing is decentralized, and if somebody has the means to produce their own chips, there's nothing we can do to stop them. If there's someone out there looking to buy 30% of the hashrate on the network, and we refuse to sell to them, they can probably just go make their own chips. And once they've crossed the R&D hurdle there's nothing stopping them from selling 20% hashrate to other interested large buyers.

If we as chip manufacturers want to stay relevant, we have to be competitive. If we close ourselves off to large buyers, we will have

significantly less capital than the competitor who does not, and we will lose the market to them. The ASIC game is very much about capital. A batch of 28nm ASICs costs millions of dollars, whether you want 100 chips or 10,000 chips, you are going to be paying on the order of millions of dollars. And if you want 16nm chips, you are going to be paying tens of millions of dollars. That means making the jump to 16nm ASICs requires tens of millions in sales.

If we close ourselves off to large buyers, we may not be able to produce 16nm chips and a competitor may be able to. That means that the competitor will have a monopoly over Sia hashrate the way Bitmain effectively lords over the Bitcoin hashrate.

If this sounds like bad situation, it's because it **is** a bad situation. But as the rest of this post has argued, it's far superior to the situation you are in when you have a GPU based algorithm. ASICs are bad, but GPUs are an outright systemic risk.

As chip manufacturers, we have to balance the goal of getting hashrate into the wide diversity of hands possible with the requirement that our chips MUST be competitive, and being competitive more or less comes down to how much money you have. We have to operate such that nobody else can have significantly more money than us, or we will become obsolete.

One of the biggest reasons that we chose to make the first batch of Sia ASICs was that we really wanted to make sure that the first batch made it into the hands of the community. The advantage that ASICs have over GPUs is enormous, and if the first ASIC chips were made by someone who intended to hold a mining monopoly, likely the only recourse would be a PoW hardfork. There's a large first mover advantage, because if you are the first ASIC you get to collect 100% of the block reward, and anybody who tries to compete with you will only be able to collect 50% (assuming they even have comparable hardware).

Why Not Proof of Stake or Proof of Capacity?

I'll start with Proof of Capacity because that's the easier one. Proof of Capacity suffers from the same exact problems as GPU mining: it's commodity hardware, and there's a LOT of it out there that could be used to 51% attack our network. Further, if the coin price goes down, the miners wouldn't care because they can just sell their hardware on the Sia network, or on Ebay, and they will still be able to use it for profit.

I'll note that this isn't a problem for the storage on Sia network, and that's because of one super key difference. When you put data onto the Sia network, you get to pick which hosts end up with your data. If there's a bad or malicious host, you have no requirement to use them. Mining is not this way though, if someone has mining resources they can force themselves onto the network. So Sia the platform enjoys protections that don't exist if we use hard drive capacity for mining.

Proof of Stake is more difficult to address, because there are a lot of highly technical reasons that it doesn't work, but there's also a lot of optimism that developments like Slasher are able to address the core problems and result with a working, secure Proof of Stake system. One of the leading Proof of Stake researchers, Vlad Zamfir, will openly tell you that a secure Proof of Stake system is very hard to achieve, that his designs are not finished, and that anyone looking to create a Proof of Stake system with security properties that rival Proof of Work has a long road ahead of them. I more or less agree, with the caveat that I think Proof of Stake is inherently insecure, and that the problems in front of it today are unsolvable in theory, let alone in practice.

Even if we assume that Proof of Stake is solvable, it has an unavoidable property that I really dislike. Once you get stake in a Proof of Stake system, by following the rules and being attentive you can guarantee that you never lose that stake. If someone were to get 51% control, they have that control forever. And it's true at smaller amounts too, if someone gets 5% control, they have that control forever. Proof of Work is different, because maintaining control requires actively burning electricity. And at any point in time, someone else can come along and burn electricity themselves, which reduces the total percentage of hashrate that you have. Essentially, with Proof of Work, you have to remain active, competitive, and invested to maintain control. And if the userbase doesn't like you, they can potentially fund their own miners to regain control.

There's a lot more I could talk about, but really any serious discussion about Proof of Stake needs to be it's own blog post. For now, I'm happy enough to say that I don't think it's the right move for Sia. We have enough hard problems that we're solving already, we don't need to add another to our plate. Especially because Proof of Work is already really, really good. Even with the centralization risks, even with the energy waste, Proof of Work is an amazing way to build consensus, and I think it's the best foundation for Sia.

Conclusion

There's a lot more that I could talk about, but we have to draw the line somewhere and this post is already very long. I can say that there are multiple factors that went into this decision that I didn't have time for

in this post, and all of them relate to keeping Sia as decentralized as possible.

The choice for ASICs is distasteful, because the disadvantages are more visible vs. other choices we could make, but I strongly believe that ASICs are far and away the best long term decision. Exploring why required a deep dive into all the components that make Proof of Work viable at all, but hopefully you've walked away for a deeper appreciation of everything that Bitcoin does well. Proof of Work is a very impressive system, and it's impressive in spite of all the miner centralization that plagues it. If you want true decentralization and trustlessness, it is the only solution that has stood the test of time.