Cell: A Scheme for Distributed Atomic Computation

Raspet, Sean
7x7

Tseng, Francis
7x7

Abstract

We propose a way to use blockchain technologies as a simulation medium, where instead of tokens there are atoms, instead of a hashing algorithm there is a physics engine, and instead of blocks there are temporal frames of a continuous simulation. As an initial use-case, we focus on cellular simulation and its applications in medical research. Much of this proposal is speculative and is intended as a starting point for further discussion.

I. Motivation

Biological models and simulations are indispensable tools for better understanding of complex biochemical interactions and mechanisms, drug discovery, and estimation of treatment effects, reducing the need for costly or risky experiments and accelerating progress in these domains. As simulations grow in resolution they provide more nuanced understandings of these systems, and a detailed enough simulation could in theory function as a "genuine" cell. However, the sheer complexity and density of interactions that constitute most biological systems of interest make it so that it is typically computationally infeasible for detailed, low-level simulations. Even higher-level simulations focused on relatively isolated parts of cell function, such as protein folding, take vast amounts of computing power, surpassing even the most powerful supercomputers. At their best, these titanic machines are capable of simulating only a few hundred atoms [1]. However, this is a rapidly developing field and recent advancements in machine learning have shown potential in predicting atomic-molecular interactions using classical computers [2]. Even so, the vast number of individual atoms interacting in a biological system such as an individual cell (a typical animal cell has been estimated to contain approximately 100 trillion individual atoms) are far beyond the limits of any individual computer at the time of writing.

The Human Genome Project, which cataloged genetic sequences of the human genome, revolutionized not only the field of genomics but medicine more broadly, giving rise to new practices such as personalized medicine, and others in tangentially-related fields such as forensics and anthropology [3]. An atomic-level cell simulation, were it feasible, might have a similarly broad-reaching impact. In addition to a general increase in the speed of pharmaceutical and medical research, a few specific possibilities include:

- Genetics research, including identifying gene targets and downstream effects for CRISPR-based therapies in the near future
- Synthetic biology, including design of de novo organisms and genes
- To predict the effects of a mutation introduced into an organism (for commercial or research purposes) e.g. predicting the trade offs and downstream effects of a gene that increases the growth rate of a target organism
- Providing an accurate foundation to other larger scale but lower resolution or more abstracted simulations (e.g. brain simulations, whole body simulations)
- Drug development, prior to clinical trials to predict the efficacy of a drug (and the likelihood of success in a costly clinical trial).
- Development of antibodies and antivenoms
- Development of antibiotics for drug resistant pathogens
- To determine the mechanism of action for drugs and therapies that have not been developed through a phenotypical approach
- Replacing animal toxicology testing in some cases

II. Related Work

A. Folding@Home

Historically, these issues of computational limits have been approached from the perspective of distributing that computation across multiple machines, with the assumption that in aggregate that power will surpass a single supercomputer, even when additional overhead of passing data across the internet is taken into account.

In the context of biological simulation, this is embodied by the Folding@Home project, an initiative based at Stanford where anyone can contribute excess computational power towards protein folding simulations and other molecular dynamics simulations [1]. As of January 2018, the Folding@Home network generates 135 petaFLOPS (floating point operations per second, a typical measurement of a supercomputer's raw computational power) [**?** , **?**]]. In contrast, the world's most powerful operating supercomputer, the Sunway TaihuLight, generates 93 petaFLOPs, with a theoretical peak of 125 petaFLOPs. However, the Folding@Home network's power may soon be bested by another supercomputer. Oak Ridge National Laboratory's Summit, a new supercomputer under construction for 2019, is expected to reach 200 petaFLOPs [4]. But the Summit too is bested – not by a supercomputer, but by another distributed computing network that is already in operation.

B. Cryptocurrency: A New Interest in Distributed Computing

Since its release in 2000, Folding@Home has never quite achieved mainstream popularity. As a matter of fact, until recently no distributed computing paradigm has gained any traction outside of academia and the software engineering profession. The recent change is in most part – if not entirely – due to the meteoric rise of cryptocurrencies such as Bitcoin.

These cryptocurrencies could be characterized as fundamentally distributed computation schemes. Speaking in idealized terms, they provide ways for large and disparate networks of people to contribute their computing power to a single cause. Given that Bitcoin could be called the ur-coin of cryptocurrencies, and that, at time of writing, it has the largest market cap at $135 billion, more than double that of the next-largest (Ethereum, at $66 billion) [5], it has the most power devoted to its mining (at time of writing, estimated to amount to 0.5% of the entire world's electricity usage by the end of 2018 [6]), and that the term "Bitcoin" is often conflated with cryptocurrency more broadly in its popular usage, the discussion here will focus on Bitcoin as a point of reference.

[1] https://folding.stanford.edu/

A core mechanism in Bitcoin's functioning is a scheme called "Proof-of-Work", which requires devoting considerable computational resources to solving arbitrary difficult problems [7]. This is part of a process called "mining", and successfully solving these problems can yield Bitcoins, meaning that it can be quite lucrative to dedicate significant computing power exclusively to mining Bitcoin. In theory anyone can use their computer to mine Bitcoin with. This mining, like all computation, requires energy, and one of the main criticisms of Bitcoin is that it uses exorbitant amounts of electricity – as cited previously, estimated to reach 0.5% of the world's total energy usage by the end of the year – and that the computation this electricity is used for is essentially wasted, since it never produces anything of value (depending on your opinion of Bitcoin) [2].

In 2013, the total computational power of the Bitcoin network was estimated to be 1,000 petaFLOPS, over seven times the current power of Folding@Home, and five times the power of the Summit supercomputer. Bitcoin has seen enormous growth in popularity since then. One estimate states that the current Bitcoin network's power is 80.7 million petaFLOPs [8], eclipsing the network's already staggering power from 2013.

That Bitcoin has managed to attract so much more computational power than Folding@Home speaks to the incentive structures around Bitcoin. We are not the first to recognize that perhaps there is something to be gained by porting Bitcoin's incentives to a project like Folding@Home.

C. FoldingCoin and CureCoin

FoldingCoin [9], and another similar coin CureCoin [10], form a layer on top of the Folding@Home network by providing economic incentives to contribute computational power. They do not appear to fundamentally alter how this network is structured. The primary differentiation from Folding@Home is that they distribute their own coins as reward for joining the network, but they do not offer any convincing way of making these tokens valuable. It seems the hope is that they will develop their own exchange value in the way that Bitcoin did, but it would be nicer if the value of these coins were more intrinsic to the project itself.

[2] These are proposals for alternatives to this Proof-of-Work method, most notably Proof-of-Stake, which requires relatively little energy and is instead designed around economic incentives. However, whereas Proof-of-Stake seeks to remove the need for heavy computation, our focus is on how that heavy computation can be redirected to meaningful purposes.

III. CELL

A. Atoms as Tokens and Computation

The CELL platform will repurpose the cryptographic architecture and distributed structure common to cryptocurrencies to generate use value through the simulation of complex physical phenomena at an atomic resolution.

The specific model of distributed computing used will greatly impact CELL's performance and feasibility. Distributed computing paradigms such as agent-based models can see a performance drop when distributed across multiple machines, often because the process of distribution itself introduces additional overhead in the form of network communications. For example, if agent A is on machine X, and agent B is on machine Y, and these machines are very far apart, communication between agents A and B would be slower than if they were located on the same machine. So the question of how a program scales in performance as it is distributed is not a straightforward one.

Bitcoin and other cryptocurrencies are known to have scaling problems, typically framed in how many transactions per second the network can process. At time of writing, Bitcoin peaked at 9 transactions per second [11], whereas Visa, the common benchmark for these comparisons, processes on the order of 10,000 transactions per second [12]. This is because these cryptocurrencies derive their security from independent verification of transactions, so in practice while these networks are *distributed*, their computation is not in practice *parallel*, and parallel computation is foundational to the performance gains from distributed computing. So we cannot use the Bitcoin distributed computing model directly for CELL. Instead, we look for inspiration from Ethereum, which has some projects focused on moving some computations "off-chain" to increase the network capacity, effectively parallelizing some parts of the network.

We look to the video game industry as another source of inspiration. In the video game industry, open-world games have risen in popularity, especially MMOs (massively multiplayer online games), where hundreds to thousands of players may occupy the same massive world. In these games there may be many, many interactions sprawled over an extremely large space, and this becomes infeasible to host on single servers. So distributed systems are needed, and issues like network overhead come into play again. Companies like Improbable [3] specialize in services (in particular, their SpatialOS product) that take care of distributing these game worlds across multiple servers and minimize this distributed overhead through a variety of strategies. One effective strategy, which is common in simulations that have a spatial component, is to localize players that are physically near each other (in-game) to the same server, under the assumption that they are more likely to interact with those around them. This means that most interactions don't require network communication to be computed.

For CELL we propose a similar approach. Whereas agent-based models describe their fundamental computational components as *agents*, ours are *Atoms*. Atoms that are near other in-simulation are also co-located on the same physical machine where possible, to minimize network overhead. This is an appropriate assumption, since atomic interactions tend to be local, especially as they form larger structures (molecules).

We combine this agent-like notion with the cryptocurrency notion of *tokens* (discrete units of exchange). The specifics of the Atom token design depend on the incentive structures; some options are proposed later in this paper.

An Atom does not represent a specific atom, e.g. a carbon atom or a phosphorus atom, but rather some abstract "stem atom" that can represent any arbitrary atom, depending on the specific needs of the simulation. This means that they are fungible from the perspective of simulation and computation. However, some of our proposed incentive structures require that they are non-fungible from the token perspective, detailed later in this paper.

B. Simulation Details

Bitcoin is built around a distributed ledger that is appended to, keeping track of all transactions (the global state of the network). CELL uses a similar ledger to keep track of the simulation state. However, this singular ledger contributes the previously mentioned transactions-per-second bottleneck. Ethereum, the second most popular cryptocurrency after Bitcoin, has a number of proposals to resolve this bottleneck, one of which is to take some computation "off-chain" or to fracture the network into sub-networks "shards". We can take a similar approach, especially given the nature of localized atomic interactions, where sub-networks manage localized states that merge into the global simulation state as needed.

As mentioned, the structure of the CELL platform would roughly correspond to Atoms as tokens, and a "block" or ledger entry as a unit of time, or temporal frame that tracks the movements and physico-chemical interactions of those Atoms.

A very detailed simulation would include the Brownian motion of molecules and an accurate simulation of temperature and the infrared vibrational patterns of molecular compounds. However, a high-resolution

[3] https://improbable.io/

"frame-by-frame" rendering of these movements would be extremely computationally intensive. Hence "cost-saving" measures such as a probabilistic rendering of the position of an atom and its vibrational movement would likely be employed for most simulations.

The location of the Atom unit would in this case be calculated as a field with areas of more and less probability radiating out from a central point. This calculation would very similar to the typical rendering of an electron cloud around an atomic nucleus or molecule system as a Gaussian distribution of probable positions. In this case the position within space of the atom (and its electron field) would also be distributed in a probability cloud. This feature is here termed a "motion cloud" to distinguish it from the "electron cloud" of typical molecular renderings.

Likewise, a working simulation would need to account for the potential chemical interactions of the molecules contained in the simulation. As mentioned, this is also a computationally intensive area, but recent developments in machine learning may allow for an accurate and less computationally intensive calculation of these interactions. The values for possible chemical interaction and covalent bonding of Atoms will be distributed in a probabilistic or "electron cloud" manner.

C. Values

Among the values that would be attributed to an Atom within the CELL platform are the following:

- X, Y, and Z spatial values to determine the position of the Atom system at a given time within the finite space of the simulation
- A Time value linking it to a frame of the simulation
- A Velocity value
- A Type value that determines the atomic element (and electron cloud and chemical bonding properties)
- A Charge value
- A set of values relating to its Electron Cloud, orbitals and bonding potential. Interactions such as covalent bonds, hydrogen bonds and ionic interactions would arise from these values.
- A set of values related to its Motion Cloud
- A set of values relating to Neighbors—-i.e. covalently bonded atoms within a molecule.
- An individual Vibration variable that is responsive to the temperature conditions of the simulation universe, and the IR vibration of the molecular system to which an Atom belongs. Depending on the nature of the simulation and the computing intensity, this value may be separate from or merged with the Motion Cloud value above. While temperature will technically be an emergent property of the simulation as a whole, this value will be meant to represent the target temperature so that interactions can be coordinated between Atoms to maintain the target temperature throughout the simulation.
- As Atoms are arranged into molecular sets, other emergent values result from the results of the overall system. These values may likewise be referenced within the unit of the Atom itself for tracking purposes or other purposes within the CELL platform.
- In addition to all of the above "physical features" of an Atom, it would also have a unique identification number, and a "public key" or other method of linking it to a particular "wallet" or account owner.

All of these properties would be components of the Physics algorithm which in the CELL platform replaces the hashing algorithm common to Bitcoin and other blockchain networks.

Ideally, the values of the Atoms would be calibrated such that properties of the molecules that are formed from them as well as subsequent activity at higher levels of the simulation would be emergent properties. Likewise biological activity taking place in the simulation would be an emergent property of the atoms and molecules present and their interactions.

Blocks (or "Frames") will be variable and adjustable to account for the different needs of a particular simulation project.

D. Advantages

Some proposed advantages of the CELL platform over other methods of biological research in no particular order:

- Two simulations can be run in parallel with the exact same cell changing only one variable—-such as a change in a particular gene or the introduction of a particular drug molecule. The two simulations can then be compared to show what changes resulted, allowing for a level of standardization and certainty beyond what is possible in a lab.
- Atoms and molecules can be tagged or tracked easily
- Does not require abstraction or theoretical assumptions to be made; allows an extremely high resolution of detail.
- The behavior of the simulation is an emergent property related to the interaction of its underlying units (Atoms). As such it may yield discoveries that are outside the realm of knowledge that has been already theorized.
- New molecules can be introduced for testing without the need for costly physical synthesis.
- Biological processes can be slowed down and zoomed in on to gain understanding at a much higher degree of detail than is currently possible with empirical observation.

- Machine learning can be applied to the data to suggest new target sites for therapies, new therapeutic molecules or other treatments that may not be intuitive to or otherwise discoverable to an expert in the field

E. Incentives

It's hard for us to confidently assert what incentive structure would be best suited for this project, and we plan to develop the structure with the feedback of potential contributors to the platform. In general, good incentive structures are extremely difficult to design. Ideally our system does not have a compounding "rich-get-richer" effect, i.e. where those with more Atoms have more resources to acquire more Atoms until the network is effectively owned by a small set of parties.

We also want to avoid issues of FoldingCoin and CureCoin where the value of these tokens aren't designed into the system itself, but instead are presumed to take on their own exchange value at some later point.

Finally, mere ownership or holding of Atoms should not guarantee accrual of value from the network, whatever the particular incentive model. Rather, it is *participation* – i.e. donation of computational capacity – that yields value from the network. If Hosts are offline, i.e. their computer power is inaccessible, then they are missing out on a stake of future rewards. This discourages practices of rent extraction.

There are two main possibilities we're considering.

First is a more traditional high-performance-computing-as-a-service model, where access to powerful computational resources (e.g. a supercomputer or a computer cluster) is leased out to research institutions. In the spirit of cryptocurrency, this service could be implemented as an automated bidding process to eliminate possibility of arbitrage. Revenue from this leasing would then be distributed according to Atom ownership. This is the simplest approach, but does not guard against rich-get-richer effects.

The second approach, which is not mutually exclusive with the first, is based on the assumption that this cell simulation will be instrumental in future drug discovery, organism design, and other medical and commercial research. Atom owners gain a share of the ownership of these discoveries, which may provide a mechanism for democratizing the fruits of medical and biological research that to our knowledge does not exist.

As an open and distributed platform, the results of any simulation would be traceable and viewable by the general public. Any company or entity wishing to employ a simulation for its research would need to give up the secrecy that typically surrounds privately-funded research, since the results of the simulation would be accessible through the public ledger of the CELL platform. This may initially dissuade certain corporations or entities from using the platform, however, given the potential economic value of an accurate simulation of, for example, a human cell, the cost-savings to (for example) a pharmaceutical company, may be well worth the trade off.

The incentive structure would likely provide for a certain degree of private ownership (as well as tradability) of Atoms that are brought into the universe. Like other blockchain-based or distributed systems, where tokens or coins can be traded and owned, Atoms may develop an intrinsic and/or speculative value that complements their immediate use value as a component of a simulation.

A related issue is that of how the ownership of a previously "mined" Atom relates to its future computational deployment. If the owner(s) of a particular Atom is offline. Here we propose a continuum between simulation matter and simulation energy, loosely correlated with the hierarchy of matter and energy in the physical world. For example the amount of stored potential energy in a single carbon atom is 12 GeV or $2e10^{-8}$ joules. In this sense, within the simulation, a carbon atom could be said to be "worth" or exchangeable for 12 GeV of energy.

From outside the simulation we would say this energy corresponds to computing work contributed by each Host (i.e. "miner") to run the simulation (i.e. the distributed computing system). However, given the very large amount of energy stored in a physical atom, the "exchange rate" of computation work and the generation of atoms would likely be on a more even ratio.

F. Open Technical Issues

1) Computational Intensity: At this point it is difficult to predict the computational resources needed to accurately simulate a complex biological process. Many trade-offs will exist between simulation speed, simulation resolution, and the completeness and accuracy of the physico-chemical interactions of the component atoms. Likewise the size and distribution of the CELL network and its incentive structure will change over time and will determine the total pool of computation on which to draw, adding another layer of variables and trade-offs to any attempt to predict the above variables at this point in time. The feasibility and usefulness of the CELL platform's simulations will help determine the adoption rate of of the CELL network, this in turn will influence the incentives of the Host (i.e. miner) participants, and this in turn will impact the quality, speed, and usefulness of the simulations.

As such we propose a gradual approach to development by initially implementing comparatively small-scale biological simulations. One specific starting point would be, for example, to model the mechanics of a ribosome assembling a protein from a particular RNA sequence. Here we would focus on determining and validating the Physics algorithm governing block (Frame) generation and how this relates to the values contained in the component Atoms. Since the ribosome has been studied and modeled we can compare the results of the simulation to the empirical data and iterate the simulation parameters as necessary.

2) Incentive Structure: At the same time, and of paramount importance we would work on the gradual building of a participating community of Hosts and an incentive structure that is arrived at through consultation with the emerging community. The incentive structure will attempt to balance community needs, future network adoption, and long term network stability and security.

Some open questions regarding the incentive structure in addition to those stated above are:

- The continuum of value between matter (Atoms) and energy (computations).
- The method of determining what simulation project the network will implement.
- The means by which the initial state of a simulation is administered or populated to the network participants.

3) Input: In addition to the functionality of the Physics algorithm, the quality of the initial input of the simulation will be of great importance. A likely strategy is to import the positions of all the atoms in a cell from a high resolution, static scan.

High-resolution scans exist of whole cells that may be suitable for an initial data state. However if the data is input from a scan, there is a difficulty in determining in what way atoms are bonded together into molecules. For example, if two molecules are immediately adjacent, the atoms of one molecule may appear to be part of the other molecule. It may be possible to develop a machine learning system that can determine what atoms are bonded to which based on a library of probable molecules and an ability to recognize these. A comparison of multiple scans of the same cell taken over a short time may also help in this process. After a successful simulation of cell is achieved it may later be possible to design cells for simulation de novo or through a hybrid approach.

IV. Further Discussion

While an accurate atomic-resolution simulation of a complete, functioning cell is admittedly still a long way off, the possibility raises some interesting philosophical questions and observations that are worth noting. Rather than attempt to answer any of these we instead present a list of some of these below and we welcome further discussion on this aspect of the project.

- If the simulation was working with excellent accuracy to a living cell, would it be considered alive?
- If so, would a simulation of a non-living thing/-material process at the same level of accuracy be considered alive or not alive?
- The cell's biological functions can happen in a discontinuous temporal manner—-e.g. In "our time" the cellular processes do not have to be computed at the same time, but if they contain continuous time values within the simulation they will function as if continuous from within that perspective.
- If it is working correctly, the cell will need to be "fed"—-nutrients will need to be supplied from outside and waste products will need to be able to to be cleared from it. As such the cell will be capable of "death" or a ceasing of biological function.

References

[1] G. Popkin, "Quantum computer simulates largest molecule yet, sparking hope of future drug discoveries," *Science*, September 2017.

[2] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the kohn-sham equations with machine learning," *Nature communications*, vol. 8, no. 1, p. 872, 2017.

[3] J. L. Fridovich-Keil, "Human genome project," *Britannica*, April 2018.

[4] D. Galeon, "The next most powerful supercomputer in the u.s. is almost complete," *Futurism*, October 2017.

[5] "Coinmarketcap," " [Online; accessed 17-May-2018].

[6] A. de Vries, "Bitcoin's growing energy problem," *Cell*, May 2018.

[7] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," " 2008.

[8] "Bitcoincharts - bitcoin network."

[9] "Foldingcoin white paper," " [Online; accessed 17-May-2018].

[10] "Curecoin white paper," " [Online; accessed 17-May-2018].

[11] "Bitcoin transaction rate," " [Online; accessed 17-May-2018].

[12] T. B. Lee, "Bitcoin needs to scale by a factor of 1000 to compete with visa. here's how to do it." *The Washington post*, November 2013.