Summary

Artificial intelligence (AI) is the single most important technology we've ever seen. But what only a relative few can see is the imminent transformation of life as we know it once artificial general intelligence (AGI) is achieved—as soon as 2027, by some accounts. Whoever is first to roll it out at scale will be granted an unprecedented level of control over the future of humanity.

The race to AGI has created a thirst for compute so insatiable that there is no conceivable scenario where demand can be met. This has sent big tech, governments, and anyone else with the resources scrambling to secure enough GPU chips, spending tens of billions in the process. For those without the resources, they are stuck relying on expensive cloud compute providers.

Decentralized physical infrastructure networks, or DePINs have emerged as an alternative. As distributed networks of aggregated small-scale infrastructure, they are designed to support rapid scale. In the compute context, this means GPU chips owned by data centers, tech companies, telecom companies, top gaming studios, and crypto mining companies

As we move towards AGI, we are confronted with a problem that stands to alter the course of humanity if left unchecked: the unfair distribution of outcomes that stems from a handful of companies controlling a technology as important as AI. The widening **AI wealth gap** between the GPU Rich and GPU Poor is tilting the balance heavily in favor of big tech.

In response, we need to increase the accessibility of on-demand compute so that AI companies can claw control away from big tech and produce the innovative outcomes that will ensure AI is developed for the good of humanity. The only way to do this is by leveraging the power of DePINs to create a distributed network of affordable compute resources accessible by all.

Aethir has built this distributed network. It aggregates enterprise-grade GPU chips into a single global network that increases the supply of on-demand cloud compute resources for the AI, gaming, and virtualized compute sectors. **Enterprise GPU owners** can unlock the revenue potential of their underutilized GPU chips, while **end users** get access to the affordable on-demand compute resources they need to power their AI training and inferencing workloads, real-time rendering applications, and other virtualized compute operations.

What differentiates Aethir from other cloud compute infrastructure is its ability to scale to meet the demands of rapid AI and cloud gaming growth. Its distributed model is designed to streamline network expansion so that it can deliver the most optimal and affordable compute resources on demand to wherever they are needed. Five key features make this possible:

- **Enterprise-grade** compute resources
- **Low latency** to support real-time rendering and AI inferencing use cases
- A **distributed** model that can scale much faster than its centralized counterpart
- Superior unit economics that make on-demand compute resources **affordable**
- **Decentralized ownership** that allows resource owners to retain control

Aethir is powering the future of gaming. It's the only cloud compute infrastructure designed specifically to support real-time rendering at scale for mobile and PC cloud gaming. It can serve gamers across geographies with a seamless, low-latency experience without the need for high-end hardware.

It's also enabling the evolution of AI interaction, through a combination of real-time rendering and AI inferencing that requires compute infrastructure not yet accessible outside of big tech. Aethir's real-time rendering infrastructure is the foundation on which innovators can revolutionize the way we interact with AI models and enable the kind of innovation that will bring AI truly into everyday life.

As a DePIN, Aethir leverages Web3 to enable incentivization and consumption tracking. Resource owners are rewarded in Aethir's native $ATH token for providing resources to the network, while blockchain is used to track resource consumption by end users and facilitate the transfer of rewards.

Aethir's cloud compute infrastructure network supports a number of critical use cases, including AI model training, AI inferencing, cloud gaming, cloud phone, and AI productization. In each case, its capability for scalable, affordable, low-latency compute can keep up with the demands of innovation in these areas.

The network consists of 3 core roles: Container, Checker, and Indexer. Containers, including the Aethir Edge, are where the actual utilization of the compute resources takes place. Checkers ensure the integrity and performance of the Containers. Where required, Indexers match end users with suitable Containers based on end-user requirements. Together, they ensure that Aethir can deliver the highest-quality compute to support rapid AI and cloud gaming growth.

Market Context

Artificial intelligence (AI) is the single most important technology we've ever seen. To truly understand Aethir's impact on AI's future, we first need to look at where AI is today, where it's going, the state of the graphics processing unit (GPU) industry, and the rise of decentralized physical infrastructure networks (DePINs).

The Future of AI

Bullish economic forecasts by today's generation of research report leaders fail to capture AI's true potential for impact. Estimates such as a US$15.7 trillion contribution to the global economy by 2030 may make headlines, but they miss the point. AI isn't about economics, it's about the future of the human race. Through that lens, AI's impact can't be understated.

This may not be so obvious today, however. The first generation of generative AI solutions—ChatGPT, Midjourney, Gemini, and others—offer glimpses of the future but are really just scraping the surface of what's possible. What only a relative few can see is the imminent transformation of life as we know it once artificial general intelligence (AGI) is achieved—as soon as 2027, by some accounts.

What will that life look like? Fundamentally different, in short. The monotony, the long commutes, the menial labor, the subsistence farming will all be relics of our current industrial age. AGI-powered agents and robots will be able to do nearly everything for us, and our interaction with them will be via real-time rendered avatars, not keyboards.

This will make a vast majority of current jobs redundant—as many as 300 million by 2030, according to a report by Goldman Sachs—but the inevitable introduction of universal basic income schemes will enable a transition to a post-work society with greater economic equality. We will, for the first time, be able to choose the life we want to live, not work solely for the sake of survival.

Given what's at stake, it should come as no surprise that a great AI arms race has begun. And there has never been a more important race. Whoever is first to achieve AGI and roll it out at scale will be entitled to enormous wealth and, more importantly, granted an unprecedented level of control over the future of humanity. The nuclear arms race and its importance to geopolitical supremacy is the closest parallel we have. The US alone spent an estimated US$5.5 trillion (in 1996 dollars) between 1940 and 1996 to entrench its

position as a superpower. Companies working on AI are well on their way to that level of expenditure.

Big tech—Google, Meta, Amazon, and Microsoft—is leading the charge. They've already spent hundreds of billions on compute resources, modular power plants, power contracts, and human talent in an attempt to control the development of AI and its component parts. The discussion is no longer about multi-billion dollar compute clusters. They're now talking trillion-dollar clusters as the foundation for the race to AGI.

Cash-rich governments have also joined the party. The UAE, for one, has its sights set on becoming an AI power. (OpenAI's Sam Altman famously sought US$7 trillion from the country to build a competitor to NVIDIA.) Not to be outdone is the US$350 billion gaming industry and its real-time rendering capabilities that are critical to evolving our interaction with AI. Real-time rendering on its own is a market that's set to grow to nearly US$4 billion by 2033.

Such intense competition means it is no longer a question of if we'll transition to a post-work future but when. A key indicator of how quickly we get there are the GPU chips that power the training and inferencing computations.

The State of the GPU Industry

GPU demand has become a new global constant. The race to win AI has created a thirst for compute resources so insatiable that there is no conceivable scenario where demand can be met. To put it in perspective, Moore's law states that hardware efficiency doubles every 18 months. One estimate suggests that GPU demand is growing 10 times over that same period.

For example, it took an estimated 3,000-10,000x ***more*** compute to go from GPT-2 (2019) to GPT-4 (2023)—a jump in intelligence, experts say, that's equivalent to evolving from a preschooler to a smart high school student. It's not difficult to imagine just how much more compute will be needed to get to AGI and beyond.

One company, NVIDIA, controls some 70% of the estimated US$65 billion GPU market because it's the only company that can produce the battle-tested GPUs coveted by AI and gaming. Predictably, due to its own resource constraints, the company isn't able to produce enough GPU chips to meet demand. Those it does produce have sparked controversy around how they are allocated, with the CEO, Jensen Huang, having to state in February 2024, "We allocate fairly."

One commentator, in November 2023, put the supply gap north of half a million chips.

This has had a far-reaching impact on the AI industry, even threatening to slow its development. It has sent anyone with a horse in the race scrambling for GPU chips. Those with means—big tech—are investing heavily to maintain their post position. Meta spent US$10 billion on 350,000 NVIDIA H100 chips; Amazon, Google, and Microsoft are all building their own proprietary chips; and the proliferation of cloud GPU providers are snapping up chips to make available on demand.

Those without the means are forced to rely on the proverbial scraps—slower GPUs and even CPUs—or expensive on-demand resources from cloud GPU providers. To no one's surprise, today's cloud GPU infrastructure is dominated by Microsoft, Amazon Web Services (AWS), Google, and even NVIDIA. But alternative players, such as CoreWeave, are also making noise and attracting billions in investment.

An important point is that GPU chips sold to companies other than hyperscalers and cloud GPU providers have no meaningful impact on compute supply. Even if they wanted to, these companies don't have the capabilities to make these resources available via cloud infrastructure. While difficult to quantify with any certainty, there are likely hundreds of millions of these GPU chips sitting underutilized (an estimated 15-25% utilization rate) inside tech companies, enterprises, data centers, and cryptocurrency mining companies. Unlocking access to these resources is enough to materially affect the supply of compute resources.

The Rise of DePIN

Decentralized physical infrastructure networks, or DePINs, are based on the idea that you can use blockchain and cryptocurrency to build distributed networks of aggregated small-scale infrastructure, such as energy, compute, wireless, AI, and sensors. On the supply side, cryptocurrency is used to incentivize owners to contribute their resources, while the blockchain is used to track consumption on the demand side and facilitate payments to the owners.

A good example is decentralized storage. People with underutilized space can rent it to a decentralized network in exchange for a fee paid by those who utilize the space. When aggregated, the total storage available can easily surpass that of data centers.

One thing that makes DePIN attractive is that instead of an entire infrastructure network being owned by a single centralized entity, each part of a DePIN is controlled by its respective owner. This shifts control away from centralized resource monopolies, who may not be incentivized to expand access; ensures more equitable resource distribution; and keeps capex and operating costs low.

But the real advantage of DePINs lies in their ability to support scale. More specifically, it's easier and cheaper to expand a network of small-scale resources than it is to build new data centers. It makes DePIN-based compute infrastructure, for example, better aligned to serve the needs of AI growth than traditional models like centralized cloud infrastructure

DePIN's popularity has caught the attention of the wider blockchain and cryptocurrency industry. Billions of dollars have flowed from investors to DePIN projects, with the top ones having raised at least $1 billion. In its State of DePIN 2023 report, Messari estimated the market cap of DePIN companies with liquid tokens to be about $20 billion, but generating *only* $15 million in annualized on-chain revenue. (For perspective, Aethir has already eclipsed US$15 million ARR in its first year). Experts are predicting DePIN to take an even bigger leap in 2024, with some suggesting that a DePIN project will be the first decentralized app with a billion users.

All told, the race for AI supremacy and resulting compute shortage point to a concerning outcome: What will our future look like if only a handful of companies have control of the way AI develops?

The Problem

Make no mistake, the demand-driven shortage of compute resources is not short on consequences. You have the higher prices, longer wait times, greater latency, and lower-quality chips typically associated with a supply gap. Then there's the talent migration to companies with better access to compute resources. And that's not even to mention the potential environmental impact.

But only one stands to alter the course of humanity: the unfair distribution of outcomes that stems from a handful of companies controlling a technology as important as AI.

The AI Wealth Gap

The way it stands now, those with the means—big tech—are in control of how AI develops, not because they are better at it, but because they control the resources. OpenAI, NVIDIA, Google, Amazon, and Meta aren't the ones suffering from chip shortages; they aren't forced to use old hardware or wait in line for access. This privileged position means they ultimately decide which underlying models get built, how they get trained, and how much they cost to use. It's a true oligarchy scenario in which companies building the AI product and service layer will have little choice of which models they have access to.

This scenario has led one commentator to dub those with means as the "GPU Rich" and everyone else as the "GPU Poor." In the context of compute, we're seeing the Rich push the resource shortage further down the chain towards the Poor. Represented by startups and well-known AI companies alike, the Poor simply don't have the means to create their own AI outcomes—despite having a fair share of the world's AI experts. Society is shutting itself out of a massive avenue of innovation. This is particularly true at the intersection of AI and gaming, where real-time rendering promises to rapidly evolve AI interaction.

What we're effectively witnessing is the creation of an AI wealth gap that is dangerous, both for the industry and humanity. For one, there are serious ethical concerns around a handful of companies dominating an industry. We only have to look to social media as a prime example. Censorship, lobbying, impunity, campaign financing, exploitation, privacy violations. The list of concerns runs long. Now imagine those same companies in control of AGI. It's not a comforting thought. Even more sinister, there is a belief that big tech will begin exploiting AGI long before the achievement is made public. It's part of the reason why the US justice department and federal trade commission are probing monopolistic practices at Microsoft, OpenAI, and NVIDIA.

Secondly, a widening AI wealth gap increases the cost of competition. The further along we go towards AGI, the costlier it will be for smaller companies to build and train competing models. Those that do show promise will simply be swallowed up by the Rich and used for their own benefit.

What we're heading towards is an over-centralized AI industry with a handful of companies acting as gatekeepers to everything from compute resources to talent. It's a blow to the very ideals of democratic capitalism and the best way to guarantee an unequal distribution of outcomes. We simply can't have this if we are going to evolve as a species. Without urgent action, we risk losing our agency in a technology that will transform our world like no other in history.

In response, we need to increase the accessibility of AI resources, to give AI companies and those working at the intersection of AI and other industries the tools to claw control away from big tech and produce the innovative outcomes that will usher us into a post-work society. We can do this by leveraging the power of DePINs to create a distributed network of affordable compute resources accessible by all.

The Solution

Aethir is best described as distributed cloud compute infrastructure. It aggregates enterprise-grade GPU chips into a single global network to increase the supply of on-demand cloud compute resources for the AI, gaming, and virtualized compute sectors.

What this means is that:

1) **Enterprise GPU Owners** can unlock the revenue potential of their underutilized GPU chips by becoming their own cloud compute provider.
2) **End Users** get access to the affordable on-demand compute resources they need to power their AI training and inferencing workloads, real-time rendering applications, and other virtualized compute operations.

Scaling Compute On Demand

What differentiates Aethir from other cloud compute infrastructure is its ability to scale to meet the demands of rapid AI and cloud gaming growth. Its distributed model is designed to streamline network expansion so that it can deliver the most optimal and affordable compute resources on demand to wherever they are needed. This is critical for real-time rendering, in particular, because it can't currently be done at scale using today's cloud compute infrastructure. It's on this foundation that the next generation of innovative AI products and cloud-optimized games will be built.

Leveling the Playing Field

Most importantly, Aethir is a direct response to the widening AI wealth gap. By providing the GPU Poor with access to affordable enterprise-grade compute resources, it has positioned itself as a force for good that helps balance the distribution of AI outcomes. It also avoids the geographical and economic favoritism typically associated with other compute providers. At the same time, it's creating new revenue streams for resource owners who wouldn't have had this opportunity otherwise.

Features

Aethir has five defining features that separate it from other cloud compute infrastructure providers:

Enterprise grade

The demands of AI and cloud gaming require high-quality cloud compute infrastructure. Aethir focuses on aggregating high-quality GPU resources, such as NVIDIA's H100 chips, from data centers, tech companies, telecom companies, top gaming studios, and crypto mining companies. This ensures a level of compute quality that end users can rely on, whether they are startups, small- and medium-sized enterprises (SMEs), or large enterprises.

Low latency

Latency is a core aspect of real-time rendering. High latency makes cloud gaming impossible, especially at any sort of scale. But this is exactly what happens with today's cloud compute infrastructure. Even if the compute resources are available, they aren't often capable of delivering on these latency requirements unless the end user is geographically close to the origin of the compute. This puts serious constraints on real-time rendering innovation.

Consider gaming in Asia. More than half of the world's gamers live there, but most are gaming on low-end devices. This means two things. First, most simply don't have access to the high-end gaming content produced by the biggest gaming companies. Second, the biggest gaming companies don't have access to the majority of Asian gamers.

Cloud gaming technology would unlock a huge amount of value in Asia simply because it abstracts compute requirements away from low-end devices and allows those devices to access high-end gaming content. Until now, this technology hasn't been deployed throughout the region due to the cost of scaling cloud compute infrastructure. This is a problem that DePINs solve. Aethir, with its distributed model, can meet the demands of cloud gaming in any specific region, providing scalable, low-latency compute to any gamer that needs it.

Distributed

Centralized compute infrastructure models are simply too slow to keep up with the demands of AI and real-time rendering. They are built on the premise of buying new chips to increase supply, something that is difficult amid a demand-driven chip shortage. As mentioned, they also struggle to meet geographic- and industry-specific requirements.

Aethir is distributed by design. It aggregates existing chips from across the world into a powerful and responsive network that can meet end users where they are. Scalability

then becomes a reality because new compute supply can be added without Aethir needing to purchase any chips.

Affordability

Aside from ethical concerns around control, centralization has another disadvantage: high cost. Data center operation and company overhead are ultimately reflected in the price paid by the consumer for compute resources. And with insatiable demand for compute driving prices ever higher, smaller companies are often completely priced out of the market.

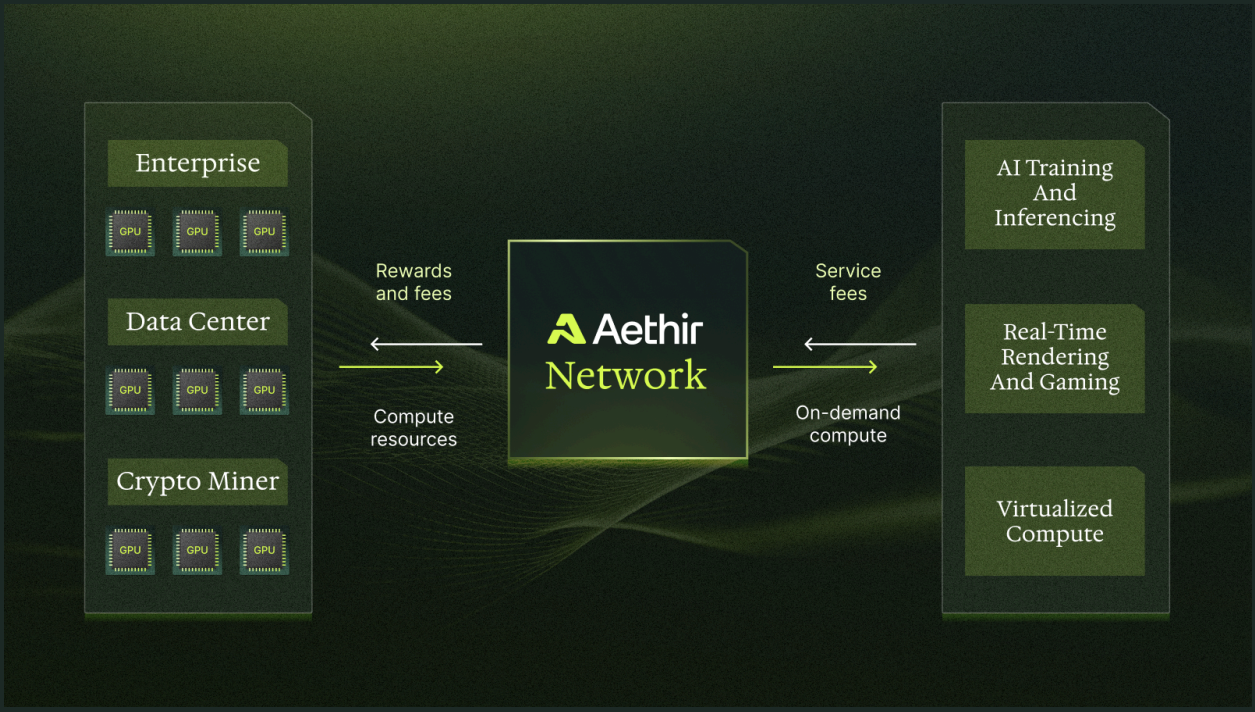
Aethir's distributed network eliminates the costs associated with legacy providers by as much as 80%. It doesn't have to pay to construct and operate data centers, nor does it carry the same level of company overhead as large corporates. Aethir can, therefore, achieve superior unit economics, enabling compute providers to compete fairly on price. This ultimately means affordable on-demand compute resources for all, at scale.

Decentralized ownership

One of the primary benefits of DePINs is that resource owners always retain ownership of the resources they contribute to the network. In the context of the billions of dollars big tech makes off ownership of our data, we can begin to understand how important it is to make decentralized ownership a core pillar of infrastructure networks.

Aethir's model guarantees resource ownership by design. Owners have full control over how often they make their resources available. And provided they meet quality control standards, they can earn a steady reward stream that accurately reflects the value of their contribution. What this means is that it's now financially viable for a resource provider to contribute to Aethir and operate as its own AWS-like cloud compute infrastructure.

How It Works



On the supply side, the process for resource owners is straightforward. They simply register their compute resources with the network, have the resource specifications evaluated and confirmed, and then stake $ATH to be eligible to service requests.

On the demand side, end users submit requests to the network which are matched with the most appropriate, high-performance resources. For gaming use cases, the end users are the gamers themselves and are matched to a high-performance resource that contains their requested gaming experience.

Upon delivery and confirmation of satisfactory request completion, rewards flow from the compute buyer and the Aethir treasury to the resource owner.

Powering the Future of Gaming

The gaming industry is massive. With more than 3 billion gamers currently and an estimated half-a-trillion-dollar market value by 2027, it sits at the forefront of innovation. Two factors holding the industry back are the download size of large online multiplayer games and the increasingly high consumer hardware requirements. A side effect of increasing graphics quality, a game like Starfield requires a 125GB download. And that doesn't even include any additional components or the bandwidth required to play the

game in real-time. In regions with slower internet speeds or data caps, playing games like this can prove challenging.

Enter real-time rendering. The future of gaming, it allows gamers to play real-time, photorealistic games directly in the cloud, storage free, without needing to download them. The problem is that current cloud compute infrastructure can't support real-time rendering at any sort of scale. There just aren't enough compute resources available at any given time, let alone affordable resources with low latency.

Aethir has the answer. It's the only cloud compute infrastructure designed specifically to support real-time rendering at scale for mobile and PC cloud gaming. By providing affordable, low-latency compute resources at scale, it can serve gamers across geographies with a seamless experience. With this foundation in place, the next step is for game studios to begin building for a cloud gaming future rendered in real-time.

Enabling AI Productization

While the underlying AI models are developing exponentially, few companies are looking at ways to productize AI for the next billion users. Our interactions with today's models are largely confined to text and voice. ChatGPT, Midjourney, among others, feel impersonal and require complex prompting and fine-tuning to extract the most valuable output. And despite having so much potential to assist us in our daily lives, AI agents are even more difficult to set up and derive value from. This isn't a roadmap for mass adoption, but it won't be like that for long.

As AI evolves, so too will the ways in which we interact with it. We'll see AI avatars—agents, assistants, and non-playing characters (NPCs)—rendered in real-time on our screens. Imagine communicating with an agent as a photorealistic avatar instead of through a chat box. It will change the way we see the technology, adding a much-needed human element and enabling more natural communication.

The issue is that this evolution of AI interactivity requires compute infrastructure that is not yet accessible outside of big tech. Aethir's real-time rendering infrastructure has the solution. By delivering low-latency, enterprise-grade compute at scale, Aethir is the foundation on which innovators can revolutionize the way we interact with AI models and enable the kind of innovation that will bring AI truly into everyday life. And once this highly interactive state is reached, the race for low-latency, real-time rendering compute will be on.

Leveraging Web3

Web3 is a core component of any DePIN. Without it, building and operating a globally distributed network of affordable compute resources that can meet the demands of AI and real-time rendering is nearly impossible. The costs of creating a centralized network, for one, would be so high that it would defeat the purpose of trying to bridge the AI wealth gap.

Web3 solutions, on the other hand, are capable of underpinning a DePIN. Blockchain maximizes the efficiency of resource consumption tracking, smart contracts reduce the need for large company overhead, and crypto streamlines the transfer of value. There are also the added benefits of transparency around cost and resource availability and an assurance that resource owners will get paid for their contributions.

More specifically, Aethir leverages Web3 for two core operations:

Incentivization

Incentives are what allow DePINs to build and scale their networks. Aside from the revenue generated from resource consumption, there can be rewards for things like availability (or liveness) and resource quality. And as demand for resources rises, it continually attracts new resource owners. Web3 makes it possible to pay these incentives, regardless of the amount, with the DePIN's own native token at a fraction of the cost of traditional payment channels.

Aethir has issued its own native ecosystem token, $ATH, to power its economy. Providers are rewarded in $ATH for proving their capacity (liveness), delivering resources, and completing workloads:

- **Proof of Capacity (PoC)** - The PoC reward is critical to the network because it rewards early contributors and deters them from taking their resources offline when there is no demand.
- **Proof of Delivery (PoD)** - The PoD reward is paid by the Aethir network. A tiered system has been implemented to ensure higher rewards are paid to owners providing the highest quality resources.
- **Service Fee** - The Service Fee is paid by the resource consumer as a price set by the network and/or resource owner.

At the same time, a slashing mechanism has been implemented to deter bad actors and safeguard consumers from supply interruptions. It works by requiring resource owners to stake $ATH prior to contributing their resources to the network. This stake is ultimately slashed in case the resource owner does something considered detrimental to the network, such as taking their resources offline in the middle of processing a request.

Consumption tracking and real-time payments

Blockchain has emerged as the perfect settlement layer for tracking and processing transactions. In Aethir's case, this means all of the resources, orders, deliveries of compute resources, service fees, and rewards. For example, each Checker node is represented by an NFT, which enables all operations associated with it to be tracked. Resource owners can leverage this data to build a transparent track record of availability and delivery that can be trusted by end users.

Blockchain and crypto have also lowered the barriers to real-time payments. Traditionally, resource owners had to accrue some minimum of revenue in order to withdraw funds. In some cases, this could take weeks or months. DePINs, on the other hand, can leverage blockchain and crypto to stream payments in near real time at virtually no cost. Aethir uses this feature to build trust with resource owners.

Use Cases

Once companies have access to scalable on-demand compute, the opportunities for innovation are limitless.

AI Model Training

The simple truth is that AI models aren't evolving without a massive scaleup in compute. In other words, we aren't getting to AGI, and we're not ensuring a fair distribution of AI outcomes, unless we can find a way to train AI models at scale. This is exactly what Aethir's cloud infrastructure is designed to do.

AI Inferencing

AI inferencing is the process through which a trained AI model can reason and draw conclusions from data it has never seen before. Every time ChatGPT is asked a question, for instance, it must do inferencing. But inferencing requires significant levels of low-latency compute to respond with the correct outcome within a reasonable amount of time. Traffic management, self-driving cars, air traffic control, and a host of other mission-critical applications require the absolute minimum of latency to be viable. It's a level of quality compute infrastructure that few companies have access to.

Aethir's distributed cloud compute infrastructure is designed to deliver low-latency compute to wherever it's needed. It can aggregate resources in specific geographic areas without needing to rely on centralized data centers far from the end user. This ultimately reduces the wait time for inferencing and enables more complex AI use cases.

Cloud Gaming

Also known as game streaming, cloud gaming is on the precipice of rapid growth. With entry-level gaming computers running at about $1,000, they are beyond the reach of hundreds of millions of gamers. Additionally, that many of the most popular mobile games require high-end mobile devices is equally exclusionary. One alternative is to rent time at an internet cafe where gaming computers are available, but costs can add up quickly over time and, in most markets, these businesses are decreasing in their availability.

Cloud gaming makes high-end gaming accessible. Game studios no longer have to worry about designing for low-end devices. They can simply abstract the resource-intensive

operations to the cloud while allowing users to play on whatever device they can afford. And provided low-latency is ensured, the experience is just like playing a game on a high-end local device. Aethir makes this possible.

Cloud Phone

With enough compute, it's possible to transform a user's mobile phone into a cloud-powered, infinitely replicable, super-powered mobile device. By downloading and installing an app, users can generate and access cloud phone instances that offload resource-heavy applications to the cloud. Advanced photo-editing software or real-time games, which typically require high-end hardware, are suddenly possible for even the lowest end of the smartphone market.

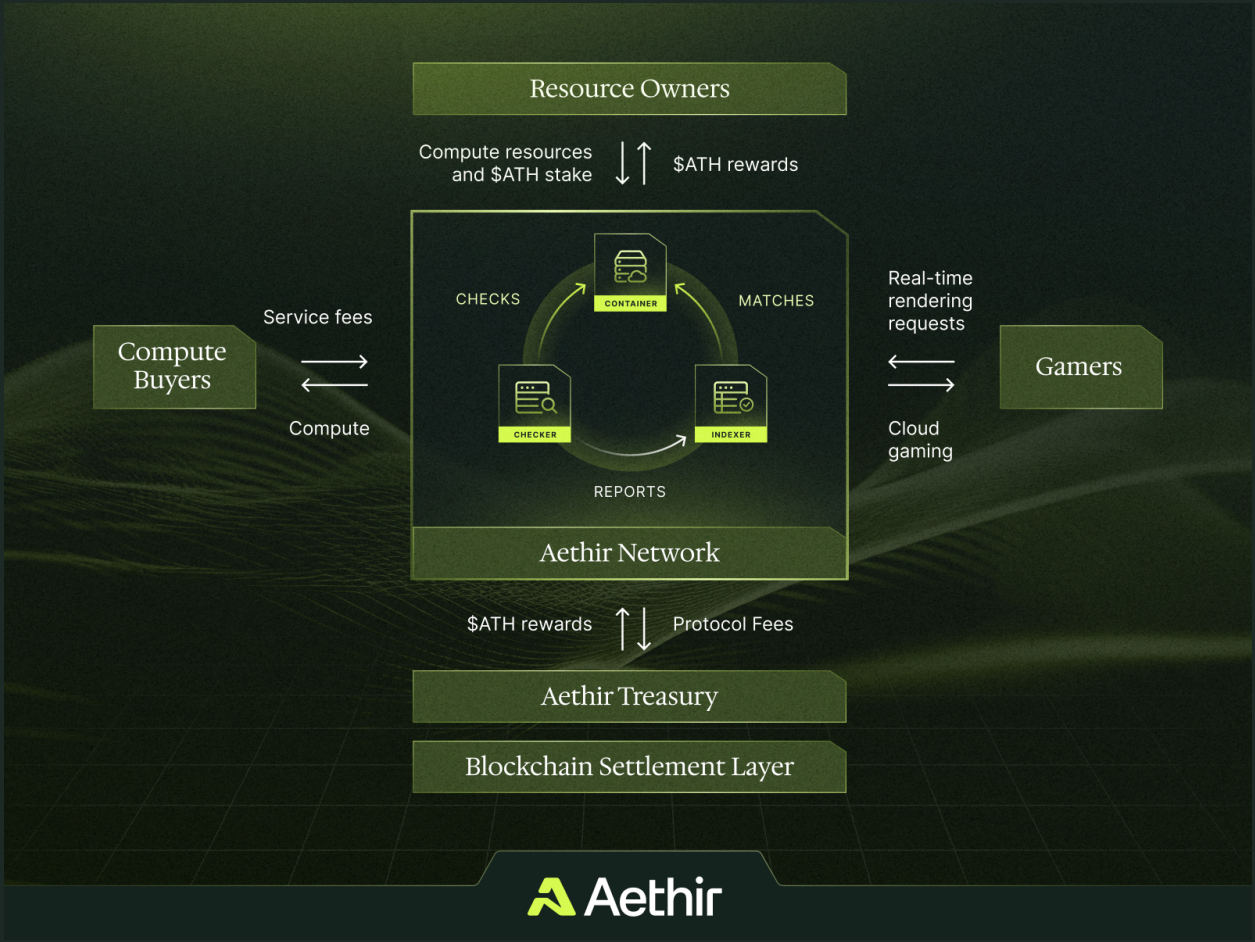
Aside from significantly enhancing device performance, battery life, and efficiency, the Cloud Phone unlocks incompatible applications that Google Play or the App Store restrict access to due to owning old or lower-end hardware. Imagine installing a cloud phone application on a low-end smartphone and, upon opening the app, you are able to use the equivalent of the latest, high-end smartphone device.

AI Productization

As AI models become smarter and more useful, the AI product market is likely to experience rapid expansion. For example, we'll see products such as remote worker agents that can be onboarded directly to your company's Slack channel and contribute to team meetings and Zoom calls. Having already been trained on your company history and processes, it's just like having a new employee with Ph.D-level knowledge.

The level of inferencing, and real-time rendering for Zoom calls, needed to reach this level of integration and performance requires *a lot* of compute. Current infrastructure is simply not designed for this type of compute requirement. Only Aethir is optimized for low-latency real-time rendering in such a way that large scale, real-time rendered AI interactions are possible.

The Technology



The Aethir stack comprises five core components:

- **Resource owners -** The GPU chip owners who provide the actual compute resources to the network for use in Containers.
- **Aethir Network -** Comprises Containers, Checkers, and Indexers which house the compute resources, provide quality assurance, and facilitate matching based on end user requirements.
- **Compute buyers and gamers -** Consumes Aethir network compute resources on demand, whether for AI training and inferencing or gaming.
- **Treasury -** Manages protocol fees and $ATH meant for protocol growth.
- **Settlement layer -** Aethir leverages blockchain technology as its settlement layer to record transactions and support scalability, efficiency, and $ATH for incentives.

Aethir Network Roles

Within the Aethir network itself, there are three roles that ensure the availability, suitability, and quality of compute resources:

Container

The Container is where the actual utilization of the compute resources takes place. It acts as a virtual endpoint for execution and/or rendering and ensures that the cloud experience is immediate and responsive—a "zero lag" experience.

Selection

The selection process differs depending on the type of compute required:

- **AI compute customers** - Containers are selected by the customer based on its performance requirements.
- **Gaming compute customers -** Containers are chosen based on their ability to provide the highest quality of service for the lowest possible cost. Capability to deliver the best possible gaming experience—factors such as frame rate, low latency and resolution—are considered.

Staking

The staking mechanism is a critical component of any DePIN because it incentivizes good behavior and penalizes bad behavior. It ensures that participants are aligned with the platform's objectives and motivated to provide optimal services. New node operators are required to stake $ATH before they can contribute resources to a Container. If a node operator is found deviating from quality control standards or disrupting the network, it’s at risk of having its stake slashed.

Rewards

Containers are rewarded in two different ways:

- **Readiness** - Containers receive compensation for maintaining a state of high readiness and for providing standby services, e.g., PoC reward.
- **Service** - Additional rewards are given for the compute resources actively used by the end user, e.g., PoD and Service Fee rewards.

Aethir Edge

The Aethir Edge is a device that functions as a standardized container node. Upon purchase, owners can use it to contribute their **consumer-grade compute resources** to the Aethir network in exchange for rewards.

In the majority of cases, consumer compute hardware is not suitable for use within enterprise-grade compute infrastructure. The range of devices is simply too broad. Aethir Edge solves this problem. Consumer-grade resource owners can contribute standardized, high-performance container nodes directly to the Aethir network, allowing Aethir to push enterprise workloads to them. For the first time, consumer compute infrastructure is being homogenized and unlocked for the demands of AI and cloud gaming.

Checker

The Checker's primary responsibility is to ensure the integrity and performance of Containers within the network. It does this by running tests at key stages of the Container's lifecycle.

Checking schedule

Checks are made at three key stages:

- **At registration** - An initial check is done when a Container applies for registration on the Aethir network to confirm specifications. Containers that pass this check are successfully registered.
- **In standby state** - Random checks are conducted on Containers in standby to ensure availability. These checks influence the container's scheduling opportunities and priority by the Indexer.
- **During delivery state** - Service data is collected and assessed to judge the actual service status. The results determine whether penalties for subpar service quality need to be levied.

Checking methods

The Checker runs its tests in two different ways:

- **Performance parameters** - Directly reading Container performance data.
- **Simulation testing** - Acting as a compute buyer to test resource consumption and analyze interactions, ensuring compliance with claimed specifications.

Proof of Capacity (Liveness)

Recognizing and rewarding the readiness of node operators is paramount. Even when there's a dormant phase without active compute contribution, Aethir recognizes and rewards the availability of node operators. This mechanism promises a baseline level of compute so that end users always get the resources they need, even during demand spikes. Proof of Capacity checks are made by Checkers every 15 minutes.

Proof of Delivery

Container performance is closely monitored to ensure quality. Checkers make sure that service requests are completed according to Aethir's quality standards. This is an essential service that results in rewards and fees being paid to the resource owner or its stake being slashed. It also has an impact on future scheduling possibilities.

Indexer

Where required, the Indexer matches end users with suitable Containers based on their requirements. For gaming use cases, the goal is to deliver an "instant play" experience so that the transition from a gamer's on-screen request to its actual delivery occurs in the shortest possible time.

Selection

When delivering gaming services, an Indexer is chosen at random for each service request to maintain decentralization. This approach helps mitigate potential fraud and minimizes signaling delays due to protocol complexity.

Matching criteria

When matching Containers with service requests, Indexers first consider the Container's status, readiness, latency, specifications, and service fee. Actual selection is then based on a combination of lowest service fee, best experience, and overall ranking in the network's evaluation index.

Conclusion

In the race to AGI and its promise of unprecedented control, we can't lose sight of the fact that a technology as important as AI needs to work for the benefit of humanity, not for that of a privileged few. This is, unfortunately, not the direction in which we're heading. The demand-driven GPU shortage, and the resulting AI wealth gap, is threatening to skew the distribution of AI outcomes heavily in favor of those that can stockpile the most chips. It presents the very real risk that the development of AI will be driven mainly by a handful of companies that already control most of our online experience.

But AI isn't the only technology affected by the GPU shortage. Half a billion gamers in Asia alone don't have access to high-end gaming content because they game on low-end devices. The very thought of cloud gaming in real-time is but a pipe dream. That would change if gaming studios could ensure that gamers had access to the low-latency compute resources needed to power cloud gaming.

What we need in this critical moment is to guarantee fair access to the one resource that will have the biggest impact on the future of AI and gaming: compute.

This is at the heart of Aethir's mission. By building a globally distributed cloud compute infrastructure network, it has positioned itself as the foundation on which AI and gaming innovation can be unleashed for the benefit of humanity.

To become a GPU provider, get access to enterprise-grade compute resources, or for more information, please visit aethir.com.