GPU 革命：Aethir 如何重塑 AI 和游戏基础设施

作者：AYLO；编译：陶朱，本站

上周，Aethir 宣布他们的 ARR（年度经常性收入）为 3600 万美元，这将使他们在今年的收入排名前 20 的加密协议之列。



现在我吸引到你的注意力了吗？很好，这篇文章值得一读。

随着人工智能快速向 AGI 发展，对计算资源的需求猛增，导致能够使用强大 GPU 芯片的人与无法使用的人之间的差距越来越大。Aethir 是一个创新的去中心化物理基础设施网络 (DePIN)，旨在实现云计算资源的民主化访问。

Aethir 的成立愿景是让按需计算更易于访问且价格更实惠，它构建了一个分布式网络，可聚合来自各种来源的企业级 GPU 芯片。该网络旨在支持人工智能、云游戏和其他计算密集型应用日益增长的需求。

在这次采访中，我与 Aethir 的联合创始人 Mark Rydon 坐下来讨论 Aethir 的独特方法。

## 鉴于最近去中心化 GPU 领域的竞争日益激烈，Aethir 与该领域的其他参与者有何不同？

这是一个非常好的问题。我将分两部分来回答。首先，我将解释我们要解决的问题，因为这是理解的关键。我相信您和您的听众都知道，全球存在计算稀缺问题。科技巨头正在争夺关键的 GPU 资源。这是一场创造越来越智能的人工智能的大规模竞赛，直到我们达到 AGI 或 ASI，然后世界上的一切都将改变。



这场竞赛的有趣之处在于，它是由一个简单的原则驱动的。如果你在生态系统中添加更多的 GPU 和数据，人工智能就会变得更聪明。这就像一条从左下角到右上角的线，随着人工智能智慧的增加。这所需的 GPU 类型至关重要。你不能在消费级 GPU 或低功耗显卡上做到这一点。人工智能竞赛中的所有大公司，无论是在训练方面还是在应用方面，都使用企业级 GPU。在过去一年半到两年的时间里，他们一直在寻找的具体型号是 Nvidia 的 H100 GPU。

关键点在于，企业公司是主要对 AI 计算有巨大需求的公司。他们正在这种计算基础设施上建立业务。因此，我们必须问他们想要什么样的计算、什么类型的 GPU，以及他们需要什么样的质量、性能和正常运行时间要求才能满足其内部指标。这就像 Netflix 需要具有高正常运行时间保证的服务器以避免服务中断一样。GPU 和任何计算提供商也是如此；它们必须满足严格的服务、质量、正常运行时间和性能要求。

不幸的是，Web3 空间中的大多数计算网络都聚合了消费者 GPU。这是构建网络的最简单方法——向人们贡献闲置游戏 GPU 的社区提供代币。这将迅速吸引大量 GPU 并建立一个对代币奖励感到兴奋的强大社区。这就是为什么当今存在的许多计算网络都聚合了消费者 GPU。

挑战在于，要想拥有真正的业务，你需要出售聚合计算。然而，消费者 GPU 网络很快就达到了低上限，因为 99.9% 的公司不想从消费者分散式网络购买计算。他们无法保证 GPU 不会在晚上关闭，或者带宽不会受到流媒体 Netflix 等家庭活动的限制。这导致企业需求与消费者 GPU 网络所能提供的内容之间存在相当大的脱节。



资料来源: Layer.gg

从第一天起，我们就决定不聚合任何消费级 GPU。连接到 Aethir 的每个 GPU 都是企业级的，通过企业网络基础设施集成，并位于适合企业工作负载的数据中心。最大的 AI 公司、电信公司和科技公司可以使用我们的网络做任何他们需要的事情，而不会牺牲性能或质量。事实上，他们获得了更高的性能和更好的整体体验。

例如，IO.net 需要克服对其庞大的消费级 GPU 网络的大量 FUD。当他们想证明他们的网络可以处理实际业务时，他们从 Aethir 租用企业 GPU。因此，IO.net 上的所有企业 GPU 均由 Aethir 提供。这是生态系统内的公共知识。

Aethir 从第一天起就致力于服务企业客户，这一点至关重要。

还有一件事：当我解释我们做什么时，这曾让 AI 领域的人们感到困惑。根据定义，分布式 GPU 意味着我们不拥有任何 GPU。我们的数据中心有大约 43,000 个 GPU，但我们不拥有任何一个 GPU。其中，我们拥有 3,000 多个 H100，这是 Web3 中迄今为止最大的 H100 集合，几乎是我们最接近的竞争对手的 10 倍。这就是为什么有这么多大型 AI 公司使用我们的基础设施，因为我们可以真正为他们提供服务。



一些 AI 公司感到困惑的一点是我们所谓的“共置机器”的重要性。如果你正在进行大规模训练，比如 OpenAI 或类似项目，并且需要 500 个或更多 H100，那么这些 GPU 必须位于同一个数据中心。你不可能在日本有一台 H100，在美国有一台，在印度有 200 台。AI 无法在分布式硬件上进行有效训练。这是其他 DePin 公司一直在努力解决的一大技术挑战，目前仍是一个未解决的问题。我认为这是一个巨大的机会，但它非常复杂。

由于 Aethir 从一开始就专注于企业，我们明白，为企业客户提供服务不仅仅意味着拥有一堆断开连接的企业级 GPU。我们还需要考虑我们网络中机器的共置。因此，Aethir 拥有大量大型共置高性能 GPU 集群。这意味着我们不仅仅是一个拥有全球企业级 GPU 的分布式网络；我们的网络拥有大型的共置集群，使我们能够为需要共置机器的公司处理那些大型人工智能工作。

## 是什么启发创建了Aethir?

所以实际上，正是云游戏网络让我们对分布式 GPU 计算感到兴奋。团队是在我住在北京大约七年的时候认识的。我搬到那里创办了我的第一家公司，并最终开始致力于扩展云游戏网络。

长话短说，我们有一个想法，即我们可以通过以分散的方式分发硬件来解决云游戏网络的性能和可扩展性挑战。从高层次来看，前提是延迟是这些网络的杀手。用户距离计算越远，由于延迟增加，用户体验就越差。这个想法是，如果你消除集中计算的动机，你就可以拥有一个更加分布式的网络。随着网络变得越来越大、越来越分散，用户更接近计算的可能性就会增加，从而减少延迟并提高性能。

中心化解决方案专注于将所有资源集中到一个位置以实现规模经济，但这并没有从用户角度增加价值。它实际上限制了网络性能。如果您要构建一个针对用户体验进行优化的网络，您会将计算分布到各处，以便用户始终靠近它。我们认为，如果我们能够解决分布和单位经济挑战，我们就可以解决阻止 Google Stadia 等服务在需要的地方部署的问题。

这就是我们的起点，我们很快意识到我们与人工智能领域的相关性，并开始在那里构建产品。



另一个值得关注的利好消息是全球游戏人口约为 33 亿。其中大多数人（约 28 亿）使用低端设备玩游戏，这意味着他们无法玩主流 AAA 游戏，同样，AAA 开发商也无法接触到这些玩家。

解锁所有这些资本的最可行解决方案是使用云游戏来消除用户的硬件要求。只需采用我们已有的技术，并使其更具成本效益即可扩展。现在，您拥有了一项可以解锁这 20 亿玩家的技术，无论他们身在何处。这正是以去中心化方式构建网络的作用。



当前的 Aethir 游戏指标

我们正在为全球数十亿游戏玩家带来与硬件分离的游戏体验，而这种方式从根本上来说不可能实现。这就是我如此看好游戏领域的原因；这是我们最初的愿景。

## 您是否认为对去中心化计算的需求会大幅增加？或者您认为我们已经到了这个地步，但客户需要更多时间来采用这项技术？

我认为，说实话，如果我们谈论的是去中心化云解决方案，或者只是谈论 Aethir，那么主要都是关于教育的。您不需要知道 Aethir 是一家去中心化云提供商，就可以与我们合作。我们与 Web2 公司合作得很愉快——90% 的客户都是 Web2 公司，他们对我们提供的服务非常满意。

纵观更广泛的人工智能生态系统，拐点在哪里？有一些疯狂的统计数据。

几周前，我读了一篇研究论文，上面说，根据对计算需求的预测增长，到 2030 年，地球上的电力将不足以满足人工智能的计算需求。这太疯狂了。

这些宏观数字表明，大量资本被部署到生态系统中。这些数字大得几乎让人无法理解。但如果你稍微放大一点，就会发现有两种类型的计算需求：训练和推理。训练是让人工智能变得更加智能的过程，例如从 ChatGPT-4 升级到 ChatGPT-5。推理是人工智能完成其工作的过程，例如回答问题。

像你我这样的人大多使用 ChatGPT 或其他大型模型，对吧？例如通过微软生态系统的 ChatGPT 或通过谷歌的 Gemini。我们的大多数互动都是与来自极少数公司的通用大型语言模型进行的。但如果我们展望未来一年，考虑到该行业的指数级增长，我猜你会在比现在更多的地方与人工智能互动。

很快我们将以更有意义、更具代理性的方式与人工智能互动。人工智能将为我们做更多的事情，例如预订机票、提供援助服务和处理客户服务电话。它将比现在多得多。

如果你关注计算上游，除非某家公司只是使用 ChatGPT API 来创建应用程序，否则他们很可能正在构建自己的 AI 产品，这意味着他们有自己的计算基础设施需求。因此，随着推理空间的增长，它将变得更加分散。目前，大多数用于训练的基础设施都来自少数大公司。虽然有些公司正在开发这些大型语言模型的新竞争对手，但推理方面的 AI 应用程序的爆炸式增长将导致计算市场更加分散。

这意味着我们将看到来自具有竞争力的价格和对初创企业和小型企业友好合同的渠道的大量需求。这可能是我看到的最迫在眉睫的转折点。

## Aethir 生态系统中的另一款产品是 A-phone。您能否向我们详细介绍一下这款产品以及其目标受众是谁？



A-Phone 直接在我们的基础设施上构建和扩展。它使用我们的云游戏技术以低延迟的方式将实时渲染传输到设备。这非常酷，因为它完全与访问有关。例如，您可以拥有一部 150 美元的智能手机，下载 Aethir 应用程序，然后打开它以访问相当于 1,500 美元的设备。本地设备的所有硬件限制都消失了，因为您拥有支持该应用程序的云功能。

您可以在 Aethir 手机上打开几乎无限的应用程序，而且不会耗尽您的电池。所有计算、处理和存储都在云端完成，基本上为您提供了一部超级手机，您可以随时调用它来运行所需的任何应用程序。

无论是游戏还是带有视频会议的教育平台，它都非常酷。它消除了硬件对人们访问内容、工具或实用程序的障碍，尤其是对于占互联网用户绝大多数的移动用户而言。

## 您认为迄今为止 Aethir 成功的关键是什么：您的技术解决方案，还是您的业务发展努力？

我认为，作为一家公司，我们重点关注两个方面。首先是我之前提到的企业元素。这意味着在早期做出一些非常艰难的决定。正如我所说，聚合消费级 GPU 要容易得多。聚合这些我们已经知道很难找到和访问的企业级 GPU 要困难得多。我们早期走的是一条更艰难的道路，这使我们在运营初期面临风险。但正因为如此，我们做了艰苦的工作，现在情况更好了。没有多少公司有决心在早期做如此冒险的事情，这对我们来说意义重大。

其次，我们一直非常专注于实际业务——实际利用率、实际合同、实际收入。从一开始，这种关注对我们来说就非常重要。这就是我们选择企业路径的原因。我们希望充分利用 Web3 技术，提供改变行业的解决方案，不仅仅是 Web3 解决方案，而是 AI 和游戏领域一流的行业解决方案。

我们的业务开发团队在说服合作伙伴加入我们的生态系统方面发挥了至关重要的作用，尤其是在早期。在技术方面，我们使连接计算资源的过程变得无缝。目前，有比我们所能接受的更多的供应希望进入我们的生态系统。在未来，我们的目标是成为一个真正无需许可、完全去中心化的生态系统，我们会实现这一目标。但一开始，我们必须务实。打开计算的闸门，让大量的 GPU 耗尽你的代币，这不是一个好的商业举措。

我们将自己视为一个供应主导的组织。我们总是试图让供应多于需求。我们不想拒绝需求，但我们也不希望供需之间出现巨大的差距。我们希望明智而稳步地发展供需关系。我们不只是为了吹嘘我们的数字而加入无限量的 GPU；这不是正确的做法。

我们计划在未来几周内发布一些重大公告，以表明我们对透明度的承诺。这对人们来说真的很有趣，并且表明 Aethir 是一家人们想要参与的公司。

## 您能向我们介绍一下 Aethir 代币吗？它如何融入生态系统以及如何产生价值？



这实际上是一个更大的版本的主题，你们很快就会看到。我现在不能再谈论这个话题了，但我可以说的是，由于需要处理代币，很多项目以前很难与大型 Web2 实体合作。

这是该领域的一个持续挑战，我们有一个非常令人兴奋和新颖的解决方案。我认为人们看到它时会非常乐观。因此，它将使我们能够为代币带来大量交易量。

我们最大的客户是 Web2 客户，我认为这不会改变。我们需要确保我们从事这项业务，并允许该价值累积到 Aethir 代币及其支持的生态系统中。这是我们的承诺，我认为下周您将看到一些关于如何实现这一点的非常有趣的东西。