


27June 2024:

In our recent community call, we discussed updates on Test Grid, announced new partnerships with Matrix One, Aurory, Alpha Neural AI, and Aurori. Community news included job openings at Nosana and highlighted upcoming events like the EthCC event, Encode Hacker House, and World AI Week in Amsterdam. Key action items include clarifying inferences run on Testgrid, onboarding new nodes based on customer needs, launching benchmark jobs, developing AI structures for gaming, and integrating Nosana into AI marketplaces. Our next community call is scheduled for July 25.

## Transcript:

**Rochelle:**
Hey, everybody. Welcome to our monthly community call. I'm, of course, Rochelle, if any of you are not familiar with me. And today we have Sean and Jesse joining in on the call with us. We are so excited that you all could attend and spend some time with us today at Nosana. So on the agenda for today's community call, we're going to give some Test Grid updates, and then I'm going to give out some community news. And we're going to talk about partnerships. A lot of partnerships have been announced in the past month. We have a whole bunch to get into with that because I know if any of you have missed some of the AMAs, you want to know, '*What do they mean?' 'How does it apply to the Grid?*' So we will cover all that soon.

So the first thing I think we'll get to talking about is Test Grid updates, both technical and functional. Sean, would you like to pop up and give some Test Grid updates?

**Sean:**
I'm Sean, for anyone who doesn't know me, and I see some familiar faces here this afternoon. Jesse is also here with us today and we're going to talk about a few interesting things. I guess I will kick off with a bit of an update on Test Grid. Jesse's going to go into the partnerships in a little while after that. And Rochelle is going to give us an update on community fun things. So let's go!

So I'm going to give a brief update on a few things that have been going on at Test Grid. And I'm going to attempt to fill in for Laurens, who unfortunately could not be here today. So bear with me. I'm not the technical wizard of the company, but I'll get by.

But before I get into this, I want to clarify a comment that I made in last month's community call. And I think it's important that we get this out of the way right now. I'd stated in that call that we had run half a million inferences in our Test Grid to date, and I need to apologize to the entire audience and community as the number that I quoted is completely incorrect. Now, that half a million number accurately reflects the *inference* jobs that have been posted and run on the Nosana network. So that's the total number of jobs posted to our nodes since we started the Test Grid participation back in December 2023. And that number has gone up in the past four weeks. So we're now over 600,000 jobs run. But my statement that this is inferences run is just totally incorrect as each job that we post for our nodes to run contains multiple inferences to be executed.




And in a lot of cases, there are ten and more inferences loaded per job. So what's the total number of inferences that we've run so far? Well, to be completely honest with you, we don't know. It's going to take a lot of number crunching and we're probably going to have to use some of your compute that you provide to figure this out. But for a more accurate guesstimate, and take it with a grain of salt—the number of inference jobs that we've run, take that number and multiply that by about a factor of, let's say, five—just to be on the safe side. Okay? So let's say we've run at least 3 million inferences. So anyway, I apologize for grossly underestimating the total inferences run. I hope you will all forgive me and I'll try to do a little bit better in the future.

Now for Test Grid: things have been running really smoothly. There are bumps along the way and we have individuals and individual nodes that encounter issues. But for 95% of them, they're running. They don't need to touch anything. They don't need to do anything, of course, until we ask them to restart their node. And then there are sometimes issues. We reopened the wait list for newly registered nodes a couple of weeks ago. We've got more than 1000 new nodes signed up on the wait list already. So we're going through those and we'll begin to onboard some of those nodes into *select markets* soon. I'm stressing the select markets comment, as we will not just throw open the doors to everyone and then everyone comes flying in because our focus shifts a little bit now onto customer partner needs.

So what GPUs do they need? (Instead of just padding our GPU numbers!) That's not something that we do from a markets perspective. We're beginning to fine-tune our markets a little bit. So what I mean here is that we've got some markets where there is a queue of nodes waiting for work. And this work is 100% test jobs. Other markets have much less of a queue or no queue at all. And it becomes important that when we start introducing customer jobs, we have nodes always available for real work so to speak. And that doesn't help us when we have all the nodes in a market constantly busy running test jobs. So we're going to accomplish this in two ways.

Obviously, we bring more nodes into the mix, especially in the markets where we need them, and we ensure a good balance between our test jobs and real customer jobs. From the technical side, I don't have a 30-minute slot to go through all of the new and improved goodness that Laurens and his team have cooked up. So I'm just going to touch on a few of the major points for those in our Test Grid. You will be familiar with our benchmarking jobs. These are heavy jobs that all nodes are running at the moment, and we've been running them, geez, I don't know, for four, five, or six weeks. And these test different LLMs and GPUs. And running these jobs gives us quite a lot of valuable information.

So it gives us insights into which GPUs can handle specific AI models the best, the performance of those GPUs and the performance of the different GPU/CPU combinations. Keeping in mind that running AI inferences, CPU is also important, we can also use these benchmarks to identify and catch those who wish to cheat the system. So we can see this now. Yesterday, we pushed a new benchmark job for a large model job for the big boy GPUs. So we rolled this out to the H100s and the A100s yesterday, and we will gradually introduce these a little bit further down the chain over the coming days. And the nicest part of this is that it is all going to be made visible to you on a node leaderboard website, which is coming very soon. So you will be able to see the performance of nodes with their GPUs and with their CPUs.

It's show-and-tell time, right? So let's go. We've also added a CUDA check at node startup to ensure that, for




each specific node, it's actually capable of running AI jobs. And again, the CPU remains important here, right? So, as we have seen, at the start of Phase 2, we had CPUs that were in the Test Grid that were simply not capable of running AI inference jobs. So this check will make sure that doesn't happen. There's also been an awful lot of improvements to the CLI for job posting and retrieval, which is really important for our customers and developers so that it is really easy for them to post jobs for our nodes to run.

We've implemented better error handling, calculating the required NOS for the job, allowing users to post jobs with the default Solana RPC showing exposed web service URLs and a number of other highly sought-after features from the dev and customer sides. That exposed web service URL is probably the most important change that we've put in, which is that we now allow our customers to post jobs that expose a port that is running in a container on a node through a fast reverse proxy. Okay, so that's the technical explanation. Think of it as a session capability, or, in real layman's terms, a pipe between the customer and the node. Your node becomes the customer's GPU to use. So now a customer can open a session with a node or a number of nodes and have exclusive access to those nodes.

So they will be able to keep that session open for whatever their desired time period is, right? So however much they pay for, they need that node for 4 hours. They get it for 4 hours. And so they can, as an example, deploy a specific LLM model to your node and expose it through an API endpoint or a website. And while that session remains open, the customer is paying and node gets paid. So that kind of thing is a wrap for me. And that last point ties directly into our next update. So I'm going to hand this off to Jesse to give you all the juice on our newly announced partnerships. All yours, Jesse.

**Jesse:**
Thanks Sean, for the updates. Hey everyone. I'm happy to be here again. A month later, time flies, but an amazing amount of updates have happened, as Sean just explained. And this has enabled us to run some amazing use cases on the Nosana Network. This month, we have announced the first client partnerships for partner projects that will be running workloads on the Nosana Network. And these specific clients all have very interesting use cases that are pretty diverse, and it's an excellent opportunity for us to demonstrate what you can do on Nosana and to give everyone a better idea, but also to put some expectations on the workloads that will be starting to run in the coming months. So, yeah, for now, I will take you guys through some of these clients that we have announced.

And I would go into the general idea behind the partnership, but also concretely what it means and what kind of workloads we will be able to run with these projects. And of course, we will tie this into what Sean just mentioned. I think one of the most important features, maybe a bit understated, is this endpoint, the peer-to-peer endpoint that is now available within the nodes. This is an extremely important capability of our network that basically allows us to use the Nosana Network anywhere where people need their GPUs. So I will start with one of the first announced partners, which is Matrix One. And to make it more concrete, their main product is called Avatar One. This is a virtual girlfriend platform. You guys might have heard of it.

And it basically means that at Avatar One, anyone can create their own virtual girlfriend and use it to have an interaction with her. So once you have spun up your virtual girlfriend, you basically get a chat experience with a beautiful cover. Right? There is this 3D girlfriend character that is able to interact with you. And the technology behind that is obviously LLMs. So with the advent of LLMs and the personalization and depth that these




models provide, these platforms are becoming very popular, as you can have a pretty realistic chat experience. Therefore, Nosana is offering Avatar One, a platform where they can quickly spin up various LLM models for their virtual girlfriend servers to interact with.

And as LLMs are developing so quickly and there's a lot of innovation happening, a platform like Nosana is extremely beneficial here because we're able to use different models and the newest setups on specific GPU cards to have a really scalable LLM endpoint that they can use for their product.

Another project that we work with a bit in the same category is Aurori, which is a Web3 gaming world and AI and gaming just go very well together for multiple reasons. Not only because AI and gamers both use GPU hardware to be able to run and use their tools, but also because AI in games is just a very important element of having realistic interactions in the gaming world. NPCs, or non-player characters, are a very important aspect of gaming. And LLMs are also really important here.

So there's many different layers in which we can collaborate with Aurori and other gaming companies. But the main one that we're developing now is to have conversations with NPCs in a specific Aurori gaming world that will be more realistic. So, to some extent, Nosana LLMs will power the NPC conversations. And we're also looking ahead, as the gaming world is more than just conversations; chat conversations with virtual characters. But later on, we're going to collaborate with Aurora to develop more in-depth AI structures for their gaming world that integrate more deeply into the environment. So the specific workloads we're talking about are again focused on LLMs.

So using the current Nosana network, we're able to spin up a diverse set of LLM endpoints that scale nicely and that can support a diverse set of LLM models, which would be perfect for them to interact with and integrate these AI elements into their gaming world.

Another very interesting project that's going to be launched soon on Nosana is Alpha Neural AI. So they're a new project. I think it's a very cool use case for the people with some AI background here who are listening. I normally compare them with Hugging Face, but it's a “decentralized Hugging Face," so it’s a very popular AI marketplace where you can look for and explore AI models and datasets. You can basically think of it as an app store where you can see all the AI models that are being used.

You can see the inputs and outputs, as well as some instructions on how to use that model. And like what Hugging Face provides, which is quite popular, you're able to try out those AI models in their user interface directly from the UI, which is a very powerful tool. So what Alpha Neural is doing is basically the same, but they will be using a decentralized compute provider, primarily Nosana, to launch their AI services from their marketplace. So Nosana will be where we're integrating Nosana as one of the compute providers. If people look up models, they will be able to launch them directly from Alpha Neural on the Nosana GPU grid.

And it is an excellent use case because I think a lot of end users of artificial intelligence will be flocking towards these marketplaces, which means that it allows us to get exposure to the users that really need access to the GPU computes to run those things. So it will really allow us to get noticed and get used by the AI users that can benefit from our product. And because it's a marketplace, meaning that people can post their own models,




and all of these models can be run on Nosana, it means that potentially Alpha Neural can provide a really diverse set of models running on Nosana.

We're basically bridging the model definition that exists within their platform to a run GPU configuration on the Nosana Network, which means that it will enable a whole lot of models to be run on our platform, which is of course very exciting. And initially, we'll probably focus on the really popular ones, such as LLaMA and the other LLMs, and maybe some image generation models that will be able to be spun up from their environment on our grid.

Another use case, the last one I will cover, is Arbius. This is something different, because they are like, I think they call themselves a peer-to-peer machine learning network, but they are their own network, where nodes actually mint or mine Arbius tokens by running artificial intelligence models. And to do that, of course, you will need GPU power.

So what we're doing together with Arbius is making Nosana a really convenient way for Arbius' nodes to spin up a miner that will be powered by Nosana GPUs and they will be targeting the A100 GPUs. So I think the main use cases from this project will be focused on the A100 market, but it's definitely possible that this will diversify to other markets as well. So I think these partnerships should give us a really good idea of what's happening on the network and what kind of workloads we will see in the coming months. Most of them will revolve around LLMs, mainly LLaMA for the time being. But even within that category, there's such a diversity of models. So we will be running a lot of smaller LLaMA models but also a lot of the 70B models.

There are rumors of the 400B+ LLaMA model coming out as well at the end of the summer, so we're excited to get our hands on that and run that in different locations as well. But the team, our AI team, has been really hard at work creating a lot of run configurations for different types of LLMs. So once we are done with testing these partners, we will have such a wide variety and easy interface to all these models that anyone can easily run their use our network and run things on the GPU grid. I'll leave it at this and hand it over to Rochelle some community news.

**Rochelle:**
So first thing, if you haven't been paying attention to our Twitter or our Discord and missed this, we are hiring! So we are looking for a Node.js developer and a Web3 front-end developer. So if you think you have those qualifications, please go to nosana.io/team and there will be a link there to view our openings, open positions, and opportunities. And if you do not have any of those skills, keep checking back regularly for other opportunities to apply to join the Nosana team because we are not stopping at these two positions. We are growing. Yes, apply!

Big announcement! A big shout-out from not only me but also every other single member of the entire Nosana team. A big shout-out to Luvpawgs, our moderator here in our community. He is amazing! We are so thankful for you! Not only are you an outstanding member of our community as part of our moderation team, but this man also supplied the entire Nosana team and some other community members with hats and swag and it's amazing! So we have the most epic swag sent to us all the way over in the Netherlands. We can't thank you enough for your contributions! We love it and you know what you do every day in the server. You are amazing




and we would not have such a good community without you! So thank you so much! We are so happy to have you as part of our family. So everyone, wherever you're at on your devices, give him a round of applause!

And equally, we are so glad for our community members! You're such a great community and a very strong community, and you inspire us to keep pushing forward every day with all the hard work we're doing in the back end, out with events, partnerships, and things. You guys all drive us forward! And so we want to extend this to you; if you want to get more involved and have some Nosana-related meetups in your local area, we would love to help you out with that! So we would encourage all of you to think about if that's something you want to do, step up, give me a shout-out, send me a direct message here in Discord, and let me know you're interested, because we would love to see you all share with everyone about our project and our product and what we're doing and our community as well, because you are just all rock stars.

And then, speaking of in-person events, the team is going to be at an EthCC event at the Encode Hacker House, and Sjoerd will also be speaking on the DePIN panel. So check that out! And if you're around, come meet us! We want to connect with more of you in person. So we're really excited for that. That's this July. So, yes, check that out.

And we are preparing for a very special Nosana event the second week of October in Amsterdam. It's going to take place during World AI Week, and we're going to share some details soon. But in the meantime, mark this on your calendar if you would like to join us and would love to go to beautiful Amsterdam in October for that event.

So I will :get to some questions. The other day, I asked you, lovely community members, if you had any questions, to pop them up here, and we would answer them if we had time. So let's see, we had a question about the estimated timeframe for when Phase 2 of Test Grid will start. And I’m pretty sure it was covered by Sean at the beginning of the call.

And as always, do we know which models are in the highest demand from Nosana's new partners? I think that was asked by MachoDrone, and I think Jesse covered it pretty well when he talked about the partnerships. Is Nosana also going to be at Solana breakpoint, and will Nosana and Jesse speak at the Solana BreakPoint conference this fall? That's a nice question.

**Jesse:**
We will be going to Singapore BreakPoint with part of the team. So details are still to be announced, but yes, we will be there. But the speaking slots—I'm actually not sure yet. I think we applied. I don't think the speaker grid schedule is out yet, so it's possible we will be speaking there, but nothing has been confirmed.

**Rochelle:**
Let's see… More questions were asked about future listings. We cannot and will not talk about future listings. As Sean puts it, “the number one rule of Fight Club is we don't talk about Fight Club.” There's always one twhoasks that every time, either seriously or jokingly. And we had a very interesting one: Is the new ZK compression from Solana something that can bring even better value to Nosana, making job fees even cheaper? Anyone from my team want to answer that one?




**Jesse:**
Great question! Yeah, it's a very cool technology. It's very new and it could potentially make things a lot cheaper, especially since it might allow us to store more on-chain or so. Yeah, there's a lot of potential there. I don't have the answers right now, but this is a very interesting technology and something we're discussing with the dev team on how it can be used within our current network. But cheaper storage costs would be an amazing addition and very useful for Nosana.

**Rochelle:**
Thank you. And then Rides would like to know if the roadmap is up-to-date and accurate and what's a realistic timeline for the mainnet launch.

**Sean:**
I can try to pick that one up. It's a very good question. The accurate answer is that we don't know today. We've actually got a big discussion on this next week. The main reason is that we don't know exactly what form Test Grid Phase 3 will take at the moment because we are covering off a lot of things in this phase that were planned in Phase 3. So we need to sit down as a group to look at this a little closer to figure out where we're at. So I can't answer the question of whether it's accurate or inaccurate today. End of story. Maybe Jesse wants to add lib a little bit there?

**Jesse:**
No, well said. Yeah, I don't think there's a good, definite answer for this part. If the roadmap is up-to-date onthe website, I'm not sure. We should maybe check. But yeah, the timelines are not set in stone right now. Test Grid Phase 2 is indeed encompassing more than initially planned. Mainnet will start when the network's finished and everything is working well. So what we're going to see in the next few months, in Test Grid Phase 2, will be a very big indicator of how far we are and will help us clarify how long it will take to add the missing pieces.

**Rochelle:**
Another question we have time to ask is: what is the protocol if there's another Solana outage?

**Jesse:**
I can answer this one. So at the moment, yeah, if there is a Solana outage, it means Noana is also down. We won't be processing any jobs if Solana is congested or down. Part of this is because we've committed to being a really decentralized network. So our job distribution, jobs being picked up and the GPU marketplace aspect are being run on-chain, unlike some other networks. We really prioritize the fact that we are decentralized and permissionless in that aspect, which means that the marketplace doesn't work if the blockchain doesn't work. And I think the protocol is that we're now really alert when this happens and we can clearly communicate it to both our clients and the nodes.

**29:54**
Jesse
We are in a test grid phase, which means that it can be expected that these outages will happen on our infrastructure blockchain side. So we will clearly communicate this across the board. And yeah, these things can happen; they might happen again, but we are also betting that Solana will get more mature and will get




more and more stable over time, which means that we should be able to rely on it, but that's our current protocol. So I hope that answers the question.

**Rochelle:**
Thank you. Yes, I'm pretty sure that answers the question there. Thank you all for joining us in on this meeting, our monthly community call. I'm loving these. I hope you guys are too. All of us here at Nosana really enjoy these and connecting with you more. I record these calls and they are available a little bit later. So you can look forward to seeing it posted. If you go to our docs page, there's a community corner, and we are starting to list all these community calls in there. Our previous one is in there if you didn't get a chance to listen in on that one. And this one will be in there too.

And I invite you all to mark your calendars, for Thursday, July 25th, at the same time and place. We'll be meeting again. So thank you all. Have a good afternoon, evening, or morning, wherever you are, and we'll speak soon. Thank you, everyone.