The First Dedicated YC GPU Cluster - With Together AI

15 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=TOHKV307PM0&FEATURE=YOUTU.BE
https://www.youtube.com/watch?v=tOhkv307Pm0&feature=youtu.be

SUMMARY

Together AAI, co-founded by Bippol, has partnered with Y Combinator (YC) to launch the first dedicated YCGPU cluster, aimed at providing AI native startups with essential computing resources. This initiative is designed to support a diverse range of companies in the YC portfolio, enabling them to efficiently build, scale, and innovate in the generative AI space.

- *Together AAI focuses on the entire lifecycle of generative AI, from model building to serving at scale.*
- *The partnership with YC allows startups to access GPU clusters without long-term commitments, enhancing flexibility and cost efficiency.*
- *The landscape of AI startups is evolving, with many industries now integrating software-first approaches, creating new challenges and opportunities.*
- *Access to compute resources is increasingly critical for startups, especially as demand for capacity grows.*
- *Together AAI emphasizes the importance of research and best practices in AI systems to optimize model training and deployment.*
- *Startups can plan their compute needs in advance, allowing for better financial management and scaling.*
- *The collaboration aims to democratize access to advanced computing resources, leveling the playing field for early-stage companies.*
- *Feedback from startups indicates high satisfaction with the support and flexibility provided by the YCGPU cluster.*

Hey everyone, I'm here with Bippol, CEO and co-founder of Together AAI and we've got some exciting news. YC and Together are partnering to bring online the first dedicated YCGPU cluster. This gives our portfolio of AI native startups easier access to the compute they need to build and scale. Vipple, thanks for being here. >> So great to be here and get Vipple, why don't you start by telling us a bit about Together AI. What exactly do you guys do? Yeah. So, you know, we started the company almost exactly four years ago and as as you know, my co-founders are professors and academics and we were quite convinced at the time that generative AI was going to be a very foundational technology and I think we also had a sense that because of the capital structure acquired we may see large concentration and and and and we really wanted to create a platform for innovation for you know the sort of a participatory side of together. So what we've built at together AI is really a cloud service that's designed for the whole life cycle of generative AI which is everything from building models to post-training open models to serving them at scale.

>> Nice. And could you tell us a little bit about how companies actually work with you? You know YC we are we acquired the first GPU cluster for YC itself to share to our companies. Tell us a little bit about the types of

companies that have worked with together in the past. Yeah, you know, we our our focus is really on AI native startups and we have over 8,000 customers today and this range from, you know, small research groups who are sort of still figuring out what they want to do and experimenting and this just such an empirical field that we we enable them with access to compute or access to infant services all the way to some of the biggest names like cursor and cognition and 11 labs and these companies companies are training models.

They are and and we're serving their model at scale for them. >> Yeah. One of the really incredible things from my perspective over the last year has been all the different types of what we would call AI native startups that have been built, right? You have companies that are training their own models from scratch where they're now able to use this GPU cluster to get access to compute, right? Especially in this environment. You have AI native services companies that are often building on top of models and increasingly building on top of open source models hosted at places like together. And then you have kind of everything in between from data providers to application layer companies. So there's this whole landscape of companies in the YC portfolio and they all have kind of very heterogeneous compute needs and so I think for us having a partner where there's different ways to work with you guys has been a big win for a lot of companies.

>> Yeah. And and I I I think that flexibility is particularly important because you you're trying to be efficient with your capital and and being able to deploy it behind the workloads you want and you want really rapid sort of iteration. I think we have sort of really helped create a platform that allows companies to you know to be able to do all these things efficiently. building a company in today's native world is it's quite different I imagine from the days you started your company >> what are you seeing what what do you think are the main challenges that you know YC startups are experiencing >> yeah I think there's a bunch of different ones all at once it's a very exciting time at YC I think one of the the really cool things is way more things look like software companies now than ever before like when I ran my company we were a pharma company this was back in 2018 At the time saying that you were a software first pharma company was sort of an absurd concept. now I think it is actually very common to think of >> industries in which things you don't necessarily think of as software companies whether it's insurance or healthcare or pharma or whatever else are software companies for the first time. which expands just like the the pie of things that people might work on in Y cominator. at the same time it means there's whole new challenges right there's whole new industries in which people are facing bottlenecks that they didn't face before. So on compute right one of the things that's very different you know eight years ago when I was running my company we could just like go to AWS and get a thousand you know instances with GPUs and yeah they were like maybe expensive or whatever but it wasn't that hard to get them like you could get them and you can get them on spot instances and we don't have to do long reservations and whatnot. I think now what we see is as people's compute needs needs go up, actually just having access to capacity is a really big problem, let alone great pricing. And I think, you know, for us with this partnership, being able to secure capacity, secure it at a really great price, have great support also available and make it so startups don't have to commit to two years, they only have to commit to a few weeks, right?

>> Makes it so that for startups that have these heterogeneous computs, Yeah. >> you can kind of solve for that. and so it kind of solves for this bottleneck that can help them access all of these weird new markets that are now for the first time software problems >> and and and do you see this becoming a differentiator from the perspective of the fund it it's you know it is becoming harder and harder and I I think I I feel as models are

getting better right the the value of tokens that are being produced is is increasing which is sort of has an impact on the value of flops and and the cost of compute and you may have this curve of you know ever expanding cost of compute.

So so finding ways of enabling companies seems important. >> Absolutely. Especially because you know YC has become probably the biggest seed funer of research style companies. You know probably very first real machine learning seed investment was literally open AI. >> it was open AAI was built as YC research when Sam was president here. and then since then we've funded all sorts of companies developing their own models in a bunch of different verticals. There's companies like Deepgram Graham making voice AI models.

There's companies that are then distributing those and building their own companies like Vappy for example or retail ory companies. there's a new one called Miso Labs just launched a few weeks ago that has an incredible new emotive waste model. in other domains like in language you have companies building new edge device models or different variants of model architectures that aren't the ones that are currently already being scaled up. And then you have this whole long chain of other modalities in which people are doing research whether it's in biology or in healthcare or kind of you name it. And so I think what's really exciting right now and actually didn't even say anything about hard techch where there's actually tons of interesting robotics companies that we're we're funding these days. And with a lot of these they are research bets. They are not companies where we expect them to commercialize in significant ways immediately. They're ones where they have to solve a hard technical problem and access to compute ends up being a big bottleneck for a lot of them. And so yeah, we we very much see securing access to compute as a really important thing to attract the best founders to do and then also to enable them to succeed both during the program and afterwards. like what's great is companies can get on this cluster or you know our future ones basically during the batch and then continue to grow with whatever their compute provider is for a very long time.

>> Yeah. Yeah. And we we are definitely seeing that it's been pretty exciting to you know track the journey of some of these companies that that we've onboarded. What was it about together that led you to do this part partnership with us? >> YC's next batch is now taking applications. Got a startup in you? Apply at y combinator.com/apply. It's never too early and filling out the app will level up your idea. Okay, back to the video.

>> There was a few things. one is we really prioritize speed. I think being able to work with a partner who is excited to get a cluster up really fast and then also build it in a way that works with startups with lots of different needs. You know, we have some companies who are just looking for a single node and they just want access now, but they're in the scaling mode where you could totally see them becoming a lot bigger. There's others who are buying 256 GPUs on day one and looking to scale to several thousand very soon.

>> And being able to offer a product that actually supports all of those use cases was a big ask. And then the second was having a great engineering and research team that could help actually enable the best practices in these companies. You know, a lot of the founders, they have really great engineering backgrounds. They have trained models in other places where there's been managed compute available to them. They haven't necessarily

had to manage their own comput cluster. And so having a partner that can help with just showing the ropes so that yeah, of course there's a lot that's managed back together directly. There's some parts that the pro that the founder themselves need to figure out, but they can make it so it's not a zero to one right away, right? And there's a lot of best practices to encode. I think that was a big part of it. And then I will say like I just lastly personally I've actually followed a lot of the research from together for many years. Like I'm friends with Z and I've you know seen some of the work that they've put out in research papers in the past. And so actually maybe something be curious if you want to share a little bit more about like how do you guys think about >> all of your research efforts you know work like flash attention for example.

Yeah. >> and all the other work you guys do and how that fits into your product offering. Like how do you balance those two things and how do companies benefit from that? when we founded together, we felt that there was going to be quite a bit of research and effort that's going to go into making models better. you know reducing loss and building better and better models and there was this entire space of production AI which it's a science and it was underststudied and it's really kind of you know linear algebra executing in these novel accelerators.

It's a new form of computer science and we felt it was going to be very important because this is you know a a very very scaled endeavor and you are deploying large amounts of capital the the cost of training and the cost of producing tokens is very high and you want to make it really efficient. So a lot of her research has focused on AI systems from you know approaches to doing attention and that continues to be a really sort of rich field of as as now you know attention is going after longer and longer context to you know architectures. We also did work on the mamba architecture which is now making its way into you know long context attention like in the neotron. I >> was going to say that the latest neatron models also seem to use it right.

>> Yeah. and compilers and so so this this is a very interesting and productive area of research for us and for the companies that work with us because we can accelerate their workloads make them can generally achieve the best unit economics possible on these workloads and and that's really having it's it's very useful to early startups but then also the scaled ones who care a lot about the you know economics of their applications. So, >> yeah, I was going to ask you on on that latter point like you know when you think about a lot of the research companies you've worked with, you've seen a lot of them scale on together platform. What should early stage companies who are starting to use the together cluster or a different comput provider think about now to prepare for what's coming both in terms of the compute shortage and how to sort of set themselves up for success as their company starts to work and get product market fit. Yeah, you know, there's so much infrastructure now, right, relative to even a couple of years ago from really frontier class opens models are are accessible. There's so much machinery and tooling for for you know training models, post- trainining them. There's rich source of data sets. So I think you can be incredibly efficient in in in terms of kind of using this machinery and deploying it and I I you know we do see many of the VIC startups we're working with are very proficient with the the current set of tools that exist. I think the the compute planning is definitely something that it's complex, right? You you when you are a seedstage company, you have a certain amount of capital and your sort of perspective of what you're going to do with it is very different from you know once you get to a next level of scale and so I I think part of what we are trying to do is build products and build price and package them in ways that allow companies to be able to scale up to the to the next stage and and I think some of the collaborations that we are doing are going to

be sort of essential in you know helping companies underwrite and I think this is where companies are very different from software companies because you you really have to think about this is going to be the biggest expense in some ways for development and and product.

>> Yeah. Yeah, I mean I think a big challenge we we see and I think we saw it before having what really motivated setting up this cluster was a lot of companies just to be like just to put into numbers you know the upfront they would need to pay in order to secure the capacity for their next two years of compute was greater than their current cash balance and so you just could not do that unless you go and raise and basically put all of your money into a compute partner >> and so a big thing we look to solve with this is you know use some of YC's scale so that we can secure access to that compute so a a an individual startup doesn't have to make a 24-month bet.

They can make a few month bet instead, which is much more doable with access to capital dollars in the way they currently flow and so they can kind of continuously scale and not have to lock up all their cash basically and make it so you know they raise at more effective terms for themselves. >> Yeah. Yeah. Absolutely. What what is the kind of feedback u that you're hearing from from these companies? >> Oh, it's been great. I mean, I think well what's what's very cool is you know there's lots of different companies who have different needs. you know some with training clusters versus inference clusters. I think they've really loved working with your guys' support team to build the right kind of workflows to work really well and understand the kind of realistic underlying sort of compute setups to actually have really good performance and availability.

and then we've made it so that they can sort of book in advance and now companies can plan out their compute several months in advance which has been very helpful for them where some of them might know that they need to do a big training run 3 months from now and they can choose to buy a relatively small amount of compute for a while and then a much larger amount and that flexibility is great for them and you know across the portfolio of companies there's still enough to service all the demand at all times. So the cluster is at 100% utilization. But for any individual company who doesn't have, you know, a fixed utilization, it's really helpful to have this sort of lever they can pull and plan out in advance.

>> Yeah. Yeah. I'm really looking forward to expanding this. I think it's working. It's it's it's a it's been a very creative way of solving this problem and I think it's working really well and and I think this is going to become essential to startups who are thinking about creating AI products. Yeah, I think you're going to be hearing from me quite a lot about how we can make this thing even bigger. I I think in general, you know, at YC, we think of ourselves as sort of a founders union.

You know, I think a lot of how we what we what we push to do is regardless of how someone's building technology, whether it's a research company or an application company or whatever else, is to give the shared resources to allow the founder to have a lot of the leverage that a much bigger company might have, whether that's in terms of raising capital and having things like demo day to support that or a GPU cluster in the way that a bigger company might have or access to certain advisors or whatever else. And yeah, I'm excited to bring compute as one of the levers we can pull to kind of enable that for people. I'm really excited for where this is

going to go.

>> Awesome. >> Cool. Thanks so much for coming. This is going to be awesome and excited to keep working with you guys. >> Thanks for having me. >>