BofA SECURITIES

IT Hardware Industry

Expert Call Series Takeaways: HBF architecture, applications, & milestones

Industry Overview

Featured Mr. Sudharsanan, former President of Nintendo

As part of our 2026 BofA Hardware Expert Call Series, we hosted Sudha Sudharsanan, former President of Nintendo Technology Development, to discuss HBF architecture, applications, and technical milestones. Mr. Sudharsanan has more than 25 years of semis & SoC experience, with particular expertise balancing compute, memory, storage, power, and cost. Transcript from call is in the back. We view Sandisk as the clearest direct HBF beneficiary, although the opportunity depends on successful technical execution.

HBF likely complements HBM, with inference as initial use

Mr. Sudharsanan does not expect HBF to replace HBM in training given NAND's slower write performance, lower endurance, and HBM's established role in supporting frequent weight updates. Instead, the initial opportunity is likely within hybrid inference systems, with HBM remaining the fast memory tier for latency-sensitive compute while HBF provides greater capacity for read-intensive data, including model weights, KV cache, prefill data, and indexes. Mixture-of-experts and dense models could both benefit, particularly as model and context requirements exceed available HBM capacity, but HBM should remain essential to producing outputs at sufficiently high performance.

Power, endurance, controllers, & packaging remain hurdles

Mr. Sudharsanan identified power consumption, thermal dissipation, write endurance, controller design and achievable bandwidth as the principal obstacles to commercial adoption. The lack of working devices means investors cannot yet validate whether proposed HBF systems can deliver targeted performance under real workloads or assess the associated cost and reliability tradeoffs. Key milestones include the introduction of an HBF controller integrated with PCIe 6.0, subsequent compatibility with accelerator interconnects such as NVLink, and packaging solutions capable of locating flash near high-power compute without creating prohibitive thermal, yield or cost challenges. Mr. Sudharsanan views Sandisk's work on a standardized host interface as directionally positive, but broader adoption will require demonstrated controller functionality, accelerator integration, effective thermal management, manufacturable packaging, and customer qualification.

Incremental products could emerge in 2027

Mr. Sudharsanan believes initial higher-bandwidth flash products based on PCIe architectures could emerge in the latter part of 2027, while more tightly integrated HBF implementations likely require a longer engineering and qualification cycle. Longer term, Mr. Sudharsanan sees potentially significant applications across PCs, gaming systems, robotics, industrial automation, and eventually smartphones. Nonvolatile HBF could preserve models, local context and application state without continuous DRAM refresh, potentially enabling faster resume times, lower standby power, and more capable on-device inference. However, commercial adoption will ultimately depend on demonstrated performance, power efficiency, endurance, and cost. Mr. Sudharsanan also expects broader edge AI adoption to begin accelerating in early 2027, although HBF is not required for the initial adoption cycle.

BofA Securities does and seeks to do business with issuers covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.
Refer to important disclosures on page 13 to 16. Analyst Certification on page 12. Price Objective Basis/Risk on page 12.
13020272

23 September 2026

Equity
United States
IT Hardware

Wamsi Mohan
Research Analyst
BofAS
+1 646 855 3854
wamsi.mohan@bofa.com

Ruplu Bhattacharya
Research Analyst
BofAS
+1 646 855 0315
ruplu.bhattacharya@bofa.com

Ryan Seungin Choi
Research Analyst
BofAS
+1 646 743 0587
ryan.choi2@bofa.com

List of abbreviations on page 11

Transcript

Operator: Hello, everyone. Welcome to the HBF Architecture, Applications and Technical Milestones with Sudha webinar. My name is Tamia and I will be your coordinator today. This call or the replay is not for media representatives. All such individuals are instructed to disconnect now. If you wish to ask a question during the webinar, please use the raise hand button if you have joined via Zoom. If you have joined us on the phone, please press star one on your telephone keypad. I would now like to hand the call over to Wamsi Mohan with Bank of America. You may proceed.

Wamsi Mohan: Yes, thank you so much. Good afternoon, everyone. Thank you for joining us. This is actually the 29th call of the BofA Hardware Expert Call Series this year. A quick disclaimer before we dive in. Conflict disclosures relating to the individual companies or securities discussed on the call today can be found on the call invitation.

Today, as we said, we will be focusing on HBF, including where HBF could fit within the AI memory hierarchy and the potential applications across training and inference. So, joining us today is Sudha Sudharsanan. He is the former President of Nintendo Technology Development. He has more than 25 years of experience developing semiconductor and system-on-chip devices, with a particular focus on balancing compute, memory, storage, power, and cost requirements, so very relevant to everything we're going to discuss. Sudha has previously held engineering roles at Nvidia and other tech companies, and currently advises AI startups and semiconductor companies on memory and storage architectures. So with that, Sudha, thank you so much for joining us here today.

Sudha Sudharsanan: Thank you. I'm looking forward to the conversation. Yes.

Wamsi Mohan: Excellent. Maybe just to kick it off, could you just walk us a little bit about your background that's relevant to HBF and the work that you've done this space, and then we can jump into Q&A.

Sudha Sudharsanan: Yes. I started my technical career as a microprocessor architect, all the way back to IBM, Digital Equipment Corporation, Sun Microsystems, LSI Logic, Nvidia, and then Nintendo. So, along the way, I moved at Nvidia, I started more on the client side of the computing, more SoC architecture for the client side, and the key part of the client side SoC architecture has always been to do with how to optimize your storage and the memory, and in terms of the performance and the power that has to be optimized, and that is a very key factor that I was able to bring to Nintendo to be able to, I was very key in shipping both Switch and Switch 2, and I was able to do that one, so it is all clearly related to that.

Along the way, we also started using the server technology, and then the server technology migration started happening all the way from the hard disk drives, the old tape drives, to hard disk drives, then to the non-storage SSD, non-storage based SSD devices. We have been using that one, now I have led teams that made the key patents on really how to manage the NAND devices, because NAND devices have their own complexities in terms of writing, in terms of wear happening, W E A R happening, and it can wear off, and how to mitigate all of those things. We have had key patents on that one way back in 2005 that all went to Nvidia later.

Wamsi Mohan: Excellent. So pretty rich background here, very relevant to what we're going to talk about. So maybe before we dive in, how are you thinking specifically about the current AI memory and storage hierarchy? If you go all the way from HBM to conventional DRAM, flash, SSDs, HDDs, how are you thinking about where AI is taking the mix of this at the highest level?

Sudha Sudharsanan: So, originally, you may recall that the DDR DRAMs were part of both the client computing, the workstations, desktops, and then servers did all the DDR memories, and then they were geared in such a way that you could add DDR DIMMS on those devices, so you could expand more the type of things. Then the graphics card

came, and the graphics card had an architecture that is different from the unified architecture. It went to non-uniform memory architecture because the graphics needed to have its own memory. And then at that time, then we can actually build a little bit higher performing in terms of the higher data rate and also higher and lower latency to be accessed by the graphics part of the graphic shaders and other graphics processing side of the things, and it was called the GDDR, graphics DDR special device. It's essentially the same DDRs, but built for a higher spec.

And then AMD actually is the one who introduced the HBM into graphics cards. They started using, they said we can actually design the processor and the processor die. Connect that one to the memory die, memory specifically. You can actually keep the DRAM devices in a much parallel way and be able to have a cache with the SRAM-based cache to be able to connect that one to any processor request at a higher, faster speed and lower latency. And originally, this was AMD actually built that one with an interposer architecture that is required for the packaging of this one, so it was packaged, introduced that one in the GPUs, mainly for the graphics processing in the workstations. Then it slowly migrated, and then, of course, Nvidia's CUDA technology started leveraging that on Nvidia's platform, and then they realized that when you go into the deep neural networks, you have to be able to access a large number of parameters, and also from layer to layer computed elements, much more with a lower latency, and at also at a much higher bandwidth, to be able to satisfy the high floating point operations that can be achieved within the GPUs. Then they started increasing that one. Then the HBM started going through their revolutions, and now HBM4 right now, and it is at multiple terabytes of this and then multiple terabytes per second bandwidth also at the same time providing. Essentially it is based on the original DRAM ideas and the large number of devices to be able to calculate that.

Now, this is well established by both when the deep learning neural networks are learning, meaning the model creation, model training aspects of it, because you constantly change the weights, you learn the new weights, you go through all the architectures, and HBM is heavily used for that purpose. Then similarly, then they use the same HBM during the inference side also. Meaning it is not necessarily learning, but it is being used for decoding or inferencing, meaning that there are the large language models sort of a transformer based or the derivatives of the transformer based models, and then they all heavily use and need to bring different parameters, and also to be able to add the context to it. Let's say that you have started decoding it, a large sequence of conversations, but you have to provide the context: what happened previously? All of them, essentially you had to keep that went into the architecture called the key value cache, meaning that whatever you have done in the past you go and fill that one. All that key value cache is also stored and retrieved from the HBMs to be able to fairly quickly access that one, and then to provide a higher tokens per second response, essentially. And not only that one, Nvidia designed the NVLink architecture that essentially provided the ability for the GPUs to be able to address the HBMs in different connected GPUs as well, meaning that you essentially increase your memory storage with a wider network, but at the same time providing a relatively low latency, even if it were in a different GPU, and that comes through the Nvidia's NVLink technology.

Then, of course there are now competitors to NVLink, UAL link technologies, and other things. AMD and other people have been developing different types of technologies similar in principle, essentially connecting these HBMs in different GPUs. To think of it, multiple different islands are connected through very high-speed network interface. Then, at the same time, SSDs have been used primarily for storing your code and primarily operating system and other code, earlier parameters before you start bringing up your network, bring up your processing, how to start computing. You have to bring that one from a non-volatile storage, and SSD has been used as the non-volatile storage. But if you look at the SSDs that are created through this NAND mechanism, NAND has a kind of a similar structure, but completely different, fundamentally completely different process technology related to the NAND, but similar structure in terms of being able to

add multiple devices. The multiple devices, the NAND went in the direction initially to use these multiple devices to actually put together large, high capacity storage devices, non-volatile, high capacity storage devices. Then recently within the past couple of years, the thoughts have come back, instead of while we do this high capacity, it is also possible for us to just like what was done with the DRAM in the HBM case, we can also use multiple devices in parallel to be able to retrieve the information also faster. That is essentially the HBF. But fundamentally, there are many different problems that are unique to the NAND. Right now it is using a one cell to store multiple bits, and the type of structures have further underlying electronic complications that they have to be mitigated, and you have to provide the controller to handle all types of things. So something, as I said earlier on, the wear leveling type of algorithms, you actually extend that type of things further, and then that means you also have to have highly capable error correction mechanisms.

And even when you provide all these things, there are many fundamental physics things that make it non-equivalent to the DRAM type of memory for the non-volatile, because when you do the write to a cell, you have to increase the higher voltage, and you will have a higher thermal restrictions, it'll generate more thermal energy out of it, so heat dissipation has to be considered. And then, if the heat increases, then the higher temperatures can have other additional problems with many, many things. It can affect neighboring cells, and it can have parasitic capacitance aspects of it. All these are fairly complex electronic problems to be solved. Although, from a conceptual point of view, it is easy, but all these problems have to be solved. And so these are the challenges. I'm sorry, I went on.

Wamsi Mohan: No, no, no, no. That's fine. That's a great overview. Like maybe to dig into a couple of these things that you spoke about. First of all, what might be helpful is when you're comparing the architecture, the different vendors seem to be pursuing different types of HBF architecture, and so do you have an opinion on what Sandisk is doing with respect to Hynix versus what Kioxia is doing, Samsung, and if you have a view on what is the maybe preferred approach that might be getting traction and what the timeline of that might be?

Sudha Sudharsanan: So what Sandisk has promised is to some extent almost a long term solution. This is the target solution in the long term. But approaches taken by Samsung or Hynix are more of, even than similarly Kioxia, more of a step-by-step approach to this one. So, as I indicated earlier on, if you want to take it to a really a high bandwidth immediately, you have to face all these problems significantly. Instead, what you do is you slowly increase this one, you learn the technology. How can you mitigate it? You add it, and then you put it in production, and then you go through this usage of this one. So I think that is the approach of both Hynix people started doing that one, and also Samsung approach is also very similar. What it means is what they are looking at is think of it that currently the SSD devices are primarily connected through a PCIe interface, PCIe5 interface, and then the PCIe6 next generation is going to provide much higher bandwidth. And then if we want to provide design closer to providing the PCIe6 bandwidth, and that the HBF can really do that one from their point of view. That may be the right step for the HBF to try out, the HBF overall idea. Instead, Sandisk is trying to do more of a no, no, no, we can do something very similar to a DDR DDR5 or DDR6 type of an interface that we can do already. It may not be all the way up to the HBM bandwidth, but at the same time you can provide it. Think of it more the DDR5 level bandwidth that you can provide. Of course, by having this parallelizing that, then it is possible for you to envision that you can provide an HBM style bandwidth also on that one. That is the approach that Sandisk wants to take. But I like the approach that Samsung Hynix take in this one, step by step.

Wamsi Mohan: Okay. I mean, isn't Hynix sort of working with Sandisk on this?

Sudha Sudharsanan: They are, they are working on it, but they have already started providing some early experiments more along the ways that I just talked about. How can

you put that one onto the PCIe5, PCI6, and they have been publishing, they have been creating devices based on that one, step by step. Yes.

Wamsi Mohan: Okay. So, as you think about the timeline of adoption of HBF, when do you think that this could go somewhat mainstream?

Sudha Sudharsanan: So I would think it is still I haven't seen any device out yet, it has to come from Sandisk per se. But from other vendors, I expect sometime in 2027, latter part of 2027, more like the architecture, like what Samsung is proposing, or the earlier publications from the Hynix people, and that type of devices can come out sometime in 2027. That may not immediately be available for "HBM" replacement or work with HBM for the AI part of it, but it is more of a hey, how you can actually increase the startup of the device, startup of everything, and then that will happen. And I think from that, latter part of 2027, I also see another potential kind of a bifurcation of that technology, that beyond the server side of it can actually have a lot of interesting usage on the edge devices, starting from desktop computers, laptop to gaming devices, and then in a few more years, even to your mobile phones.

Wamsi Mohan: So, I think at least Sandisk seems to have spoken about these kinds of applications of HBF. So, I'm curious, in HBF, architectural proposals that various companies have shared, there are a multitude of these. There is sort of an HBF only, or there can be an HBF plus HBM kind of shared architecture. There is sort of HBM with caching in HBF, and then there is disaggregated kind of separate. So, if you think about the progression of HBF, how do you think of this in terms of if we think about these four, or if you have to add any more, how would you think that the technology evolution would happen over the next few years?

Sudha Sudharsanan: So let me first go to this, particularly for the hyperscaler part of it, applications part of it.

Wamsi Mohan: Sure.

Sudha Sudharsanan: So when it is very difficult, I do not see the HBF can replace HBM in the training case, or model building cases that the HBF is very difficult to enter right now.

Wamsi Mohan: So can we double click on that? Why you see that as the reason?

Sudha Sudharsanan: The main reason is that it is the write endurance is bad. Not only write endurance, write performance itself is bad when compared to HBM. Then I'll come back to a few other things and where they have to show, but those are some very critical points. And then HBM is has been shown to work, and it is working, and people are making money off of it because people are using it.

Wamsi Mohan: Sure.

Sudha Sudharsanan: I mean, money off of it, because people are producing tokens, and then real productivity is being seen out of that one. Not just producing, but coming up with new models that can produce intelligent tokens out.

Wamsi Mohan: Yes.

Sudha Sudharsanan: So that means HBM, it is very difficult to move HBM's premier position out of it because it has shown that it's doing that part. So the HBF that means can only enter from right now the architectures and the exploration has been primarily limited to the inference side of the things. But any hyperscaler building just for inference, they are not making a smart decision. You cannot tell anyone, oh, you can do this one, hyperscaler you can do these things just for your inferencing only, because they need to be able to build their networks to be able to train as well as to use it, and that is where then people have been saying, "Oh, you can mix these things to be able to do certain things. And even the inferencing side of the things, they are mixing. Oh, in some cases that you

need to be able to have fast memory attached, so that you can also write faster and read faster memory from the HBM side of the things. But you can, for a large segment of that need, you can just do that one as a read-only memory type of thing, or read more often. And for that, the HBF current, proposed HBF bandwidth and the latency seem to satisfy that. That is the reason that people are creating, can we build this hybrid type of networks?

Wamsi Mohan: Right. So as you think about on the inference side, should we think that basically in inference on a predominantly read-based solution, you're carrying model weights, the KV cache can be large, and you're retrieving this at somewhat faster speed, but you have larger context windows to offset this? So, as you think about that, how do you see the opportunity in inference for these solutions? I mean, we've some idea of how many terabytes in these Nvidia systems of HBM can be attached. Like, how do you think about HBF in relation to that?

Sudha Sudharsanan: So the relevancy of that one is that difficult to say. The reason for that one is the models are still improving. Meaning that it is going to continue improving for the next, I cannot even say when this is going to some extent to a reasonably saturated level where the large percentage of the users are going to use these specific models only. When that becomes a little bit clear, then you can say you can bring just the HBF network only. We can do better in faster inference for a cheaper cost. That promise at that time will come true. Now, having said all of these things, we still do not know what is the power consumption. How is this going to be created? What is the exact, because of the devices, they are not in practice. It has not really gone into implementations, into productions. So, you do not know what are the limitations. You only have proposed numbers, and people are estimating these things just with the proposed numbers, not necessarily seeing that one running it on a real system, and that is that is perhaps is the biggest unknown at the moment.

Wamsi Mohan: Okay. So, as you think about that, as we sort of fast forward a little bit and think about potentially where this might land in practical terms, which accelerator vendors are actively evaluating HBF at the moment?

Sudha Sudharsanan: It is not publicly clear the HBF part is out there yet or not, except for the experimental ones that Samsung and Hynix people have created. I have not seen anything to the extent of where you can actually measure the thermal limitations, power consumptions, and the bandwidth and the latency, it is not seen. But people are assuming certain numbers out of it, people are providing some numbers in terms of you can do this many tokens per second, that type of things, that computations, but it is largely hypothetical at that time because they don't have any evidence for any of these things at the moment.

Wamsi Mohan: So you're saying that today there is practically no measured data, but there is a lot of simulation data. People are making their assumptions based on the simulation data on what the bandwidth could be, what the latency could be, what the power metrics could be, and then based off of that, they're making some projections on what the size of the market potentially can be?

Sudha Sudharsanan: Yes, exactly.

Wamsi Mohan: Okay. So as you think about these limitations that we're discussing, where do you see the biggest hurdles in practical terms? Is the power constraint going to be the largest thing? Is it going to be the right endurance of these devices or is it going to be more around the bandwidth or latency? I mean, it feels like we're creating this for a specific type of use case, which is inference. And so, in that context, what do you think are the biggest hurdles that you see?

Sudha Sudharsanan: All of the hurdles that you have mentioned: power consumption, bandwidth, right endurance, and the controller algorithms that can actually, practicable controller algorithms that can mitigate these things, both from the physics side of the

things as well as the algorithm side of things, no evidence right now exists out there to be able to judge this.

Wamsi Mohan: Okay. So, you're saying basically it's like uncharted territory, and this is going to take some time to see actual two samples and dyes and like put it into production and then actually do the measurement?

Sudha Sudharsanan: Yes, I mean, you may have to go back all the way to the, I forgot the HBM1 era, maybe 2017/18 time period. That is when the first HBM GPUs came out. AMD started with that one, and they had to actually do a lot of the work before they publicized that one, but they did a lot of the hard work. But then it was not at this performing capacity at that time, but it slowly improved. Then you can improve the versions, and then they understand how to work with this – it'll take lot of engineering effort, and not much evidence is seen of any of this one, I have only seen papers so far. But maybe some of the experiments that Samsung and Hynix have been doing, those are the initial experiments. But they are also not going with the bandwidth talks that Sandisk has been providing.

Wamsi Mohan: I know we've not seen actual observable data, but if you think about the simulation data that's out there, how much performance degradation occurs when the weights are stored in HBF instead of HBM?

Sudha Sudharsanan: So the simulation data is based on what the promised numbers are, what they are targeting right now. With that one, the HBF looks decent. Decent meaning that there are models that there have been many different – there was a recent article in the Hot chips, where was that, the week of 24th.

Wamsi Mohan: A few weeks ago, yes.

Sudha Sudharsanan: Few weeks ago, Hot chips, it was provided, but that number was based on just what the Sandisk numbers are, what they are targeting. But with that, they have done a decent comparison. The decent comparison is that the mixture of the HBM/HBF systems will have advantages, inference side of the things and specific scenarios that you talked about earlier where there's this mixture of it is good. So, you have to realize that you cannot have, even on the inference side of the thing, HBM will be needed because you do have to compute the final output through that has to go through the LLM, and at that time the HBMs will significantly improve the performance. Even if you can handle the KV cache and the indexing, all those things with the HBF and the prefills, the HBM is still needed for you to actually produce the results fairly quickly and progression of the LLM. So that means that the hybrid systems are the ones that will work well.

Wamsi Mohan: So what model architectures will benefit the most from HBF? I mean, are these dense LLMs? Are they mixture of experts? Are they agentic systems? How should we think about that?

Sudha Sudharsanan: I think I have seen papers, analysis that the mixture of experts can actually benefit out of this one, of the HBF, and also the dense systems can also benefit. But as soon as the number of parameters go really high and then they go beyond the capacity limits of the HBF or HBM, HBF specifically in this one, beyond the proposed capacity limits of it, that is when you will have some interesting new architectural implementation approaches have to be taken. Meaning that let's say that you have, I don't know, they are promising up to two terabytes of this HBF and then if the models and the prefills and your context levels exceed that one, then you have to have some mechanism to be able to, think of it that windowing mechanisms that you have to use to be able to not include some of the context information at that point.

Of course, when you go to the really true models, there are some models that are more of the state space type of models. Think of it that they are working internally, they have it was the old signal processing point of view, it is similar to the infinite impulse response systems, meaning there is a loop inside that will keep the state internally looping around. That means with that type of a looping mechanism, it has been shown

you can actually have a much longer contextual length with your models. And there are some models, one of the examples of that model is there is a Mamba model, and the Mamba based models, and people have been trying out those new models also, but that is actually used. In fact, some of that is in fact used in the recent – the name escapes me – the recent model that was released by the Open Source, the Chinese Open Source community released a recent model based on that one.

Wamsi Mohan: You mean DeepSeek?

Sudha Sudharsanan: Not DeepSeek, after DeepSeek. So DeepSeek introduced the mixture of experts, part of it also, but then there are other models that are relying more on the—.

Wamsi Mohan: Okay. So let me ask you a couple more, and then I will open it up to the audience as well. I just wanted to ask you, in terms of when you think about what is sort of the milestones that you would watch to make sure that this technology is getting commercialized appropriately and potentially getting used, what kind of milestones would you say are the most important ones to watch over here?

Sudha Sudharsanan: I think two things to look for. Is there an HBF controller that is coming out first, the HBF controller. The first thing I would like to see potentially is the HBF controller tied with the PCIe6 controller at HBF? You know, beyond the PCIe control, and in that case, think of it the extension at the NVMe type of things. That is, of course, with a higher bandwidth with the HBF controller with the PCIe6, that is what I would observe for. And the next one is the HBF control is really connecting to NVLink type? Can it actually interface with the NVLink? That'll be the next challenge. If they can do it, then it is essentially kind of being treated as one more element, one more HBM, with certain limitations and properties associated with that one, then it can be incorporated into the larger memory system, multiple of these HBMs and HBF that are connected through the NVLink.

Wamsi Mohan: Can I ask you a question on that? So, as we think about hitting this memory wall, there is an alternative sort of architecture around overcoming that with faster interconnect technology, particularly optical. So, as you think about the parallel development of this, is there an intersection or a good way to think about the roadmap of interconnect in relation to the roadmap of HBF?

Sudha Sudharsanan: Yes, there is a way to do that one provided that the HBF can interface to these interconnects. That is the other hardware element that has to be. Think of it for all of these HBMs, it goes through an opto couplers, optical connectors, that type of an interface that has to be built, so that has from the HBF controllers also that has to be built, that is one way to think about it. But beyond that, I do have to give one credit to Sandisk. They recently described that they are actually building an interface for the HBF with the OCP 3.0 network. That is, the OCP is more of a think of it your SoC system on chip that includes both the CPU and the GPU directly connect, how is it right now connected to the DDR memory or PCIe? And that type of an interface, they have done that one, they have come up with an implementation, or they are working on an implementation that will be connected to this OCP network. So, I think that is the right step for them to take it to devices directly. That is one other right step beyond the NVLink type of connection.

Wamsi Mohan: Okay, that's helpful. Maybe I'll just pause here. I have some more questions, but before that, let me just see. Operator, do you want to give the instructions again on any questions that might be there in the audience?

Operator: Absolutely. If you are in the event via the Zoom interface and would like to ask a question, press the raise hand button, and you should see a prompt on your screen to unmute and speak when it's your turn. If you are dialed in via a telephone and would like to ask a question, please signal by pressing star one on your telephone keypad. If you are using a speakerphone, please make sure your mute function is turned off to allow your signal to reach our equipment. When it is your turn to speak, you will hear a message

saying your line is now open and may proceed with your question. Again, press star one to ask a telephone question.

Wamsi Mohan: And as that's happening, if anyone wants me to ask on their behalf, just shoot me a Bloomberg, and I'm happy to ask anything on your behalf as well if you have any questions here for Sudha.

Wamsi Mohan: Okay. So, I have a few more that I want to go through. So, when you think about the architectural evolution, so in the case of some of these players that have both HBM and pursuing HBF, obviously they're making money on HBM today, do you see them leading the charge in some ways on a maybe architecture that is more HBM with HBF cached kind of architecture, or it really doesn't matter as much because it's early enough that like there's then if HBF succeeds, it's going to be a big enough market that it's not going to really matter?

Sudha Sudharsanan: So my perspective on this one, specifically I think maybe Samsung is in both camps, and they will be in an interesting situation. Samsung and Hynix, both of them are in both camps in some sense, and it will be interesting, but they may take the right engineering approach. But people like Micron would have very little interest in promoting HBF. In fact, they will work hard to actually point out that that is not the way to go. In fact, you may recall that the HBF kind of technology with trying to put that one with some DRAM/SRAM integration was tried by Micron and Intel. It was called the 3D Optane Technology. Maybe it has died about a decade ago, but they did try this type of things in the past in some form. But they will be the ones initially to kind of put this one down, try to put this one down. That's my personal perspective. But people are who are on this, my thinking is that Samsung and Hynix initially will push this one for edge devices, because Samsung and Hynix are big players in both desktops, workstations, gaming systems and then ultimately the mobile phones, so they will have interest in bringing that one out to them.

Wamsi Mohan: Do you have a view, as you're looking at running these models, maybe open models on at the edge, what do you think is the time frame when that happens? Because we hear companies like HP talk about these AI PCs that effectively can run models with, obviously, you don't pay a big token overhead in these situations. So, when do you think that becomes realistic in terms of time frame?

Sudha Sudharsanan: So I know there are several companies that I have consulted to, and I know that is a very, very active field, both using everything from Apple hardware to the AI PC hardware. How you can actually bring these large models to this usable for your own data, for your own purposes? It is happening in everything from industrial automation, robotics, where you need the responses quickly. You don't have the time to go to the server and come back. And also, at the same time, you don't want to send the data to the server. That is the other thing. Also, that is happening with law offices, financial planners, and that is a very active. And I expect to see the take-off of that one early 2027 itself, I expect to see that one.

Wamsi Mohan: Okay. And in that context, do you really need HBF to be at any certain stage, or it doesn't matter?

Sudha Sudharsanan: Okay, but I can tell you, if you think about the HBF in the edge devices, I can envision computers that don't necessarily, computers and devices, that can be started much faster, and that won't consume significant energy when they are sleeping or turned off, and that can be done if the HBF promise comes out through. And I think that is where the HBF's biggest in the long term run will happen on the edge devices. And then where you directly deal with consumers.

Wamsi Mohan: Can you just expand on that a little bit? So, as you think about the long term opportunity there, what needs to happen, and why do you think that that is a material market?



Sudha Sudharsanan: It's fairly simple. Right now, when you have a device, let's say that you have a consumer device, your phone, when you put your laptop or MacBook, close the lid, it's sleeping, and the sleeping and the information is actually stored in the DRAM and the DRAM refreshes happen and the battery is used for that purpose. Think of it that you don't have to do this much of energy, and then you can be stored in your non-volatile HBF, and then you can quickly kick off the HBF, you will be able to get the same experience. And in fact, many more programs can be active for you than the limited programs that they have active on the DRAM. So, the user experience will be significantly better, but at a lower energy cost while you are sleeping, or while the device is sleeping.

Wamsi Mohan: I see. Okay. So, do you think that this is going to come first to laptops, desktops versus phones, or do you think it's the other way around?

Sudha Sudharsanan: I think it will be experimented on the laptops and desktops first. The reason is they will work with the PCIe like Samsung and Hynix working on the PCIe6 interfaces, and that is initially to be able to try out on them, and then it will migrate to the phones. But I can be wrong, and you never know, one of the Chinese phones, the Oppo, or something that can come out with this technology. But the only thing I don't see is that I have not seen the evidence in terms of the power, and then performance maintenance, and that is where I'm reluctant to say it is coming to your phones type of things, and it can.

Wamsi Mohan: Okay, understood. Maybe just to wrap it here, let me just ask you something on – sorry you were going to make a point?

Sudha Sudharsanan: No, I just wanted to make a point about the recursive models, the Kimi K3 implements very similar to a Mamba architecture I told you about, and that doesn't necessarily require the standard long KV cache length type of things.

Wamsi Mohan: So there's actually a question here from an investor. Let me just ask this. So, the question, "It appears Sandisk showed that they're working on phones, cars, robotics already on the potential longer-term applications for HBF Next Gen on the edge. The question is, if we take Meta Muse, for instance, which is prime to like the number one app and was named in the HBF consortium, can you elaborate on how HBF would enhance edge performance?"

Sudha Sudharsanan: If it is specifically with respect to the AI models, then it is the standard way that we are talking about in terms of improving the context lengths and in terms of using the KV cache and in terms of inferencing in those terms that keeping enough already precomputed information as in the HBF, and that is how it will be used, and that's how it will be leveraged. Then your local information also can be included into those indexes, and that can also be stored in the HBF.

Wamsi Mohan: Okay, got it. Then maybe just to wrap, I just want to ask you a little bit around packaging technologies. Does the industry need like an HBF equivalent of a pretty established HBM packaging ecosystem?

Sudha Sudharsanan: Yep, it does need a packaging technology. That is needed. And right now it is based on somewhat expensive interposer type of technology, so core SAR type of technology that is not cheap enough. Another technology improvement that has to happen is can you actually bring all the way to organic packaging, or maybe take along the way of what Intel has been doing with various architecture that they have been doing.

Wamsi Mohan: Okay. Okay. Yeah. Look, we covered a lot of ground here. I think we're at the top of the hour, so appreciate everyone joining. If you have any follow-ups, definitely reach back out to us, and we'll try our best to get those answered. And Sudha, thank you so much for all your insight, your time. We really appreciate it.

Sudha Sudharsanan: Okay, thank you very much, and enjoyed the talk. Thanks.

Wamsi Mohan: Thank you so much

Abbreviations

3D: Three-Dimensional
AI: Artificial Intelligence
AMD: Advanced Micro Devices
BofA: Bank of America
BofAS: BofA Securities
CPU: Central Processing Unit
CUDA: Compute Unified Device Architecture
DDR: Double Data Rate
DDR5: Fifth-Generation Double Data Rate
DDR6: Sixth-Generation Double Data Rate
DIMM: Dual In-Line Memory Module
DRAM: Dynamic Random-Access Memory
GDDR: Graphics Double Data Rate
GPU: Graphics Processing Unit
HBM: High-Bandwidth Memory
HBF: High-Bandwidth Flash
HDD: Hard Disk Drive
HP: HP Inc.
IBM: International Business Machines
IT: Information Technology
KV: Key-Value
LLM: Large Language Model
MoE: Mixture of Experts
NAND: Not AND
NVLink: Nvidia High-Speed GPU Interconnect
NVMe: Non-Volatile Memory Express
OCP: Open Compute Project
OS: Operating System
PC: Personal Computer
PCIe: Peripheral Component Interconnect Express
PCIe 5.0: Fifth-Generation Peripheral Component Interconnect Express
PCIe 6.0: Sixth-Generation Peripheral Component Interconnect Express
PCs: Personal Computers
PO: Price Objective
R&D: Research and Development
SoC: System-on-Chip
SRAM: Static Random-Access Memory
SSD: Solid-State Drive
SSDs: Solid-State Drives
UALink: Ultra Accelerator Link
UCIe: Universal Chiplet Interconnect Express

Price objective basis & risk

Sandisk Corporation (SNDK; C-1-9; $1887.04)
Our PO of $2500 is based on approximately 10x C27E EPS of $255 which is in line with SNDK's global memory peers' average due to similar profitability.

Upside risks: 1) we are earlier than expected in the NAND upcycle, 2) stronger than expected expansion in the NAND market on sales of AI enabled consumer products, 3) faster than expected eSSD market share gains, and 4) faster than expected recovery in NAND pricing.

Downside risks: 1) sharp drop in NAND prices due to oversupply, 2) Competition and over expansion from Chinese suppliers such as YMTC, 3) slower than expected adoption of AI enabled consumer products, and 4) share loss in eSSD market.

Analyst Certification

I, Wamsi Mohan, hereby certify that the views expressed in this research report accurately reflect my personal views about the subject securities and issuers. I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or view expressed in this research report.