EDB POSTGRES AI

EXPERIENCE THE FUTURE OF AI: THE SOVEREIGN DATA AND AI FACTORY

Chapter 2. The Potent Combination: Postgres + NVIDIA for Accelerated AI

Simon Lightstone (EnterpriseDB): So we're talking about the Postgres® data and AI factory. This is a combination of Supermicro servers, NVIDIA's GPUs, and EDB software. It's a turnkey system. It makes it so simple. Once it's set up in your data center, which we do for the customer, it's just deploying a highly available Postgres cluster. You can do that in a few clicks, rather than having to go through all these processes. And, of course, everything you deploy works with AI. It gives you all the power of the NVIDIA GPUs and the Supermicro servers to let you query an inference from your database dataset, using whatever you'd like. You can use RAG, or you can use other separate workflows, and basically bring it all together to deliver these solutions in a totally private environment.

So, Somik, why don't you talk a bit more about the hardware that's being run to the table by Supermicro, what we have in store.

Somik Behera (Supermicro): Yeah, we have loved working with EDB on this project, on what we're building today. We take EDB's market-leading AI database and AI software, marry it together with Supermicro's building-block architecture—which can be customized for any workload and optimized together with NVIDIA's accelerated graphics computing units, GPUs—so that you can not only have AI available, sovereign, in your data center but you can have it AI fast, right?

And we do that together because we have this capability over the last 30 years of building engineered systems together with the work of software design partners, being able to kind of rack, stack, and integrate the software image in our factories and provide that distribution, manufacturing, and distribution to companies globally. Anywhere you are, we can deliver that in a matter of days. We're excited to bring those capabilities and support EDB's growth—their AI software, their AI database—and leverage NVIDIA's accelerated computing to drive this AI revolution to AI boom for every enterprise in every corner of the world.

Lightstone: Very cool.

Nave Algarici (NVIDIA): Absolutely. It's been great to work with EnterpriseDB on this solution. And what we've built to that end is something called NeMo Retriever, which is a set of microservices that are built for accelerating information-retrieval applications, from extracting data from enterprise documents to indexing and storing them into the vector database for easy retrieval. And then accurately retrieving that for real-time applications, if it's RAG and if it's others. And by packaging that in the microservices, we enable EnterpriseDB to take that and build that into their platform and build on top of that all the amazing capabilities. And we really see the integration, and the better-together of the platform, is what brings the key values of the solution. And by working on that full stack, we're really able to optimize it end to end and build something valuable for our customers.

Behera: Yeah. And I would like to add that, you know, just like NVIDIA has their hardware and software capabilities, Supermicro brings the same to bear to enable EDB to succeed globally. So we have capabilities like SuperCloud Composer that can manage its entire hardware, entire rack, compute network storage, anything else—power, cooling if you need it—the entire data center, the entire data center building-block solution. And then we marry that together with SuperCloud Orchestrator and the SuperCloud suite of software, to kind of provide this turnkey supply chain experience globally, anywhere, at any time.

Lightstone: Yeah, and that's sort of why it's such a good fusion of three, right? So from the EDB side, we are the Postgres experts. Postgres is the most popular database today. If you look at surveys out there from Stack Overflow…. And, you know, databases are actually pretty hard to set up and manage, and, of course, they just can't go down, right? So you have this marriage of the database—you can deploy a high-availability database with just a few clicks—with the engineered system. And then you have the ability to use the AI and AI models, powered by NVIDIA and Supermicro, to do a lot more with that data, to essentially answer questions, have agentic workflows, and all that good stuff.

And by combining all these things and adding EDB software, you get to build chatbots in months rather than years, right? You just get this really accelerated workflow that lets you take the best of what every company has to offer and then just deliver real production, right?

And I think real production is one very big area where we all work together to deliver for the customers here. Because this is no longer an experimental thing, where AI—you're just building small experiments and maybe building a small chatbot. These days we're talking about very serious, mission-critical applications, using it for emergency services,

using it for rapid decisions for fraud detection. These systems can't just go down anymore, the way that perhaps many years ago people thought of AI as an experimental asset. This is now a real production, and that's what this system is designed to do very easily.

Algarici: One of the values of working together with EDB and Supermicro is by taking the NVIDIA NIM and putting it inside the platform, and really ensuring the partnership and integration of the two companies in building the solution, we've been able to really optimize this pipeline end to end on workloads that really matter for enterprises at scale. We talked about bringing stuff into production, ensuring that the experience that you have when you build the application, to scale the application, is the same, and you get the sync quality at scale. So that comes from bringing in the data, extracting all the relevant information accurately, working together with the vector database and the database in general to ensure the data is up to date and is available in real time, and then retrieving that accurately.

Lightstone: So talking about, specifically, scale, I think that that's a great story, because with this offering with the engineered system, you can start with a smaller number of servers. That'll fit inside a rack or two, depending on your initial configuration. But then you can just add as you go. You can just add more compute nodes with more GPUs, or, if you just need more database workloads, you can add those.

So you're no longer being forced to make this huge commitment. It's fully modular, you add to that rack, and ultimately you can meet your goals. If you want to plan in the future, you don't have to buy everything in advance. You can just start with what you need now, and then add the GPU power and the Supermicro power as you need.

Behera: Yeah, and we have totally been excited working together with NVIDIA's leading NIM software, as well as EDB's Postgres, as well as AI software, to which we can fuse this together with our building-block server architecture, and the building-block architecture gives you that benefit, right? You buy the minimum you need, and you scale on demand, just the way you are at cloud—but, except it's 90% cheaper. It's on your own data and you can do it today with a turnkey solution with one number to call, one throat to choke, right? And we can do it anywhere globally, right? Wherever you are, we will meet you. And that's what's so exciting about this partnership.

Lightstone: Yeah, agree. And we're saving customers a lot of money. I mean, one of the big complaints that we hear from customers is the cost of I/O. Sure, cloud might be cheaper for sort of the small stuff. As soon as it gets to serious amounts of I/O, those costs just start to snowball, and they're recurring. It's not like they're buying one high-

powered Supermicro disk system and then they're using it over and over. They have to keep paying for that month after month.

But with this system, you make that one-time purchase, which of course can also be leased—however they want to do it—and then that's it. You're essentially cutting that tax of having to pay for your performance every single month.

With the engineered system, you're going to get about 30x faster return on results from your data.

Behera: Wait, that's not 30%. That's 30 *times* the speed of traditional, non-accelerated computing.

Lightstone: Yes, yeah. Because when you consider all the goodies that we've put into the engineered system—the ability to take on analytics workloads, the ability to really be optimized—and you've seen the benchmarks, you're talking about 6x acceleration just from working with Supermicro. We're talking about a very, very big jump in terms of what we can do. And all that is baked right in, tested, and validated for top performance with NVIDIA.

So you're getting a lot more value than the individual parts. You're actually getting something that works well together and is really optimized as one package.

Algarici: Yes, absolutely. And we reviewed each part of the pipeline to make sure that all the bottlenecks are addressed by accelerated computing. So if it's embedding your data and storing it into the vector database, you're seeing 2x to 3x throughput improvement on that. If it's extracting your documents—extracting information from your documents and bringing that into your enterprise application—and if it's retrieving and re-ranking them, we're seeing 1.5x to 2x improvement in throughput of retrieval.

So you get your answers quicker, you generate insights faster, and you scale your deployment even more.

Lightstone: Yeah, and also the system is modular, right? Which is very important for our customers. They want the ability to be able to just add to it later. They want the ability to scale up very, very quickly should they want to do that, right? So you really get the best of both worlds.

Behera: Yeah, and DCBBS, our data center building-block architecture, is unique in that way. And we love to support these next-generation accelerated workloads, where we

can give you that benefit. Enterprises can start small and then expand. In fact, with this 30x speed and performance improvement you're talking about, maybe they don't need to expand that fast. Or maybe they're going to think of use cases none of us have thought about and it's going to explode.

But either way, leveraging the Supermicro hardware, our data center building-block architecture, you're prepared. You're ready to tackle any agentic workload out there.

Lightstone: Yeah. In fact, to make sure—because it's not just about the small workloads, of course; we're dealing with huge customers of multiple racks—and in that case, every component of the system is scalable.

So, for example, we have the main Hybrid Management, which is going to manage your whole Postgres estate. And that cluster of three systems initially—you can expand on that cluster to ensure that you're always going to be having enough compute power to monitor it. Or you can expand your compute cluster specifically for your databases by adding in compute nodes. Or you can also add in AI-advanced, accelerated compute nodes packed to the brim with the best NVIDIA technology to make sure that you're you're going to be able to scale up whatever you need for the AI workloads that you've got.