ARTICLE

Financial Services Firms Will Bank On Homegrown AI Training

By Timothy Prickett Morgan, *The Next Platform*

Every company in every industry in every geography on Earth is trying to figure out how they are going to train AI models and tune them to help with their particular workloads.

While the hyperscalers, cloud builders, and HPC supercomputer centers of the world are doing a lot of large-scale theoretical work evolving AI models – and these days, specifically for chain of thought generative AI models and agentic architectures that will interlink a mesh of independent models to do analysis about or cause action in the real world – it is the world's financial services companies that will be at the forefront of commercializing these technologies, just as they have been with earlier technology waves in information technology.

Just think about how quickly we moved from automated teller machines to the Internet and its digital revolution to have full mobile banking from our smartphones and laptops.

But here is the conundrum. Financial services firms move fast and move first, but parts of the financial services industries are more heavily regulated than others, and that regulation generates an impendence – but not a barrier – to AI adoption.

"Banks, insurance companies, and other financial institutions tend to be a leading adopter of emerging technologies," explains Scott Hebner, principle analyst for AI at theCUBE Research. "Since they operate in these highly regulated environments – it's real time, with super high dimensional problem sets, and it's other people's money – they have to tread much more carefully than every other industry with maybe the exception of healthcare. But financial services tend to be on the leading edge of things, and they tend to solve the biggest problems first, and then those things spread into other industries."

Hebner says that FSI companies were among the first to use predictive AI models, and they have also been early experimenters with generative AI. The predictive AI models comb massive datalakes and try to correlate statistical probabilities across huge classes of data; once done, you can take an input and predict and outcome based on past history. FSI companies are also building their own foundation models, or working with the big model

builders like OpenAI, Anthropic, and Google, and it is not entirely clear yet where GenAI models can help.

For instance, trading systems have to be trusted, which means they have to be repeatable, explainable, and – because of regulations – reportable. Proprietary trading firms (those who are playing with their own money) and hedge funds (those that trade assets of a selected few investors) can take more risks than the trading desks of commercial banks such as JPMorgan Chase, Bank of America, and Wells Fargo and investment banks such as Goldman Sachs, Morgan Stanley, and Credit Suisse. These companies, as well as asset management firms, private equity firms, and venture capital firms will all be looking to use GenAI to figure out what to invest in and when to get out. Insurance companies will be using GenAI to help process claims – for damages to our homes, autos, and such – as well as to make their own investments to hedge against the future.

What is absolutely clear that GenAI can immediately help with regulatory reporting, which is massive and which needs a fast turnaround, across the FSI spectrum. This is the low-hanging fruit that might help save enough money for FSI companies to do larger investments in AI down the road. The trick is, again, that AI models have to be shown to be trustworthy at culling data and summarizing it without inaccuracies.

How – And Where – Will FSI Companies Do GenAI Training?

While we know that financial services firms will be among the early adopters of commercialized GenAI and still heavy users of traditional, statistical AI, what is not clear as yet is where the processing and storing for GenAI training will be done. FSI companies could just license models and run them in cloud if they wanted to do the easiest thing, but that would also mean giving up a lot of control and paying a heavy premium for both the models and the systems to train them.

One need only look at the cost of a buying an Nvidia GPU and running one in the cloud for three years. Last year, a "Hopper" H100 SXM GPU – one that works in an eight-way shared GPU memory system – with 80 GB of HBM memory from Nvidia probably cost on the order of $22,500. If you wrapped a system around it, the portion of the system one H100 SXM used might cost another $25,000 with storage, networking, and host computing; power and cooling over three years would be on the order of $6,250 **at 19 cents per kilowatt-hour in the New York Metropolitan area**. That's $53,750 for an H100 SXM and its share of the system for three years of operation.

On Amazon Web Services and Microsoft Azure, pricing for their respective P5 and ND-H100-v5 instances is the same at $323,202 per H100 for three years using on-demand

pricing and $141,868 per H100 using three-year reserved instance pricing. It costs 2.6X to 6X as much to run GPUs in the cloud, assuming 100 percent utilization of those GPUs both on premises and in the cloud, as it does to buy them. (This is a simplified comparison that does not take into account datacenter space and people costs. But you see the gap.)

Given the profit margins the big clouds are getting for their GPU instances, you can see why FSIs will inevitably want to put their AI models either in their on premises datacenters or a co-location facility that has fast links to their datacenter and also to one or more of the big clouds where other applications might be running on less expensive and expansive infrastructure.

While cost is a big factor in deciding where to buy or rent AI training infrastructure, security concerns and regulatory compliance reasons are two others, and a third is that for machinery that is going to do double duty for training and inference, the latency for inference response times is critical and argues for having the inference being done in the same place as the applications that are being augmented with AI.

“In the FSI segment, the one thing that we are hearing loud and clear from many customers is that they don’t want to put their GPU-centric workloads on the cloud,” Vik Malyala, senior vice president of technology and AI at Supermicro, tells *The Next Platform*. “They are just beginning to see how their models are working and if the use cases are panning out, and they don’t want to go to a cloud. It is too expensive, and they are still trying to figure out how to optimize the links between the AI platform and their own data platforms and data patterns. Many are retraining models in-house or in co-location facilities.”

Which brings up another reason to be hesitant about training AI models – or retraining one that is licensed or is available open source – in the cloud. And that is security for both the model and the data that is used to create it.

Given the uniqueness and diversity of their applications, we think financial firms will have lots and lots of models as well as heftier chain of thought models, which “think” more deeply about more complex problems that they are presented with – and that take a lot more infrastructure to train. These models that FSIs develop will be just as proprietary as the trading algorithms they created in the past, and given that models will often be trained on customer data and that AI models and their weights are a new secret sauce, we think many FSIs will want to have sovereignty over their AI training systems (and ultimately, also on the inference systems that integrate with or actually are their applications) and their data. If you have someone’s AI model weights, you essentially have their model.

But setting up a datacenter to run GPU-accelerated systems for AI training is not trivial in the financial centers of the world, and getting the power to do the processing and to cool the machinery is not trivial, either. And getting access to GPUs is also difficult, but moreso for companies that are not hyperscalers and cloud builders, right now. And all of these factors are dependent. If you can't get the power and space to run an AI supercomputer, then Nvidia and AMD will not allocate them to you or the company building your systems because they can only get to book revenue when you accept delivery. GPU makers want people cranking away as soon as the GPU systems hit the floor.

For one thing, and AI training system is a supercomputer, where as a high frequency trading system – another kind of HPC used by FSIs – is more like a very single node doing one thing at a time. You might have dozens or hundreds of trading nodes, but they are doing their own thing. An AI training system is hundreds to thousands of nodes doing one big job that is orchestrated across the memory and compute of the thousands to tens of thousands of GPUs that comprise the system. For it to work properly, it needs not just high bandwidth memory on its compute engines, but high bandwidth networking between the GPUs inside of a node with load/store memory semantics so it looks like one big GPU, and then high bandwidth networking across the nodes that has low latency and as little jitter as possible.

This AI supercomputer is not like the CPU-only clusters that FSI datacenters run applications and databases upon, which might have 10 kilowatts or 15 kilowatts per rack. A rack half full of HGX H100 eight-way GPU nodes will burn 40.8 kilowatts (and it is only half full because without liquid cooling this is too much heat in a rack), and with the "Grace" CG100 and "Blackwell" B200 GPUs that are shipping now, Nvidia pairs two Blackwell GPUs to a single Grace CPU and puts 36 nodes in a rackscale system with 72 GPUs with fully shared, all-to-all memory links that burns 120 kilowatts. **Nvidia's roadmap says that by 2027 with the "Rubin Ultra" GPUs**, Nvidia will have four reticle-limited GPU chiplets per socket for 576 per rackscale system with over 600 watts of power consumption.

This is not something you can drop into most metropolitan datacenters. At least not yet.

The good news is that most FSI companies don't have to do this today, since they are doing a lot of proofs of concept and only early deployments for very precise workloads. In one case, just to give an example, an FSI company that is a credit card provider is losing several hundred million dollars a year in credit card fraud and has requisitioned a few Grace-Hopper nodes from Supermicro to test out a new fraud detection application with AI hooks in it.

Typically, says Malyala, FSI customers start out with a mix of 8, 16, or 32 H100 nodes (meaning 64, 128, or 256 H100 GPUs). There are some FSI customers who are buying 128

nodes or 256 nodes using B200 GPUs (which means 1,024 to 2,048 GPUs), Malyala reckons that they are actually training their own models. FSI companies, as you know, are very secretive about what they are doing and how well it is working. Malyala adds that for both training and inference, there is a lot of interest in FP8 and FP4 precision processing and that FP16 and BF16 formats "are history."

The reason is simple: You can get almost as good of an answer with data that is half to a quarter of the precision, which means you can train 2X to 4X faster. And therefore, customers who want FP4 performance and all of the recent hardware assists that Nvidia has created for transformer models customers are willing to wait a few months to get Blackwell GPUs so they can push training performance even further.

Supermicro has done a lot of Grace-Hopper and Grace-Grace seed units into the financial services sector, and the GH200 NVL2 (with two interconnected Hopper GPUs) and the GB200 NVL4 (with four interconnected Blackwell GPUs) have been snapped up by some FSI customers, and this suggests they are using these nodes for inference, but FSI customers could also be using them for training. (Supermicro is not privy to the details.)

One of the reasons why these GH200 NVL2 and GB200 NVL4 systems are popular is that Supermicro has designed them for metropolitan datacenters where there is no liquid cooling available and the power is capped at 15 kilowatts per rack. Many of the banks, hedge funds, and proprietary traders park these systems in co-location facilities.

One important thing that separates the banks from the hedge funds and proprietary trading companies is the Volcker Rule, which is part of the Dodd-Frank regulations that were put into place in the wake of the Great Recession to curb the proprietary trading by commercial banks. This rule went into effect in 2015, just as the first wave of the AI revolution with traditional, statistical AI was taking off. The banks could not use deposit money to trade anymore, but the hedge funds and proprietary trading companies can because they have their explicit permission to do so. Hence the rise of these companies and the enormous wealth they have created from the risks that they take.

The big commercial banks are using GenAI for the kind of textual work that large language models are good for – regulatory compliance and reporting, and they are toe-dipping a bit on smaller machines in many cases. Banks are also using GenAI applications to manage their loan portfolios: it can read loan agreements and figure out from news feeds when something has gone wrong and when collateral needs to be called in.

Some hedge funds are only now experimenting using GPU-accelerated systems running GenAI to boost the intelligence of algorithmic trading. Some of Supermicro's hedge fund

customers are buying big CPU systems based on AMD “Turin” Epyc 9005 CPUs and Intel “Granite Rapids” Xeon 6 CPUs systems to do their trades.

But Nvidia, which has a bird’s eye view of the financial services market, tells us that some of these trading firms buy clusters with 1,000 or more GPUs and they also buy enormous time series financial datasets – datasets that are not available on the Internet but from the capital markets themselves, hundreds of terabytes of data each day – to have GenAI models figure out trading patterns instead of hand-coded algorithms created by quants, who learned the hard way that interest rates are mean-reverting, commodities are cyclical, and equities follow random walks.

There is an important thing going on here. With GenAI, trading companies are figuring out how to trade smarter, not just to trade faster like the high frequency trading that took off in the 1990s, executing trades in seconds, and a decade later, you had to goose your systems with FPGAs and blazing faster CPUs and networks to get it down to microseconds to compete. Because it is hard to go *faster*, trading companies need to pick the moment of buying and selling their stocks and bonds *better,* and that means replacing simpler regression models with something more sophisticated and malleable.

This is why trading companies are at the forefront of the GenAI revolution, and it is also why some of them are buying relatively large GPU clusters. And they are buying rather than renting because it is hard to find available GPU capacity from the big clouds and even among the specialist GPU clouds. So trading companies cram GPU systems into their datacenters if they can and park them in a co-location facility if they can’t.

Among the bigger clusters that Supermicro is selling into FSI companies, about 20 percent of them are liquid cooled, and eventually, we think liquid cooling will be normal, reverting to a mean that was established back in the 1960s and 1970s when the mainframe was the hot system – literally – for commercial applications.

As GenAI takes off in FSI, most systems will be liquid cooled because a dense rack with 144, 288, or 576 GPU chiplets with a single shared memory space will drive down inference costs. And while training systems may be more spread out, the efficiencies that come through liquid cooling – and the better total cost of ownership or rental – will win out. So the financial services industry has to plan for that future when a training node or an inference node – really a rackscale system – might be pushing up against 1 megawatts of power, and all of that heat has to be extracted back out of the system and it has to go somewhere. Liquid cooling is the most efficient way to do this, and the heat can then be put to other uses within a metropolitan area.

In other words, financial services firms will be among the first to widely deploy AI supercomputers and to adopt agentic AI architectures to create new applications and new ways of cultivating our money so it grows. And eventually, every industry will follow suit, as they have in past computing revolutions, and supercomputing will be. . . *normal*.