ARTICLE

For Financial Services Firms, AI Inference Is As Challenging As Training

By Timothy Prickett Morgan, *The Next Platform*

A decade ago, when traditional machine learning techniques were first being commercialized, training was incredibly hard and expensive, but because models were relatively small, inference – running new data through a model to cause an application to act or react – was easy. Almost an afterthought.

As the early commercializers of traditional machine learning and now generative AI, companies engaged in the financial services industry – commercial and investment banks, trading companies, insurance companies, and the like – are coming up against the somewhat surprising fact that GenAI inference is in many ways more difficult – and varied – than training. Models have to be fit for purpose, and they also sometimes have to fit inside the smartphones and laptops we carry around, or the edge devices in bank branches, to provide low-latency processing. Sometimes, AI models are doing something so complex that it requires big compute, memory, and storage, which means it has to run back in a datacenter and an application has to compensate for the low latency this will cause.

By necessity, FSI companies will have a wide spectrum of inference running on a diverse set of compute engines and the storage that feeds it – and storage cannot be an afterthought anymore, as it has been with other kinds of high performance computing in the past. Storage systems will maintain context to avoid computation wherever and whenever possible and thus make inference cheaper than it might otherwise be.

And thus, it is important for all industries to learn from early adopters like FSI companies. The pity is that FSI companies are very secretive about what they are doing with AI and how they are implementing it to be reliable and affordable.

Well, maybe it is not a bad thing that they are secretive. After all, these companies handle our most precious data and financial assets. Maybe we don't want them to be too chatty about what they are doing...

In any event, across the broader spectrum of the global economy, every organization will not have to train AI models, but every organization absolutely will be running a variety of inference. Some of it will run on CPUs with vector and tensor engines, some of it on GPUs, and some of it on FPGAs or custom ASICs.

Let's talk about the workloads that are being enhanced with AI inference in the financial services industry and take a look at some early examples from JPMorgan Chase and Wells Fargo.

"It comes down to three different categories where the use of AI is being explored," Michael Watson, general manager of field application engineering at Supermicro, tells *The Next Platform*. "The first is traditional quant finance, where you're looking at investment risk, actuarial assessments for insurance, and back testing of algorithms used to make trades or investments against historical data to make sure they work but do not overfit to that historical data. The next category of AI exploration and deployment is in underwriting and analytics, and using alternative data streams is something that's newer in the sense that you can look at news and sentiment, satellite imagery, video feeds, and other data to get a sense of what the market is thinking and doing using AI. This category also includes operational applications like fraud detection, claims and margin settlement, and intelligent automation of documents like contacts, claims, and so forth. And then the last category is the customer experience, where companies have conversational AI, natural language processing, chatbots, personalization, recommendation engines, and agentic advisories. These are a lot of the things that Supermicro and Nvidia are pitching to the financial community right now."

Another one that we will toss into the works, and which is completely internal, that financial services companies are absolutely interested in and deploying in varying degrees are code assistants and code porting tools to help them write new code or modernize old code at their heart of their core banking systems, many of which are still written in COBOL and running on IBM mainframes.

Watson says that a lot of financial services companies are very interested in deploying GenAI, and particularly with Retrieval Augmented Generation, which pipes internal a mix of structured, semi-structured, and unstructured data into a pretrained model as context to perform inference to drive tasks. Having a conversational overlay to query and analyze documents is a killer app, and not just for internal uses, but for customer facing ones. That said, FSI companies are still working out the return on investment for all of these GenAI use cases.

Taking A Cautious Approach To AI

JP Morgan, the investment banking arm of financial services giant JPMorgan Chase, has been using traditional machine learning and deep learning for risk management and fraud detection for years, but only **launched its first GenAI tool, called IndexGPT, back in July 2024**. The Quest IndexGPT tool is based on OpenAI's GPT-4 large language model and is one of the many tools in its Quest framework, which helps institutional clients build sophisticated and thematic indices for stocks. Specifically, the IndexGPT model is used to generate key words for any topic investors want to track – AI, cloud computing, cybersecurity, sports, renewable energy, whatever – and then those key words are used to comb the Internet for those keywords and bring together current information related to those keywords. It automates the kind of work that investors used to do by hand.

JP Morgan has over 5,000 indices covering all major asset classes – stocks, bonds, real estate, commodities, derivatives, and non-traditional assets like private equity and venture capital – but this is the first customer-facing one that is based on an LLM. The keyword generation, says JP Morgan, is better than the prior methods it used, and results in a more accurate representation of themes. The AI-generated indices were made available to selected customers last May and were made available through the Bloomberg and Vida trading platforms.

By the way, JP Morgan is very clear that the indices and keywords are static and GenAI is not being used to dynamically update the components of any index. It is reasonable to expect that such indices will eventually be dynamic and IndexGPT will be used by investors to create their own personalized indices with their own keywords. Why not? But for that to happen, we presume the cost of inference has to come way down.

Bank of America and Wells Fargo both have created financial assistants, called **Erica** and **Fargo**, respectively, that are driven by AI models.

BofA was a very early adopter of AI, launching Erica in 2018. It took four years to do one billion customer interactions through Erica, and reaching the second billion in April 2024 only took 18 months. We presume more than another billion interactions have been done in the past year. Through last year, more than 42 million clients had used the Erica tool, which monitors and manages service subscriptions, helps customers understand their spending and pay their bills, and does basic customer service help like looking up routing numbers. **Erica has handled 2.6 billion client interactions and has 20 million active users**, and was used during the coronavirus pandemic to automate password resets, activate devices, and perform other customer support functions.

The interesting thing about Erica is that it uses machine learning and natural language processing, but it does not use large language models and GenAI token generation to act. Rather, Erica has a proscribed set of responses to questions that BofA customers ask. It is only available through mobile banking as well. We presume the machine learning behind Erica is accelerated by GPUs.

With the Fargo app launched in early 2022, Wells Fargo added a concierge-like interface to its mobile banking, and it most definitely is using GenAI techniques. Interestingly, Fargo does speech to text transcription locally on the mobile phone using a tiny LLM, and then another LLM in the app scrubs anything you say of personally identifiable information, or PII using security lingo. A separate layer of the Fargo app then calls out to a remotely running Google Gemini Flash LLM, which has been tuned with the relevant information about banking at Wells Fargo.

Wells Fargo is partnering with Google Cloud to run the models behind Fargo, and in fact it started out using Google's PaaLM 2 LLM, but also has partnerships with Microsoft Azure and also uses Google's Gemini Pro, Anthropic's Claude Sonnet, OpenAI's o3, and Meta Platforms' Llama models for its GenAI back-ends. The choice of model often comes down to the size of context window for the application and the response time needed.

Fargo handled 21.3 million interactions with users in 2023, and use exploded by more than an order of magnitude to 245.5 million interactions in 2024. That's a lot more inferencing, and the growth is probably not going to abate.

Hence, the intense focus on driving down the cost of inference among the FSI companies.

The Ever-Embiggening Inference Node

Before the GenAI wave hit at the end of 2022, "classical" AI models could be trained on hundreds to thousands of GPUs and the resulting model could usually fit, along with its weights, into the memory of a single GPU. With GenAI, the size of the training clusters scaled linearly with the number of input tokens thrown at them, and as model weights got larger, foundation models got larger and got higher scores on tests given to them. It wasn't long before training runs were done on tens of thousands of GPUs for several months and inference needed systems with four, eight, or sixteen GPUs to hold the resulting models and weights.

With chain of thought reasoning models, the expectation is that it will take on the order of 100X more compute to have interlinked models break down a problem and only activate

certain of its dozens to hundreds of its constituent models and thoughtfully consider a complex problem. It takes longer to do chain of thought reasoning, even when you throw a lot of iron at it, but you get better answers. To get faster answers requires machinery optimized for the simultaneous running of many different smaller models working in concert. Something akin to the GB200 NVL72 rackscale system that is now shipping in volume and its future GB300 NVL72 that will be shipping later this year.

The GB300 NVL72 is a reference design that is made by Supermicro and other OEM and ODM system makers, and has 72 "Blackwell" B300 GPU accelerators (which means 144 distinct GPU chiplets) and delivers a combined 1.1 petaflops of dense FP4 inference compute.

Next year, **Nvidia will deliver its 88-core "Vera" upgrade to the 72-core Grace Arm server CPU and marry it to its "Rubin" GPU socket**, which will also have a pair of GPU chiplets. This system, called the VR200 NVL144, will offer 3.6 exaflops of FP4 oomph for inference and 1.2 exaflops at FP8 precision for training.

These are both single rack, shared memory systems tailor-made for chain of thought inference. This is not the kind of inference a low-end GPU card can do, or a vector or tensor engine on a CPU can do.

Here in 2025, the state of the art in AI inference for GenAI models is obviously a long, long way from fitting models into a single GPU, even if programmatically speaking, these two rack-scale systems look like a giant, single GPU to AI applications. There will still be plenty of models that will be run in machines with two, four, or eight GPUs – and this will certainly be the case for financial services companies with datacenters in close proximity to large metropolitan areas where power and power density are limited and liquid cooling is not available.

But in the long run, the kinds of machines that Nvidia is delivering in 2025 and 2026 for very complex inference will be more widely adopted. There will be alternatives, with AMD and possibly Intel GPUs and memory interconnects from them based on UALink that can create rackscale or even rowscale machines for doing very large inference workloads.

It remains to be seen how aggressively FSI companies will adopt chain of thought models and the big iron it will require. A lot depends on how financial institutions demonstrate that they can trust the output of the larger and more sophisticated GenAI models. Banks will be understandably cautious, being heavily regulated as they are, while hedge funds and proprietary trading firms will stay on the bleeding edge.

Storage Cannot Be An Afterthought

One of the bad habits from corporate computing and high performance computing is that storage in a system is often an afterthought. With AI inference, storage is important, and when it comes to speeding up inference and driving down its cost, key-value caches and context window caches are the hot new items.

With these, the underlying storage in an AI system stretches the memory space of the cluster with the DRAM and flash in a storage cluster to help lighten the inference load on GPU memory by storing the state of key-value vectors used to generate tokens so these key-values do not have to be recomputed with each token generated.

Moreover, persistent memory can store the context in a query so that repetitive tasks based on the same context do not have to recompute their answers and hammer the GPUs to answer a query that has already been posited.

“You could have a really crowded multi-tenant system, and you can extend the context length outside of the AI system to persistent memory with, for example, Vast Data Platform with NFS over RDMA,” explains Jeff Denworth, co-founder at Vast Data, one of the handful of flash storage upstarts that is tackling the multi-faceted data issues with AI workloads. “Now you don’t have to go bonkers on GPU memory. And if you had to evict a user because they were idle, you don’t have to recompute the entire session once they come back. This is important because the cost of computing grows quadratically with the length of the context window. And so people are storing state on the network and pulling it back in when they need it.”

Another thing that has to be considered with AI inference is how data gets orchestrated to be at the right GPU for processing at the right time. Hammerspace solves this problem by creating global metadata for all of the storage the AI system has access to, and then treating local flash storage inside of the GPU server nodes as a Tier 0 distributed file system. The Hammerspace Data Platform orchestrates the movement of data into this Tier 0, which then feeds inference jobs.

“We can do that because we have a global file system,” Molly Presley, senior vice president of global marketing for Hammerspace, explains. “And all those NVM-Express devices we call Tier 0 are just part of a big storage pool to us.”