Large Memory Servers for Real Time AI Inferencing in Financing Markets

We are quickly moving from the fun stage of AI models to serious enterprise class AI applications. In the world of finance, AI can be highly beneficial in a number of ways, including providing real time trend predictions, fraud prevention and credit scoring to name a few. Perhaps one of the most significant additions to the world of AI inferencing, especially for enterprise class data centers, is the migration from batch thinking to real time in terms of data processing. AI models like ChatGPT train on massive data sets that not only take weeks to process, but are typically based on historical data or are incapable of incorporating data sets without rerunning the models.

AI, Memory and Financial Applications

One of the critical issues for finance, in particular, is performance or low latency access to data records and other applicant related data. Responses are needed within milliseconds, not seconds, hours or days. Real time AI presents some significant challenges for the financial world, where data has to be highly accurate, up to date and responses are needed “instantly” in order to make more informed decisions at most levels. Real time data includes “online” data, i.e. what the user has been researching on your site and other readily available online data based on recent activity. This is in contrast to static features that do not change very often, such as name, gender, address, job, age and so on.

All of this leads to a hybrid model which requires both model generation and real time data stored in feature stores in fast in-memory databases closely coupled with AI interference and local model generators. Translating raw data, in particular, requires substantial feature engineering to ensure that data is optimized to create a fast predictive model.

New Technologies Making Real Time AI Possible

To help with the transition to real time, enterprise AI is adding in-memory database servers to reduce data access latencies and adding new capabilities such as feature stores for online prediction services, vector oriented databases capable of serving several millions of requests per second, and integral front end intelligent caches for repeat queries in order to meet some pretty stringent real time response goals.

Local Memory Makes all the Difference

One clear metric emerges in most cases where a large number of users and datasets are involved, and that's memory. Today, when attempting to increase the size of an in-memory data base, the answer once a single CPU's worth of memory has been maxed out, has been to add more CPUs and memory. This is problematic, in that as soon as this starts to spill over to multiple server boxes, then network latency and overhead quickly start to degrade the performance of data accesses.

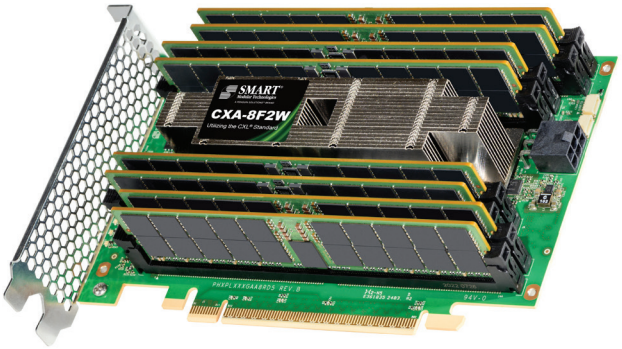
CClearly, the larger you can make the memory in a single server running the database, the less chance there is of having to suffer delays while the application "reaches" over the network.

Enter CXL®, or Compete Express Link. CXL is designed to solve the problem of running out of memory in a single server by allowing memory to be added by the CPU's peripheral I/O bus, more commonly known as PCIe. Starting with PCIe Gen 5, a new protocol layer has been added that is tightly coupled with the CPU's memory architecture allowing additional memory or memory modules to be added to a mainstream single or dual socket server.

Just How Much Memory Can I Put in a Standard Server?

A modern PCIe Gen 5 AMD or Intel CPU can support either 12 or 8 RDIMM memory modules attached to each processor in the system.

So for a 2-socket AMD server for example, this equates to 24 RDIMMs. Using 256GB RDIMMs, this provides a total of 6TB of memory in a server. Given 256GB RDIMMs are using expansive stacked memory dies, it's more common to find either 64GB or 96GB RDIMMs in most servers given their significantly better cost point. So 1.5-2.3TB is more typical for servers.



Enter Penguin CXL Memory Expansion

The upcoming family of Penguin big memory server solutions are capable of supporting one or more double wide PCIe Gen 5 CXL 8-DIMM add-in-cards. For example, in our 4U dual socket servers, we can now support up to a massive 22TB of memory in a single Penguin server at the high end, 6-11TB for more mainstream applications that don't wish to use expensive memory modules. Either way, with CXL, the option is yours.

A Comparison of Total Memory Options in Penguin's Big Memory Servers

Server Capacity	CXL Cards	Total DIMMs	64GB DIMMs	96GB DIMMs	128GB DIMMs	256GB DIMMs
4U Dual Socket AMD	8x 8-DIMM	88	5.6TB	8.4TB	11.3TB	22.5TB
1U Single Socket	4x 8-DIMM	44	2.8TB	4.2TB	5.6TB	11.3TB

SMART® Modular Technologies
A PENGUIN SOLUTIONS™ BRAND

Learn more at : **www.penguinsolutions.com**

Headquarters/North America:
T: (+1) 800-956-7627 • T: (+1) 510-623-1231
F: (+1) 510-623-1434 • E: info@smartm.com

Latin America:
T: (+55) 11 4417-7200 • E: sales.br@smartm.com

Asia/Pacific:
T: (+65) 6678-7670 • E: sales.asia@smartm.com

EMEA:
T: (+44) 0 7826-064-745 • E: sales.euro@smartm.com

Customer Service:
T: (+1) 510-623-1231 • E: customers@smartm.com