SMART®
Modular Technologies
A PENGUIN SOLUTIONS® BRAND

White Paper M-WP019

Big Memory Requirements for Next Generation Fintech Machines

Building Efficient and Scalable Fintech Computers using CXL®

February 2025

Table of Contents

Abstract

Fintech computers are adopting larger amounts of system memory as they need to handle increasing amounts of alternative data such as social media, satellite images, and so on, in order to support AI/ML enhanced services or near real-time data analysis for decision-making or automated trading. Traditional methods of scaling memory have involved novel RAM plus SSD tiering schemes or increasing CPU sockets in servers, but these are quickly becoming inadequate to meet the demands of modern trading systems and predictive, AI/ML assisted trading.

In this paper, we discuss the impact on fintech of a new memory interconnect standard known as Computer Express Link or CXL®. We explore how it is becoming essential to have a strategy for cost effectively increasing the amount of memory to move many of these systems from batch processing to real time recommendation, natural and large language-oriented processing engines. We explore how CXL changes the landscape in terms of scaling memory, allowing fintech system builders to greatly increase the responsiveness of their systems while adding the increased data capacity for their In-Memory storage.

Fintech and Quants

First, a few definitions.

Fintech, short for financial technology, refers to technology used to help automate or enhance financial related decisions or reporting. Example industries utilizing fintech are banks, brokerage or hedge funds, which implement increasingly sophisticated high frequency trading systems, background checks, fraud risk detection, personal financial management, business transactions, investment decisions, etc. Such systems have been in use for many years in the financial world and have evolved more recently to take advantage of natural language processing enabled by AI and machine learning (ML).

A Quant, short for quantitative analyst, is a professional who uses advanced math, statistics and machine learning/AI to analyze and predict financial market trends. Investment banks, hedge funds, commercial banks, insurance companies and financial information consultants are increasingly employing quants to enhance decision-making and drive improved results in financial returns for their customers.

Fintech and Predictive Models

For Quants and Fintech, extensive use is made of alternative data sets tied to automated and direct trading environments where humans are no longer involved in the direct purchase or sale of securities. The benefits are that these fintech systems can make very fast financial trades (hence the term high frequency trading or HFT) and may also take advantage of micro-variations in stock or commodity prices. With the addition of AI/ML, these systems are now being augmented to provide predictive services or recommendations.

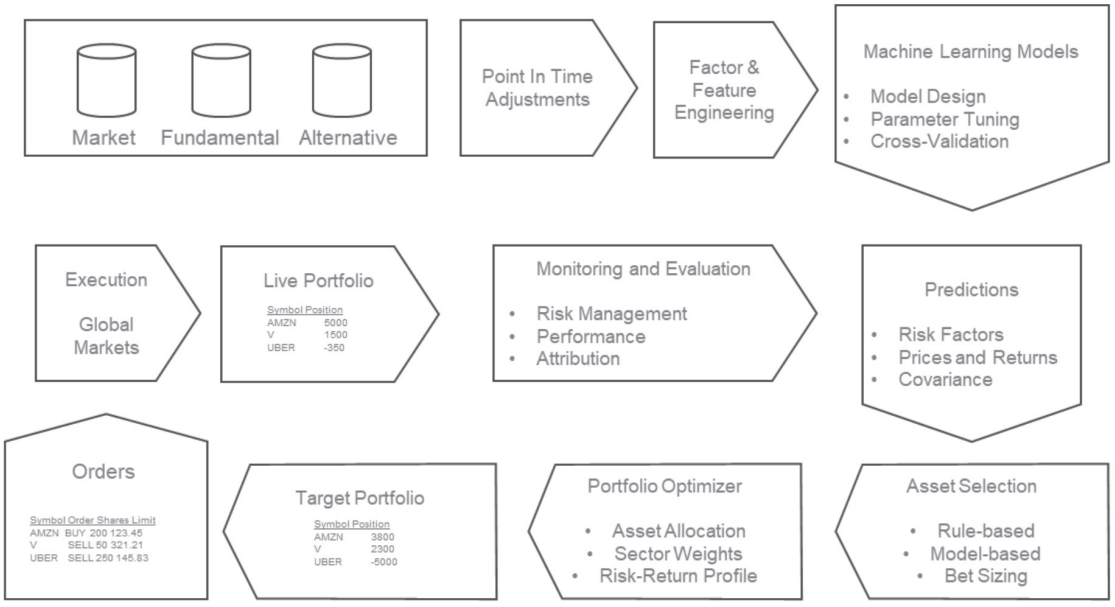
As competitiveness increases among financial institutions, modern fintech systems rely increasingly on ingest of real time market data from various web and alternate sources combined with real time streams from securities data from financial exchanges. AI and machine learning are used to create predictive models that can help quantify risk in commodity trading, high frequency trades or credit worthiness decisions for loans, etc.

The types of alternative data types now processed in fintech environments include:

- Social media data, reviews and postings from individual users
- Business analytics data such as commercial transactions and supply chain trend data
- Sensor data i.e. satellite images, security cameras, movement data via cell phone towers

For example, Citadel, an alternative investment management firm, relies on weather prediction from sensor data to help predict the price of commodities and determine potential risks factors that may affect these commodities in the future[1].

The adoption of AI/ML in the world of quants is part of the unique product offering provided by each of the financial institutions, so little is available in the public domain on what specific algorithms or approaches are used. However, some publications do cover the general types of processing that is performed. For example, for those interested in learning more, a ML4T Workflow is outlined in the publication *Machine Learning for Algorithmic Trading*[2]. This industry continues to evolve as AI/ML models and natural language models are increasingly adopted along with real time streaming data-bases and processing schemes to provide enhanced services and automation.



Example Machine Learning Flow

(source: Machine Learning for Algorithmic Trading 2nd Edn, Stefan Jansen)

[1] Taking Weather Prediction to a New Level. https://www.citadel.com/what-we-do/

[2] Machine Learning for Algorithmic Trading 2nd Edn, Stefan Jansen published in 2020

Memory and Fintech Computers

As more real time data is ingested into and processed by a fintech computer, having more memory available to the CPU means it can process data faster (i.e. more responsive), handle more simultaneous users, and of course, handle more complex models with the assistance of hardware accelerators such as FPGAs and GPUs.

Until now, the traditional methods for increasing the memory size in a system have been limited to the following:

1. Utilize 3D stacked die DRAMs, e.g., 256GB or 512GB RDIMMs to create a large amount of memory per CPU
2. Adopt a RAM-SSD tiering scheme to create a larger virtual memory
3. Add more CPUs plus RDIMMs through the use of 4 or 8-way socket servers
4. Add more networked 2-socket servers via a low latency network to create a cluster.

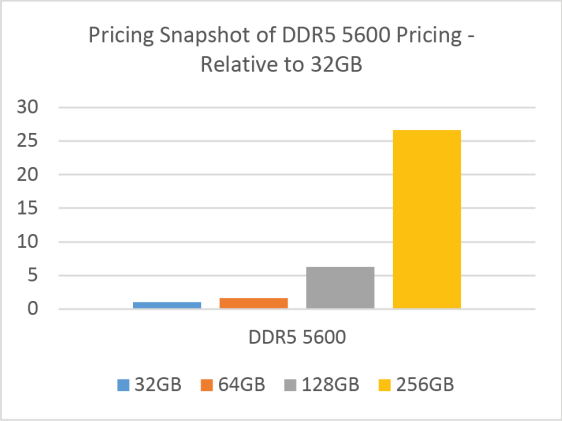
A mix of the above are often used depending on the size of jobs and numbers of users supported by the fintech machine or cluster.

Let us look at each approach and compare how CXL offers a complementary or alternative approach for fintech or any large memory server application. Note, not all applications are created equal, so the following serve more of a general comparison.

3DS DRAMs

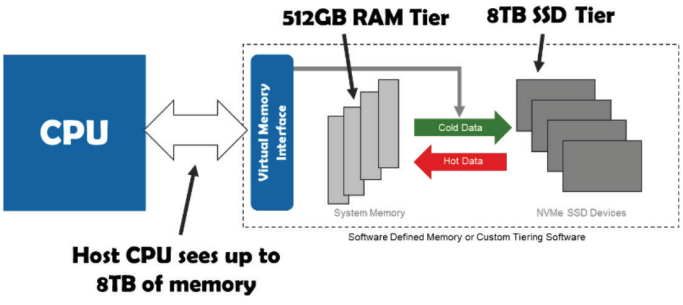
3D-stacked DRAM is a technically innovative way to solve capacity limitations of monolithic die versions of DRAM. However, due to manufacturing yields of the stacking process, the price of 256GB RDIMMs are currently more than 3-4x the cost of their prior generation, especially for higher speeds.

This can drive up overall system costs by 100-200% depending on memory requirements given memory is a significant cost component of a fintech computer.



RAM-SSD Tiering or Software Defined Memory

A practical way to increase the size of memory is to use a software-based scheme to create a virtual memory space for the application to access that consists of RAM plus solid state disks or SSDs. An individual SSD can be as large as 32TB, currently, with 64 and 128TB versions in the pipeline. So it's relatively easy to create 8, 16 or 32TB virtual memory systems that are a hybrid of RAM and SSDs, with intelligent software that ensures hot or frequently accessed data is operating out of the RAM "tier" of the virtual memory, while cold or infrequently accessed data is stored in the SSD portion.

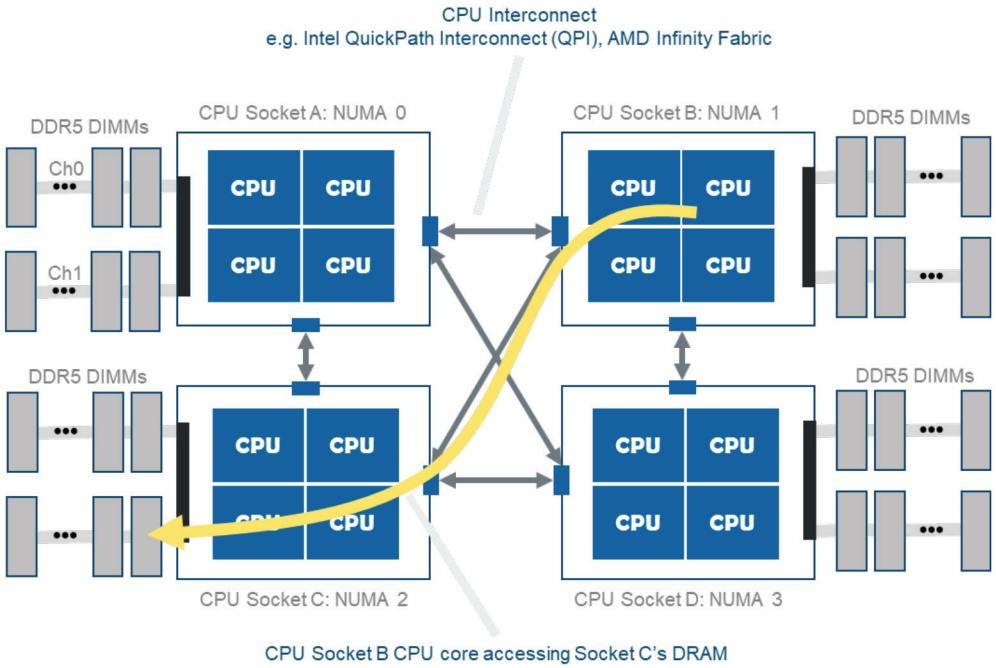


Such systems have served the industry well over the past few years as memory requirements grew. However, as AI/ML models grew combined with the rapidly changing random data sets, such tiered systems become less effective and, in fact, start to slow things down as data "spills" over to the much slower SSD tier. SSDs, while fast, are still up to 1000 times slower from a data access point than main memory.

Scale up Memory Using More CPUs

Another practical way to scale up memory has been to simply opt for 4 or 8-way socket systems. Modern CPUs have sophisticated “fast-paths” between CPUs to allow memory attached to each CPU socket in the system to be accessed via any of the other CPUs, i.e., each CPU shares its memory with neighboring CPUs. The memory, in this case, is referred to a NUMA or Non-Uniform Memory Access. Memory accesses to local memory are the fastest, and when accessing a neighboring CPU’s memory via one or more NUMA “hops”, accesses may take up to twice as long from a latency perspective, but still dramatically faster than going to SSDs or accessing data over a network for example.

Again, one of the challenges with scaling up memory this way is that we are often over provisioning the CPU power, i.e., these expensive CPUs are often used simply as data cops or memory expansion devices. Not an efficient use of expensive CPUs.



Scaleout Memory Using Low Latency Networks

One other practical method of scaling up memory and compute at the same time is to use a scaleout network such as Infiniband and a cluster of lower cost 2-socket servers. For AI/ML applications, these 2-socket servers are often loaded with two or more GPUs to accelerate AL/ML model generation or perform inference tasks in near real time.

[3] Visit https://computeexpresslink.org/ for more information.

Compute Express Link®

CXL is a new industry standard for connecting memory and other acceleration device buffers to a CPU via the PCIe bus versus the more conventional memory bus. For those wanting to learn more about CXL in general, the CXL Consortium who produced the standard, provide a rich set of use cases and articles.

CXL was first introduced by AMD with their EPYC Genoa in late 2022 and Intel with their 4th Generation Xeon Sapphire Rapids CPUs in early 2023. However, while the CPUs themselves supported CXL 1.1 functionality at that time, it has taken two years for the rest of the eco system to catch up and provide add-in CXL products and BIOS support that take advantage of the new standard and provide deployable commercial products. Server vendors are now providing support that is more complete for CXL as well as quickly migrating to v2.0 of the CXL 2.0 standard. At the time of writing, AMD has just launched their EPYC Turin class of CPUs and platforms that support CXL 2.0 and Intel has launched their Granite Rapids family of CPUs also with CXL 2.0 support and we expect the first CXL 3.1 based solutions to come to market by late 2025, early 2026.

Mainstream tier 1 server vendors are actively working toward release of products in 2025 supporting CXL, along with supporting peripherals such as PCIe add-in-card formfactor CXL RDIMM expansion cards and to a lesser extent, E3.S 1T and 2T removable CXL expansion modules.

CXL also has a rich set of features for support cache-coherent cache protocols that are expected to also help address many network accelerators in terms of enabling their local data buffers to integrate tightly with the CPU cache hierarchy. Not a topic for this paper, but we expect this to become another important of CXL especially for fintech applications in the not so distant future.

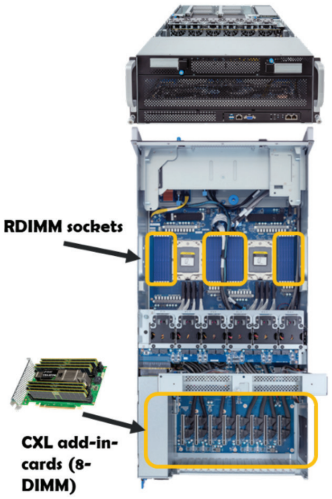
Fintech and CXL®

The key question is how can fintech community take advantage of CXL?

For those in the fintech community simply wishing to maximize the amount of memory a single CPU socket can address, CXL 1.1 or CXL 2.0 based systems are perfectly adequate. By using some of the PCIe slots in the server for memory, up to 1TB can be added to an individual x16 slot for example, soon to be 2TB and higher. If using an E3.S server, it is also possible to add between 128GB to 1TB using a server with E3.S CXL enabled drive bays.

For a single compute environment with a small number of users, a 2-socket server today is hardware limited to around 6TB worth of RDIMMs, assuming the most expensive 256GB 3DS DIMMs are used. This is because a server typically comes with no more than 8-12 RDIMM sockets per CPU in the system.

MAX 2-SOCKET SERVER MEMORY CAPACITY	AMD TURIN CLASS SERVER (12 PER CPU)	INTEL GRANITE RAPIDS SERVER (8 PER CPU)
64GB RDIMMS	1.536TB	1.024TB
96GB RDIMMS	2.304TB	1.536TB
128GB RDIMMS	3.072TB	2.048TB
256GB RDIMMS	6.144TB	4.096TB



When adding CXL memory devices, the amount of memory will depend on the available PCIe slots in the system that support CXL natively or E3.S CXL capable front bays. For large capacity CXL cards such as the SMART Modular 8-DIMM add-in card, servers designed for multiple GPUs are typically best adopted as they are already designed for the proper airflow and power connectors.

As an example, a 2-socket server can support up to 4 x16 PCIe slots per CPU, allowing an additional 64 DIMMs to be added to the server via CXL. This provides an additional 8TB using 128GB RDIMMs. So now the total memory for the server has been increased by the following amounts:

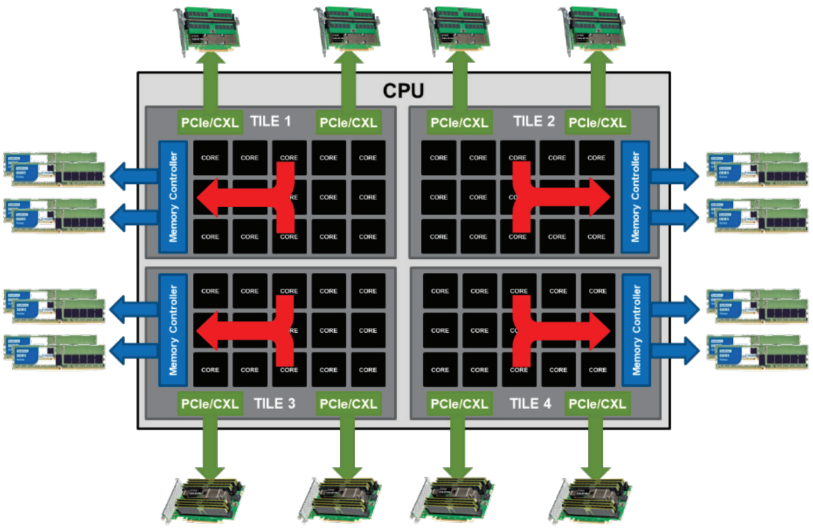
As we can see, the total available memory has now been increased to around double for the case shown above.

MAX 2-SOCKET 8-SLOT CXL SERVER MEMORY CAPACITY	AMD TURIN CLASS SERVER (12 DIMMS + 32 CXL DIMMS PER CPU)	INTEL GRANITE RAPIDS SERVER (8 DIMMS + 32 CXL DIMMS PER CPU)
64GB RDIMMS	2.816TB	2.560TB
96GB RDIMMS	4.224TB	3.840TB
128GB RDIMMS	5.632TB	5.120TB
256GB RDIMMS	11.264TB	10.240TB

Benefits of Adopting CXL®

There are several benefits to adopting CXL in memory hungry applications such as AI/ML enabled fintech systems.

Performance



The first benefit is improvement in system performance. One of the challenges faced by fintech systems adopting real time data streams for AI/ML processing is how to deal with increasing amounts of data ingested and held in memory for fast processing. If insufficient memory is available to hold the required data, the operating system memory management will page swap the contents of memory to persistent storage, such as SSDs, to make room for the new memory allocation requests. When page swapping occurs, performance often drops significantly leading to stalls in the applications worst case, or a slowdown of memory accesses to SSD speeds, i.e., up to 1,000 times slower. Worse still, memory access contention will occur between new incoming data and data swapped to SSD that decreases the performance further. In the extreme case, the operating system may decide to kill the running process to avoid a system crash that for financial trading is catastrophic leading to potential monetary losses.

A secondary performance benefit comes from collision avoidance. Modern CPUs have multiple cores that access memory by shared memory controllers. If multiple of these CPUs are attempting to access the main memory at the same time, the memory requests end up being queued up which often pushes the actual data access latencies of main memory up causing multiple cores to stall waiting for data. The CXL memory is on a different, albeit slower, highway: the PCIe bus which is accessed via a totally different data path by the CPU, and hence, it is often the case that the accesses to CXL memory can be faster than main memory under these heavy load conditions.

Lastly, by providing more memory to a single CPU can result in avoiding utilizing a low latency network e.g., Infiniband that can add significantly more latency as data has to be split up across several servers instead of residing in a single server's memory.

Cost

The second major benefit is that CXL makes memory costs more scalable on a $/Gbit basis. As we saw earlier, the only way to increase server memory is to use 3D stacked memory, which today is more than 2-3x the cost per Gbit than monolithic-die, or we have to add more CPUs. Both are inherently expensive.

CXL allows the system architect to trade-off using 3DS DRAM versus adding more monolithic DRAM via CXL DIMM slots. The cost of the CXL card is significantly lower than the cost of either 3DS memory or adding more CPUs, and can results in significant system cost savings.

Conclusion

Fintech, like other memory hungry applications, demands new ways to reliably and cost effectively scale up system memory directly attached to the CPU to address processing needs driven by AI/ML and real time alternative data sources. CXL provides an industry standard way to increase memory and increase performance in many cases by avoiding memory page swaps to SSD and providing another memory bandwidth channel outside of the standard CPU memory controllers. This enables fintech companies, and specifically those that require large In-Memory databases, to continue to scale their systems to exploit broader alternative data sources and gain competitive edges in the financial investment markets.

SMART® Modular Technologies
A PENGUIN SOLUTIONS® BRAND

For more information, please visit: **www.smartm.com**

**Product images are for promotional purposes only. Labels may not be representative of the actual product.*

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

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

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

Customer Service
T: (+1) 978-303-8500 • E: customers@smartm.com