AI update May 13, 2024

AAPL: Apple CPUs in the Cloud vs. NVDA GPUs

LYNX EQUITY STRATEGIES

KC Rajkumar, CFA, PhD
(408) 425-5680
KC@LynxEq.com

The arc of Apple Silicon over several generations of evolution intersects the need for power-efficient computation of Gen AI workloads at just the right time in the guise of Apple's 3nm-based M4 silicon, and its follow-ons at the 2nm process node. We were on-board the potential for M3/Mac in Gen AI usage prior to its launch (link). We pointed out the potential for M4/iPads for multi-modal Gen AI applications prior to its launch (link).

The potential for Apple Silicon to break into data centers was always there in our view, but such a dramatic break from AAPL's traditionally client-centric focus required a formal announcement from the Company for us to take the idea seriously. But now with media reports coming from two major outlets, we consider the Company has essentially made a soft announcement of its intention to allow the traditionally client-focused Apple Silicon into Gen AI data centers.

With Apple Silicon in data centers quarterbacking the delivery of Gen AI services to the 2+billion Apple devices, we think the Company is taking a major step in ***diversifying away*** *from its dependence on unit volume of iPhones and other Apple products, and instead leverage the power-efficient Apple Silicon hiding inside those devices to delivery high-margin Gen AI services.*

The idea of Apple silicon going head-to-head with formidable GPUs from NVDA will likely be met with a healthy dose of skepticism, until one starts looking at the technical details available in the public domain. We are not playing for Apple Silicon to target the GPT4 class of trillion parameter models. We do not think there is a need to hit that benchmark, if the focus is on mass-market AI applications based on smaller ***open-source class of LLMs*** *such as Llama3, Claude3 and sundry Hugging Face models.*

In comparing Apple Silicon to NVDA's GPU-based solutions, the key is not merely in a raw comparison of model size and TFLOPS, but also to address the issue of ***power consumption****. As we noted in our recent AAPL preview (link), the two companies branched off in different directions many years ago, with Apple's focus on lower power consumption. And today, when Gen AI cloud capacity expansion is being gated by the availability of utility power, water and ESG considerations, we think the journey Apple Silicon has taken over several generations could find resonance with data center operators.*

On a processor-to-processor basis, we think ***Apple silicon could already be competitive to NVDA's H-series*** *and possibly even the B-series class of GPUs. In conjunction with power consumption considerations, Apple silicon could be more than competitive. We provide four technical reasons below. A key detail – our checks show that Apple Silicon for data center servers could have the same packaging as Apple Silicon for client devices – no need for liquid cooling.*

--XX--

Apple Silicon's within-module memory – higher than NVDA's: We think investors are comfortable with the idea that the size of within-module 'unified' memory goes a long way in determining the processor

performance. NVDA's H100 SGX has 80GB of within-module unified memory, and it looks like the B100 more than doubles it to **175GB** per module (8 GPUs in a server gets you **1.4TB** per server).

Apple's M2 Max has 96GB, the M2 Ultra doubles it to 192GB. The M3 Max has 128GB. While an Ultra version of M3 has not been launched, its memory is likely 2x the Max at **256GB**. We do not have data on the memory of the just-unveiled M4 for iPad. However, according to a Bloomberg article, M4 desktop could be as high as **500GB**. Although the article does not specify, let's assume this is an M4 Ultra. There is no reason an Apple server would not place **multiple M4 Ultras** on a high-speed bus. Just four M4 Ultra gets you **2TB**, higher than NVDA's yet-to-be launched 8GPU B100 server.

Apple Silicon – server packaging is likely identical to client packaging: We think Apple's server module packaging could be the **same as Apple's client module**. This means 1) no need for special package design to accommodate liquid cooling ala NVDA's GPUs and 2) no need for the cumbersome and expensive CoWoS packaging ala NVDA's GPU.

We think Apple Silicon could avoid expensive 2.5D packaging such as TSM's CoWoS process, and with it, avoid CoWoS' high cost, low supply capacity and low yield. We think packaging of Apple Silicon could be a **key differentiator** vs. the GPU solution from NVDA and AMD. Why do we think so?

In terms of compute horsepower, and within the limits of publicly available data, the M4 in its Ultra version could be on par with NVDA's H100. NVDA specifies H100's performance as **34** trillion FLOPS at 64-bit precision. At the Apple event last week, Apple management stated the M4/iPad's performance at **38** trillion FLOPS, but without stating the floating-point precision.

So, 1) if the M4 client processor has comparable FLOPS to an NVDA H100 GPU chip, 2) the M4 Ultra packs more DRAM than an NVDA B100 GPU chip, and 3) Apple Silicon for client devices is based on standard packaging which does NOT involve cumbersome CoWoS packaging technology, then we suspect Apple Silicon for servers could use the exact same package as client modules. If Apple Silicon for servers avoid the 2.5D CoWoS packaging process and instead Apple uses the same 1D packaging it has been using for client processors, then that is a **substantial advantage in cost and in ease of supply**.

Apple Silicon – cuts out a major memory copy operation vs. GPU servers: Investors tend to forget that in the Windows/Intel framework within which NVDA operates, a GPU is considered a peripheral device by the operating system. This is a legacy from the old WinTel days.

What this means is that when data is transferred from one discrete GPU module to another module, the transfer is done via a **discrete main memory,** with the discrete CPU on the server board acting as traffic cop. What does this mean? The data being transferred is first read from the within-module GPU memory, then transferred across the PCIe bus and written on the discrete main memory, then read from the main memory by the CPU before the CPU sends the data to the receiving GPU module.

However, if the main processing unit is a CPU module, as is the case with Apple Silicon, there is no need for a main memory. The data to be sent for one CPU module is read from the within-module memory and then sent across the system bus directly to the receiving CPU module. The operation of writing the data on a main memory is eliminated.

Apple Silicon architecture likely eliminates the redundant memory-write operation, thus saving system power and reducing system latency.

Apple server racks- standard air-cooled CPU racks vs. water-cooled GPU racks: If Apple can use client processors in its servers as well and if client processors today are air-cooled (in fact the Apple Macbooks do not even need a fan for air-cooling), Apple servers could avoid liquid-cooling and instead use standard air-cooling available in traditional data centers. And this could deliver a huge advantage in terms of capex and opex to Apple data centers

Net/Net: The idea of Apple silicon going head-to-head with formidable GPUs from NVDA will likely be met with a healthy dose of skepticism from investors. We provide technical reasons above as to why we think Apple Silicon is likely comparable or even superior to NVDA GPUs for the kind of LLMs we think Apple Services are likely to focus on. With Apple Silicon in the cloud quarterbacking the delivery of Gen AI services to the 2+billion Apple devices, we think the Company is taking a major step in diversifying away from its dependence on unit volume of iPhones and other Apple products, and instead leverage the power-efficient Apple Silicon hiding inside those devices to delivery high-margin Gen AI services.

As investors mull over the potential for Apple Silicon in client devices, quarterbacked by Apple Silicon in the cloud, for the delivery of Gen AI Services, we expect a slow run-up in the stock into the WWDC event next month. We reiterate our $220PT.

Disclosures and Disclaimers

Lynx Equity Strategies, LLC is an independent equity research provider. The Company is not a registered broker dealer or investment adviser. No employee or member of the Company, or immediate family member thereof, exercises investment discretion over securities of any issuer analyzed in this report two days prior and/or two days after this report is issued. It participates in "Alpha Capture Systems" that seeks investment or trading ideas from the sell-side and may pay for the participation based on relative performance

Limitations of Information

Copyright, Trademarks, Intellectual Property