Earnings preview Aug 28, 2024

NVDA: Optimists vs. Skeptics

LYNX EQUITY STRATEGIES

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

*NVDA investors wade today into one of the more **confusing** earnings events in recent memory. During NVDA's 'quiet period', investors have been pummeled with a profusion of lightly sourced information with regards to product delays, alternate product SKUs, detailed failure analysis and so on and so forth. The supporting cast of analysts, channel partners and Asia media sources appear **confident** that none of these concerning details should put a dent to NVDA's revenue ramp profile, such is the strength of the end demand. As their confidence can only come from NVDA itself, we suspect the company had messaged the gist of what is to be said on the call today through channel partners.*

*As such, we expect the message on the call today to be aligned with that from the supporting cast - a minor delay in the revenue ramp of Blackwell **to be offset** by higher-than-expected demand for Haswell SKUs. With the stock back to its previous high, we can safely assume this view is mostly **priced into** the stock. What could nudge the stock into a higher orbit? For the stock to launch into a higher orbit, management needs to communicate that the demand for Haswell **more than offsets** the revenue delay from Blackwell ramp. In one form or another, we expect the management to communicate just that. The first reaction to the earnings call may well be to the upside. But that could be a head fake.*

--xx--

NVDA has become a 'battle ground' stock: The stock came under severe pressure in mid-July as the **Skeptics** questioned the wisdom of vast amounts of capex spent by hyperscale players chasing what the skeptics claimed to be less than optimal ROI. A few weeks later in early August, the hefty revenue outlook from SMCI encouraged the **Optimists** to return, driving the stock back up. The resolution of the battle, we suspect, will emerge only in the days and weeks following the earnings event today as investors mull over the details of the call and as management resumes meeting with investors.

We are in the Skeptics camp: We do not support the idea that increased shipment of H100/H200 could make up for the delay in Blackwell shipment. All things being equal, we think it appropriate to adjust NVDA's data center revenue ramp **downward** vs consensus estimates until Blackwell gets into volume shipment, which could be well into NVDA's FY26

Our view - 1) the constrained supply of **H100** earlier on in the year may well have turned into excess supply as demand stabilizes, as seen ease of availability, falling rental/token pricing, an emerging secondary market, 2) **H200** may have good demand, but could suffer from low yield and heat-related failures (link) and 3) there is no way to assess with any precision the actual delay in volume shipment of **Blackwell**; it could be more than just a few months. Our view - heat-related failures could be endemic to GPU/interposer modules, which gets worse with higher HBM content. Blackwell has higher HBM content than Haswell.

A caveat: Despite potential issues with the H200, we suspect that ODM/OEM partners **may absorb sales from NVDA** and place the volume on their inventory. NVDA management may claim sales of H200 GPUs have picked up, but we are not yet seeing signs of H200 racks being absorbed into data centers in a big way. A rather curious article in Digitimes last week related to Foxconn caught our attention. The article talks about 'complex AI server trading models' and the moving away from the simple 'buy and sell' model to more complex models, which we suspect involves leasing the GPUs to end customers, rather than outright sales. Such trading models usually result in **pulling in future 'real' demand** and in build-up of channel inventory.

H100 shortage may have turned into surplus: H100 lead times have come in. The question is whether increased supply of H100 is keeping pace with increased demand. We do not think so. We think **supply may be outstripping end demand**. We are seeing several signals that help us draw the conclusion.

Three months ago, we noted that HGX H100 server rental, as reported by GPU-specialized data centers had dropped to ~$2.25 per hour vs. ~$4.75 earlier in the year (link). Our latest checks show that pricing has dropped further, now running below $2, at levels where smaller DCs may well be pricing server rental **below operating cost**.

VC platforms in Silicon Valley which bought GPU rental time contracts at wholesale prices earlier this year for the use by AI startup clients, we understand have begun to offload time slots, due to reduced demand from their client companies.

Blocks of H100 GPUs have begun to appear on the **secondary market** in Silicon Valley as tactical buyers of the highly valued chips now face reduced demand and falling value of the chips. We are aware of a secondary market for H100 in **Hong Kong**, which is now reporting declining price.

As for the hyperscale players, there no longer are constraints to the availability of MSFT's CoPilot; users can gain access at will. We are also hearing scattered reports of some hyperscale CSPs, **slowing down or even pausing** purchasing of H100.

H100 – why the surplus? In short, because hyperscale CSPs may have, to paraphrase an expression from the Google earnings call, converged on a set of base capabilities i.e. trained models, sooner than NVDA had anticipated. And this releases a large amount of H100 installed base capacity formerly allocated to training now to inferencing.

Two quarters ago, NVDA management commented that inferencing was running at ~40% of the workload on NVDA GPUs. The implication was that the training workload was still more than a majority of the workload. Training workloads required a huge commitment of GPU cluster size and continuous usage time, thereby sopping up large amount of GPU (mostly H100) capacity.

Today, we think training workload usage of H100 capacity may now be no more than 20-30%, thereby releasing 70%-80% of installed capacity of H100 for inferencing.

The peak in training workloads is done with: Six months ago, umpteen LLMs were in the process of being trained. Of late though, in recent months, we think CSPs have settled on a handful of fully trained LLMs that they plan to take into production, i.e. to run inference workloads at scale.

So, for instance Google's Gemini 1.5 is frozen and so is MSFT's internal LLM used for CoPilot. At META, with its recent launch of Llama3 405bn model (which took 3 months to train), we think META has a production-worthy family of models in place. At AMZN, which had experimented with a menagerie of LLMs, we think AWS has converged on a handful of LLMs, including Llama3. At all these major players, we think training days are over. All the publicly available data sets have been consumed. The little there is left to train is focused on a declining supply of fresh data sources.

Independent LLM vendors too may be largely done with modeling, and even if there are not, they may have run out of resources. Stability AI is on the way out, due to solvency issues, as reported in the media. Anthropic may be largely done with modeling with the release of Claude3 a few months ago. Even OpenAI may be finding it difficult to persist with training ever larger models given its relatively small revenue base.

If the peak in training workloads on NVDA GPUs is behind us, we think it releases a large installed base of GPUs to inference workload, thereby increasing H100 availability and driving down pricing of server rental, price per token and price of the GPU itself.

If H100 is in surplus, why are hyperscale CSPs raising capex? Are they raising capex due to unmet infrastructure capacity for current demand? We do not think so. We think they are raising capex to 1) meet future demand for AI services they anticipate will emerge, 2) to spend on shell construction and power/water utilities, 3) on inference-optimized GPU and 4) at mid-sized data centers distributed geographically worldwide.

So, for instance, of the $19bn capex MSFT announced for the June quarter, the bulk of the cash spending allocated for the near future we think will go towards land/shell acquisition worldwide. The monies allocated for infrastructure spending we think will be spent much further down the road depending on actual demand.

But the key is this – we think future infrastructure spending will be on inference-optimized GPU. In the case of MSFT, we think they will hold out for a GPU with a higher density of HBM than that on the H100. We think they will hold out for the B200 or denser versions.

We do not think MSFT is likely to populate new shells with H100/H200, only to rip out the racks in less than a year and install Blackwell racks. That would simply not be a capital efficient approach.

While MSFT waits for Blackwell, we think they are likely to aggregate existing capacity of H100 at third party data centers.

Blackwell could be a winner, but when? We think hyperscale CSPs are likely to stick to the plan of allocating fresh infrastructure spending to inference-optimized, higher memory density GPU such as

Blackwell. We do believe that there will be good demand for Blackwell, as and when supply becomes available.

What is a realistic timeline for Blackwell to ramp up in production quantity? Before taking delivery at production scale, we would expect hyperscale players to take delivery of a limited volume of test servers and run field testing for several months. If initial shipment is delayed by say one quarter, as media reports seem to suggest, we would expect another 1-2 quarters before Blackwell is shipping in volume. This may delay **revenue ramp deep into NVDA's Fy26**.

H200/B200A - Could they ship in the interim? Of roughly equal HBM density (144GB), these GPUs could be better suited for inference workloads than the lower density H100.

Going back nearly two months, there have been reports of heat-related failure with the H200. Our checks show that there have been instances of customers rejecting test racks due to **GPU failures** in the field.

It is no secret that these GPUs release intense heat due to the ~1000W of power consumption per chip. At the root of the problem, we suspect that NVDA has **not standardized** the exact solution for heat extraction, leaving the problem up to the ODMs/OEMs to solve. There could be important differences in the cooling systems adopted by the various vendors resulting in failures at some vendors.

The upshot being that hyperscale players are unlikely to take delivery unless they are convinced that NVDA and its channel partners have settled upon a common solution. Until then, we expect shipment of H200/B200A to be constrained.

Net/Net: We do not support the idea that increased shipment of H100/H200 could make up for the delay in Blackwell shipment. All things being equal, we think it appropriate to adjust NVDA's data center revenue ramp **downward** vs consensus estimates, until Blackwell gets into volume shipment, which could be well into NVDA's FY26.

But this may not be how it turns out. NVDA may still indicate higher than expected shipment of H100/H200 in the medium term. Were this to happen, we think investors would need to worry about a **build-up in channel inventory**, a negative situation.

We think the first reaction to the earnings call could be to the upside. But that may be a **head fake**. In the days and weeks ahead as investors mull over the details on the NVDA call and as earnings reports from hardware/software vendors such as DELL/HPE and CRM/MDB emerge, we would expect NVDA investors to step aside and look for **better points of entry**.

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

This report has been prepared for distribution to only qualified institutional or professional clients of Lynx Equity Strategies, LLC (the "Company"). The contents of this report represent the views, opinions, and analyses of its authors. The information contained herein does not constitute financial, legal, tax or any other advice. All third-party data presented herein were obtained from publicly available sources which are believed to be reliable; however, the Company makes no warranty, express or implied, concerning the accuracy or completeness of such information. In no event shall the Company be responsible or liable for the correctness of, or update to, any such material or for any damage or lost opportunities resulting from use of this data. Nothing contained in this report or any distribution by the Company should be construed as any offer to sell, or any solicitation of an offer to buy, any security or investment. Any material received should not be construed as individualized investment advice. Investment decisions should be made as part of an overall portfolio strategy and you should consult with a professional financial advisor, legal and tax advisor prior to making any investment decision. Lynx Equity Strategies, LLC shall not be liable for any direct or indirect, incidental or consequential loss or damage (including loss of profits, revenue or goodwill) arising from any investment decisions based on information obtained from Lynx Equity Strategies, LLC. Reproduction and Distribution Strictly Prohibited. No user of this report may reproduce, copy, distribute, sell, resell, transmit, transfer, license, assign or publish the report itself or any information contained therein. This report is not intended to be available or distributed for any purpose that would be deemed unlawful or otherwise prohibited by any local, state, national or international laws or regulations or would otherwise subject the Company to registration or regulation of any kind within such jurisdiction.

Copyright, Trademarks, Intellectual Property