SECTOR UPDATE

2026. 1. 7

Tech Team

Jongwook Lee
Team Leader
jwstar.lee@samsung.com

Kyoungbeen Kim
Research Associate
kyoungbeen.kim@samsung.com

Scan to go to
Research Center report database

Tech (OVERWEIGHT)

Jensen Huang highlights memory bottlenecks

- Vera Rubin is in mass-production, Nvidia has confirmed, while concurrently highlighting the importance of memory storage management via BlueField-4.
- Inference processes for agentic AI and physical AI should cause a surge in the number of memory bottlenecks, indicating co-growth not only for HBM but also for DRAM and NAND.

WHAT'S THE STORY?

Nvidia CES keynote speech—AI investment cycle to persist in 2026: The AI infrastructure investment cycle will remain strong throughout 2026, Nvidia's CEO Jensen Huang said during his CES 2026 keynote speech on Jan 6. He described the current phase not simply as the creation of new demand, but as a broad modernization of computing infrastructure. Huang explained that the massive investments are being driven by the structural limitations of existing data centers, which are ill-equipped for handling workloads from LLMs, agentic AI, and physical AI. This structural mismatch calls for a redesign and incremental investments across servers and storage systems. The advent of numerous open models, along with Alpamayo (a physical inference model designed for autonomous driving), underscores the ecosystem's ongoing expansion. Meanwhile, mass-production of Nvidia's next-generation GPU platform Vera Rubin has begun, and new server designs and memory architectures should be launched down the road.

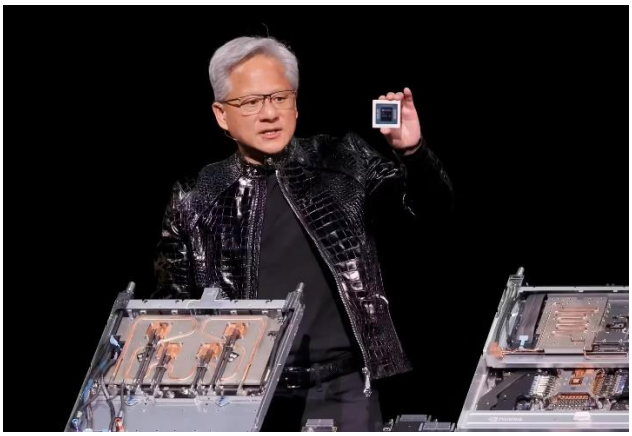
Trench architecture central to AI memory in 2026: The most significant implication of the keynote speech is that Nvidia has officially adopted GPU-centric trench architecture in its AI servers. Alongside Vera Rubin, Nvidia has introduced an AI memory storage system powered by the BlueField-4 DPU, marking an important turning point—DRAM and NAND have been fully integrated into the cache hierarchy, moving away from the traditional HBM-centric structure. With the rise of agentic AI, the number of compute calls and sequence lengths during inference has increased dramatically. Jensen Huang noted that the bandwidth and capacity of memory for handling KV cache have become a technical bottleneck. To address this, Nvidia is integrating a continuous memory hierarchy of HBM - LPDDR (DRAM) - NVMe (NAND), and using BlueField chips to directly connect and control the GPU. This enables the processing of large-scale KV cache. In this way, the software-hardware integrated architecture—including DOCA (Data-center On-a-Chip Architecture) for programming DPUs, GDS (GPU Direct Storage) for connecting storage to the GPU, and the BlueField chip (DPU + Ethernet)—is likely to function as another structural moat, much like CUDA.

Memory to remain key focus of data centers for now: The growth of memory content per AI server should outpace the growth of GPU computing in the near future. Nvidia's message suggests that maximizing capacity to accommodate KV cache when adopting the Vera Rubin system is a top priority. Agentic AI and physical AI workloads should be inference heavy, with a higher proportion of memory-resident data compared to training processes—a development that should lead to an increase in the amount of memory and storage connected per GPU. This should not only drive demand for HBM but also for general DRAM and high-capacity NVMe NAND. 2026 should mark the starting point for a sharp increase in bit demand for data center memory.

Nvidia: BlueField-4 powers inference context memory storage platform

Source: Nvidia, Samsung Securities

Jensen Huang's keynote speech at CES 2026



Source: Press release, Samsung Securities

Nvidia investor meeting Q&A summary

Q) What does the mass-production of Vera Rubin signify?

Vera Rubin is currently in mass-production, and it is expected to be made available in 2H26. There is a lead time of approximately 9 months.

Q) What is the strategic significance of the Groq licensing announcement?

While Nvidia architecture remains mainstream, Groq's technology excels in areas requiring extremely low latency, such as smart glasses.

Q) What is the scope, timing, and revenue model for autonomous driving and robotics markets?

The autonomous driving-related data center market is already in the billions of dollars and is expected to grow further. Autonomous trucks and robotaxis are projected to become massive businesses within the next decade. It is important for the ecosystem to thrive, even if Nvidia chips are not used.

Q) Why did Anthropic choose Nvidia semiconductors?

Initially funded by Google and Amazon, Anthropic relied heavily on Google and Amazon. However, considering performance, it expanded to other vendors. Nvidia's platform is advantageous for generating high-quality tokens. Nvidia is used by nearly all AI models globally.

Q) What is the status of DRAM pricing, supply, and power supply issues?

Nvidia is the only global chip company that purchases DRAM in large quantities. It collaborates closely with its HBM partners. Additionally, Nvidia is involved in upstream supply-chain management, including partnerships with semiconductor equipment makers (*eg*, AMAT, ASML), securing land, and establishing electricity grids. Directly managing the supply chain ensures a more stable supply.

Q) What is Nvidia's China strategy?

Products for China are managed through a separate supply chain to avoid impacting other regions. It is currently waiting for its license applications to be processed by the US and Chinese governments.

Q) Is DGX Cloud a service that competes with customers?

Initially, CSPs were unfamiliar with AI architectures, so we built our own systems and attracted over 20,000 AI-native companies to CSPs. In a way, we acted as the sales force for CSPs through DGX Cloud. Now, the importance of DGX Cloud as a rental business is declining.

Q) What is the memory/storage controller business?

The inference volume at AI factories is so large that KV cache has become extremely heavy. To handle this, a new layer of storage directly connected within the computing fabric is needed. BlueField 4's DOCA software performs this role. It should grow like CUDA.

Q) How fast will Vera Rubin be mass-produced, and how profitable will it be?

Vera Rubin involves developing all component chips from scratch. It reduces server assembly time to just two hours and improves NVLink maintenance efficiency. The value of adopting Vera Rubin is high, and profitability should be maintained.

Q) Should AI models consolidate into a few winners or should they remain fragmented?

General-purpose AI is smart enough, but vertical agents are needed in specific domains. Many companies should use open-source models to build their own AI.

Q) When should the inflection point for agentic AI and physical AI arrive?

Agentic AI and physical AI are already here. Their use should become more ubiquitous.

Q) What should be the role of Neo Cloud?

In a fast-changing technological landscape where securing power is a challenge, specialized regional players have emerged. They are important partners for Nvidia.

Q) How sustainable is your profitability?

Profitability is directly tied to the value we provide. Our value is overwhelming in terms of reducing model training time, lowering inference costs, and increasing profitability at the data center level. When a client is building a gigawatt-scale data center valued at USD50b, the client would gladly pay a premium if Nvidia's products help generate more revenue (vs that generated by competitors' products).

Q) What is the next game-changer? Will token demand continue to grow?

The next challenges are expansion of context length and continuous learning. Thanks to agent systems, token generation has increased 50x.

Q) Tell us more about your R&D investments and M&A philosophy

We make investments across the board—including in chips, models, and applications. We do not distinguish between R&D activities and external investments. If there is a team that shares our vision and moves quickly in technology, we will consider acquiring them.

Q) What is the possibility of reducing your dependency on HBM via SRAM and proprietary technologies?

Products like CPX that reduce reliance on HBM may be more efficient for specific tasks. However, the challenge is that AI workloads continue to change. Since data center power is limited, a unified architecture that delivers top performance across all models should ultimately be more cost-effective in terms of total cost of ownership (TCO).

Compliance notice

- As of 1/6 2026, the covering analyst(s) did not own any shares, or debt instruments convertible into shares, of any company covered in this report.
- As of 1/6 2026, Samsung Securities' holdings of shares and debt instruments convertible into shares of each company covered in this report would not, if such debt instruments were converted, exceed 1% of each company's outstanding shares.
- This report has been prepared without any undue external influence or interference, and accurately reflects the views of the analyst(s) covering the company or companies herein.
- All material presented in this report, unless specifically indicated otherwise, is under copyright to Samsung Securities.
- Neither the material nor its content (including copies) may be altered in any form, or by any means transmitted, copied, or distributed to another party, without prior express written permission from Samsung Securities.
- This memorandum is based upon information available to the public. While we have taken all reasonable care to ensure its reliability, we do not guarantee its accuracy or completeness. This memorandum is not intended to be an offer, or a solicitation of any offer, to buy or sell the securities mentioned herein. Samsung Securities shall not be liable whatsoever for any loss, direct or consequential, arising from the use of this memorandum or its contents. Statements made regarding affiliates of Samsung Securities are also based upon publicly available information and do not necessarily represent the views of management at such affiliates.
- This material has not been distributed to institutional investors or other third parties prior to its publication.

Global Disclosures & Disclaimers

General

This research report is for information purposes only. It is not and should not be construed as an offer or solicitation of an offer to purchase or sell any securities or other financial instruments or to participate in any trading strategy. This report does not provide individually tailored investment advice. This report does not take into account individual client circumstances, objectives, or needs and is not intended as recommendations of particular securities, financial instruments or strategies to any particular client. The securities and other financial instruments discussed in this report may not be suitable for all investors. The recipient of this report must make its own independent decisions regarding any securities or financial instruments mentioned herein and investors should seek the advice of a financial adviser.

This report may not be altered, reproduced, distributed, transmitted or published in whole or in part for any purpose. References to "Samsung Securities" are references to any company in the Samsung Securities, Co., Ltd. group of companies.

Samsung Securities and/or other affiliated companies, its and their directors, officers, representatives, or employees may have long or short positions in any of the securities or other financial instruments mentioned in this report or of issuers described herein and may purchase and/or sell, or offer to purchase and/or sell, at any time, such securities or other financial instruments in the open market or otherwise, as either a principal or agent. Any pricing of securities or other financial instrument contained herein is as of the close of market for such day, unless otherwise stated. Opinions and estimates contained herein constitute our judgment as of the date of this report and are subject to change without notice.

The information provided in this report is provided "AS IS". Although the information contained herein has been obtained from sources believed to be reliable, no representation or warranty, either expressed or implied, is provided by Samsung Securities in relation to the accuracy, completeness or reliability of such information or that such information was provided for any particular purpose and Samsung Securities expressly disclaims any warranties of merchantability or fitness for a particular purpose. Furthermore, this report is not intended to be a complete statement or summary of the securities, markets or developments referred to herein.

Samsung Securities does not undertake that investors will obtain any profits, nor will it share with investors any investment profits. Samsung Securities, its affiliates, or any of its and their affiliates, directors, officers, employees or agents disclaim any and all responsibility or liability whatsoever for any loss (direct or consequential) or damage arising out of the use of all or any part of this report or its contents or otherwise arising in connection therewith. Information and opinions contained herein are subject to change without notice. Past performance is not indicative of future results. Foreign currency rates of exchange may adversely affect the value, price or income of any security or financial instrument mentioned in this report. For investment advice, trade execution or other enquiries, clients should contact their local sales representative. Any opinions expressed in this report are subject to change without notice and may differ or be contrary to opinions expressed by other business areas or groups of Samsung Securities. Any analysis contained herein is based on numerous assumptions. Different assumptions may result in materially different results. Samsung Securities is under no obligation to update or keep current the information contained herein. Samsung Securities relies on information barriers to control the flow of information contained in one or more areas or groups within Samsung Securities into other areas or groups of Samsung Securities. Any prices stated in this report are for information purposes only and do not represent valuations for individual securities or other financial instruments. Samsung Securities makes no representation that any transaction can or could have been effected at those prices and any prices contained herein may not reflect Samsung Securities' internal books and records or theoretical model-based valuations and may be based on certain assumptions. Different assumptions by Samsung Securities or any other source may yield substantially different results. Additional information is available upon request.

For reports to be distributed to US:

Securities research is prepared, issued and exclusively distributed by Samsung Securities Co., Ltd., an organization licensed with the Financial Supervisory Service of South Korea. This research may be distributed in the United States only to major institutional investors as defined in Rule 15a-6 of the U.S. Securities Exchange Act of 1934, as amended, and may not be circulated to any other person otherwise. All transactions by U.S. investors involving securities discussed in this report must be effected through Samsung Securities (America) Inc., a broker-dealer registered with the U.S. Securities & Exchange Commission and a member of the Financial Industry Regulatory Authority/SIPC, and not through any non-U.S. affiliate thereof. The analysts listed [on the front of this report] are employees of Samsung Securities Co., Ltd., or a non-U.S. affiliate thereof, and are not registered/qualified as research analysts under applicable U.S. rules and regulations and may not be subject to U.S. restrictions on communications with covered companies, public appearances, and trading securities held by a research analyst account.

For reports to be distributed to UK:

This report is not an invitation nor is it intended to be an inducement to engage in investment activity for the purpose of section 21 of the Financial Services and Markets Act 2000 of the United Kingdom ("FSMA"). To the extent that this report does constitute such an invitation or inducement, it is directed only at (i) persons who are investment professionals within the meaning of Article 19(5) of the Financial Services and Markets Act 2000 (Financial Promotion) Order 2001 (as amended) of the United Kingdom (the "Financial Promotion Order"); (ii) persons who fall within Articles 49(2)(a) to (d) ("high-net-worth companies, unincorporated associations etc.") of the Financial Promotion Order; and (iii) any other persons to whom this report can, for the purposes of section 21 of FSMA, otherwise lawfully be made (all such persons together being referred to as "relevant persons").

Any investment or investment activity to which this report relates is available only to relevant persons and will be engaged in only with relevant persons. Persons who are not relevant persons must not act or rely on this report.

For reports to be distributed to Korea:

This report is for private circulation only, not for sale, and is issued and distributed only to persons permitted under the laws and regulations of Korea.

For reports to be distributed to Singapore:

This report is provided pursuant to the financial advisory licensing exemption under Regulation 27(1)(e) of the Financial Advisers Regulation of Singapore and accordingly may only be provided to persons in Singapore who are "institutional investors" as defined in Section 4A of the Securities and Futures Act, Chapter 289 of Singapore. This report is intended only for the person to whom Samsung Securities has provided this report and such person may not send, forward or transmit in any way this report or any copy of this report to any other person.

Analyst certification

The views expressed in this research report accurately reflect the personal views of the analyst(s) about the subject securities or issuers and no part of the compensation of such analyst(s) was, is, or will be directly or indirectly related to the inclusion of specific recommendations or views in this research report. The analyst(s) principally responsible for the preparation of this research report receives compensation based on determination by research management and senior management (not including investment banking), based on the overall revenues, including investment banking revenues of Samsung Securities Co., Ltd. and its related entities and has taken reasonable care to achieve and maintain independence and objectivity in making any recommendations.

Copyright © 2010 Samsung Securities Co., Ltd. All rights reserved. This report or any portion hereof may not be reprinted, sold or redistributed without the prior written consent of Samsung Securities America Inc

Samsung Securities

SAMSUNG SECURITIES

Samsung Electronics Bldg., 11, 74-gil,
Seochodae-ro, Seocho-gu, Seoul, Korea 06620
Tel: 02 2020 8000 / www.samsungpop.com

Family Center: 1588 2323

Voice Of Customer: 080 911 0900

For more information,
please call our sales representatives:

LONDON

Samsung Securities Europe Limited
1st Floor, 30 Gresham Street, London EC2V 7PG UK
Tel. 44-207-776-4311
Fax. 44-203-837-9219

NEW YORK

Samsung Securities America Limited
1330 Avenue of the Americas. 10th Floor. New York,
NY 10019
Tel: 1-212-972-2454
Fax: 1-212-972-2704

HONG KONG

Samsung Securities (Asia) Limited
Suite 4511, Two International Finance Center,
8 Finance Street, Central, Hong Kong
Tel: 852-3411-3608
Fax: 852-2114-0290

BEIJING

Samsung Securities Beijing Representative Office
Rm. 910, The Exchange Building No 118 JianGuo Lu, Chao Yang District, Beijing, China
Tel: 86-10-6522-1855 (extension 7891)
Fax: 86-10-6522-1855 (extension 7889)

TOKYO

Samsung Securities Tokyo Representative Office
#106-8532 19F, Roppongi T-Cube 3-1-1,
Roppongi Minato-ku Tokyo, Japan
Tel: 81-3-6333-2952
Fax: 81-3-6333-2953

Member of
Dow Jones
Sustainability Indices
Powered by the S&P Global CSA