SEMICONDUCTORS & IT HARDWARE

Samsung Memory Tech Day Takeaways

24 OCTOBER 2023 at 06:35*

Karl Ackerman, CFA
(+1) 415 633 1698
karl.ackerman@us.bnpparibas.com

Jerome Ramel
(+1) 415 659 2809
jerome.ramel@us.bnpparibas.com

David O'Connor
(+1) 415 659 1888
david.oconnor@us.bnpparibas.com

Yang Pu, CFA
+(1) 646 408 4363
yang.pu@us.bnpparibas.com

tech@exanebnpparibas.com

We had the privilege of attending Samsung's Memory Day in San Jose, an invite-only event that showcased Samsung's technology roadmaps and vision to support AI workloads. Overcoming the "memory wall" will require the use of CXL, greater use of high-bandwidth memory (HBM), and new memory technologies such as LHM (low power, high bandwidth memory) and PIM (processing in memory). There is a battle brewing in HBM, and Samsung (NR) appears well positioned.

DRAM is the primary beneficiary of AI, supported by CXL memory pooling

Samsung and Microsoft both spoke about how we are in the early days of the AI revolution, which will drive significant demand for HBM and high-capacity DDR5 DRAM but only modest demand for SSDs and HDDs. CXL is aimed to address the "memory wall" or the widening gap between LLM computational needs and the memory infrastructure it relies on. Samsung noted that memory demand for AI/ML accelerators will grow at a 50%+ CAGR to reach 11% of total DRAM demand by 2028. Similarly, CXL should grow at a 275% CAGR from '24 to '28 to reach 11% of the DRAM market.

Here comes Samsung in HBM (High-bandwidth memory)

Hynix remains in the lead on HBM, which is used in compute accelerators and GPUs. Samsung is shipping HBM3 now and in quals with HBM3E. Samsung's HBM3E is on par with Micron's and above Hynix's. Samsung's HBM4 is very competitive supporting 48GB at 2.0TB/s bandwidth w/ 2.6pJ/bit.

DRAM roadmap indicates high-NA EUV in mid 2025

1-beta (12nm) is ramping today, and 1-c (11nm) could ramp in mid-2024. High-NA EUV patterning appears to ramp with 1-d in late '25/early '26. Samsung will introduce a 3D structure for sub-10nm DRAM capable of a 100Gb die, which may occur as HKMG transitions to FinFET for 0bnm (2029?).

NAND roadmap showcases wafer bonding in V10 (2026?)

V8 (238L 3D NAND) was Samsung's second node to use logic under the array; V9 will ramp mid '24, and V10 will use wafer bonding as Samsung takes a page from YMTC, Kioxia and WDC's play book.

Key valuation metrics

	Stock Rating	Price (LC)	TP / Upside		Mkt cap (USDm)	P/E (x) CY23e	P/E (x) CY24e	EV/EBITA (x) CY23e	EV/EBITA (x) CY24e	ESG Rating
AMD	(+)	100.0	150	50%	161,766	85.2	37.4	63.3	28.5	Average
Applied Mat.	(=)	134.2	140	4%	112,867	17.3	16.5	14.5	13.9	Leader
ASML	(+)	552.6	810	47%	230,423	28.5	27.9	24.2	23.3	Leader
Intel	(-)	33.9	28	-17%	141,324	NC	51.4	NC	52.3	Average
KLA	(=)	462.6	550	19%	62,165	20.6	20.2	16.9	16.5	Average
Lam Research	(=)	599.6	570	-5%	79,495	20.3	21.4	17.3	18.0	Average
Micron Tech.	(+)	66.9	75	12%	65,318	0.0	NC	NC	NC	
Nvidia	(+)	429.8	745	73%	1,059,212	44.1	26.6	37.7	21.4	Average
Seagate	(=)	65.3	58	-11%	13,624	NC	24.2	42.7	18.7	
Western Digital	(+)	41.8	58	39%	13,765	0.0	0.0	NC	37.4	

Prices at 23 October 2023

**** Date and time (London Time) on which the investment recommendation was finalised. It may differ from the date and time of broad dissemination on the website. See Appendix (on p8) for Analyst Certification, Important Disclosures and Non-US Research Analyst disclosures. This research report is published and distributed by BNP Paribas Securities Corp., a FINRA member firm and broker-dealer registered with the SEC.***

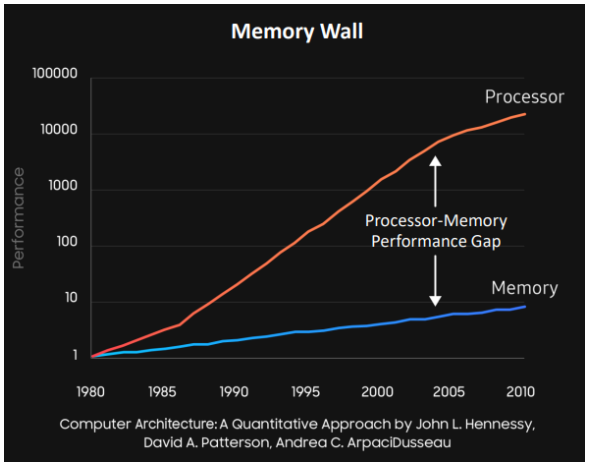
New DRAM technologies are being developed to support AI workloads

It is no surprise that inference and training workloads are memory intensive. As model parameters continue to grow at 10x a year and should exceed 1T parameters for GPT-4, AI accelerator performance is increasingly bottlenecked by the ability to store and retrieve training and inference data from memory. At the Open Compute Project Summit earlier last week, Microsoft highlighted how computational demand of transformer models has grown at 750x every 2 years while memory bandwidth has been just 1.6x/2yrs and interconnect bandwidth has been just 1.4x/2yrs.

This "memory wall" problem arises when processors must wait clock cycles to retrieve data from memory. Essentially, while the computing power of processors has advanced exponentially over the past 30 years as Moore's Law predicted, the performance of memory chips – although much faster than ever before – has advanced in a more linear way. This creates infrastructure scaling problems and underutilization of system resources.

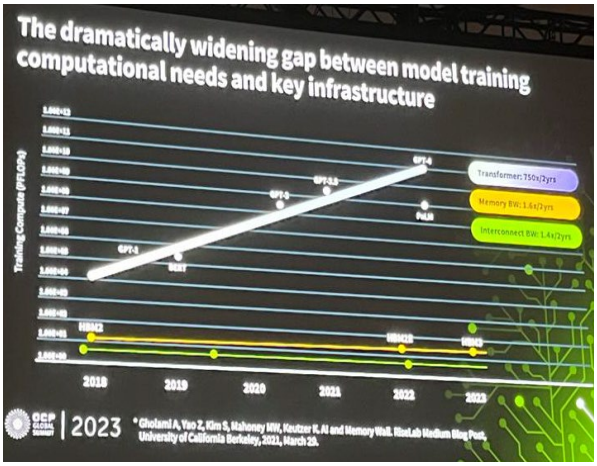
Figure 1: Model parameters of AI training systems are growing faster than improvements in infrastructure

Memory I/O has not been able to scale at the same rate as processor performance



Source: Samsung

AI models depend on advancements in networking, compute, and memory



Source: Open Compute Summit presentations

DRAM will be the primary beneficiary of AI/ML adoption

To overcome this "memory wall" problem, new forms of memory need to be developed. Samsung and Microsoft both spoke about how we are in the early days of the AI revolution, which will drive significant demand for high-bandwidth memory (HBM) and high-capacity DDR5 DRAM but only modest demand for SSDs and HDDs. This is relatively consistent with what semicap and memory IC providers have indicated earlier this year: an AI server is supposed to drive 8x higher DRAM content and 3x higher NAND content than a "traditional" server.

3D XPoint "died" in 2021 when Intel and Micron abandoned their efforts, which has left a gap between CPU cache (SRAM) and HBM. At the conference, Samsung introduced LHM, or low power high bandwidth memory. They also discussed upcoming HBM products such as PIM (process in-memory) and several products that support the CXL transport protocol.

LHM for AI Inferencing

Similar to Samsung's introduction of Wide IO in circa 2015, LHM offers high bandwidth and low-density cache memory. Essentially, the first chip on LHM (in 2026?) will run at the same speed of Samsung's HBM3 (820GB/s) but will be just 2GB in size (vs. 16GB) and can run at only 1.4 picojoule (pJ/Bit) that's roughly 4x lower power than HBM!

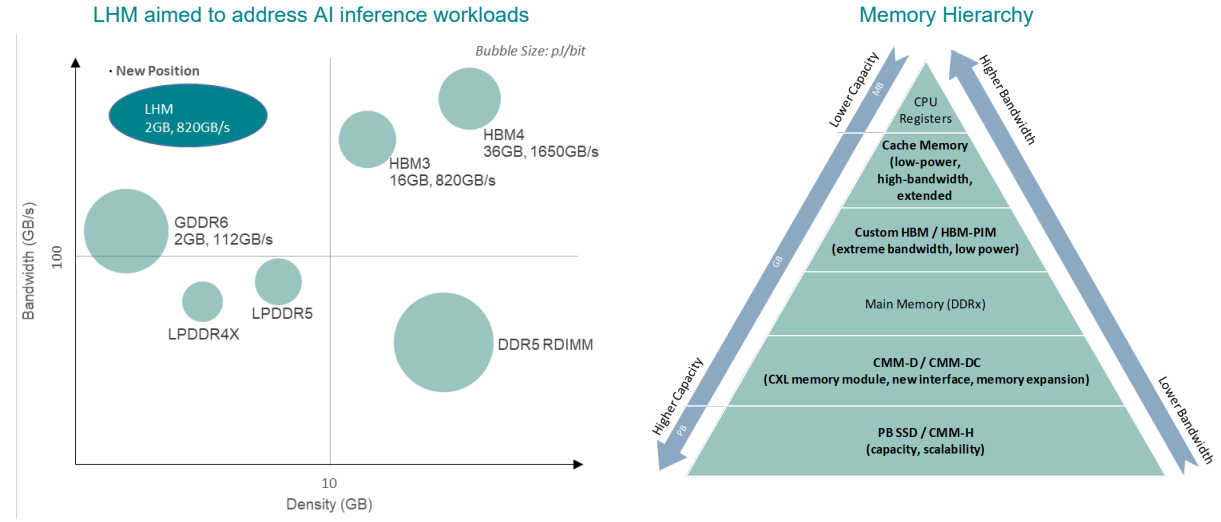
Samsung will be able to achieve this using hybrid bonded and 2.5D packaging technology. In other words, LHM is being pitched for AI inference given its low power and higher bandwidth capabilities vs. GDDR and LPDDR DRAM offerings.

PIM (Processing In Memory) for AI applications

Samsung also discussed the use of PIM in its future DRAM roadmap. The idea is that the IO counts and thermal constraints are becoming a limitation for higher memory bandwidth of AI applications. PIM overcomes these constraints by calculating target operations in parallel inside DRAM. This results in improved energy efficiency and higher bandwidth since the AI processor will receive only final data from the LPDDR DIMM or HBM-PIM rather than sending all data to the processor host for AI model calculations. Samsung indicated that LPDDR-PIM could be commercialized in 2026.

Similarly, Samsung noted that the use of HBM-PIM can drive a 2.25x increase in energy efficiency and 2.3x gain in bandwidth over "regular" HBM. This technology is expected to be commercialized in 2027.

Figure 2: New forms of memory needed to address the "memory wall"



Source: BNP Paribas Exane estimates

Understanding the need for CXL

CXL will add another tier in the memory hierarchy below the main system memory (DRAM found in a server) to address growing demand for different tiers of memory capacity and bandwidth. CXL is needed to support the planned increase in I/O traffic by adding DRAM capacity and bandwidth data-intensive, high performance computing applications.

Samsung's CXL solutions have a naming convention of "CMM" or CXL Memory Modules. Samsung believes CMM can be used for capacity expansion, bandwidth expansion, tiered memory, and memory pooling/sharing for in-memory databases, HPC, AI training and inference, and general purpose virtualization applications.

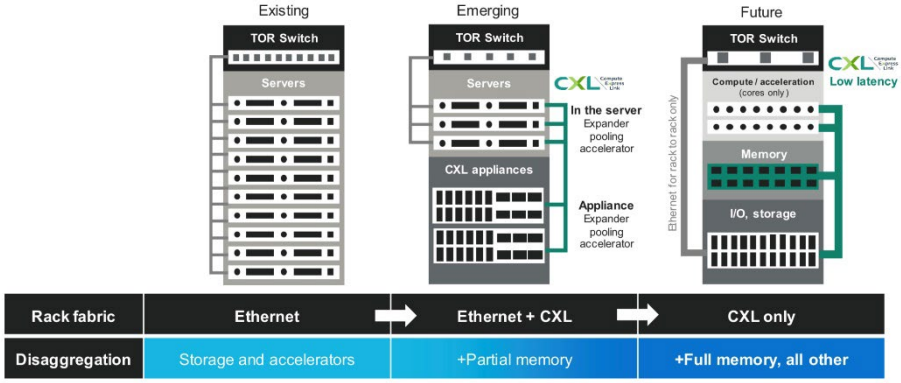
Compute express link (CXL) allows low-latency high-bandwidth interconnects

Compute Express Link (CXL) is a new open standard communication design based on PCIe PHY that allows low-latency and high-bandwidth interconnects between CPUs, GPUs, FPGAs, and other devices. It enables devices to share memory and cache resources directly with each other and eliminates the need for data to be transferred over the interconnect.

CXL 1.0 was first introduced in 2019 and upgraded to CXL 2.0 in 2020, both based on PCIe 5.0 physical layers. The latest version CXL 3.0 was released in Aug 2022 and is based on PCIe 6.0 PHY and PAM-4 coding, supporting up to 128GB/s bi-directional communication over a 16-lane link. It's worth noting that most products out today by vendors are for CXL 1.0 and 2.0 and that product supporting CXL 3.0 will likely not be out until PCIe 6.0 in 2026.

AI and HPC will drive the need for higher-performance accelerators and low-latency memory solutions, and it will be the test bed for CXL solutions. What's interesting to us is that the CXL technology could potentially allow multiple devices to coherently share memory spaces and make it possible to create a scalable disaggregate memory pool for GPU devices. This could essentially change the architecture of integrated system designs and significantly improve the efficiency of processing workloads.

Figure 3: CXL allows the creation of a disaggregate memory pool



Source: Marvell

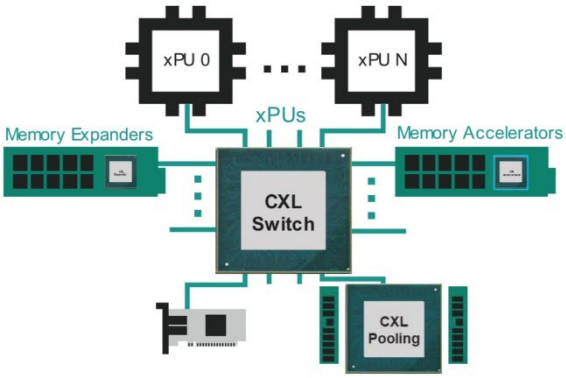
The implementation of CXL is still at very early stage, but the industry has coalesced to push for the next generation interconnects. The CXL Consortium was formed with all major technology companies, including AMD, Broadcom, Cisco, Google, Intel, Marvell Meta, Microsoft, Micron, Samsung, SK Hynix, Nvidia, Western Digital and many others.

In 2021, CXL 1.1 support was announced for the latest Intel Sapphire Rapids and AMD Zen 4 EPYC Genoa and Bergamo CPUs, and deployments have already occurred in 2023. Micron anticipates CXL 2.0 will be rolled out in late 2024, and the market will accelerate further when CXL 3.0 rolls out afterwards – perhaps late 2025/early 2026 – which enables memory pooling.

At Samsung's Memory Tech Day, the company noted that memory demand for AI/ML accelerators will grow at a 50%+ CAGR to reach 11% of total DRAM demand by 2028. Similarly, CXL should grow at a 275% CAGR from '24 to '28 to reach 11% of the DRAM market. At Micron's last analyst day, the company suggested CXL could exceed $20B+ by 2030.

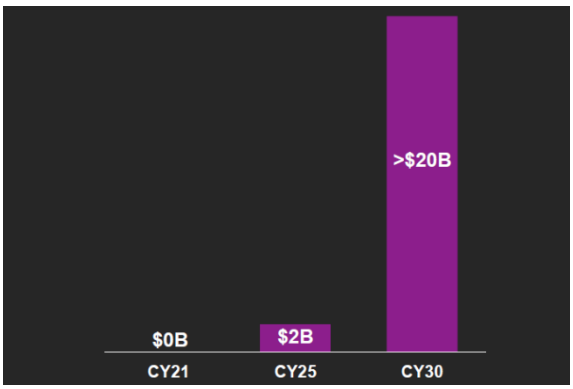
Figure 4: CXL is the next generation of interconnects

CXL allows cache-coherent interconnects across device memory



Source: Marvell

CXL TAM could grow to over $20B by 2030



Source: Micron

DRAM Roadmap

We believe Samsung will remain the #1 provider of DRAM globally, and we think Samsung is making good progress on 1-beta (12nm) node today. We estimate that Samsung could be shipping 1-c (11nm) in mid-2024. We think that High-NA EUV patterning should ramp with 1-d in late '25/early '26. In our view, the use of high-NA and no mention of 3D DRAM on the roadmap should be viewed positively for ASML.

Figure 5: Our estimate of Samsung's DRAM roadmap

	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Tech node	2znm		1xnm		1ynm		1znm		1anm	1bnm		1cnm	1dnm		0anm	0dnm	
Capacity Speed	8Gb		2.4Gbps	2.6Gbps			3.2Gbps		4.8Gbps	24Gb 6.4Gbps	32Gb	7.2Gbps	9.6Gbps		48Gb 10.8Gbps		
Products	DDR4		LPDDR4x		GDDR6 HBM2		DDR5 LPDDR5 HBM2E			LPDDR5x GDDR6+ HBM3	GDDR7 HBM3E CXL-MXP	HBM3-PIM AX DIMM	HBM4	CXL-PNM	DDR6		
Patterning	Double Patterning		Quadruple Patterning				EUV					Advanced EUV					

Source: BNP Paribas Exane estimates

New sever server platforms from Intel and AMD support more memory channels and DIMM slots per server CPU. The largest benefit of Sapphire Rapids launching will be the support of DDR5 DRAM, PCIe 5.0 SSDs, CXL 1.1, and up to 64GB of HMB2.

Figure 6: Our estimate of Samsung's DDR5 Roadmap

		2022	2023	2024	2025	2026	2027	2028
Compute Needs	x86 CPU Core	40 - 64 cores	56 - 128 cores		128 - 192 cores		192+ cores	
	CPU Mem. Ch.	8 Channels	8/12/16 Channels				12/16 Channels	
	CPU TDP (max.)	280W	350W+		400W+		Beyond	
	Rack Power (Volume.)	4.4 - 8.8 kW	8.8 - 15+ kW					
	DIMM Config	2SPC /	1SPC or mixed 1&2SPC / 7+mm				1SPC	
System Requirement & DRAM Solution	Process	1anm		1bnm	1cnm	1dnm	0anm	
	DRAM Generation	DDR5					DDR6 (TBD)	
	Bandwidth	4800Mbps	5600Mbps	6400Mbps	7200/8000Mbps	8000Mbps+	9600Mbps+	TBD
	Capacity comp.	16/24Gb		16/24/32Gb				
	Capacity DIMM	Max. 512GB		Max 1TB				

Source: BNP Paribas Exane estimates

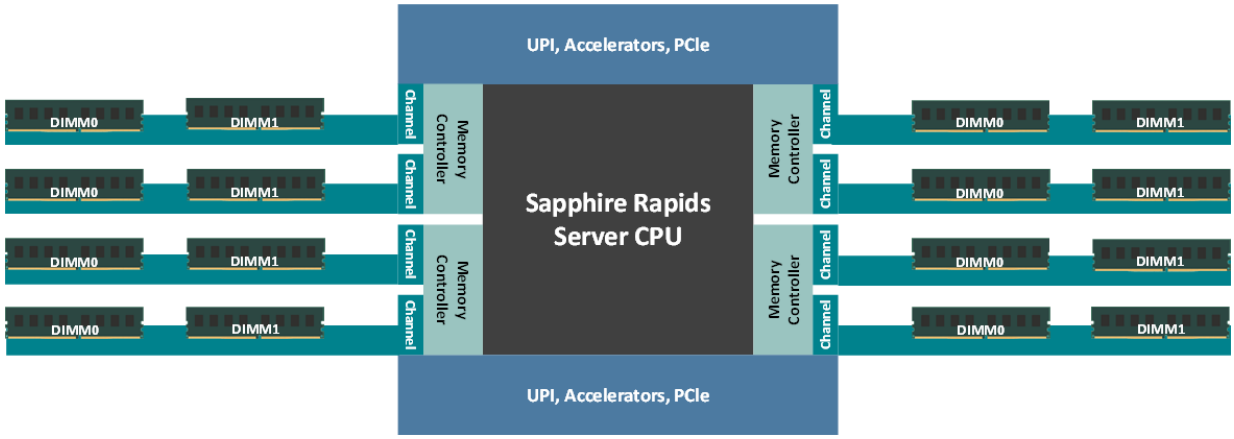
Understanding The Benefits Of DDR5

Historically, server system performance has been bottlenecked by different components of the system architecture. Over the last decade, MPU vendors have increased overall processing performance via process node transitions, enhanced instruction set architecture (ISA) compilation, and utilization of multi-core architectures. The problem, though, is that there is a widening gap between the performance of CPU cores and the DRAM DIMMs ("dual in-line memory modules", the primary form factor for server memory) that 'feed' and transport data to them each clock cycle. Much of this can be explained by DRAM core clock rates falling behind improvements in CPU core clock rates. This creates infrastructure scaling problems and underutilization of system resources.

There are two ways to remedy system performance gated by memory: increase the total system memory capacity via volume, and/or increase the transmission speed via bandwidth of the memory itself commensurate with the processor it supports. The former is done by optimally allocating server DRAM DIMMs across a server's memory channels. Increasing the bandwidth is done by upgrading the clock speed of the DRAM, such as increasing from DDR4 (3.2Gb/s) to DDR5 (initial speed of 4.8Gb/s with a path to 6.4Gb/s). Thus, DDR5 provides 50% better performance out of the gate with capacity to run at twice the effective bandwidth of DDR4. Of course, DDR5 offers reduced power consumption vs. DDR4 making it more suitable for all applications as a system's main memory configuration.

DRAM speeds and densities should continue to expand as memory requirements in servers grow to help serve a growing number of CPU cores. Higher compute capacity and bandwidth arises from user-generated content of video (Youtube, Netflix), gaming, big data analytics, high performance computing and virtualization. Similarly, future server platforms such as INTC's Granite Rapids (3nm) will support even more DIMM channels (12 vs. 8). Combined, improved DRAM speed and capacity per server will allow DRAM to continue to grow within server architectures and provide Micron, Samsung and Hynix ample opportunity to grow over the next few years.

Figure 7: Server DRAM "Feeds" The CPU

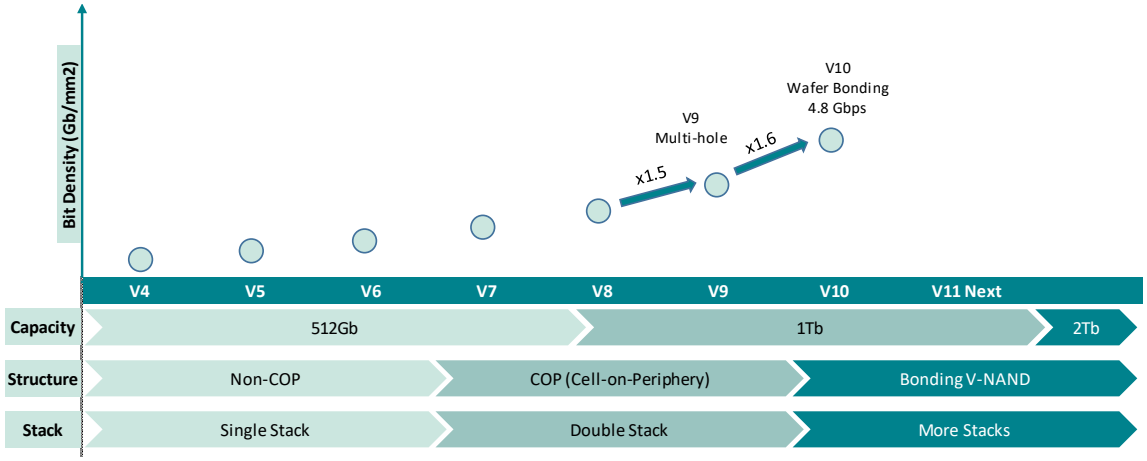


Source: BNP Paribas Exane

NAND Roadmap

Samsung's V7 node, or its 176-layer 3D NAND die, was Samsung's first node to use logic under the array. Samsung's 238L die (V8) continues to drive bit density, though the use of multi-hole technology in the upcoming V9 node is expected to drive a 50% improvement in bit density. For V10, we expect Samsung to introduce wafer bonding that will drive another 60% improvement in bit density vs. V9. YMTC, Kioxia and WDC have all since adopted wafer bonding to drive improvements in bit density and cost across their NAND roadmaps.

Figure 8: Our estimate of Samsung's NAND roadmap



Source: BNP Paribas Exane estimates

DISCLOSURE APPENDIX

Analyst Certification

We, Karl Ackerman, David O'Connor, Yang Pu, Jerome Ramel, (authors of or contributors to the report) hereby certify that all of the views expressed in this report accurately reflect our personal view(s) about the company or companies and securities discussed in this report. No part of our compensation was, is, or will be, directly, or indirectly, related to the specific recommendations or views expressed in this research report.

Non-US Research Analyst Disclosure

The research analysts at Exane SA (including its branches in the United Kingdom, the European Economic Area or Switzerland) named below were involved in preparing this research report. They are not associated persons of BNP Paribas Securities Corp. and thus they are not registered or qualified in the U.S. as research analysts with the Financial Industry Regulatory Authority (FINRA) or the New York Stock Exchange (NYSE). These non-U.S. analysts are not subject to FINRA Rule 2241 concerning communications with a subject company, public appearances and trading securities held by a research analyst account.

None

BNP Paribas SA and Exane SA Affiliates Involved in Preparation of the Research Report

Exane SA is authorised by the Autorité de contrôle prudentiel et de résolution ("ACPR") and regulated by the Autorité des Marchés Financiers ("AMF") in France.
Exane SA branches based in the European Economic Area are authorized by the ACPR and regulated by the local relevant regulator.
Exane SA Geneva Branch is authorized and regulated by the Swiss Financial Market Supervisory Authority ("FINMA")
Exane SA, London Branch is deemed authorised and regulated by the Financial Conduct Authority. Details of the Financial Services Contracts Regime, which allows EEA-based firms to operate in the UK for a limited period to carry on activities which are necessary for the performance of pre-existing contracts, are available on the Financial Conduct Authority's website.
BNP Paribas Securities Corp. is regulated by FINRA (CRD# 15794) and the Securities and Exchange Commission in the United States.

Research Analyst Compensation

The research analyst(s) responsible for the preparation of this report receive(s) compensation based upon various factors including overall firm revenues, which may include investment banking activities and/or sales and trading activities.

Disclosure of the report to the company/ies

Companies	Disclosures
NONE	

1 – Sections of this report, with the research summary, target price and rating removed, have been presented to the subject company/ies prior to its distribution, for the sole purpose of verifying the accuracy of factual statements.
2 – Following the presentation of sections of this report to this subject company, some conclusions were amended.

The latest company-specific disclosures, valuation methodologies and investment case risks for all other companies covered by this document are available on http://cube.exane.com/compliance.

LONDON
Branch of Exane SA
10 Harewood Avenue
London NW1 6AA
UK
Tel: (+44) 207 039 9400
Fax: (+44) 207 039 9440

PARIS
Exane SA
6 Rue Ménars
75002 Paris
France
Tel: (+33) 1 44 95 40 00
Fax: (+33) 1 44 95 40 01

FRANKFURT
Branch of Exane SA
Senckenberganlage 19
60325 Frankfurt
Germany
Tel: (+49) 69 42 72 97 300
Fax: (+49) 69 42 72 97 301

MADRID
Branch of Exane SA
Calle Emilio Vargas, 4
28043 Madrid
Spain
Tel: (+34) 91 114 83 00
Fax: (+34) 91 114 83 01

MILAN
Branch of Exane SA
Diamante Tower
Piazza Lina Bo Bardi 3
20124 Milan
Italy
Tel: (+39) 02 89 63 17 13
Fax: (+39) 02 89 63 17 01

STOCKHOLM
Branch of Exane SA
Hovslagargatan 3
111 48 Stockholm
Sweden
Tel: (+46) 8 5663 9820
Fax: (+46) 8 5063 9751

NEW YORK
BNP Paribas Securities Corp.
787 Seventh Avenue
New York, NY 10017
USA
Tel: (+1) 212 841 2000

SAN FRANCISCO
BNP Paribas Securities Corp.
180 Montgomery Street
San Francisco, CA 94104
USA
Tel: (+1) 415 772 1300

All Exane research documents are available to all clients simultaneously on the Exane website (http://cube.exane.com). Most published research is also available via third-party aggregators such as Bloomberg, Refinitiv, Factset and Capital IQ. Exane is not responsible for the redistribution of research by third-party aggregators.

Important notice: Please refer to our complete disclosure notice and conflict of interest policy available on http://cube.exane.com/compliance

This research is produced by one or more of EXANE SA (including Exane SA branches, based in the United Kingdom, in the European Economic Area and in Switzerland) and BNP Paribas Securities Corp. EXANE SA is authorized by the Autorité de Contrôle Prudentiel et de Résolution ("ACPR") and regulated by the Autorité des Marchés Financiers ("AMF"). Exane SA branches based in the European Economic Area are authorized by the ACPR and regulated by the local relevant regulator. Exane SA Geneva Branch is authorized and regulated by the Swiss Financial Market Supervisory Authority ("FINMA"). Exane SA, London Branch is deemed authorised and regulated by the Financial Conduct Authority. Details of the Financial Services Contracts Regime, which allows EEA-based firms to operate in the UK for a limited period to carry on activities which are necessary for the performance of pre-existing contracts, are available on the Financial Conduct Authority's website.

BNP Paribas Securities Corp. is registered and regulated by the Financial Industry Regulatory Authority ("FINRA") and the U.S. Securities and Exchange Commission. In accordance with the requirements of Financial Conduct Authority COBS 12.2.3R and associated guidance, of article 313-20 of the AMF Règlement Général, and of FINRA Rule 2241, Exane's policy for managing conflicts of interest in relation to investment research is published on Exane's web site (cube.exane.com). Exane also follows the guidelines described in the code of conduct of the Association Francaise des Entreprises d'Investissement ("AFEI") on managing conflicts of interest in the field of investment research. This code of conduct is available on Exane's web site (http://cube.exane.com).

This research is solely for the private information of the recipients. All information contained in this research report has been compiled from sources believed to be reliable. However, no representation or warranty, express or implied, is made with respect to the completeness or accuracy of its contents, and it is not to be relied upon as such. Opinions contained in this research report represent BNPP's and/or Exane's current opinions on the date of the report only. BNPP and/or Exane are not soliciting an action based upon it, and under no circumstances is it to be used or considered as an offer to sell, or a solicitation of any offer to buy.

While BNPP and/or Exane endeavour to update its research reports from time to time, there may be legal and/or other reasons why BNPP and/or Exane cannot do so and, accordingly, BNPP and/or Exane disclaim any obligation to do so.

This report is provided solely for the information of professional investors who are expected to make their own investment decisions without undue reliance on this report and BNPP and/or Exane accept no liability whatsoever for any direct or consequential loss arising from any use of this report or its contents.
(i) this research is intended for general circulation; (ii) this research does not take into account the specific investment objectives, financial situation or particular needs of any particular person; and (iii) advice should be sought from a financial adviser regarding the suitability of the investment product, taking into account the specific investment objectives, financial situation or particular needs of any person in receipt of the research, before the person makes a commitment to purchase the investment product.

This report may not be reproduced, distributed or published by any recipient for any purpose. Any United States person wishing to obtain further information or to effect a transaction in any security discussed in this report should do so only through BNP Paribas Securities Corp., which has distributed this report in the United States and, subject to the above, accepts responsibility for its contents.

For Institutional and Accredited Investors in Singapore who receive this report, please contact BNP Paribas, acting through its Singapore Branch (company registration number: S71FC2142G; address: 10 Collyer Quay, 34/F Ocean Financial Centre, Singapore 049315; tel: (65) 6210 1288; fax: (65) 6210 1980) for all matters and queries relating to this report.

BNPP has acquired VERNER INVESTISSEMENTS, the parent company of Exane. VERNER INVESTISSEMENTS is wholly owned and controlled by BNPP, which holds 100% of the share capital and voting rights of VERNER INVESTISSEMENTS.