AI Hardware: The Second Wave (2025-2027)

11 September 2024

Brett Simpson, Analyst
Arete Research Services LLP
brett.simpson@arete.net
+44 (0)20 7959 1320

Nam Hyung Kim, Analyst
Arete Research, LLC
nam.kim@arete.net
+1 424 228 7914

Janco Venter, Analyst
Arete Research Services LLP
Janco.venter@arete.net
+27 71216 3220

Key Summary Takeaways

We see a second wave of AI growth coming in the 2025-2027 period. Training clusters are scaling up, while we see AI inference demand layering in as new architectures drive material reduction in cost per token. See slides 3-12.

The speed of technology change is frightening in the 2025-2027 period. Back-side power, 102T Ethernet, serdes to 400G, Next generation COWOS, UALink, wafer scale technology, HBM4, use of photonics, etc. See slide 4.

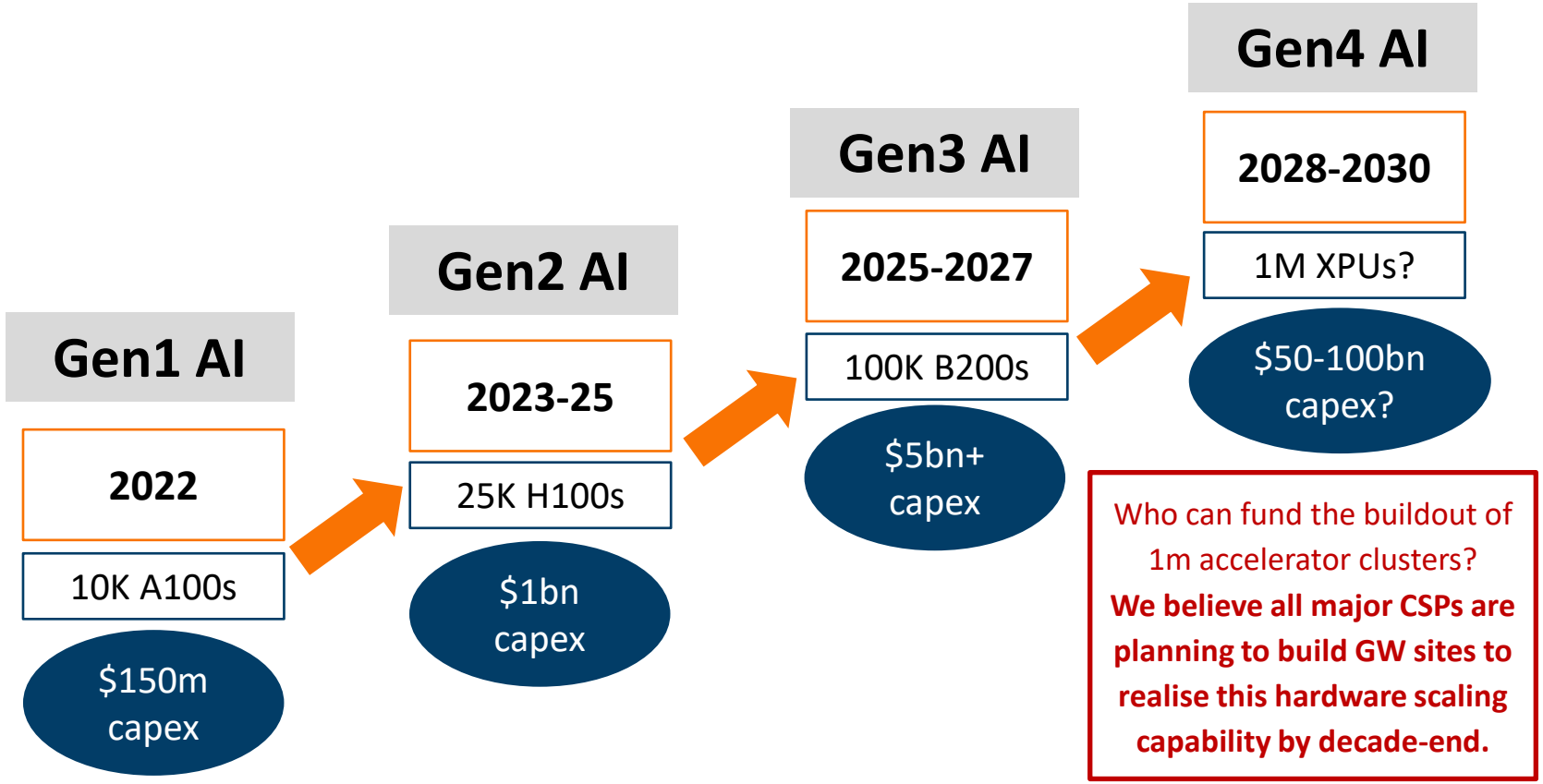
We review industry capacity plans through 2027 for both foundry and HBM, and see a fresh round of LTAs ahead. HBM4 pricing likely to jump 50% on a per GB basis. COWOS could reach 100k wpm by YE27. See slides 5-7.

We look at the challenges around monetization and the opportunities ahead from enterprise, consumer, SaaS co-pilots, and government. Hyperscaler capex this year is broadly capacity add next year. See slides 9-11.

With smartphones and PCs embracing EDGE AI starting in '25, we look at the opportunities and challenges in scaling this up over the next few years. See slide 12.

The Exponential R&D Arm's Race Continues

Cluster Sizes continue to scale exponentially the next five years to support next gen foundation models. Leading-edge model makers need access to this capability at each generation.



We are heading into third-generation Gen AI clusters – but yet to see many of the flagship models from the Gen2 era. Still training-centric market structure today.

Unique Pace of Technology Change through 2027

GPU/XPU	AI accelerators shrinking to TSMC A16 process with backside power. Interposer sizes to jump to 8x mask (6800mm sq) in ’27. System on wafer (SoW) starts ’27?
Switching	Low-latency Ethernet platforms in development – 102T switching commercially available by 2027 with major advances around integration and packaging ahead.
Interconnect	400G serdes on UALink by 2027? Cluster sizes surge next 3-4 years: 1m accelerator cluster = 3m back-end connections? Or will photonic interposers disrupt?
HBM	20HI HBM4E (80GB per HBM chip) by ’27 vs 24GB per chip today - means memory per accelerator will surge. With the push to lower precision (FP2?) ample cost reduction for inference ahead.
Other	Rising chiplet adoption, cache coherent CPUs, wafer scale platforms, FP2 precision, new lasers, AEC cables, liquid cooling, nuclear-powered datacenters

Pace of tech change frightening – so much disruption ahead. The industry needs to convince TSMC to build major capacity through 2027. This planning process has already started.

The Importance of Planning Capacity to 2027

It takes 2 years to build a fab and another year to equip it. New Fabs for ’27 are already planned in.

2nm will see AI accelerators ramp early, given power savings. Intel and Samsung capacity plans still in flux.

It costs around $4bn per 10k wpm to bring on new fab equipment at 2nm. Production time c.130 days.

It takes two years to build a new datacenter. Cloud capacity plans ebb and flow. Capex to sales jump contentious.

AI demand for inference is challenging to forecast. Cyclical risks are rising across AI supply chain.

Table 1: Foundry Capacity Plans, YE27

2nm Capacity Plans - 2027	Wafers Per Month	Annualised		
TSMC Peak Capacity	180k	2,160k	Prior nodes peaked at 120-130k wpm	
Intel Peak Capacity	50k	600k	Arizona and Ohio ramp plan	
Samsung Peak Capacity	30k	360k	SF2 capacity ramp	
Total Industry Capacity 2027	260k	3,120k	just over 3m wafers in CY27	
2nm Capacity Breakdown	**Good Die/Wafer**	**nits (m)**	**Wafers (k/year)**	**% 2nm capacity**
PC Processor	250	250	1000	32.1%
CPU server chips	110	22	200	6.4%
Smartphone AP	500	600	1200	38.5%
Implied Capacity for AI Accelerators	36	20.0	720k	23.1%

Source: Arete Research estimates.

Balance sheets appear challenged at Intel and Samsung. Burden is largely on TSMC to build out 180k wpm or more for its next node (2nm, A16, A14) through 2027.

Accelerator Outlook to 2027: 12m Units, 680k Wafers

Table 2: Accelerator Forecast Build, 2022-2027

Accelerator Market (Units)	2022	2023	2024E	2025E	2026E	2027E
NVIDIA	780k	1,800k	3,400k	4,600k	6,000k	7,000k
Broadcom	350k	600k	1,200k	1,300k	1,600k	1,800k
AMD	0k	0k	400k	800k	1,400k	2,000k
Marvell	0k	0k	185k	525k	750k	950k
Other	0k	0k	100Kk	200k	350k	500k
Total Accelerator Units	**1,130k**	**2,400k**	**5,285k**	**7,425k**	**10,100k**	**12,250k**
Die/Accelerator	1.0	1.0	1.1	1.8	2.0	2.0
Total Die (k)	1,130	2,400	5,814	13,365	20,200	24,500
Good die per wafer	36	36	36	36	36	36
Accelerator Wafer Capacity (k wafers)	**31.4**	**66.7**	**161.5**	**371.3**	**561.1**	**680.6**

Source: Company data, Arete Research estimates.

Why is Foundry Planning Critical For This Period?

- **Scale of Growth is Unique?** Industry planning 5m accelerators in CY24. Projecting 10.5m units (i.e. a doubling) in 2026. Die per accelerator doubling over this period. Implies a quadrupling of die area from '24-'26.
- **TSMC COWOS capacity plan is lifeblood of industry**. Capacity was 12.5k wpm exiting 2023 – could reach 50k wpm by YE25 (a 4x Jump). Could TSMC support 100k wpm by YE27? We assume 680k wafer output for CY27 (i.e. <60k wpm) – capacity planning across industry fundamental. TSMC more dominant inside AI supply chain than NVIDIA.
- **Hopper saw super-normal returns. BoM of c. $4k.** Industry driving for more participants reduces power base. Cost or ownership declines will be sharp to incentivise deployment of inference. Benchmarking around cost per token needed.
- **Rising Risks for Semis Industry.** Almost 90% of incremental gross margin gains of semis industry this cycle is coming from just five companies - all the key AI supply chain players. Never seen such concentration. If there's an air pocket – will be painful.

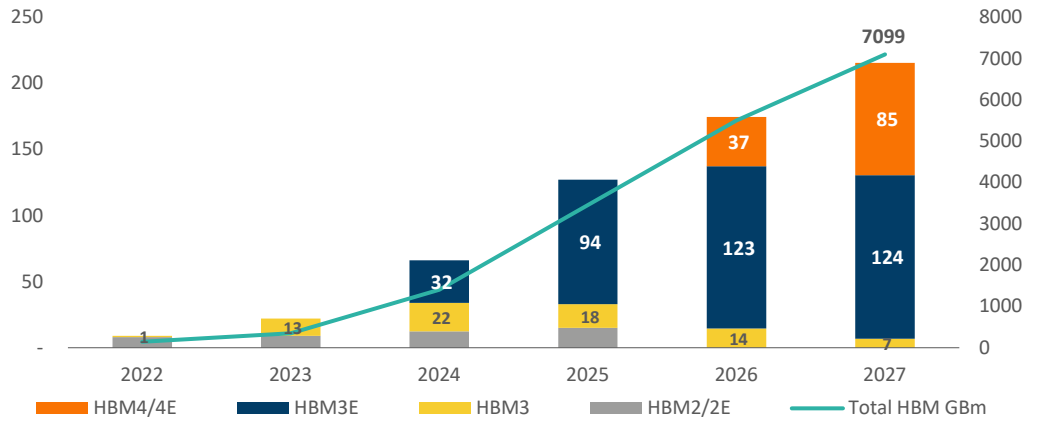
Our estimates imply stellar AI capacity plan in 2025-2027 period. If we see demand air pocket for AI compute, industry goes into downturn.

Memory HBM Supply: Aligning With TSMC Plan

Big Transitions Ahead: HBM3E 12hi and beyond

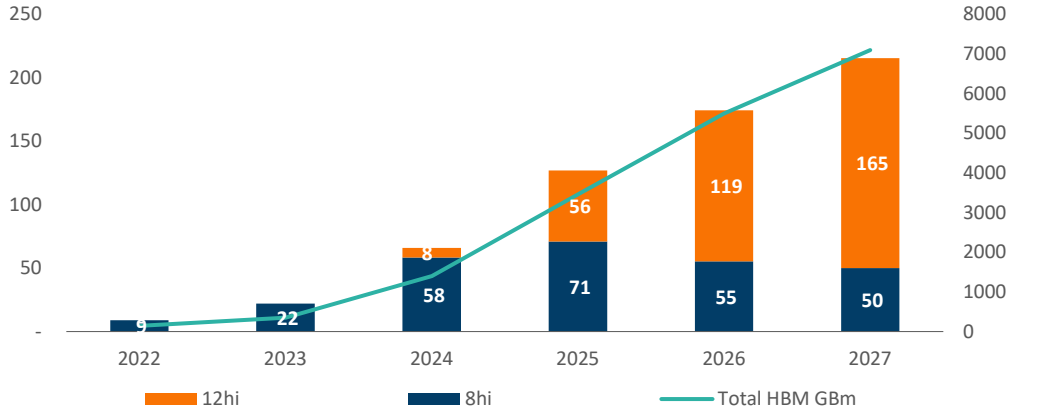
- **We forecast HBM sales to reach $18.2 billion in '24 and anticipate a 4x increase by '27**, driven by HBM3E 12hi adoption in '25 and HBM4 in '26**.**
- **We see four dimensions to growth for the HBM market ahead:**
  - **1) ASPs jump with each node.** The transition to HBM3E is expected to boost HBM ASP by ~25% per GB. HBM4 is expected to command an additional ~50% price premium over HBM3E.
  - **2) Increased GB per layer:** A 50% increase in GB per layer for HBM3E vs HBM3. HBM4E will again improve on HBM3E/HBM4.
  - **3) Rising layer count:** 12Hi adoption takes off in 2H24 with main adoption in ’25 when Blackwell ramps. HBM4 is expected to be introduced by YE25, with adoption in ’26. HBM4E should see 16hi (and possibly higher), further increasing GB/unit.
  - **4) More HBM units are designed around each accelerator.**
- **HBM3E 12hi should reach 8 million units in ’24 and then jump 7x to 57 million in ’25 before commercialization of HBM4 12hi in ’26. Fig.1**
- **Moving beyond ’27, HBM4E is expected to feature 16 to 20 layers.** SK Hynix recently expressed optimism that hybrid bonding could allow stacking more than 20 layers without exceeding 775 micrometers height.

Fig 1: HBM Units by Generation and Total mGB



Source: Arete Research

Fig 2: HBM 12hi vs 8hi unit supply



Source: Arete Research

Memory Still The Most Underestimated Part of Supply Chain Next 3-4 Years. Market Size just over $20bn in ’24, rising over 4x to $88bn by 2027?

AWS CSP Capex Plans: More AI Disclosure Please

New datacenter projects reveal a pattern that CSPs are planning massive new projects out to '27.

FROM AFP NEWS May 22, 2024

Amazon To Invest 15.7 Bn Euros In Spain

Reuters

Amazon to invest $11 billion in Indiana to build data centers April 25, 2024

DATA CENTER Frontier

AWS Eyes 960 MW for Newly Acquired Nuclear Power Data Center in Pennsylvania March 6, 2024

Amazon's AWS to double down on Singapore with additional $9 billion cloud investment

CNBC May 07, 2024

DCD AWS to invest $15bn in cloud computing in Japan January 19, 2024

EqualOcean

Amazon Invests $5.3 Billion in Saudi Data Center Mar 13, 2024

CRN **Amazon's AWS to invest over $5 bln to boost cloud computing in Mexico** 27 Feb 2024

DCD AWS confirmed to be behind $10bn Mississippi data center development January 26, 2024

In 2024 alone, AWS has already announced $70bn+ in new datacenter projects – before factoring in the new GW project in Pennsylvania. This is a marked inflection vs prior years.

AI CSP: Reviewing the Business Model

AI CSPs are moving to min contract value of 1k H100s for 3-4 years

At $2.50/hr per GPU, that equates to $22m per year in IaaS rental revenue

Equipment capex for 1k H100 GPUs around $40-45m. Capex to sales year 1 of 200%!

Depreciation period of 4-6yrs, i.e. up to $10m/yr. Implies Depn to sales of 40%

1K H100 GPU cluster consumes 1.4MW of power at system level

CSPs tell us they have many challenges in driving their business model forward:

- **The upfront capex burden is brutal.** Depreciation costs structurally jump. Payback period of around 2yrs.
- **Will inference run on same infrastructure?** What value-add services can CSPs upsell to drive monetization higher?
- **Upfront engineering.** NVIDIA moving to 1yr product cadence requires early access and fast qualification time.
- **Equipment turnaround**. It takes 6-9 months from GPU sell-in to "in-service". Capex in '24 is capacity in '25.
- **Bottom line.** Will GPUs even last 4-6yrs? What is demand for aging GPUs, after initial 3-4 yr contracts?
- **Datacenter capacity planning and funding.** Scale of new datacenter buildouts is really challenging.
- **Location, Location, Location.** Power costs vary. Getting $40/MW is great value. Power in Europe 5x higher.

G2000 corporates all need this type of infrastructure in time.

Assessing Demand: Enterprise, Consumer, Government

Investors' no. 1 concern is over AI monetization: Elevated investment cycle and uncertain adoption curve. Risk of air pocket ahead?

Consumer AI	**Subscription-led monetization is a well-trodden path. Netflix, Spotify, Prime, energy bills, telecom, etc. billions of users paying $20+ per month will be lucrative.**
Enterprise AI	**Few corporates spending over $10m/year at present, but this should inflect. Today, less than 1% of IT budgets is allocated to AI. On 3yr view, we expect >10% for G2000 corporates.**
SaaS Copilots	**Enterprise software (a $1tn market) readying AI co-pilots as an upsell to core offerings. Can they bolt on 10-20% upsell on 3-5 yr view?**
Government AI	**Global government IT spend reached $700bn in 2023, growing double digits yoy. In our view, there is no government department that will not see meaningful productivity savings from implementing AI across their business next 3-5 years.**

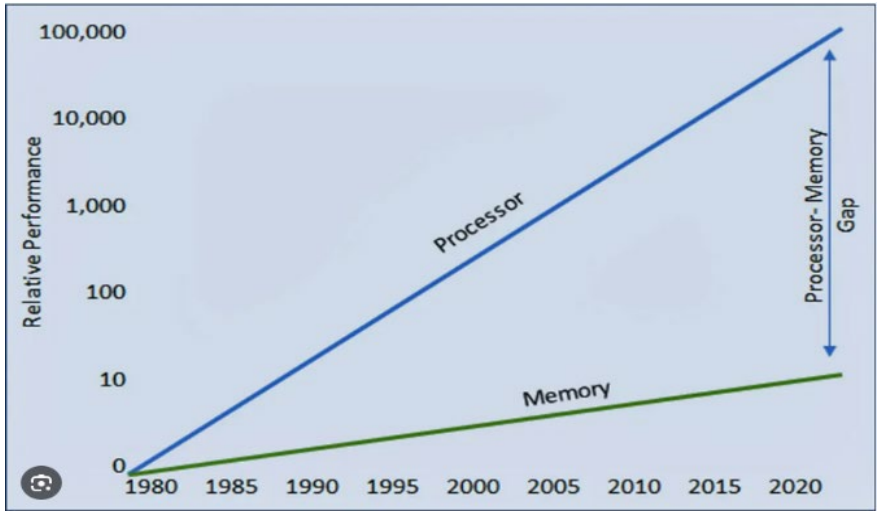
Hyperscalers are the only players that are deploying AI in any scale today. The industry needs to broaden this out in 2025/2026 to avoid a period of disappointment ahead.

On-Device EDGE AI: Starts in 2025

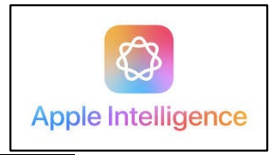
PC and Smartphones "AI Ready"

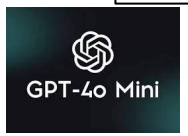
- **The market is readying for a big replacement cycle in PC and Smartphones.** Globally, 600m PCs are 5+ years old. 300m refurbished smartphones were sold in 2024. Out with the old, in with the new.
- **Small models still create power consumption challenges, even with dedicated GPU/NPU compared to smartphones.**
- **The industry faces significant integration issues** – multiple OEMs, multiple model makers, multiple chip architectures. Building out a compelling voice experience for AI is really challenging.
- Despite advancements in various memory technologies, there is a significant "memory wall" **challenge – i.e., performance disparity between processor speed and memory access grows (Fig 3). Can the industry create an LP-HBM adaptation for client-side devices?** JEDEC standards are not dealing with the issues here.
- **NPU compute to scale fast.** Minimum TOPS thresholds will jump. Vertical players such as Apple can move fast. Key area of differentiation.
- **We forecast 300-350m AI smartphones to ship in 2025E (25-30% of the market) and 60-90m AI PCs (20-25% of market). Half of PCs and smartphones to be AI ready in 2026E?**

Fig 3: Memory Wall – Industry Needs To Bridge Gap



Source: EDN Asia





Semis content jumping 10-15% per year to support rising specs. Period of inflation ahead for 2025-2027.

Thank you for listening

Want to see more of our work on AI? Scan the QR code and we will reach out to you directly when we launch later this year.

Disclosures

Regulation AC – The research analyst(s) whose name(s) appear(s) on the front cover of this report certify that: all of the views expressed in this report accurately reflect their personal views about the subject company or companies and its or their securities, and that no part of their compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this report.

Overall Industry Risks: Algorithm changes can take longer than expected, ETH prices could rise enough to offset block reward and difficulty changes in the near term, and a new GPU mineable cryptocurrency driving significant GPU demand could emerge. A deteriorating global economicenvironment could impact the semiconductor industry, rapidly creating significant oversupply, underutilization of fabs, declining ASPs or the write-off of inventory. During '09, semis sales fell 10% (ex-memory). Competition in all sectors is intense. Equally, in some segments the availability of leading-edge capacity promptly is also a problem. The smartphone space is a dynamic market with dozens of players making products that require complex hardware and software integration skills; it is hard to predict how any one vendor might fare with any particular model, though the space for so called " devices is limited by the struggle to differentiate "flat black slabs", i.e., standard touchscreen devices largely running on Android OS.

Primary Analyst(s) Coverage Group: Brett Simpson –Advanced Micro Devices, Broadcom, Credo Technology Group, Intel, Marvell Technology Group, MediaTek, NVIDIA, Qorvo, Qualcomm, Skyworks Solutions, TSMC; Nam Hyung Kim –Lenovo, LG Display, LG Energy Solution, Micron Technology, Rambus, Samsung Electronics, Samsung SDI, SK Hynix, Sony, Universal Display, Western Digital; Jim Fontanelli –Analog Devices, ASM International, ASML Holding, BE Semiconductor, Infineon, NXP Semiconductors, Soitec, STMicroelectronics, Wolfspeed; Blair Botha –Ambarella, GlobalFoundries, United Microelectronics Corp.

Potential Conflicts: Advanced Micro Devices, Broadcom **–** An analyst or a member of an analyst's household owns equity securities in this company.

For important disclosure information regarding the companies in this report, please call +44 (0)207 959 1300, or send an email to allison.kraver@arete.net.

Rating System: Buy (B), Neutral (N) and Sell (S) – A Buy-rated stock is projected to outperform the analyst's industry coverage universe and rise in price over the next 12 months. A Neutral-rated stock is projected to perform in line with the analyst's industry coverage universe over the next 12 months. A Sell-rated stock is projected to underperform the analyst's industry coverage universe and decline in price over the next 12 months. Being assigned a Buy or Sell rating is determined by a stock's absolute return potential, related investment risks and other factors, which may include share liquidity, debt refinancing, estimate risk, economic outlook of principal countries of operation, or other company, political, regulatory, competitive, technological or industry considerations. A stock's absolute return potential represents the difference between the current stock price and the target price over a period as defined by the analyst, and may also include dividends or other forms of capital return forecast due to be paid over the target price period, if the analyst considers that they may be material.

Distribution of Ratings – As of 30 June 2024, 57.9% of stocks covered were rated Buy, 8.8% Sell and 33.3% Neutral.

Global Research Disclosures – This globally branded report has been prepared by analysts associated with Arete Research Services LLP ("Arete LLP"), Arete Research, LLC ("Arete LLC"), and Arete Research Asia Ltd. ("Arete Asia"), as indicated on the cover page hereof. This report has been approved for publication and is distributed in the United Kingdom and European Economic Area (EEA) countries by Arete LLP (Registered Number: OC303210, Registered Office: 10 Queen Street Place, London, EC4R 1AG), which is authorized and regulated by the UK Financial Conduct Authority ("FCA"); in North America by Arete LLC (1309 Walmsley Ave, Dallas, TX 75208, United States), a wholly owned subsidiary of Arete LLP, registered as a broker-dealer with the Financial Industry Regulatory Authority ("FINRA"); and in Asia and Australia by Arete Asia (CE No. ATS894, Registered Office: 3573, Lv 35, Infinitus Plaza, 199 Des Voeux Road Central, Sheung Wan, Hong Kong), which is authorized and regulated by the Securities and Futures Commission in Hong Kong. Additional information is available upon request. Reports are prepared using sources believed to be wholly reliable and accurate but which cannot be warranted as to accuracy or completeness. Opinions held are subject to change without prior notice. No Arete director, employee or representative accepts liability for any loss arising from the use of any advice provided. Please see www.arete.net for details of any interests held by Arete representatives in securities discussed and for our conflicts of interest policy. Please contact Arete Research Services LLP at +44 207 959 1300 in respect of any matters arising from, or in connection with, this document.

Disclosures

U.S. Disclosures – Arete provides investment research and related services to institutional clients around the world. Arete receives no compensation from, and purchases no equity securities in, the companies its analysts cover, conducts no investment banking, market–making or proprietary trading, derives no compensation from these activities and will not engage in these activities or receive compensation for these activities in the future. Arete restricts the distribution of its investment research and related services to institutional clients only. This report may be prepared in whole or in part by research analysts employed by non–US affiliates of Arete LLC that are not registered as broker dealers in the United States. These non–US research analysts associated with Arete LLP and Arete Asia are not licensed as research analysts with FINRA or any other U.S. regulatory authority. Additionally, these analysts may not be associated persons of Arete LLC and therefore may not be subject to Rule 2241 restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.

Singapore Disclosures – This document is distributed in Singapore only to institutional investors (as defined under Singapore's Financial Advisers Regulations ("FAR")) in reliance on Regulation 27(1)(e) of the FAR read in conjunction with Section 23(1)(f) of the Financial Advisers Act, Chapter 110 of Singapore. This document does not provide individually tailored investment advice. Subject to the foregoing, the contents in this document have been prepared and are intended for general circulation. The contents in this document do not take into account the specific investment objectives, financial situation or particular needs of any particular person. The securities and/or instruments discussed in this document may not be suitable for all investors. You should independently evaluate particular investments and strategies and seek advice from a financial adviser regarding the suitability of such securities and/or instruments, taking into account your specific investment objectives, financial situation and particular needs, before making a commitment to purchase any securities and/or instruments. This is because the appropriateness of a particular security, instrument, investment or strategy will depend on your individual circumstances and investment objectives, financial situation and particular needs. The securities, investments, instruments or strategies discussed in this document may not be suitable for all investors, and certain investors may not be eligible to purchase or participate in some or all of them. This document is not an offer to buy or sell or the solicitation of an offer to buy or sell any security and/or instrument or to participate in any particular trading strategy. This document may not be reproduced or provided to any person in Singapore without the prior written permission. The use or reliance on any information in this document is at your own risk and any losses which may be suffered as a result of you entering into any investment are for your account and Arete Research Services LLP and its affiliates shall not be liable for any losses arising from or incurred by you in connection therewith. You will conduct your own evaluation and consult with your own legal, business and tax advisors to determine the appropriateness and consequences of any investment and you will make any investment pursuant to an independent evaluation and analysis of the consequences of the same in reliance only upon your own judgment and not in reliance upon this document and/or any views, representations (whether written or oral), advice, recommendation, opinion, report, analysis, materials, information or other statement by Arete Research Services LLP or any of its affiliates, agents, nominees, directors, officers or employees. Arete Research Services LLP and its affiliates do not hold out any of its affiliates, agents, nominees, directors, officers or employees as having any authority to advise you, and Arete Research Services LLP and its affiliates do not purport to advise you on any investment. You will evaluate and accept all of the risks associated with an investment in any investment. Accordingly, Arete Research Services LLP and its affiliates is entitled to rely on your own independent evaluation and analysis. Any investment will be made at your sole risk and Arete Research Services LLP and its affiliates are not and shall not, in any manner, be liable or responsible for the consequences of any investment.

Asian Disclosures – The contents of this document have not been reviewed by any regulatory authority in Asia. You are advised to exercise caution and if you are in doubt about any of the contents of this document, you should obtain independent professional advice. Whilst considerable care has been taken to ensure the information contained within this document is accurate and up-to-date, no warranty is given as to the accuracy or completeness of any information and no liability is accepted for any errors or omissions in such information or any action taken on the basis of this information. The information may not be current and Arete Asia has no obligation to provide any updates or changes.

Australian Disclosures – Australian investors should note that this document will only be distributed to a person in Australia if that person is: a sophisticated or professional investor for the purposes of section 708 of the Corporations Act of Australia; and a wholesale client for the purposes of section 761G of the Corporations Act of Australia. This document is not intended to be distributed or passed on, directly or indirectly, to any other class of persons in Australia. No analysts who prepared this document hold an Australian financial services license. The information in this document has been prepared without taking into account any investor's investment objectives, financial situation or particular needs. Before acting on the information, the investor should consider its appropriateness with regard to their investment objectives, financial situation and needs. This document has not been prepared specifically for Australian investors. It may contain references to dollar amounts that are not Australian dollars; may contain financial information that is not prepared in accordance with Australian law or practices; may not address risks associated with investment in foreign currency denominated investments; and does not address Australian tax issues. **To the extent that this document contains financial product advice, that advice is provided by Arete Research Asia Limited. Arete Research Asia Limited is exempt from the requirement to hold an Australian financial services license under the Corporations Act with respect to the financial services it provides. Arete Research Asia Limited is regulated by the Securities & Futures Commission of Hong Kong under Hong Kong laws, which differ from Australian laws.**

Disclosures

General Disclosures – This report is not an offer to sell or the solicitation of an offer to buy any security or in any particular trading strategy in any jurisdiction. It does not constitute a personal recommendation or take into account the particular investment objectives, financial situations, or need of the individual clients. Clients should consider whether any advice or recommendation in this report is suitable for their particular circumstances and, if appropriate, seek professional advice. The price and value of the investments referred to in this report and the income from them may fluctuate. Past performance is not a guide to future performance, future returns are not guaranteed, and a loss of original capital may occur. Fluctuations in exchange rates could have adverse effects on the value or price of, or income derived from, certain instruments. As with all investments, there are inherent risks that each individual should address.