# AI Hardware Summit:
# Key Takeaways

September 2022

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

## KEY INSIGHTS

- **The AI hardware industry continues to evolve at a break-neck pace.** We think there are a dozen or more of 10+ Exaflop machines planned for rollout in '23 to support bleeding-edge AI. The need for major hardware upgrades is clear as we enter an era where model sizes are measured in tens of trillions of parameters. Adoption of transformer models by the wider industry is underway, which we think will lead to a substantial growth story in NLP inference (particularly at the edge) in coming years.
- **Given this backdrop, we think AI hardware market growth will continue to impress.** While 2022 saw big upgrades (particularly at Meta) and there are real concerns over the cost of energy that may slow down some projects, we believe there is a fundamental need for big spec upgrades (e.g., look at the EOS supercomputer from NVIDIA with H100).
- **We think it's clear that we will not see a meaningful challenger to NVIDIA in training over the next few years.** Start-ups will get more traction with their new platforms, but need hyperscaler support to really scale up. We were generally disappointed with Intel/AMD proof-points in AI compute (particularly in meeting hyperscaler's massive training clusters) and are concerned that 2023 will be another "development" year for both companies.

We attended this year's AI Hardware Summit to try to gather insights into enterprise and hyperscaler activity in large-scale AI, with a particular focus on the adoption of new training clusters and the read-through into how alternatives to NVIDIA are progressing. This is more of a technical event rather than focussing on strategic insights around AI hardware, but there were impressive keynote presentations from Meta and Azure (among others) as well as a raft of insights from corporate users embracing new AI workloads, particularly around NLP. Our main takeaways are five-fold:

1. **First, it's clear that the use of large transformer models is moving well beyond the major hyperscalers,** and we see the adoption curve jumping sharply in coming years, thanks to the industry finding novel ways to reduce the cost of training by developing and sharing pre-trained models. There is a large and growing industry adapting, tuning and optimising models that are tens and hundreds of billions of parameters in size, without the need for hundreds of data scientists. This is an efficient way for many organisations to reap the rewards that AI can offer, without ground-up training. We think this is going to help spur industry adoption.
2. **Second, Hyperscalers continue to operate at the bleeding edge with research model sizes that are 10-100x larger than the wider community.** Training times here are extending beyond what is reasonable. The message from hyperscalers to the chip industry is clearly to "up their game" as Moore's Law is not keeping pace with industry developments. Systems for training starting next year will be measured in tens of exaflops, and we see 10x jumps in system bandwidth and models heading to tens (if not hundreds) of trillions of parameters. There is little question a major re-architecting of datacenters to support large-scale AI is underway, and it is evident to us that it is not just the need for accelerator upgrades but major advances in network I/O and memory are needed to evolve hardware to keep up with the industry trends at leading-edge. We do not see much competition to Nvidia's EOS supercomputer in performance, which looks set to act as a blueprint for many exaflop systems getting rolled out in 2023 and beyond.
3. **Third, while the leading-edge training market is entering a new phase of performance, the deployment of all this is still in the early phases. This is inevitably going to lead to an explosion in inference deployments at the edge in the next few years,** in our view, particularly around NLP workloads. AI is moving beyond being largely a cloud-based

training phenomenon, and we expect to see a raft of accelerated servers deployed in conjunction with the growth in cloud operations.

4. **Fourth, with energy costs rising sharply of late (particularly in Europe), power is a *hot* topic.** The industry has an opex problem with energy that could stunt growth near-term – power-hungry instances were already expensive, but some projects could get pushed out given cost of power hikes. There is clearly a need for real innovation here, particularly with new accelerator cards and network I/O consuming vastly more power and generating far more heat than ever. New exaflop systems will all need water-cooling.
5. **Fifth, alternatives to NVIDIA are making progress with new training systems, and while its hardware specs look impressive, we think they still need a few years to mature their capabilities, particularly in software, APIs, SDKs, as well as enterprise reach.** Hyperscalers have begun to qualify alternatives (AMD at Azure, Habana at AWS, etc.), but these are not on a scale that NVIDIA should be concerned about. We were generally disappointed with progress made by Intel and AMD at this event, particularly in tangible software stack development. Start-ups (such as Sambanova, Cerebras, Graphcore, etc) are all now generating revenue (some more than others) and are in a race to evolve beyond $100m in annual revenues – we think traction with hyperscalers in '23 is a must. The inference market is still relatively nascent, but we were impressed with the role Qualcomm was taking, particularly in its modular approach to AI, leveraging its smartphone volumes into server-class systems like the AI 100.

## Company Takeaways: Major Developments Look Encouraging

We summarise below some of the main sessions from the event along with recent AI compute news.

- **Meta** has had a big year in AI development. At the beginning of the year, its RSC compute cluster would be among the top 5 supercomputers in the world with over 6k A100 GPUs, but this is expected to reach 16k GPUs by YE22, with H100 upgrades to follow. Meta also uses a 5.4k A100 GPU cluster from Azure. So they have substantially upgraded their hardware platform to drive major AI advances. It now supports the translation of 200 native languages between each other (i.e., not using English as an intermediary) on its NLP platform, and has made its OPT-175B model (open pre-trained transformer model) available on open source to share for free. Meta has also taken the decision to integrate its AI framework (Pytorch) into The Linux Foundation. These are big strategic moves from Meta that could give Microsoft and Open AI's GPT-3 pre-trained models a run for their money. Meta is also running model sizes for its Recommendation engine of up to 10 trillion parameters and has also made big advances in content understanding, using NLP and computer vision to manage objectionable content. Overall, Meta is running thousands of training jobs concurrently, it trains a whopping 6bn images a day, and it is now processing 200bn translations and 2tn predictions each day. The scaling of models by 10x-100x in future years will mean these clusters only jump in size. Meta was clear in its message to the hardware community, that advances in accelerator compute is one thing, but they need to see equally advances in memory and networking to keep up with the pace of change in model sizes and datasets.
- **Microsoft Azure** was the first hyperscaler to build a massive cloud-based AI training cluster and with new models scaling so fast (e.g., MT-NG 530G), Microsoft believes compute capacity needed to run models doubles every 3.5 months – but with hardware tracking Moore's Law, there is a performance gap that looks structurally challenging. Microsoft continues to scale its AI infrastructure. Its agreement with Meta for a dedicated cluster of 5.4K NVIDIA A100 GPUs earlier this year is a blueprint for what lies ahead. This builds on Azure's custom 10K GPU cluster that was built to support Open AI's GPT-3 training among other large models. While it continues to drive performance with NVIDIA, Azure recently committed to building a training cluster with AMD using MI-250 GPUs, although this size of this is not yet disclosed. Azure has also launched GPT-3 as a service, which it sees as gaining traction with wider cloud customers who are looking for turnkey solutions.

- **Google** is ramping its new TPUv4 chips, four years after its TPUv3 launched in 2018, with plans to build a 9 exaflop system later this year (based on Bfloat16) using over 32k interconnected TPUs (8 pods) with each TPU chip boasting 10x more interconnect bandwidth than the prior version. They did not present at the show, but we think its working with Broadcom on this ASIC project as it has all prior TPUs.

- **NVIDIA** will soon be finished building its EOS supercomputer, with 18.4 Exaflops of compute (at FP8) using its new H100 GPU, Grace CPU, NV switches and Bluefield 3 DPUs (claiming 3x performance uplift vs. A100 using FP16, or 6x performance uplift on FP8). This compares with its current Selene supercomputer built two years ago, which offered up to 2.8 Exaflops at FP16. This means NVIDIA's new EOS machine is a significant upgrade (which we think can scale well beyond 4k GPUs in the initial EOS rollout) and should act as a blueprint for any company looking to train trillion-plus parameter models. While this has not been publicly benchmarked, NVIDIA expects training times to be 9x faster compared with the fastest A100 systems, and its new NV link 4.0 claims to offer 7x more interconnect bandwidth than the prior generation. Training large language models often took months with top AI supercomputers based on A100, and we heard that performance gains started to decline as more GPUs were added to clusters. With the H100, NVIDIA aim to resolve these issues – something we think will lead to healthy upgrade cycle. With most US hyperscalers looking to finalise qualification of H100 in 1H23, we expect commercial ramp to see widespread adoption in CY23.
- **AMD and Intel** have announced AI training roadmaps – AMD are focussing on MI Instinct GPUs while Intel are supporting Habana AI chips (rather than its GPUs) but we have yet to see large training clusters for AI being adopted by hyperscalers. Azure with AMD have committed to supporting the MI250-based GPU (in conjunction with Genoa CPU chips) while AWS has committed to Habana's new Gaudi 2 chips (most likely paired with Sapphire Rapids) but as yet neither chipmaker has announced internal POD systems for these products, and we think several generations of software development are needed before these players realistically have a mature-enough roadmap to have a meaningful impact on this market. So while Intel and AMD have exascale systems to ramp in 2023 for HPC, we need to see what they can do in AI to compete with clusters such as EOS. Thus far, we have been underwhelmed by AI developments from both companies and are growing concerned that 2023 will be another year without real commercial progress.
- **Sambanova** is making progress with new announcements both on the systems side of the business (with the new Cardinal SM30 system) as well as its services business, with subscription offerings for enterprises covering a raft of "foundation" models. On the system side, Sambanove has combined two of its existing chips into a new approach so that this twin-socket supports twice the compute capacity, twice the local memory capacity, and twice the memory bandwidth of the first generations of machines. This is giving the company a much broader capability in building its support for larger models, without new tape-outs. While Sambanova's offerings are not yet available from any hyperscaler, and we do not have much in the way of public pricing to compare with Nvidia, we think it already enjoys the most customer traction of any of the start-ups thus far.
- **Cerebras** announced its new wafer scale cluster, based on its second-generation platform – CS-2. This means Cerebras plans to support a GPT3 model on a single-CS2 system, with the potential to be able to train 100tn parameter models on one system. So Cerebras created high-performance fabric interconnect – to connect memory X storage appliance to CS2 accelerators. It is called SwarmX and it uses tree topology. While Cerebras have had commercial success with some of the top pharma companies and HPC national research labs, it is clearly positioning its wafer scale cluster into the major cloud players.
- **Hugging Face**, the AI model provider, has seen a tripling in model downloads in the last 12 months, and its new Bloom model (launched this summer) supports 196bn parameters and is beginning to gain real traction with enterprises. With thousands of companies actively using Hugging Face's pre-trained libraries and model optimisations on a day to day basis (including US hyperscalers as its most active users), Hugging Face is becoming a key player in the transformer eco-system.

In terms of major AI enterprise end-users at the event, we interviewed a handful of players that shared with us some of their compute requirements and hardware plans. Overall, we were pleasantly surprised by the number of enterprises actively engaged in training transformer models (mainly in the cloud) and planning major NLP inference deployments in coming years. We summarise our interviews below.

- We spoke with a **major US fast-food chain** with over 7,000 restaurants that expects to roll out GPU servers in each of its restaurants as new NLP-based services and computer vision applications get rolled out across its stores. So while this company continues to scale its training plans in the cloud, there are substantial investments in NLP inferencing systems to come, a market still relatively nascent today. This business also was shifting from Windows to Android and X86 to Qualcomm Snapdragon as part of this digital transformation project. We believe that inference rollouts like this are going to be more commonplace at the EDGE – an area we think will scale sharply over the next few years.

- We talked with a **major European pharmaceutical company** about its experience working with alternative hardware to NVIDIA, but it still found a lack of maturity, particularly at the interface level. Even with platforms such as Amazon Sagemaker, the ease and stability in transitioning from training to tuning to model deployment all on NVIDIA GPUs makes for high barriers to entry. So while AWS supports Habana, we see limited instance support, and large language model optimisations and APIs are still relatively niche.
- **Petrobras**, Brazil's state-owned oil company, has been significantly upgrading its compute capacity and should have a new AI supercomputer called Pegaso installed by YE22 with a cluster of 2k NVIDIA A100s. This will mean Petrobras will have doubled its compute capacity in each of the last four years, as it drives up sub-surface mapping and simulation. We think this level of compute scaling is typical of what we are seeing elsewhere in the energy sector at present.
- **In EU, Cineca and Atos** are also building what will be among the world's largest AI supercomputers (called Leonardo) with a cluster of 14k A100 GPUs, for drug discovery and weather modelling. We see Atos driving water cooling adoption with new H100 systems from NVIDIA, with plans to support 150kw per rack, using its new DLC platform. Atos also plans to provide water-cooling support for **Graphcore's** up-and-coming "Good" supercomputer, supporting 10 exaflops of compute.

**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.

**Required Disclosures**

**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 economic environment 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, Intel, Marvell Technology Group, MediaTek, NVIDIA, Qorvo, Qualcomm, Skyworks Solutions, TSMC; Nam Hyung Kim – Lenovo, LG Display, LG Energy Solution, Micron Technology, 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.

For important disclosure information regarding the companies in this report, please call +44 (0)207 959 1300, or send an email to michael.pizzi@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 2022, 63.6% of stocks covered were rated Buy, 12.1% Sell and 24.3% 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 (15 Broad St., Boston, MA 02109), 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: 3822, Lv 38, 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.

**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.**

**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.