HSBC Global Research

Free to View
Disruptive Technology - Equities
China | Thematic research

August 2023
www.research.hsbc.com

By: Frank He (S1700517120005), Bingyi Zheng (S1700521060001), Steven Wang (S1700523030001)

China AI hardware

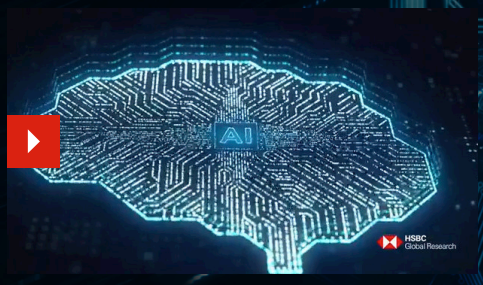
A manual for eight key sub-sectors

AI is driving up demand for more advanced computing power

We profile eight hardware sub-sectors that stand to benefit most ...

... and the opportunities and challenges from localisation

This is a Free to View version of a report with the same title published on 4-Aug-23. Please contact your HSBC representative or email AskResearch@hsbc.com for more information.



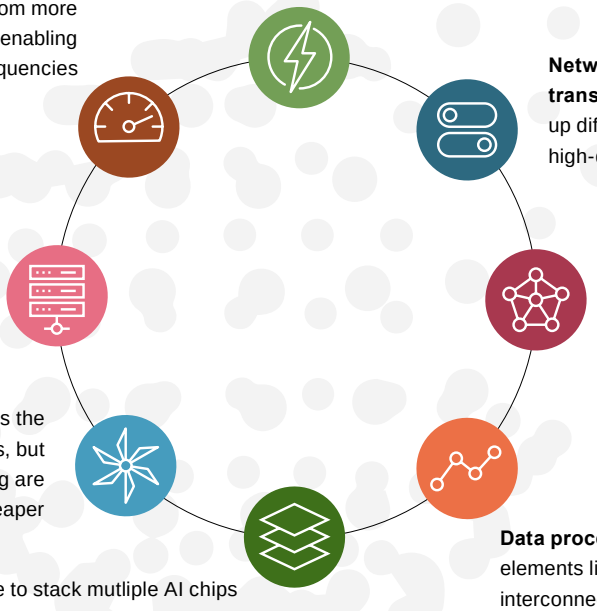
Play video

Sector snapshot: China AI hardware

Which hardware sectors are best positioned in the "Generative AI" ecosystem?

Accelerators: today's workhorse is the graphics processing unit (GPU), but application-specific integrated circuits (ASICs) are emerging as a favourite as they lower power consumption and accelerate workloads

Printed circuit boards (PCB): these connect electronic components to one another and stand to benefit from more powerful AI chips and servers by enabling higher speeds and frequencies



Networking switches and optical transceivers: equipment that connects up different components to execute high-demand workloads, as for AI training

Server manufacturing: given the need for collaboration across multiple companies in the supply chain, an AI-specific servers are both complex and costly

Interconnectors and interfaces: growing computing power is increasing the need for more sophisticated interconnector technology to support low latency and high throughput

Cooling system: air cooling is the traditional way to cool down servers, but various forms of liquid cooling are emerging as they're better and cheaper

Data processing units (DPUs): made up of elements like a core processor, high-speed interconnect and memory. Their core function is to offload some of the workload from computer processing units

Advanced packaging: able to stack mutliple AI chips and components vertically, but challenges remain around power distribution and heat dissipation

USD98bn

The PCB market is expected to grow from USD82bn in 2022 to USD98bn 2027e

49% CAGR

The global DPU market is set to increase from USD7.5bn in 2022 to USD24.5bn in 2025e

USD25.8bn

The advanced packaging market is expected to grow from USD9.2bn in 2022 to USD25.8bn in 2028e

Opportunities and challenges from localisation

	Challenges	Opportunities
Hardware	US restrictions on China's procurement of advanced GPUs and accelerators from overseas suppliers	We see opportunities in AI hardware systems, such as chip architecture, processing nodes, interconnect bandwidth and memory capacity
Software	There remains a large gap between China and the global leaders when it comes to the development of large language models (LLMs)	More than 79 LLMs are now under development in China

Source: CCID, Yole Development, HSBC Qianhai Securities

China AI hardware

- AI is driving up demand for more advanced computing power …
- … so we map out the eight hardware sub-sectors that stand to benefit most
- We also detail opportunities and challenges from localisation

We see opportunities across the AI hardware supply chain

As generative AI applications – like OpenAI's ChatGPT and Google's Bard chatbots – grow bigger and better, we expect investment in the hardware that runs these tools to head higher. Our upbeat view on demand for AI computing infrastructure rests on two catalysts: (1) major cloud and technology companies are racing to develop and train their own large language models; (2) early movers in AI, such as OpenAI, are rushing to support their growing user base by making smarter inferences and training more advanced models.

The focus of this note is to provide a manual to the eight key sub-sectors that we expect to benefit most from technology upgrades as the industry races to meet rising demand for generative AI. We detail each sector's development roadmap, growth profile, and related companies.

1. Accelerators. These are the chips that speed up the processing of information-intensive operations like AI. The most common type today is the graphics processing unit (GPU), an all-rounder that is fairly robust. But another type, application-specific integrated circuits (ASICs), are emerging as a favorite as they can lower power consumption and accelerate certain workloads. Meanwhile, different types of memory technology are also being rolled out such as near-memory computing (processing close to where data resides) and in-memory computing (running computer calculations entirely in the computer memory). We also look at the so-called software stack for developing AI models – such as programming frameworks, toolkits and compilers – which also plays a critical role in accelerating performance.

2. Networking and optical transceivers. A typical data centre consists of servers, networking equipment and storage. Networking is a vital part as it connects up different components to form computing clusters which execute high-demand workloads like AI training and inferencing. However, given AI workloads require ultra-high bandwidth, low latency and zero signal loss, we expect strong demand for high-speed 25.6T/51.2T switches and 800G optical transceivers which transmit signals. Linear-drive pluggable optics (LPO) are emerging as a new technology as these can reduce power consumption in switches and transceivers, have lower latency speeds, and are less costly. And we believe even more advanced co-packaging optics (CPO) may become more common in two to three years. Finally, Google's optical circuit switches are also an alternative approach as they cut the total cost of ownership.

3. Interconnect IP and interface IC. The growing amount of computing power needed for AI applications is increasing the need for more sophisticated interconnect technology to support low latency speeds and high throughput. In particular, interconnect intellectual property (IP) is needed for a wide range of chip connections. We think memory expansion and pooling which allows dynamic memory allocation and therefor increases memory usage efficiency – is the key trend, as evidenced by the success of Nvidia's NVLink and NVSwitch technology though others continue to push a different Compute Express Link (CXL) based interconnect platform.

4. Data Processing Units (DPUs). These are increasingly popular devices inside data centre servers as they combine networking technology with AI and computing power. Their principle function is to offload some of the workload (like networking, storage and security management) from CPUs, which then frees up CPUs to focus on other important tasks. DPUs consist of a core processor, acceleration engine, network interface, a high-speed interconnect and memory. The global DPU market is set to increase from USD7.5bn in 2022 to USD24.5bn in 2025, implying a 49% CAGR, according to CCID. While cloud service providers generally have their own in-house DPU products, we see opportunities for independent DPU vendors, mainly in the telecoms and enterprise market.

5. Advanced packaging. As AI chips integrate more and more transistors to improve their computing power, it has become essential to use more advanced packaging. One form of technology is 2.5 dimensional packaging with TSMC's technology the most common choice for most AI chip makers, including Nvidia. 3D packaging – stacking dies vertically, and also led by TSMC – offers more integration but there are still some technical challenges around power distribution and heat dissipation. While China-based semiconductor companies are unable to seize front-end opportunities (like designing and manufacturing GPU chips, an area currently dominated by Nvidia, AMD and TSMC) as they lag behind in manufacturing technology, we expect China-based outsourced semiconductor assembly and test players to win back-end packaging orders for Nvidia and AMD's AI chips.

6. Cooling systems. One unwanted byproduct of high-performing servers is an awful lot of heat. Unless cooled, server performance can suffer and worse, catch fire. Blowing cool air across or around a server is the traditional approach. However, this isn't enough for modern high-performance computing which generates a formidable amount of thermal energy, since air is not an effective heat transfer medium. And too many cooling fans causes a lot of noise. Hence, liquid cooling – which can involve wrapping a loop containing cold water around the hottest components like CPUs and GPUs to absorb their heat – is increasingly being used by cloud server providers as it is power and cost efficient. BlueOcean says air cooling a 10kW server rack costs USD7.02 per watt and liquid cooling costs USD6.98, with the costs falling when the consumption power of a server rack increases to 20kW and 40kW respectively. Liquid cooling is the best method to cool server racks consuming 50kW or higher power.

7. Servers. The server manufacturing industry is particularly complex given the need for collaboration across the supply chain. We split the production process into 12 steps and detail the leaders along the way. We also conducted a bill of materials analysis and find AI servers are 17-39x more expensive than general-purpose servers. We also see signs China is catching up in servers given ByteDance and Alibaba have placed massive GPU orders with Nvidia (source: Pandaily,14 June 2023).

8. Printed circuit boards (PCB). This sub-sector is likely to benefit from growing demand for high-performance AI servers and upgraded switches as it enables devices to better share information and communicate with each other. In a typical AI server (such as Nvidia's popular DGX), we estimate the PCB content value be four to five times more than a regular server driven by **(1)** high value substrate – which connects chips to the board – for GPUs; **(2)** high-density-interconnects (HDI) in advanced types of accelerator modules; and **(3)** multi-layer PCBs for the GPU unit baseboard. Copper clad laminate (CCL), one of the main materials used in PCB, also has an opportunity to increase its content value as there's rising demand for CCL as this can enable higher speeds and higher frequencies.

Could AI just be all hype?

While we're bullish on AI, we are very aware of the hype surrounding past technologies like the metaverse, virtual reality and autonomous driving which have failed to deliver, so far, all that they promised. It's always been difficult to forecast precisely how far and fast emerging technology will emerge. For AI though, we do see real benefits for a wide range of users. The key question is how deeply generative AI applications can influence daily work and life, and whether there will be any killer applications. We are closely watching the development of multi-modal models, which can accept both text and images as inputs, and also Artificial General Intelligence (AGI) which is a system were computers have human-level cognitive abilities.

China's localisation opportunities and challenges

In our final chapter, we assess the current state of China's AI sector. On the software front, China is fairly competitive in terms of developing application-specific AI models, but there's a big gap between where China is in large language models (LLMs), or foundation models, and where the global leaders are. Still, there are plenty of reasons to be hopeful. In the first half of 2023, China tech companies released various LLMs and AI applications. In fact, 79 LLMs were launched in China by the end of May 2023 (source: Reuters, 30 May 2023).

On the hardware front, given there are US restrictions on China procuring certain GPUs and accelerators from overseas suppliers, we see ample localisation opportunities ahead in AI hardware. In particular, we believe China can improve in areas like chip architecture, processing nodes, interconnect bandwidth and memory capacity.

Foundation models are accelerating computational requirements

Finally, a quick word on the big picture and how fast the industry is changing. According to Nvidia, computational requirements are now set to increase eight times every two years for models excluding transformers (deep learning architecture), but that ratio surges to a staggering 275x for transformer-based AI models. That means today's computation power will double, on average, every three-and-a-half months. However, the problem is Moore's Law states that semiconductor transistors on a chip double only every two years. This means the industry needs to develop more advanced computing devices to catch up with rapidly evolving AI models.

This is a Free to View version of a report with the same title published on 4-Aug-23. Please contact your HSBC representative or email AskResearch@hsbc.com for more information.

The full note takes a deep dive into each of the eight subsectors mentioned above, along with a look at China's localisation opportunities and challenges.

Disclosure appendix

The following analyst(s), who is(are) primarily responsible for this document, certifies(y) that the opinion(s), views or forecasts expressed herein accurately reflect their personal view(s) and that no part of their compensation was, is or will be directly or indirectly related to the specific recommendation(s) or views contained in this research report: Frank He, Bingyi Zheng and Steven Wang

This document has been issued by the Research Department of HSBC.

HSBC and its affiliates will from time to time sell to and buy from customers the securities/instruments, both equity and debt (including derivatives) of companies covered in HSBC Research on a principal or agency basis or act as a market maker or liquidity provider in the securities/instruments mentioned in this report.

Analysts, economists, and strategists are paid in part by reference to the profitability of HSBC which includes investment banking, sales & trading, and principal trading revenues.

Whether, or in what time frame, an update of this analysis will be published is not determined in advance.

For disclosures in respect of any company mentioned in this report, please see the most recently published report on that company available at *www.hsbcnet.com/research*.

Additional disclosures

1 This report is dated as at 4 August 2023.

2 All market data included in this report are dated as at close 2 August 2023, unless a different date and/or a specific time of day is indicated in the report.

3 HSBC has procedures in place to identify and manage any potential conflicts of interest that arise in connection with its Research business. HSBC's analysts and its other staff who are involved in the preparation and dissemination of Research operate and have a management reporting line independent of HSBC's Investment Banking business. Information Barrier procedures are in place between the Investment Banking, Principal Trading, and Research businesses to ensure that any confidential and/or price sensitive information is handled in an appropriate manner.

4 You are not permitted to use, for reference, any data in this document for the purpose of (i) determining the interest payable, or other sums due, under loan agreements or under other financial contracts or instruments, (ii) determining the price at which a financial instrument may be bought or sold or traded or redeemed, or the value of a financial instrument, and/or (iii) measuring the performance of a financial instrument or of an investment fund.

5 This document may be a translation of a document authored in another language. If so, and if there is any discrepancy between versions, the original-language version shall prevail.

Disclaimer

Issuer of report
HSBC Qianhai Securities Limited
.

This document has been issued by HSBC Qianhai Securities Limited ("HSBC Qianhai Securities"), which has based this document on information obtained from sources it believes to be reliable but which it has not independently verified. Neither HSBC Qianhai Securities nor any member of its group companies ("HSBC") make any guarantee, representation or warranty nor accept any responsibility or liability as to the accuracy or completeness of this document and is not responsible for errors of transmission of factual or analytical data, nor is HSBC liable for damages arising out of any person's reliance on this document. The information and opinions contained within this document are based upon publicly available information at the time of publication, represent the present judgment of HSBC Qianhai Securities and are subject to change without notice. This document may contain hyperlinks to external websites for convenience of its recipients and HSBC is not responsible for any content therein.

The Hongkong and Shanghai Banking Corporation Limited owns 90% and Qianhai Financial Holdings Co., Ltd. ("QFH") owns 10% of shares in HSBC Qianhai Securities, which prepared and/or contributed to this document. HSBC Qianhai Securities has established policies and procedures reasonably designed to prevent QFH from exercising direct or indirect influence over the content of HSBC Qianhai Securities research products and the choice of companies that will be the subject of research products. Furthermore, HSBC Qianhai Securities has established additional policies and procedures reasonably designed to prevent any person or entity, whether from within HSBC Qianhai Securities, QFH or otherwise, from influencing the activities of HSBC Qianhai Securities' research analysts or the content of research products.

This document is not and should not be construed as an offer to sell or solicitation of an offer to purchase or subscribe for any investment or other investment products mentioned in it and/or to participate in any trading strategy. It does not constitute a prospectus or other offering document. Information in this document is general and should not be construed as personal advice, given it has been prepared without taking account of the objectives, financial situation or needs of any particular investor. Accordingly, investors should, before acting on it, consider the appropriateness of the information, having regard to their objectives, financial situation and needs. If necessary, seek professional investment and tax advice.

The decision and responsibility on whether or not to purchase, subscribe or sell (as applicable) must be taken by the investor. In no event will any member of the HSBC group be liable to the recipient for any direct or indirect or any other damages of any kind arising from or in connection with reliance on any information and materials herein.

Past performance is not necessarily a guide to future performance. The value of any investment or income may go down as well as up and you may not get back the full amount invested. Where an investment is denominated in a currency other than the local currency of the recipient of this document, changes in the exchange rates may have an adverse effect on the value, price or income of that investment. In case of investments for which there is no recognised market it may be difficult for investors to sell their investments or to obtain reliable information about its value or the extent of the risk to which it is exposed. Some of the statements contained in this document may be considered forward looking statements which provide current expectations or forecasts of future events. Such forward looking statements are not guarantees of future performance or events and involve risks and uncertainties. Actual results may differ materially from those described in such forward-looking statements as a result of various factors.

This document is for information purposes only and may not be redistributed or passed on, directly or indirectly, to any other person, in whole or in part, for any purpose. The distribution of this document in other jurisdictions may be restricted by law, and persons into whose possession this document comes should inform themselves about, and observe, any such restrictions. By accepting this document, you agree to be bound by the foregoing instructions. If this document is received by a customer of an affiliate of HSBC, its provision to the recipient is subject to the terms of business in place between the recipient and such affiliate.

This document is intended to be distributed in its entirety. Unless governing law permits otherwise, you must contact a HSBC Group member in your home jurisdiction if you wish to use HSBC Group services in effecting a transaction in any investment mentioned in this document.

Certain investment products mentioned in this document may not be eligible for sale in some states or countries, and they may not be suitable for all types of investors. Investors should consult with their HSBC representative regarding the suitability of the investment products mentioned in this document.

HSBC and/or its officers, directors and employees may have positions in any securities in companies mentioned in this document. HSBC may act as market maker or may have assumed an underwriting commitment in the securities of companies discussed in this document (or in related investments), may sell or buy securities and may also perform or seek to perform investment banking or underwriting services for or relating to those companies and may also be represented on the supervisory board or any other committee of those companies.

From time to time research analysts conduct site visits of covered issuers. HSBC policies prohibit research analysts from accepting payment or reimbursement for travel expenses from the issuer for such visits.

HSBC Qianhai Securities Limited is regulated by the China Securities Regulatory Commission and is qualified to engage in Securities Investment Advisory Business in China [91440300MA5EPLHG1B].

[1218584]