Read carefully the questions and adequately justify your answers.
Consultation of class or other materials is not allowed.
The use of simple, non-alphanumeric, non-programmable calculators is permitted, including scientific calculators. The use of graphing calculators is not allowed.
Answer Group 1, Group 2 and Group 3 on separate sheets.

Nvidia's GPUs – Processors for AI development

Adapter from: https://www.economist.com/business/2024/05/19/can-nvidia-be-dethroned-meet-the-startups-vying-for-its-crown and https://www.economist.com/business/2024/03/20/can-anything-stop-nvidias-jensen-huang – last consulted in 12 June 2024

Demand for Nvidia's GPUs, Artificial Intelligence (AI) modellers' favourite type of processor, is so insatiable that they are in short supply. Nvidia announced the launch later this year of a new generation of superchips, named Blackwell, that are many times more powerful than its existing GPUs, promising bigger and cleverer AIs. Thanks to AI, spending on global data centres was $250bn last year and is growing at 20% a year.

Access to GPUs, and in particular those made by Nvidia is vital for any company that wants to be taken seriously AI. Analysts talk of companies being "GPU-rich" or "GPU-poor", depending on how many of the chips they have.

GPUs do the computational heavy lifting needed to train and operate large AI models. Yet, oddly, this is not what they were designed for. The initials stand for "Graphics Processing Unit", because such chips were originally designed to process videogame graphics. It turned out that, fortunately for Nvidia, they could be repurposed for AI workloads.

Might it be better to design specialist AI chips from scratch? That is what many companies, small and large, are now doing in a bid to topple Nvidia. Dedicated AI chips promise to make building and running AI models faster, cheaper or both. Any firm that can mount a credible threat to the reigning champion will have no shortage of customers, who dislike its lofty prices and limited supplies.

Ordinary processing chips, like those found inside laptop and desktop computers, are in essence designed to do one thing after another. GPUs, by contrast, contain several thousand processing engines, or "cores", which let them run thousands of versions of the same simple task (like drawing part of a scene) at the same time. Running AI models similarly involves running lots of copies of the same task in parallel. Figuring out how to rewrite AI code to run on GPUs was one of the factors that triggered the current AI boom.

Yet GPUs have their limitations, particularly when it comes to the speed with which data can be shuffled on and off them. Modern AI models run on large numbers of interconnected GPUs and memory chips. Moving data quickly between them is central to performance. When training very large AI models, some GPU cores may be idle as much as half of the time as they wait for data.

Cerebras' response to this is to put 900,000 cores, plus lots of memory, onto a single, enormous chip, to reduce the complexity of connecting up multiple chips and piping data between them. Its CS-3 chip is the largest in the world by a factor of 50. On-chip connections between cores operate hundreds of times faster than connections between separate GPUs, Cerebras claims, while its approach reduces energy consumption by more than half, for a given level of performance, compared with Nvidia's most powerful GPU offering.

Other startups in this area include Hailo, based in Israel; Taalas, based in Toronto; Tenstorrent, an American firm using the open-source risc V architecture to build AI chips; and Graphcore, a British company that is thought to be about to sell itself to SoftBank, a Japanese conglomerate. Big tech firms are also building AI chips. Google has developed its own "tensor processing units" (TPUs), which it makes available as a cloud-computing service. Amazon, Meta and Microsoft have also made custom chips for cloud-based AI; OpenAI is planning to do so as well. AMD and Intel, two big incumbent chipmakers, make GPU-like chips already.

One danger for the newcomers is that their efforts at specialisation could go too far. Designing a chip typically takes two or three years, says Christos Kozyrakis, a computer scientist at Stanford University, which is "a huge amount of time" given how quickly AI models are improving. The opportunity, he says, is that the startups could end up with a chip that is better at running future models than Nvidia's less specialised GPUs are. The risk is that they specialise in the wrong thing.

Another challenge is that Nvidia's software layer for programming its GPUs, known as CUDA, is a de facto industry standard, despite being notoriously fiddly to use. "Software is king," says Mr Kozyrakis of Stanford, and Nvidia has a significant advantage, having built up its software ecosystem over many years. AI-chip startups will succeed only if they can persuade programmers to rejig their code to run on their new chips. They offer software toolkits to do this, and provide compatibility with the major machine-learning frameworks. But tweaking software to optimise performance on a new architecture is a difficult and complex business—yet another reason Nvidia is hard to dislodge.

The biggest customers for AI chips, and the systems built around them, include model-builders (such as OpenAI, Anthropic and Mistral) and tech giants (such as Amazon, Meta, Microsoft and Google). It may make sense for such companies to acquire an AI-chip startup, and keep its technology to themselves, in the hope of besting the competition. Instead of trying to compete with Nvidia, chip startups could position themselves as acquisition targets.

But there are dangers upstream the supply-chain for Nvidia as well. You need only to recall the supply-chain problems of the pandemic, as well as the subsequent Sino-American chip wars, to see that dangers lurk. Nvidia's current line-up of GPUs already faces upstream bottlenecks. South Korean makers of high-bandwidth memory chips used in Nvidia's products cannot keep up with demand. TSMC, the world's biggest semiconductor manufacturer, which actually churns out Nvidia chips, is struggling to make enough of the advanced packaging that binds GPUs and memory chips together. Moreover, Nvidia's larger integrated systems contain around 600,000 components, many of which come from China. That underscores the geopolitical risks if America's tensions with its strategic rival keep mounting.

GROUP 1 (5 points)

1. Please explain the difference between segmentation and targeting. (3 points)

   who do we reach   how do we reach

2. Based on the information provided, what is Nvidia's target market strategy? Please justify your answer. (2 points)

GROUP 2 (8 points)

3. According to the information provided, perform an analysis of the industry competition using Porter's 5-forces framework and taking Nvidia's perspective (please identify the relevant information influencing each of the forces, discussing its impact for Nvidia; identify strengths and weaknesses of Nvidia; identify, whenever possible actors included in each of the forces). (3 points) Buyers, Suppliers, Barriers to enter, rivalry, alternatives

   SWOT

value proposition
solution
unique features
Key benefits

4. Using only the information provided in the text, characterize the product-market fit of Cerebras in detail (including solution, unique features, key benefits, customer archetype, job to be done, and current workflow). (3 points)

customer segment
customer arch
job to be done
current workflow

5. Given the challenges faced by startups such as Cerebras, particularly considering the challenge of their "efforts of specialization could go too far", and based on the customer development process, how would you recommend that such a startup should proceed to reduce that risk? (2 points)

GROUP 3 (7 points)

explain ratios

6. Based on the information about the balance sheet and the income statement of Nvidia presented below: i) calculate the financial ratios for 2022 and 2023; ii) perform a financial analysis of the company based on the sales growth and the calculated financial ratios. Considering your analysis would you invest in Nvidia?

INCOME STATEMENT	**2023**	**2022**
Sales	60 922	26 974
Costs of Goods Sold	16 621	11 618
Gross Profit	44 301	15 356
Operational expenses	11 329	11 132
EBITDA	32 972	4 224
Net Income	29 760	4 368

(values in million dollars)

BALANCE SHEET	**2023**	**2022**
Current Assets	44 345	23 073
Non-current Assets	21 383	18 109
Total Assets	65 728	41 182
Current Liabilities	10 631	6 563
Non-current Liabilities	12 119	12 518
Total Liabilities	22 750	19 081
Equity	42 978	22 101

(values in million dollars)

FINANCIAL RATIOS				
Category	Ratio		**2022**	**2023**
Profitability	Return on Assets	(net income/total assets)		
	Return on Equity	(net income/total equity)		
Efficiency	Asset turnover	(net sales/total assets)		
Liquidity	Current ratio	(current assets/current liabilities)		
	Working Capital	(current assets - current liabilities)		
Leverage	Equity to assets ratio	(total equity/total assets)		
	Debt to Equity	(total liabilities/total equity)		