Question 3 Hansard page 11

Senator DAVID POCOCK: Earlier you said that Australia doesn't have the capacity to do large language model AI.

Dr Hajkowicz: I wouldn't say we don't have that capacity, but the challenge is getting GPUs, graphics processing units. They are AI accelerator chips that will let us train up a really big, powerful AI models. There are not many of them in the world. They're made by Nvidia and AMD. The Nvidia H100 is currently the gold standard. Everyone wants it.

Senator DAVID POCOCK: So you're saying there are no Australian companies—

Dr Hajkowicz: It's hard for me to know, because I don't know what's happening, but I suspect that getting access to those is one of the key barriers that Australian AI makers would face.

Senator DAVID POCOCK: Why doesn't the CSIRO know what's happening globally?

Dr Hajkowicz: Because of commercial privacy. If an Australian company is buying these things, they may not signal it to the marketplace. They may not want to tell anyone.

Senator McGRATH: How much do they cost?

Dr Hajkowicz: I'm afraid I'm going to get the number wrong. They're not cheap. Can I take that on notice? They're not cheap, and it's expensive. The United States National Artificial Intelligence Research Resource, the NAIRR, which has been created is about making three things available to the US AI ecosystem to unleash what they can do across the entire ecosystem. It's about giving them access to GPUs, high-performance computing; it's about giving them access to data; and it's about skills and capability uplift so they can do it. There's that view. And the European AI factories that have been announced are creating the same sorts of capabilities. That's the policy mechanism they're using to increase the rate of AI building and AI making in those places.

Answer

Nvidia does not officially disclose the pricing of its H100 graphics processing unit (GPU) products as numerous factors influence the price, such as the volume of the batch and overall volumes that a particular client procures from Nvidia. Nvidia also does not sell its GPUs directly to the market, but through Original Equipment Manufacturers (OEMs) such as Dell and Hewlett Packard Enterprise (HPE) who package them into server offerings.

Noting the above, the approximate cost for a Nvidia H100 GPU is AUD$40,000-50,000. Please also note that the Nvidia H100 GPU is just one component in a larger server system. To be usable, an H100 would need to be procured as part of a server. These servers often contain multiple GPUs. This system would then need to be housed in an appropriately configured and secured data centre facility and collocated with appropriate data storage.

There are also many other suppliers of GPUs and other types of accelerator chips that are useful for AI development. It is important to acknowledge many other suppliers exist and as such pricing can and does vary widely.

In sizing an "AI infrastructure" there are a number of factors that would influence the size and cost. These include the size of the model (number of parameters) to be trained, and the time to train a model that can be tolerated by projects/users.

Another consideration that would influence size is the approach to model development – this could be training sovereign foundational models vs sourcing pre-trained foundational models from commercial or international partners and then fine tuning those models on sovereign datasets. There is a significant difference in infrastructure cost between these two approaches to model development.

Dependent on the above considerations, multiple servers housing multiple H100 GPUs may be formed into a high-performance computing (HPC) cluster. The size of the cluster and number of GPUs it houses will have significant implications for the time taken to train or fine tune an AI model. A system could range anywhere from 1 GPU to many 1000's.

The US NAIRR and EU AI Factory initiatives are targeting serving the needs of a broad swathe of research, innovation and industry and are ultimately intended for use by many research and academic institutions, as well as industry and commercial entities. These sovereign investments are in the multi-billion-dollar investment range and can be expected to deliver sizeable GPU systems that will house many thousands to tens of thousands of GPUs, along with associated datacentre, data storage and other ancillary equipment, including the skilled workforce to manage and deliver such a system.

At the extreme scale, foundational models such as ChatGPT have demanded very large sums of investment from OpenAI and its investment partners. Whilst not publicly disclosed, it has been estimated that an HPC cluster comprising of 10,000 GPUs was used to train recent OpenAI ChatGPT models.

An Nvidia DGX server is a common and broadly adopted system for AI. Each DGX server houses 8 H100 GPUs and can be procured for approximately AUD$450,000. For the above OpenAI ChatGPT example, this would require an HPC cluster of 1,250 DGX servers, at a cost of around $563m. A rough approximation of ancillary costs would take the needed investment closer towards one billion dollars.

A moderated and less costly approach would not involve "training" models, but sourcing pre-trained foundational models from commercial, open source, or international partners and then fine tuning those models on sovereign datasets. This approach of "tuning" would not require an infrastructure of the scale outlined above and could be achieved with 50 DGX servers, and ancillary items at a cost of $100 million.

Whilst there is currently no definitive view of the relevant global availability and investments undertaken, there are some efforts underway to track sovereign investments into AI infrastructure. The OECD Expert Group on AI Compute and Climate[1] is one such effort. The top 4 countries where Nvidia recognised revenue in the quarter ending October 2023 include the USA (34.77%,) Taiwan (23.91%), China (including Hong Kong) (22.24%), and Singapore (15%). Significant national investments in AI infrastructure have also been publicly announced by the USA, Canada, UK, Europe, Japan, Vietnam and Indonesia.

[1] Expert Group on Compute & Climate - OECD.AI https://oecd.ai/en/network-of-experts/working-group/1136