# Topic

"NVIDIA 2024年第二季度財報公佈：創紀錄的收入、數據中心增長、AI創新、供應鏈管理、新GPU發布、企業合作夥伴、未來投資策略、網路解決方案、DGX雲端動能以及軟體業務擴張"

# Summary

NVIDIA在2024年第二季度的財報中報告了創紀錄的135.1億美元收入，較前一年有顯著增長，主要由於雲端服務提供商和大型消費者互聯網公司的需求驅動。財務長Colette Kress對於對中國出口增加的監管可能引起的潛在憂慮表示，他們不預期這將對財務結果產生即時的重大影響。她還強調了該公司與MediaTek的合作，以及開發新的NVIDIA GPU晶片產品線，包括即將推出的L40s。執行長Jensen Huang強調了該公司對其產品的強烈需求，以及由於在AI、數據中心技術、遊戲和汽車工業的持續創新和合作，有可能繼續增長。他還討論了他們的H100產品的複雜性和奇蹟，以及該公司與VMware的合作，將生成AI帶給全球的企業。黃更進一步闡述了在加速計算和生成AI的未來投資策略。他還討論了該公司的網路解決方案和DGX Cloud的成功。此外，黃強調了該公司軟體業務的增長以及NVIDIA AI Enterprise的潛力。

# CFO's explanation about short-term financial situation

Colette Kress強調了該公司創紀錄的第二季度收入為135.1億美元，環比增長88%，年增101%。數據中心的收入為103.2億美元，環比增長141%，年增171%。她還提到，該公司預計供應量將在明年每個季度都有所增加。儘管對中國的出口可能會有更多的規定，但Kress保證他們不預期這將對他們的財務結果產生即時的重大影響。她還強調了遊戲收入的強勁增長，環比增長11%，年增22%。Kress還提到該公司與MediaTek合作開發汽車SOCs和一個新的NVIDIA GPU晶片產品線，包括即將推出的L40s。她澄清說，大型CSPs在第二季度的收入中貢獻了略多於50%，其次是消費者互聯網公司，然後是企業和高性能計算。

# CEO's description about the company's outlook

黃仁勳強調了公司對其產品的強烈需求，特別是來自雲服務提供商和大型消費者互聯網公司的需求，這暗示了一個充滿希望的未來。他還強調了持續的創新以及世代交替升級數據中心基礎設施以加速計算和AI，這表明了持續增長的潛力。黃仁勳還強調了公司在AI、數據中心技術、遊戲和汽車行業的合作夥伴關係和創新，包括與Snowflake、微軟、VMware和聯發科的合作，以及新AI模型和應用的開發。他還提到即將上市的DGX-GH200系統和Bluefield 3 DPU的成功。黃仁勳進一步解釋說，由於行業正在轉向加速計算和生成AI，公司對需求的持續性充滿信心。他還討論了他們的H100產品的複雜性和奇妙之處，以及他們的軟體生態系統在他們成功中的重要性。他進一步闡述了公司與VMware的合作，將生成AI帶給全球的企業。黃仁勳還談到了未來在加速計算和生成AI上的投資策略，強調了從通用計算轉向這些新技術的轉變。他還討論了公司的網路解決方案和DGX Cloud的成功。此外，黃仁勳強調了公司軟體業務的增長以及NVIDIA AI Enterprise的潛力。

# The issues of market concern

在此次會議中，主要市場關注的問題是對NVIDIA向中國出口的可能增加的規定。雖然該公司並不預期這將立即影響其財務結果，但它承認長期的限制可能導致美國產業永久失去在全球最大市場之一競爭的機會。其他的關注點包括全球對遊戲的終端需求，該需求在去年的緩慢增長後已經恢復成長，以及由於GPU驅動筆記本的日益普及，整體遊戲收入實際情況的變化。該公司還解決了對供應鏈的擔憂，表示他們預計在接下來的季度和下一財政年度內繼續增加供應。此外，人們還關注在未來幾年的需求狀況是否可持續，以及是否會有足夠的應用或使用案例，讓客戶在AI和加速計算上的投資獲得合理的回報。

## 介紹NVIDIA第二季度財報電話會議

### Abstract

NVIDIA的第二季度財報電話會議討論了該公司創紀錄的第二季度收入，達到135.1億美元，較上季度增長88%，較去年同期增長101%。會議還強調了該公司的數據中心收入，達到103.2億美元，較上季度增長141%，較去年同期增長171%。該公司的增長受到雲服務提供商和大型消費者互聯網公司需求加速的推動。會議還討論了對NVIDIA對中國出口可能增加的規定，表示雖然他們不預期這將立即對其財務結果產生實質影響，但長期限制可能導致美國行業在全球最大的市場之一中競爭的永久性機會損失。

### Investment insight

NVIDIA在第二季的出色表現顯示出對其產品的強烈需求，尤其是在數據中心領域。該公司的成長是由雲服務提供商和大型消費者互聯網公司的需求加速驅動的。然而，對NVIDIA向中國的出口可能增加的規定可能對公司的長期成長構成風險。

### Keywords

"NVIDIA，第二季度財報公佈，創紀錄的收入，數據中心收入，雲端服務提供商，消費者網路公司，出口規定，中國。"

### 中文逐字稿

```您今天的會議主持人。現在，我想歡迎大家參加NVIDIA的第二季度業績電話會議。今天的會議正在錄音。為了防止任何背景噪音，所有線路都已被靜音。在演講者發表講話後，將進行問答環節。如果您在此期間想提問，只需在您的電話鍵盤上按下星號鍵，然後按下數字一。如果您想撤回您的問題，再次按下星號一。謝謝。Simona Jankowski，您可以開始您的會議。謝謝。大家下午好，歡迎參加NVIDIA的2024財年第二季度業績電話會議。今天與我一起參加會議的NVIDIA成員包括Jensen Huang，總裁兼首席執行官，以及Colette Kress，執行副總裁兼首席財務官。我想提醒大家，我們的通話正在NVIDIA的投資者關係網站上現場直播。直播將在我們討論2024財年第三季度財務結果的電話會議結束後仍可重播。今天通話的內容是NVIDIA的財產。未經我們事先書面同意，不得複製或轉錄。在此通話中，我們可能會根據當前的期望做出前瞻性聲明。這些聲明可能受到許多重大風險和不確定性的影響，我們的實際結果可能會有重大差異。有關可能影響我們未來財務結果和業務的因素的討論，請參閱今天的業績公告，我們最近的10-K和10-Q表格，以及我們可能在8-K表格上向證券交易委員會提交的報告。我們所有的聲明都是在今天，2023年8月23日，根據我們目前可獲得的信息做出的，除非法律要求，我們不承擔更新任何此類聲明的義務。在此通話中，我們將討論非GAAP財務指標。您可以在我們的CFL評論中找到這些非GAAP財務指標與GAAP財務指標的對照，該評論已在我們的網站上發布，現在讓我把通話轉給Colette。謝謝Simona。我們這個季度表現出色。創紀錄的第二季度收入為135.1億美元，環比增長88%，同比增長101%，超出我們110億美元的預期。首先讓我從數據中心開始。創紀錄的收入為103.2億美元，環比增長141%，同比增長171%。數據中心計算收入同比近乎翻了兩番，主要受到雲服務提供商和大型消費者互聯網公司對我們HGX平台的需求加速推動，該平台是生成AI和大型語言模型的引擎。包括AWS、Google Cloud、Meta、Microsoft Azure和Oracle Cloud在內的主要公司，以及越來越多的GPU雲服務提供商正在大量部署基於我們Hopper和Ampere架構的Tensor Core GPU的HGX系統。網絡收入同比幾乎翻了一番，主要受到我們端到端InfiniBand網絡平台的推動，該平台是AI的黃金標準。對NVIDIA加速計算和AI平台的需求非常大。我們的供應合作夥伴在提高產能方面表現出色，以支持我們的需求。我們的數據中心供應鏈，包括擁有35,000個零件和高度複雜的網絡，已經在過去的十年中建立起來。我們還為製造過程中的關鍵步驟開發並合格了額外的產能和供應商，例如CoWAS封裝。我們預計供應將在明年的每個季度增加。從地理位置來看，數據中心的增長在美國最強，因為客戶將他們的資本投資導向AI和加速計算。中國的需求在我們的數據中心收入的歷史範圍內，包括計算和網絡解決方案，占我們數據中心收入的20%到25%。現在，讓我花一點時間來討論最近關於我們對中國出口可能增加規定的報告。我們認為目前的規定正在達到預期的結果。鑑於我們的產品在全球的需求強勁，我們不預期如果採取額外的數據中心GPU出口限制，將對我們的財務結果產生立即的重大影響。然而，從長期來看，如果實施禁止我們的數據中心GPU銷售到中國的限制，將導致美國行業在世界上最大的市場之一中競爭和領先的機會永久性地丟失。我們的雲服務提供商在本季度對HDX系統的需求非常強烈，因為他們正在進行一次世代轉換，升級他們的數據中心基礎設施，以迎接加速計算和AI的新時代。NVIDIA的HDX平台是我們在硅系統、互聯網、網絡、軟件和算法方面近二十年全棧創新的結晶。現在，AWS、Microsoft Azure和幾家GPU雲服務提供商已經普遍提供了由NVIDIA H100 Tensor Core GPU驅動的實例，其他的也將很快跟進。消費者互聯網公司也推動了非常強烈的需求。他們在專為AI打造的數據中心基礎設施上的投資已經產生了顯著的回報。例如，Meta最近強調，自從推出Reels和AI推薦以來，Instagram上的使用時間已經增加了超過24%。企業也在競相部署生成AI，推動了在雲中消耗NVIDIA驅動實例的強勁需求，以及對現場基礎設施的需求。無論我們是在雲中還是現場，通過合作夥伴還是直接服務客戶，他們的應用程序都可以在NVIDIA AI企業軟件上無縫運行，並可以訪問我們的加速庫、預訓練模型和API。我們與Snowflake建立了合作夥伴關係，為企業提供加速創建使用他們自己的專有數據的定制生成AI應用程序的途徑，所有這些都在Snowflake數據雲中安全地進行。有了NVIDIA MIMO平台來開發大型語言模型，企業將能夠從snowflake數據雲中製作自定義的LLM，用於高級AI服務，包括聊天機器人搜索和摘要。幾乎每個行業都可以從生成AI中受益。例如，像Microsoft剛剛宣布的那樣，AI副駕駛可以提高超過十億辦公室工作人員和數千萬軟件工程師的生產力。數百萬的法律服務、銷售、客戶支持和教育專業人員將可以利用在他們領域內訓練的AI系統。AI副駕駛和助手將為我們的客戶創造新的數千億美元的市場機會。我們在營銷、媒體和娛樂中看到了生成AI的一些最早的應用。世界上最大的營銷和通信服務組織WPP正在使用NVIDIA Omniverse開發一個內容引擎，使藝術家和設計師能夠將生成AI整合到3D內容創建中。WPP的設計師可以使用從技術提示創建圖像，同時負責訓練生成AI工具和來自NVIDIA合作夥伴，如Adobe和Getty Images的內容，使用NVIDIA Picasso，一個用於視覺設計的自定義生成AI模型的鑄造廠。視覺內容提供商Shutterstock也正在使用NVIDIA Picasso來構建工具和服務，讓用戶可以在生成AI的幫助下創建3D場景背景。我們與ServiceNow和Accenture合作推出了AI燈塔計劃，加快企業AI能力的開發。AI燈塔Unite將ServiceNow的企業自動化平台和引擎與NVIDIA的加速計算以及Accenture的諮詢和部署服務結合起來。我們還與Hugging Face合作，簡化企業創建新的和自定義AI模型。Hugging Face將為企業提供一種新的服務，讓他們可以訓練和調整由NVIDIA VGF Cloud驅動的先進AI模型。而且，就在昨天，VMware和NVIDIA宣布了一個重大的新企業產品，稱為VMware Private AI Foundation with NVIDIA，這是一個完全集成的平台，包括AI軟件和加速隊列來自NVIDIA，以及多雲軟件，專為運行VMware的企業。VMware的數十萬企業客戶將能夠訪問到需要定制模型和運行生成AI應用程序的基礎設施、AI和雲管理軟件，例如智能聊天機器人助手、搜索和摘要。我們還宣布了新的NVIDIA AI企業就緒服務器，配備了新的NVIDIA L40S GPU，專為行業標準數據中心服務器生態系統和Bluefield-3 DPU數據中心基礎設施處理器而設計。L40S不受CoWAS供應的限制，並正在向全球領先的服務器系統製造商出貨。L40S是一種通用數據中心處理器，專為高容量數據中心擴展而設計，以加速最計算密集的應用程序，包括AI訓練和嬰兒期、3D設計和視覺化、視頻處理和NVIDIA Omniverse工業視覺化。NVIDIA AI企業就緒服務器完全優化了VMware雲基礎設施和私有AI基礎設施。將有近100種配置的NVIDIA AI企業就緒服務器將很快從全球領先的企業IT計算公司，包括Dell、HPE和Lenovo，提供。結合我們基於ARM的Grace CPU和Hopper GPU的GH200 Grace Hopper Superchip已經進入全面生產，並將在本季度在OEM服務器中提供。它也正在向多個超級計算客戶出貨，包括洛斯阿拉莫斯國家實驗室和瑞士國家計算中心。NVIDIA和SoftBank正在合作開發一個基於GH200的平台，用於生成AI和5G、6G應用。我們的race-hopper超級芯片的第二代版本，配備最新的HBM3E內存，將在2024年第二季度提供。我們宣布了GGX GH200，這是一種新類型的大內存AI超級計算機，用於巨型AI語言模型、推薦系統和數據分析。```

### 原文逐字稿

your conference operator today. At this time, I'd like to welcome everyone to NVIDIA's second quarter earnings call. Today's conference is being recorded. All lines have been placed on mute to prevent any background noise. After the speaker's remarks, there'll be a question and answer session. If you'd like to ask a question during this time, simply press the star key followed by the number one on your telephone keypad. If you'd like to withdraw your question, press star one once again. Thank you. Simona Jankowski, you may begin your conference. Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the second quarter of fiscal 2024. With me today from NVIDIA are Jensen Huang, President and Chief Executive Officer, and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's investor relations website. The webcast will be available for replay until the conference call to discuss our financial results for the third quarter of fiscal 2024. The content of today's call is NVIDIA's property. It can be reproduced or transcribed without our prior written consent. During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent form 10-K and 10-Q, and the report that we may file on form 8-K with the Securities and Exchange Commission. All our statements are made as of today, August 23rd, 2023, based on information currently available to us except is required by law we assume no obligation to update any such statement. During this call we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFL commentary which is posted on our website and with that let me turn the call over to Collette. Thanks Simona. We had an exceptional quarter. Record Q2 revenue of $13.51 billion was up 88% percent sequentially and up 101 percent year on year and above our outlook of 11 billion. Let me first start with data center. Record revenue of 10.32 billion was up 141 percent sequentially and up 171 percent year on year. Data center compute revenue nearly tripled year on year, driven primarily by accelerating demand from cloud service providers and large consumer internet companies. for our HGX platform, the engine of generative AI and large language models. Major companies, including AWS, Google Cloud, Meta, Microsoft Azure, and Oracle Cloud, as well as a growing number of GPU cloud providers are deploying in volume HGX systems based on our Hopper and Ampere architecture Tensor Core GPU. Networking revenue almost doubled year on year, driven by our end-to-end InfiniBand networking platform, the gold standard for AI. There is tremendous demand for NVIDIA, accelerated computing and AI platforms. Our supply partners have been exceptional in ramping capacity to support our needs. Our data center supply chain, including HDF with 35,000 parts and highly complex networking, has been built up over the past decade. We have also developed and qualified additional capacity and suppliers for key steps in the manufacturing process, such as CoWAS packaging. We expect supply to increase each quarter through next year. By geography, data center growth was strongest in the US as customers direct their capital investments to AI and accelerated computing. China demand was within the historical range of 20 to 25% of our data center revenue, including compute and networking solutions. At this time, let me take a moment to address recent reports on the potential for increased regulations on our export to China. We believe the current regulation is achieving the intended result. Given the strength of demand for our products worldwide, we do not anticipate that additional export restrictions on our data center GPUs, if adopted, would have an immediate material impact to our financial results. However, over the long-term, restrictions prohibiting the sale of our data center GPUs to China, if implemented, will result in a permanent loss of an opportunity for the US industry to compete and lead in one of the world's largest markets. Our cloud service providers drove exceptional strong demand for HDX systems in the quarter as they undertake a generational transition to upgrade their data center infrastructure for the new era of accelerated computing and AI. The NVIDIA HDX platform is culminating of nearly two decades of full stack innovation across silicon systems, interconnect, networking, software, and algorithms. Instances powered by the NVIDIA H100 Tensor Core GPUs are now generally available at AWS, Microsoft Azure, and several GPU cloud providers, with others on the way shortly. Consumer internet companies also drove the very strong demand. Their investments in data center infrastructure purpose-built for AI are already generating significant returns. For example, Meta recently highlighted that since launching Reels and AI recommendations have driven a more than 24% increase in time spent on Instagram. Enterprises are also racing to deploy generative AI, driving strong consumption of NVIDIA-powered instances in the cloud, as well as demand for on-premise infrastructure. Whether we serve customers in the cloud or on-prem, through partners or direct, their applications can run seamlessly on NVIDIA AI enterprise software with access to our acceleration libraries, pre-trained models, and APIs. We announced a partnership with Snowflake to provide enterprises with accelerated paths to create customized generative AI applications using their own proprietary data, all securely within the Snowflake data cloud. With the NVIDIA MIMO platform for developing large language models, enterprises will be able to make custom LLMs for advanced AI services, including chatbot search and summarization right from the snowflake data cloud. Virtually every industry can benefit from generative AI. For example, AI copilot, such as those just announced by Microsoft, can boost the productivity of over a billion office workers and tens of millions of software engineers. Millions of professionals in legal services, sales, customer support, and education will be available to leverage AI systems trained in their field. AI copilot and assistants are set to create new multi-hundred billion dollar market opportunities for our customers. We are seeing some of the earliest applications of generative AI in marketing, media, and entertainment. WPP, the world's largest marketing and communication services organization, is developing a content engine using NVIDIA Omniverse to enable artists and designers to integrate generative AI into 3D content creation. WPP designers can create images from tech prompts while responsibly trained generative AI tools and content from NVIDIA partners, such as Adobe and Getty Images, using NVIDIA Picasso, a foundry for custom generative AI models for visual design. Visual content provider Shutterstock is also using NVIDIA Picasso to build tools and services that enable users to create 3D scene backgrounds with the help of generative AI. We partnered with ServiceNow and Accenture to launch the AI Lighthouse program, fast-tracking the development of enterprise AI capabilities. AI Lighthouse Unite will service now enterprise automation platforms and engines with NVIDIA Accelerated Computing and with Accenture Consulting and Deployment Services. We are collaborating also with Hugging Face to simplify the creation of new and custom AI models for enterprises. Hugging Face will offer a new service for enterprises to train and tune advanced AI models powered by NVIDIA VGF Cloud. And just yesterday, VMware and NVIDIA announced a major new enterprise offering called VMware Private AI Foundation with NVIDIA, a fully integrated platform featuring AI software and accelerated queuing from NVIDIA with multi-cloud software for enterprises running VMware. VMware's hundreds of thousands of enterprise customers will have access to the infrastructure, AI, and cloud management software needed to customize models and run generative AI applications, such as intelligent chatbot assistance, search and summarization. We also announced new NVIDIA AI Enterprise Ready Server featuring the new NVIDIA L40S GPU built for the industry standard data center server ecosystem and Bluefield-3 DPU data center infrastructure processor. L40S is not limited by co-op supply and is shipping to the world's leading server system makers. L40S is a universal data center processor designed for high volume data center scaling out to accelerate the most compute-intensive applications, including AI training and infancy, 3D design and visualization, video processing, and NVIDIA Omniverse industrial visualization. NVIDIA AI Enterprise Ready servers are fully optimized for VMware, cloud foundation, and private AI foundation. Nearly 100 configurations of NVIDIA AI Enterprise Ready servers will soon be available from the world's leading enterprise IT computing company, including Dell, HPE, and Lenovo. The GH200 Grace Hopper Superchip, which combines our ARM-based Grace CPU with Hopper GPU, entered full production and will be available this quarter in OEM servers. It is also shipping to multiple supercomputing customers, including Los Alamos National Lab and the Swiss National Computing Center. And NVIDIA and SoftBank are collaborating on a platform based on GH200 for generative AI and 5G, 6G applications. The second generation version of our race-hopper super chip with the latest HBM3E memory will be available in Q2 of calendar 2024. We announced the GGX GH200, a new class of large memory AI supercomputer for giant AI language models, recommendator systems, and data analytics.

## "NVIDIA NVMe交換系統和DGX-GH200系統"

### Abstract

NVIDIA已經推出其NVMe交換系統，該系統允許所有256個Grace Hopper Superchips作為一體運作，這是一個重大的進步，因為前一代只能連接八個GPU。預計DGX-GH200系統將在今年年底前提供。該公司還宣布了NVIDIA Spectrum X，這是一個旨在優化AI工作流的乙太網路加速平台。由於GeForce RTX 40系列GPU在筆記本電腦和桌面電腦上的應用，NVIDIA的遊戲收入連續增長了11%，年增22%。

### Investment insight

NVIDIA在AI和GPU技術上的進步，以及其強勁的財務表現，顯示出成長和獲利的潛力。該公司的新產品和技術，如NVMe交換系統和NVIDIA Spectrum X，可能驅動未來的收入增長。此外，NVIDIA在遊戲領域的強勁表現，特別是其GeForce RTX 40系列的GPU，暗示了該公司充滿希望的前景。

### Keywords

"NVIDIA，NVMe交換系統，DGX-GH200系統，Grace Hopper超級晶片，NVIDIA Spectrum X，AI工作流程，GeForce RTX 40系列GPU，遊戲收入。"

### 中文逐字稿

你是一位資深的中文翻譯員，會將以下的英文內容以忠於原文不多加延伸跟解釋的方式翻譯成繁體中文，並且熟知 Azure 以及 OpenAI 對於內容審查的所有標準，因此當在你產生新的文字前，請先確認新產生的詞彙是否會違反 Azure 及 OpenAI 的審核標準，若有任何一點可能違反的機會，請立即停止文字生成並置換成其他相似但不違反的 Azure 及 OpenAI 政策的詞句。  
  
這是新的NVIDIA NVMe交換系統的首次使用，使所有256個Grace Hopper Superchips能夠一起工作。與前一代只能連接8個GPU的10億次跳躍相比，這是一個巨大的飛躍。預計DGX-GH200系統將在今年年底前上市。Google Cloud、Meta和Microsoft將是首批獲得訪問權的公司。強大的網絡增長主要由InfiniBand基礎設施驅動，用於連接HDX GPU系統。由於其端到端優化和網絡內計算能力，InfiniBand的性能是傳統以太網或AI的兩倍以上。對於價值數十億美元的AI基礎設施，InfiniBand增加的吞吐量的價值達到數億美元，並為網絡付費。此外，只有InfiniBand可以擴展到數十萬個GPU。這是領先的AI實踐者的首選網絡。對於希望優化其AI性能的以太網為基礎的雲數據中心，我們宣布了NVIDIA Spectrum X，這是一個旨在優化AI工作流的以太網的加速網絡平台。Spectrum X將Spectrum 4以太網交換機與Bluefield 3 DPU結合在一起，實現了比傳統以太網更好的整體AI性能和功耗效率。Bluefield 3 DPU是一個重大的成功。它正在與主要的OEM進行資格認證，並在多個CSP和消費者互聯網中快速推廣。現在轉向遊戲。遊戲收入為24.9億美元，環比增長11%，同比增長22%。增長主要由GeForce RTX 40系列筆記本和桌面GPU推動。最終客戶需求穩定，並與季節性一致。我們認為全球最終需求在去年的放緩後已經恢復增長。我們有一個大的升級機會在我們面前，因為我們的安裝基礎中有47%已經升級到RTX，並且約有20%的GPU具有RTI 3060或更高的性能。筆記本GPU在重要的返校季節中實現了強勁增長，由RTI 4060 GPU帶動。NVIDIA的GPU驅動的筆記本電腦越來越受歡迎，其出貨量現在超過了來自世界各地的幾個地區的桌面GPU。這可能會稍微改變我們整體遊戲收入的實際情況，使Q2和Q3成為一年中較強的季度，反映了NetLock的返校和假日建設計劃。在桌面上，我們推出了GeForce RTX 4060和GeForce RTX 4060 Ti GPU，將Ada Lovelace架構的價格降低到299美元。RTX和DLSS遊戲的生態系統持續擴大。新增了35款支持DLSS的新遊戲，包括《暗黑破壞神4》和《巴爾德之門3》等大熱門遊戲。現在有超過330款RTI加速的遊戲和應用。我們正在將生成AI帶入遊戲。在Computex上，我們宣布了NVIDIA Avatar Cloud Engine，或稱ACE，為遊戲提供了一個定制的AI模型家族服務。開發人員可以使用此服務為非玩家角色帶來智能。它利用了許多NVIDIA Omniverse和AI技術，包括NEMO、Riva和Audio2Face。現在轉向專業視覺化。收入為3.75億美元，環比增長28%，同比下降24%。Ada架構的推出在Q2驅動了強勁增長，最初在筆記本工作站中推出，桌面工作站的刷新將在Q3到來。這些將包括強大的新RTX系統，最多可配備四個NVIDIA RTX 6000 GPU，提供超過5800萬兆浮點運算的AI性能和192GB的GPU內存。它們可以配置為NVIDIA AI企業版或NVIDIA Omniverse企業版。我們還宣布了基於AIDA世代的三款新的桌面工作站GPU，包括NVIDIA RTX 5400和4000，相比上一代，RT核心吞吐量提高了2倍，AI訓練性能提高了2倍。除了傳統的工作負載，如3D設計和內容創建，生成AI、大語言模型開發和數據科學的新工作負載正在擴大我們在專業視覺化中的RT考古學的機會。Jensen在本月早些時候在SIGGRAPH的主題演講中的一個主要主題是圖形和AI的轉換。這就是NVIDIA Omniverse的定位。Omniverse是OpenUSD的原生平台。OpenUSD是一種通用的交換方式，正在迅速成為3D世界的標準，就像HTML是2D互聯網的通用語言一樣。Adobe、Apple、Autodesk、Pixar和NVIDIA共同組成了OpenUSD聯盟。我們的使命是加速OpenUSD的開發和採用。我們宣布了新的和即將推出的Omniverse雲API，包括run USD和chat USD，以將生成AI帶入open USD工作負載。轉向汽車，收入為2.53億美元，環比下降15%，同比增長15%。穩健的同比增長主要由基於可再生驅動的自駕平台的推出驅動，Oren SOC與一些新能源汽車製造商合作。環比下降反映了整體汽車需求的下降，特別是在中國。我們宣布與MediaTek建立合作夥伴關係，為汽車內的駕駛員和乘客帶來新的體驗。MediaTek將開發汽車SOC並整合NVIDIA GPU晶片的新產品線。該合作涵蓋了從豪華到入門級的廣泛車輛範疇。轉向P&L的其餘部分，GAAP毛利率擴大到70.1%，非GAAP毛利率為71.2%，主要由於數據中心銷售的增加。我們的數據中心產品包括大量的軟件和複雜性，這也有助於推動我們的毛利率。GAAP營業費用環比增長6%，非GAAP營業費用環比增長5%，主要反映了薪酬和福利的增加。我們以股票回購和現金股息的形式向股東返還了約34億美元。我們的董事會剛剛批准了額外的250億美元的股票回購，以補充我們在Q2結束時剩餘的40億美元的授權。讓我轉向2024財年第三季度的展望。我們的數據中心平台對AI的需求巨大，並且在行業和客戶中都有廣泛的基礎。我們的需求能見度延伸到明年。我們在未來幾個季度的供應將繼續增長，因為我們降低了週期時間，並與我們的供應合作夥伴合作增加產能。此外，新的L40S GPU將有助於滿足來自雲到企業的許多類型工作負載的不斷增長的需求。對於Q3，總收入預計為160億美元，上下浮動2%。我們預計環比增長主要由數據中心驅動，遊戲和ProViz也將有所貢獻。預計GAAP和非GAAP毛利率將分別為71.5%和72.5%，上下浮動50個基點。預計GAAP和非GAAP營業費用將分別為約29.5億美元和20億美元。預計GAAP和非GAAP其他收入和費用將分別為約10萬美元的收入，不包括非關聯開發的收益和損失。預計GAAP和非GAAP稅率將分別為14.5%，上下浮動1%，不包括任何一次性項目。CFO評論和我們的IR網站上提供了更多的財務細節。最後，讓我強調一下金融社區的一些即將到來的活動。我們將參加8月30日在芝加哥舉行的杰弗裡斯科技峰會，9月5日在舊金山的高盛會議，9月6日的Evercore半導體會議，以及9月7日在紐約的花旗科技會議，以及9月11日的BMA虛擬AI會議。我們的業績電話會議討論2024財年第三季度的結果定於11月21日星期二。操作員，我們現在開放問答環節。你能為我們保留問題嗎？謝謝。謝謝。現在，我想提醒大家，如果要提問，請按電話鍵盤上的星號，然後按數字一。我們要求您每人限問一個問題。我們將暫停一會兒，以編制問答名單。我們將從TD Cowan的Matt Ramsey提出的第一個問題開始。你的線路現在已經打開。是的，非常感謝。下午好。一個令人驚訝的結果。Jensen，我想問你一個關於大模型推理這個迅速出現的應用的問題。我認為大多數投資者都很清楚，你們在訓練市場上的份額幾乎是壟斷的。許多較小的市場，較小的模型推理工作負載在過去都是由ASIC完成的，而隨著這些GPT和其他非常大的模型，這種新的工作負載在大模型推理上的加速非常快，我認為你們的Grace Hopper superchip產品和其他產品都非常適合這種工作。但是，你能否談談你如何看待推理市場在小模型推理和大模型推理之間的區分，以及你們的產品組合如何用於這種工作呢？謝謝。是的，謝謝。所以讓我們先退一步。這些大語言模型非常驚人。它做了幾件事，當然。它能夠理解非結構化的語言。但在其核心，它學習了人類語言的結構。並且它已經編碼或在其中壓縮了大量的人類知識。它通過學習的語料庫來學習。發生的事情是你創建這些大語言模型，並且你創建的越大越好。然後你從中衍生出更小的模型版本，基本上是教師學生模型。這是一個叫做蒸餾的過程。所以當你看到這些較小的模型，很可能是它們從更大的模型中衍生出來的，或者從更大的模型中蒸餾出來的，或者從更大的模型中學習的，就像你有教授，和老師，和學生，等等。你將來會看到這種情況。所以你從一個非常大的模型開始，它具有很大的通用性和泛化性和所謂的零次能力。所以對於許多你沒有特別訓練它的應用和問題或技能，這些大語言模型奇蹟般地具有執行它們的能力。這就是它如此神奇的原因。另一方面，你希望在各種計算設備中都有這些能力。所以你做的是你把它們蒸餾下來。這些較小的模型可能在特定的技能上有出色的能力，但是它們不具有很好的泛化性。

### 原文逐字稿

This is the first use of the new NVIDIA NVMe switch system, enabling all of the 256 Grace Hopper Superchips to work together as one. A huge jump compared to a prior generation connecting just eight GPUs over 10 billion. DGX-GH200 systems are expected to be available by the end of the year. Google Cloud, Meta, and Microsoft among the first to gain access. Strong networking growth was driven primarily by InfiniBand infrastructure to connect HDX GPU systems. Thanks to its end-to-end optimization and in-network computing capabilities, InfiniBand delivers more than double the performance of traditional Ethernet or AI. For billions of dollar AI infrastructures, the value from the increased throughput of InfiniBand is worth hundreds of millions and pays for the network. In addition, only InfiniBand can scale to hundreds of thousands of GPUs. It is the network of choice for leading AI practitioners. For Ethernet-based cloud data centers that seek to optimize their AI performance, we announced NVIDIA Spectrum X, an accelerated networking platform designed to optimize ethernet for AI workflows. Spectrum X couples the Spectrum 4 ethernet switch with the Bluefield 3 DPU, achieving 1.5x better overall AI performance and power efficiency versus traditional ethernet. Bluefield 3 DPU is a major success. It is in qualification with major OEMs and ramping across multiple CSPs and consumer internet. Now moving to gaming. Gaming revenue of $2.49 billion was up 11% sequentially and 22% year-on-year. Growth was fueled by GeForce RTX 40 series GPUs for laptops and desktops. End customer demand was solid and consistent with seasonality. We believe global end demand has returned to growth after last year's slowdown. We have a large upgrade opportunity ahead of us, as 47% of our installed base have upgraded to RTX, and about 20% of a GPU with an RTI 3060 or higher performance. Laptop GPUs posted strong growth in the key back-to-school season, led by RTI 4060 GPUs. NVIDIA's GPU-powered laptops have gained in popularity, and their shipments are now outpacing desktop GPUs from several regions around the world. This is likely to shift the reality of our overall gaming revenue a bit, with Q2 and Q3 as the stronger quarters of the year, reflecting the back-to-school and holiday build schedule for NetLock. In desktop, we launched the GeForce RTX 4060 and the GeForce RTX 4060 Ti GPU, bringing the Ada Lovelace architecture down to price points as low as $299. The ecosystem of RTX and DLSS games continued to expand. 35 new games added to DLSS support, including blockbusters such as Diablo 4 and Baldur's Gate 3. There's now over 330 RTI-accelerated games and apps. We are bringing generative AI to games. At Computex, we announced NVIDIA Avatar Cloud Engine, or ACE, for games, a custom AI model family service. Developers can use this service to bring intelligence to non-player characters. It harnesses a number of NVIDIA omniverse and AI technologies, including NEMO, Riva, and Audio2Face. Now moving to professional visualization. Revenue of $375 million was up 28% sequentially and down 24% year-on-year. The Ada Architecture ramp drove strong growth in Q2, rolling out initially in laptop workstations, with a refresh of desktop workstations coming in Q3. These will include powerful new RTX systems with up to four NVIDIA RTX 6000 GPUs, providing more than 5,800 teraflops of AI performance and 192 gigabytes of GPU memory. They can be configured with NVIDIA AI Enterprise or NVIDIA Omniverse Enterprise. We also announced three new desktop workstation GPUs based on the AIDA generation, the NVIDIA RTX 5400 and 4000, offering up to 2x the RT core throughput and up to 2x faster AI training performance compared to the previous generation. In addition to traditional workloads such as 3D design and content creation, new workloads in generative AI, large language model development, and data science are expanding the opportunity in pro-visualization for our RT archaeology. One of the key themes in Jensen's keynote at SIGGRAPH earlier this month was the conversion of graphics and AI. This is where NVIDIA Omniverse is positioned. Omniverse is OpenUSD's native platform. OpenUSD is a universal interchange that is quickly becoming the standard for the 3D world, much like HTML is the universal language for the 2D internet. Together, Adobe, Apple, Autodesk, Pixar, and NVIDIA formed the Alliance for OpenUSD. Our mission is to accelerate OpenUSD's development and adoption. We announced new and upcoming Omniverse cloud APIs, including run USD and chat USD to bring generative AI to open USD workloads. Moving to automotive, revenue was $253 million, down 15% sequentially and up 15% year-on-year. Solid year-on-year growth was driven by the ramp of self-driving platforms based on renewable drive. Oren, SOC with a number of new energy vehicle makers. A sequential decline reflects lower overall automotive demand, particularly in China. We announced a partnership with MediaTek to bring drivers and passengers new experiences inside the car. MediaTek will develop automotive SOCs and integrate a new product line of NVIDIA GPU chiplets. The partnership covers a wide range of vehicle segments from luxury to entry-level. Moving to the rest of the P&L, gap gross margins expanded to 70.1% and non-gap gross margins to 71.2%, driven by higher data center sales. Our data center products include a significant amount of software and complexity, which is also helping drive our gross margin. Sequential GAAP operating expenses were up 6% and non-GAAP operating expenses were up 5%, primarily reflecting increased compensation and benefit. We returned approximately 3.4 billion to shareholders in the form of share repurchases and cash dividends. Our board of directors has just approved an additional 25 billion in stock repurchases to add to our remaining four billion of authorization as of the end of Q2. Let me turn to the outlook for the third quarter of fiscal 2024. Demand for our data center platform for AI is tremendous and broad based across industries and customers. Our demand visibility extends into next year. Our supply over the next several quarters will continue to ramp as we lower cycle times and work with our supply partners to add capacity. Additionally, the new L40S GPU will help address the growing demand for many types of workloads from cloud to enterprise. For Q3, total revenue is expected to be 16 billion, plus or minus 2%. We expect sequential growth to be driven largely by data center with gaming and ProViz also contributing. Gap and non-gap gross margins are expected to be 71.5% and 72.5%, respectively, plus or minus 50 basis points. Gap and non-gap operating expenses are expected to be approximately 2.95 billion and 2 billion respectively. Gap and non-gap other income and expenses are expected to be an income of approximately 100,000, excluding gains and losses from non-associated development. Gap and non-gap tax raises are expected to be 14.5% plus or minus 1%, excluding any discrete items. Further financial details are included in the CFO commentary and other information available on our IR website. In closing, let me highlight some upcoming events for the financial community. We will attend the Jeffrey's Tech Summit on August 30th in Chicago, the Goldman Sachs Conference on September 5th in San Francisco, the Evercore Semiconductor Conference on September 6th, as well as the Citi Tech Conference on September 7th, both in New York, and the BMA Virtual AI Conference on September 11th. Our earnings call to discuss the results of our third quarter of fiscal 2024 scheduled for Tuesday, November 21st. Operator, we will now open the call for questions. Could you please hold for questions for us? Thank you. Thank you. At this time, I'd like to remind everyone in order to ask a question, press star then the number one on your telephone keypad. We ask that you please limit yourself to one question. We'll pause for just a moment to compile the Q&A roster. We'll take our first question from Matt Ramsey with TD Cowan. Your line is now open. Yeah, thank you very much. Good afternoon. A remarkable result. Jensen, I wanted to ask a question of you regarding the really quickly emerging application of large model inference. I think it's pretty well understood by the majority of investors that you guys have very much a lockdown share of training market. A lot of the smaller market, smaller model inference workloads have been done on ASIC use in the past and with many of these GPT and other really large models, there's this new workload that's accelerating super duper quickly on large model inference and I think your Grace Hopper super chip product and others are pretty well aligned for that. But could you maybe talk to us about how you're seeing the inference market segment between small model inference and large model inference and how your product portfolio is used for that. Thanks. Yeah, thanks a lot. So let's take a quick step back. These large language models are pretty phenomenal. It does several things, of course. It has the ability to understand unstructured language. But at its core, what it has learned is the structure of human language. And it has encoded or within it, compressed within it, a large amount of human knowledge. that it has learned by the corpuses that it studied. What happens is you create these large language models, and you create as large as you can. And then you derive from it smaller versions of the model, essentially teacher-student model. It's a process called distillation. And so when you see these smaller models, it's very likely the case that they were derived from, or distilled from, or learned from larger models, just as you have professors, and teachers, and students, and so on. And you're going to see this going forward. So you start from a very large model, and it has a large amount of generality and generalization and what's called zero-shot capability. And so for a lot of applications and questions or skills that you haven't trained it specifically on, these large language models miraculously has the capability to perform them. That's what makes it so magical. On the other hand, you would like to have these capabilities in all kinds of computing devices And so what you do is you distill them down. These smaller models might have excellent capabilities and a particular skill, but they don't generalize as well.

## "大型語言模型與零樣本能力"

### Abstract

這段文字討論了數據中心轉變為加速計算和生成式AI的過程，這被認為是當今計算領域的兩大最重要趨勢。演講者也討論了他們的軟體生態系統的重要性，以及它為他們的平台帶來的價值。他們提到了NVIDIA AI企業版，這是一個擁有4500個軟體套件和庫以及10000個依賴項的運行時。演講者還強調了架構、安裝基礎和覆蓋範圍在他們公司成功中的重要性。

### Investment insight

演講者預期在接下來的幾個季度以及下一個財政年度中，他們的供應量將持續增加。他們也提到了他們數據中心的成長，以及他們的HGX系統在其中所扮演的重要角色。演講者還討論了他們的軟體生態系統的重要性，以及它如何助力他們在市場上的成功。他們認為，他們的架構、安裝基礎和覆蓋範圍是他們公司的關鍵元素。

### Keywords

大型語言模型、零樣本能力、加速運算、生成式AI、數據中心、NVIDIA AI企業版、軟體生態系統、架構、安裝基礎、覆蓋範圍、HGX系統、數據中心增長。

### 中文逐字稿

他們並沒有所謂的良好零擊中能力。因此，他們都有自己獨特的能力，但你從非常大的模型開始。好的，接下來我們將轉到Vivek Arya，他是V of A Security的代表。你的線路現在已經打開。謝謝。我只是想快速澄清和提問。我能問你明年預計會有多少增量供應上線嗎？你認為它會增長20%、30%、40%、50%？因為你說它每季度都在增長。然後Jensen，我想問你的問題是，當我們看全球超級計算機的支出，這個數字並沒有真正增長那麼多。那麼，是什麼讓你有信心他們可以繼續從這個數字中切出更多的部分給生成型AI？請告訴我們你對這種需求在未來一到兩年內的可持續性有什麼看法。所以，如果我看你對Q3數據中心的隱含展望，12、13億。這對於已經AI加速的服務器有多少的說法是什麼？這個數字會去哪裡？所以，請給我們一些信心，你看到的增長在未來一到兩年內是可持續的。所以，謝謝你對我們供應的問題。是的，我們確實預計在未來幾個季度以及下一財政年度繼續增加我們的供應。至於百分比，這不是我們這裡有的東西。這是跨越許多不同供應商，許多不同部分建立一個HVAC和我們許多其他新產品即將上市的工作。但我們對我們與供應商的支持以及我們與他們一起改善供應的長時間都非常滿意。世界上大約有一萬億美元的數據中心安裝在雲端和企業以及其他地方。這一萬億美元的數據中心正在轉變為加速計算和生成型AI。我們看到兩個平台轉變正在同時發生。一個是加速計算，原因是因為這是最具成本效益，最節能，並且現在進行計算的最有效的方式。所以你看到的，然後突然，由生成型AI啟用，由加速計算啟用，生成型AI出現了。這個令人驚嘆的應用現在給每個人兩個理由去轉變，從一般目的計算，傳統的計算方式，到這種新的計算方式，加速計算。這大約是一萬億美元的數據中心，每年大約有四分之一萬億美元的資本支出。你看到世界各地的數據中心正在將資本支出集中在今天的計算兩個最重要的趨勢上，加速計算和生成型AI。所以我認為這不是一個短期的事情。這是一個長期的行業轉變，我們看到這兩個平台轉變正在同時發生。接下來，我們轉到Stacy Ravgon，他是Bernstein Research的代表。你的線路已經打開。嘿，伙計們。謝謝你接受我的問題。我想知道，Colette，你能告訴我數據中心在本季度，甚至在指導中，像系統對比GPU，像DGX對比H100，我真正想問的是價格或內容或你想定義的地方與實際驅動未來增長的單位有多少。你能給我們一些顏色嗎？當然，Stacy。讓我在本季度幫助你。我們的HGX系統是我們數據中心以及我們看到的數據中心增長的一個非常重要的部分。這些系統包括我們的Hopper架構的HGX，但也包括我們的Ampere架構。是的，我們仍然在市場上銷售這兩種架構。現在，當你考慮到這一點，這對於系統作為一個單位，當然，它正在大幅度增長，這正在推動收入的增加。所以這兩件事都是數據中心內部收入的驅動力。我們的DGXs總是我們將銷售的額外系統的一部分。這對於企業客戶和我們在消費者互聯網公司中看到的許多其他不同類型的客戶來說都是很好的機會。那裡的重要性也與我們與我們的DGXs一起銷售的軟體結合在一起。但這是我們正在進行的銷售的一部分。其餘的GPU，我們有新的GPU即將上市，我們談論的是L40s，它們將在未來繼續增長。但再次，我們在上一季度內的收入最大的驅動力肯定是H60。而且Stacy，如果我可以補充一些東西。你說的是H100。我知道你心裡的想法。但H100是35,000個部分，重70磅，幾乎有一萬億個電晶體的組合。它需要一個機器人來建造，好吧，需要很多機器人來建造，因為它重70磅。並且需要一台超級計算機來測試一台超級計算機。所以這些東西都是技術奇蹟。並且製造它們真的很繁重。所以我認為我們稱之為H100，好像它是從一個晶圓廠出來的一個晶片，但H100s真的像HGX和合成世界的超級計算機一樣出去，它們真的是相當大的系統組件，如果你願意的話。接下來，我們轉到Mark Lipickis，他是Jefferies的代表。你的線路現在已經打開。嗨，謝謝你接受我的問題，並祝賀你的成功。Jensen，你在市場上的成功似乎很大一部分是因為你提供了軟體生態系統以及晶片和硬體平台。我有兩部分問題。我想知道你能否幫助我們理解你的軟體生態系統的演變，關鍵元素，並且是否有辦法量化你在這個維度上的領先優勢，比如你投入了多少人年來建立它？然後第二部分，我想知道你是否願意與我們分享你對NVIDIA平台的價值是硬體差異化還是軟體差異化的百分比的看法。謝謝。是的，Mark，我非常感謝你的問題。讓我看看我是否可以使用一些指標。所以我們有一個叫做NVIDIA AI Enterprise的運行時。這是我們軟體堆疊的一部分。這是，如果你願意，幾乎每個公司都使用的機器學習的端到端運行時，從數據處理，訓練任何你喜歡的模型，任何你喜歡的框架，推理和部署，將其擴展到一個數據中心。它可以是一個超級計算機數據中心的擴展，也可以是一個企業數據中心的擴展，例如在VMware上。你可以在我們的任何GPU上做這個。我們在現場有數億個GPU，並且在雲中有數百萬個GPU，幾乎在每一個雲中。並且它在單個GPU配置以及多GPU每計算或多節點中運行。它還有每個GPU的多個會話或多個計算實例。所以從每個GPU的多個實例到多個GPU，多個節點到整個數據中心規模。所以這個運行時叫做NVIDIA AI Enterprise有大約4,500個軟體包，軟體庫，並且有大約10,000個彼此之間的依賴關係。並且該運行時，如我所說，為我們的安裝基礎，為我們的堆疊不斷更新和優化。這只是一個例子，讓加速計算工作需要什麼。那個代碼組合和應用組合的數量真的相當瘋狂。我們花了二十年才到這裡。但我會將我們公司的特點，如果你願意，可能有幾個。我會說第一個是架構。我們架構的靈活性，多功能性和性能使我們能夠做我剛剛說的所有事情，從數據處理，訓練，推理，從推理之前的數據預處理，到推理之後的數據後處理，將語言標記化，以便你可以用它來訓練。工作流程比僅僅訓練或推理要強烈得多。但無論如何，那是人們關注的地方，沒問題。但是當人們實際使用這些計算系統時，它是相當多的應用。所以我們架構的組合使我們能夠提供最低的擁有成本。原因是因為我們加速了這麼多不同的東西。我們公司的第二個特點是安裝基礎。你必須問自己，為什麼所有的軟體開發者都來到我們的平台？原因是軟體開發者尋求一個大的安裝基礎，以便他們可以接觸到最大數量的終端用戶，以便他們可以建立一個業務或獲得他們所做的投資的回報。然後第三個特點是覆蓋範圍。我們今天在雲中，無論是公共雲，公共面向的雲，因為我們有這麼多的客戶使用它，這麼多的開發者和客戶使用我們的平台。CSPs很高興將其放在雲中。他們用它來內部消費，開發和訓練並運行推薦系統或搜索或數據處理引擎等等，一直到訓練和推理。所以我們在雲中，我們在企業中。昨天，我們有一個非常大的公告。這真的值得一看。

### 原文逐字稿

They don't have what is called as good zero shot capability. And so they all have their own unique capabilities, but you start from very large models. Okay, next we'll go to Vivek Arya with V of A Security. Your line is now open. Thank you. Just had a quick clarification and a question. Could I ask if you could please clarify how much incremental supply do you expect to come online in the next year, you think it's up 20, 30, 40, 50%. So just any sense of how much supply because you said it's growing every quarter. And then Jensen, the question for you is when we look at the overall hyperscaler spending, that pie is not really growing that much. So what is giving you the confidence that they can continue to carve out more off that pie for a generative AI? Just give us your sense of how sustainable is this demand as we look over the next one to two years. So if I take your implied Q3 outlook of data center, 12, 13 billion. What does that say about how many servers are already AI accelerated? Where is that going? So just give us some confidence that the growth that you are seeing is sustainable into the next one to two years. So thanks for that question regarding our supply. Yes, we do expect to continue increasing, ramping our supply over the next quarters as well as into next fiscal year. In terms of percent, that's not something that we have here. It is a work across so many different suppliers, so many different parts of building an HVAC and many of our other new products that are coming to market. But we are very pleased with both the support that we have with our suppliers and the long time that we've spent with them improving the supply. The world has something along the lines of about a trillion dollars worth of data centers installed in the cloud and enterprise and otherwise. And that trillion dollars of data centers is in the process of transitioning into accelerated computing and generative AI. We're seeing two simultaneous platform shifts at the same time. One is accelerated computing, and the reason for that is because it's the most cost effective, most energy effective, and the most performant way of doing computing now. And so what you're seeing, and then all of a sudden, enabled by generative AI, enabled by accelerated computing, generative AI came along. And this incredible application now gives everyone two reasons to transition, to do a platform shift from general purpose computing, the classical way of doing computing, to this new way of doing computing, accelerated computing. It's about a trillion dollars worth of data centers, call it a quarter of a trillion dollars of capital spend each year. You're seeing that data centers around the world are taking that capital spend and focusing it on the two most important trends of computing today, accelerated computing and generative AI. And so I think this is not a near-term thing. This is a long-term industry transition, and we're seeing these two platform shifts happening at the same time. Next, we go to Stacy Ravgon with Bernstein Research. Your line's open. Hey, guys. Thanks for taking my question. I was wondering, Colette, if you could tell me how much of data center in the quarter, maybe even the guidance, like systems versus GPU, like DGX versus just the H100s, what I'm really trying to get at is how much is pricing or content or wherever you want to define that versus units actually driving the growth going forward. Can you give us any color around that? Sure, Stacy. Let me help within the quarter. Our HGX systems were a very significant part of our data center as well as our data center growth that we had seen. Those systems include our HGX of our Hopper architecture, but also our Ampere architecture. Yes, we are still selling both of these architectures or in the market. Now, when you think about that, What does that mean from both the systems as a unit, of course, that is growing quite substantially, and that is driving in terms of the revenue increases. So both of these things are the drivers of the revenue inside data center. Our DGXs are always a portion of additional systems that we will sell. Those are great opportunities for enterprise customers and many other different types of customers that we're seeing even in our consumer internet companies. The importance of there is also coming together with software that we sell with our DGXs. But that's a portion of our sales that we're doing. The rest of the GPUs, we have new GPUs coming to market that we talk about the L40s, and they will add continued growth going forward. But again, the largest driver of our revenue within this last quarter was definitely the H60. And Stacy, if I could just add something. You say it's H100. And I know you know what your mental image in your mind. But the H100 is 35,000 parts, 70 pounds, nearly a trillion transistors in combination. It takes a robot to build, well, many robots to build, because it's 70 pounds to lift. And it takes a supercomputer to test a supercomputer. And so these things are technology marvels. And the manufacturing of them is really intensive. So I think we call it H100 as if it's a chip that comes off of a fab, but H100s go out really as HGX and synthetic worlds hyperscalers, and they're really quite large system components, if you will. Next, we go to Mark Lipickis with Jefferies. Your line is now open. Hi, thanks for taking my question and congrats on the success. Jensen, it seems like a key part of the success, your success in the market is delivering the software ecosystem along with the chip and the hardware platform. And I had a two part question on this. Um, I was wondering if you could just help us understand the evolution of your software ecosystem, the critical elements, and is there a way to quantify your lead, uh, on this dimension, like how many person years you've been invested in building it? And then part two, I was wondering if you would care to share with us your view on the, what, what percentage of the value of the NVIDIA platform is hardware differentiation versus software differentiation. Thank you. Yeah, Mark, I really appreciate the question. Let me see if I could use some metrics. So we have a runtime called NVIDIA AI Enterprise. This is one part of our software stack. And this is, if you will, the runtime that just about every company uses for the end-to-end of machine learning, from data processing, the training of any model that you like to do on any framework you like to do, the inference and the deployment, the scaling it out into a data center. It could be a scale-out for a hyperscale data center, could be a scale out for enterprise data center, for example, on VMware. You could do this on any of our GPUs. We have hundreds of millions of GPUs in the field and millions of GPUs in the cloud, in just about every single cloud. And it runs in a single GPU configuration as well as multi GPU per compute or multi node. It also has multiple sessions or multiple computing instances per GPU. So from multiple instances per GPU to multiple GPUs, multiple nodes to entire data center scale. So this runtime called NVIDIA AI Enterprise has something like 4,500 software packages, software libraries, and has something like 10,000 dependencies among each other. And that runtime is, as I mentioned, continuously updated and optimized for our install base, for our stack. And that's just one example of what it would take to get accelerated computing to work. that the number of code combinations and type of application combinations is really quite insane. And it's taken us two decades to get here. But what I would characterize as probably are the elements of our company, if you will, are several. I would say number one is architecture. The flexibility, the versatility, and the performance of our architecture makes it possible for us to do all the things that I just said, from data processing, to training, to inference, from pre-processing of the data before you do the inference, to the post-processing of the data, tokenizing of languages so that you could then train with it. The amount of the workflow is much more intense than just training or inference. But anyways, that's where people focus and it's fine. But when people actually use these computing systems, it's quite a lot of applications. And so the combination of our architecture makes it possible for us to deliver the lowest cost of ownership. And the reason for that is because we accelerate so many different things. The second characteristic of our company is the install base. You have to ask yourself, why is it that all the software developers come to our platform? And the reason for that is because software developers seek a large install base so that they can reach the largest number of end users so that they could build a business or get a return on the investments that they make. And then the third characteristic is reach. We're in the cloud today, both for public cloud, public-facing cloud, because we have so many customers that use it, so many developers and customers that use our platform. CSPs are delighted to put it up in the cloud. They use it for internal consumption, to develop and train and operate recommender systems or search or data processing engines and whatnot, all the way to training and inference. And so we're in the cloud, we're in enterprise. Yesterday, we had a very big announcement. It's really worthwhile to take a look at that.

## ```VMware 與 NVIDIA 的合作與影響力```

### Abstract

這段文字討論了VMware和NVIDIA的合作，以及他們努力將生成式AI帶給全球的企業。它還強調每六個月就會推出新的架構和產品，並由全球的OEM和ODM廣泛分發。該文還討論了L40S，這是為了微調預訓練模型而設計的，並在超大規模數據中心中部署。該文進一步討論了對加速計算和生成式AI的需求，以及從通用計算轉變的趨勢。

### Investment insight

VMware和NVIDIA之間的合作，新產品和架構的引入，以及向加速計算和生成型AI的轉變，都呈現出潛在的投資機會。這些技術的需求量很大，公司正在將他們的資本投資轉向這些技術。L40S也有了一個很好的開始，企業和超級計算機都在爭相部署它。

### Keywords

VMware，NVIDIA，生成式AI，L40S，加速運算，超大規模數據中心，原始設備製造商，原始設計製造商，架構，產品，需求，投資。

### 中文逐字稿

```VMware 是全球企業的作業系統。我們已經合作了好幾年。我們將攜手將生成式 AI 帶到全球的企業，一直到邊緣。因此，覆蓋範圍是另一個原因。由於覆蓋範圍，全球的系統製造商都急於將 NVIDIA 的平台放入他們的系統中。因此，我們從全球的 OEM 和 ODM 等等獲得了非常廣泛的分銷，這都是因為我們的覆蓋範圍。最後，由於我們的規模和速度，我們能夠維持這個非常複雜的軟硬體、網路和計算的堆疊，跨越所有這些不同的使用模型和不同的計算環境。我們能夠在加速我們的工程速度的同時做到所有這些。我們似乎每兩年就推出一種新的架構。現在我們大約每六個月就推出一種新的架構，一種新的產品。因此，這些特性使得生態系統能夠在我們之上建立他們的公司和業務。因此，這些組合使得它特別。接下來，我們將轉到 Citi 的 Atif Malik。你的線路已開通。嗨，謝謝你接受我的問題，並在結果上做得很好。我對你們談到的 Covox L40S 有一個問題。你有沒有想法 L40S 可以幫助解決多少供應緊張的問題？如果你能談談這個產品的增量利潤或毛利貢獻就更好了。謝謝。Atif，讓我來回答你的問題。L40，L40S 實際上是為了一種不同類型的應用而設計的。H100 是為了大規模語言模型和處理非常大的模型和大量的數據而設計的。因此，這不是 L40 的重點。L40 的重點是能夠微調模型，微調預訓練的模型，並且它會做得非常好。它有一個變壓器引擎，它有很多性能，你可以在一台伺服器中獲得多個 GPU。它是為超大規模擴展設計的，意味著它很容易將 L40S 伺服器安裝到全球的超大規模數據中心。它採用標準機架，標準伺服器，所有的東西都是標準的。因此，它很容易安裝。L40S 還配有軟體堆疊，以及 Bluefield-3，以及我們與 VMware、我們與 Snowflakes 和 ServiceNow 以及許多其他企業合作夥伴所做的所有工作。L40S 是為全球的企業 IT 系統而設計的。這就是為什麼 HPE、Dell 和 Lenovo，以及其他約 20 家系統製造商將與我們合作，建立大約 100 種不同配置的企業伺服器，將生成式 AI 帶到全球的企業。因此，L40S 實際上是為了一種不同類型的擴展而設計的，如果你願意的話。當然，它是大規模語言模型。當然，它是生成式 AI，但它是一種不同的使用案例。因此，L40S 開始得很好。全球的企業和超大規模者都非常渴望部署 L40S。好的，接下來，我們將轉到 Morgan Stanley 的 Joe Moore。你的線路現在已開通。太好了，謝謝你。我想這些數字中最令我驚訝的是還有多少需求未被滿足。與你們的一些客戶交談，你知道，儘管這些數字很好，你在幾個季度內將你的收入增加了三倍以上。在某些情況下，人們對於他們所得到的東西的需求是多倍的。所以你能談談這個問題嗎？你認為有多少未滿足的需求？你談到了對明年的可見度。你是否能看到何時會達到供需平衡？是的，我們對今年和明年有很好的可見度。我們已經在與領先的 CSP 和數據中心建設者計劃下一代基礎設施。需求，最簡單的方式來思考需求是，世界正在從通用計算轉向加速計算。這可能是最簡單的方式來思考需求。公司提高吞吐量、提高能源效率、提高成本效率的最好方式是將他們的資本預算轉向加速計算和生成式 AI。因為這樣做，你將從 CPU 上卸載大量的工作負載，你的數據中心中的可用 CPU 將得到提升。因此，你現在看到的公司正在做的是認識到這裡的轉折點，認識到這種轉變的開始，並將他們的資本投資轉向加速計算和生成式 AI。因此，這可能是最簡單的方式來思考我們面前的機會。這不是一個單一的應用程序驅動需求，而是一個新的計算平台，如果你願意，一種新的計算轉變正在發生。全世界的數據中心都在對此做出反應，並以全面的方式進行轉變。好的，接下來我們將轉到 Goldman Sachs 的 Tashia Hari。你的線路現在已開通。嗨，謝謝你接受問題。我有一個關於 Colette 的快速澄清問題，然後另一個問題是給 Jensen 的。Colette，我想上季度你說過 CSP 是你們數據中心收入的約 40%，消費者互聯網 30%，企業 30%。根據你的講話，聽起來 CSP 和消費者互聯網可能佔你們業務的更大比例。如果你能澄清或確認這一點，那將非常有幫助。然後 Jensen，我有一個問題給你。鑑於你作為 AI 主要推動者的地位，以及你對客戶項目的廣泛參與和可見度。我很好奇你對是否會有足夠的應用或用例讓你的客戶獲得合理回報的信心有多大。我問這個問題是因為有一種擔憂，即在未來幾年可能會有一種需求暫停。我很好奇是否有足夠的廣度和深度來支持你的數據中心業務的持續增長。謝謝。好的，謝謝你，Tashia，關於我們數據中心業務中我們有哪些類型的客戶的問題。我們將我們的計算和網路結合在一起來看待它。我們的大型 CSP 在第二季度貢獻了我們收入的略多於 50%。下一個最大的類別將是我們的消費者互聯網公司。然後最後一部分將是我們的企業和高性能計算。所以，Shia，我不願意猜測未來。所以我將從計算機科學的第一原則來回答這個問題。人們已經認識到，通用計算已經不再是，並且使用通用計算進行大規模計算已經不再是最好的前進方式。它的能源成本太高，它太昂貴，並且應用程序的性能太慢。最後，世界有了一種新的方式來做這件事。它被稱為加速計算，而推動它的是生成式 AI。但是，加速計算可以用於數據中心中已經存在的所有種類的不同應用。通過使用它，你可以卸載 CPU，節省大量的錢，節省一個數量級的能源，並且吞吐量高。這就是行業真正反應的東西。未來，投資數據中心的最好方式是將資本投資從通用計算轉向，並專注於生成式 AI 和加速計算。生成式 AI 提供了一種新的提高生產力的方式，一種新的提供給你的客戶新服務的方式，而加速計算可以幫助你節省錢和節省電力。應用程序的數量是，嗯，很多。很多開發者，很多應用程序，很多庫。它已經準備好部署了。所以我認為全世界的數據中心都認識到這一點，這是未來部署數據中心資源、部署資本的最好方式。這對於全球的雲來說是真實的，你正在看到一整批新的 GPU 專業，GPU 專門的雲服務提供商。其中一個著名的是 CoreWeed，他們做得非常好。但你現在在全世界都可以看到區域性的 GPU 專門服務提供商。這是因為他們都認識到同樣的事情，即未來投資資本的最好方式是投入加速計算和生成 AI。我們也看到企業想要這樣做。但是為了讓企業能夠做到這一點，你必須支持管理系統、操作系統、安全性，以及企業的軟體定義數據中心方法，這就是 VMware。我們已經與 VMware 合作了好幾年，使 VMware 能夠支持不僅是 CPU 的虛擬化，而且是 GPU 的虛擬化，以及 GPU 的分散式計算能力，支持 NVIDIA 的 Bluefield 進行高性能網路，以及我們一直在研究的所有生成式 AI 庫，現在將作為一個特殊的 SKU 由 VMware 的銷售團隊提供，這個銷售團隊，我們都知道，非常大，因為他們接觸到了全球幾十萬的 VMware 客戶。這個新的 SKU 將被稱為 VMware Private AI Foundation。這將是一個新的 SKU，使得任何企業都能擁有一個最先進的 AI 數據中心，並能夠參與生成式 AI。所以我認為這個問題的答案是難以確定下一個季度會發生什麼，但我認為現在趨勢非常明確，我們正在看到一個平台轉變。接下來，我們將轉到 UBS 的 Timothy Arcuri。你的線路現在已開通。非常感謝。你能談談你的網路解決方案與你出貨的計算的附加率嗎？換句話說，你的計算中有一半以上、一半以下的部分是與你的網路解決方案一起出貨的，這是你可以用來優先分配 GPU 的東西嗎？謝謝。```

### 原文逐字稿

VMware is the operating system of the world's enterprise. And we've been working together for several years now. And we're going to bring together, together we're going to bring generative AI to the world's enterprises, all the way out to the edge. And so reach is another reason. And because of reach, all of the world's system makers are anxious to put NVIDIA's platform in their systems. And so we have a very broad distribution from all of the world's OEMs and ODMs and so on and so forth because of our reach. And then lastly, because of our scale and velocity, we were able to sustain this really complex stack of software and hardware and networking and compute across all of these different usage models and different computing environments. And we're able to do all this while accelerating the velocity of our engineering. It seems like we're introducing a new architecture every two years. Now we're introducing a new architecture, a new product just about every six months. And so these properties make it possible for the ecosystem to build their company and their business on top of us. And so those, in combination, makes it special. Next, we'll go to Atif Malik with Citi. Your line's open. Hi, thank you for taking my question and great job on the results. I have a question on the Covox L40S that you guys talked about. Any idea how much of the supply tightness can L40S help with? And if you can talk about the incremental profitability or gross margin contribution from this product? Thank you. Atif, let me take that for you. The L40, L40S is really designed for a different type of application. H100 is designed for large-scale language models and processing just very large models and a great deal of data. And so that's not L40's focus. L40's focus is to be able to fine-tune models, fine-tune pre-trained models, and it'll do that incredibly well. It has a transformer engine, it's got a lot of performance, you can get multiple GPUs in a server. It's designed for hyperscale scale-out, meaning it's easy to install L40S servers into the world's hyperscale data centers. It comes in a standard rack, standard server, and everything about it is standard. And so it's easy to install. L40S also is with the software stack around it, and along with Bluefield-3, and all the work that we did with VMware and the work that we did with Snowflakes and ServiceNow and so many other enterprise partners. L40S is designed for the world's enterprise IT systems. And that's the reason why HPE, Dell, and Lenovo, and some of 20 other system makers building about 100 different configurations of enterprise servers are going to work with us to take generative AI to the world's Enterprise. And so L40S is really designed for a different type of scale-out, if you will. It's, of course, large language models. It's, of course, generative AI, but it's a different use case. And so the L40S is off to a great start. And the World's Enterprise and hyperscalers are really clamoring to get L40S deployed. Okay. Next, we'll go to Joe Moore with Morgan Stanley. Your line is is now open. Great, thank you. I guess the thing about these numbers that's so remarkable to me is the amount of demand that remains unfulfilled. Talking to some of your customers, you know, as good as these numbers are, you sort of more than tripled your revenue in a couple of quarters. There's a demand in some cases for multiples of what people are getting. So can you talk about that? You know, how much unfulfilled demand do you think there is? And you talked about visibility extending into next year. Do you have line of sight into when you'll get to see supply demand equilibrium here? Yeah, we have excellent visibility through the year and into next year. And we're already planning the next generation infrastructure with the leading CSPs and data center builders. The demand, the easiest way to think about the demand is the world is transitioning from general purpose computing to accelerated computing. That's the easiest way to think about the demand. The best way for companies to increase their throughput, improve their energy efficiency, of their cost efficiency is to divert their capital budget to accelerated computing and generative AI. Because by doing that, you're gonna offload so much workload off of the CPUs that the available CPUs in your data center will get boosted. And so what you're seeing companies do now is recognizing the tipping point here, recognizing the beginning of this transition and diverting their capital investment to accelerated computing and generative AI. And so that's probably the easiest way to think about the opportunity ahead of us. This isn't a singular application that is driving the demand, but this is a new computing platform, if you will, a new computing transition that's happening. And data centers all over the world are responding to this and shifting in a broad-based way. Okay, next we go to Tashia Hari with Goldman Sachs. Your line is now open. Hi, thank you for taking the question. I had one quick clarification question for Colette and then another one for Jensen. Colette, I think last quarter you had said CSPs were about 40% of your data center revenue, consumer internet 30%, enterprise 30%. Based on your remarks, it sounded like CSPs and consumer internet may have been a larger percentage of your business. If you can kind of clarify that or confirm that, that would be super helpful. And then Jensen, a question for you. Given your position as the key enabler of AI, the breadth of engagements and the visibility you have into customer projects. I'm curious how confident you are that there will be enough applications or use cases for your customers to generate a reasonable return on their investments. I guess I asked the question because there is a concern out there that there could be a bit of a pause in your demand profile in the out years. Curious if there's enough breadth and depth there to support a sustained increase in your data center business going forward. Thank you. Okay, so thank you, Satoshi, on the question regarding our types of customers that we have in our data center business. And we look at it in terms of combining our compute as well as our networking together. Our CSPs, our large CSPs, are contributing a little bit more than 50% of our revenue within Q2. And the next largest category will be our consumer internet companies. And then the last piece of it will be our enterprise and high performance computing. So, Shia, I'm reluctant to guess about the future. And so I'll answer the question from the first principle of computer science perspective. It is recognized for some time now that general purpose computing is just not and brute forcing general purpose computing, using general purpose computing at scale is no longer the best way to go forward. It's too energy costly, it's too expensive, and the performance of the applications are too slow. And finally, the world has a new way of doing it. It's called accelerated computing, and what kicked it into turbochargers is generative AI. But accelerated computing could be used for all kinds of different applications that's already in the data center. And by using it, you offload the CPUs, you save a ton of money, an order of magnitude in cost and order of magnitude in energy and the throughput is high. And that's what the industry's really responding to. Going forward, the best way to invest in a data center is to divert the capital investment from general purpose computing and focus it on generative AI and accelerated computing. Generative AI provides a new way of generating productivity, a new way of generating new services to offer to your customers, and accelerated computing helps you save money and save power. And the number of applications is, well, tons. Lots of developers, lots of applications, lots of libraries. It's ready to be deployed. And so I think the data centers around the world recognize this, that this is the best way to deploy resources, deploy capital going forward for data centers. This is true for the world's clouds and you're seeing a whole crop of new GPU specialty, GPU specialized cloud service providers. One of the famous ones is CoreWeed and they're doing incredibly well. But you're seeing the regional GPU specialist service providers all over the world now. And it's because they all recognize the same thing, that the best way to invest your capital going forward is to put it already computing and generate AI. We're also seeing that enterprises want to do that. But in order for enterprises to do it, you have to support the management system, the operating system, the security, and software-defined data center approach of enterprises, and that's called VMware. And we've been working several years with VMware to make it possible for VMware to support not just the virtualization of CPUs, but the virtualization of GPUs, as well as the distributed computing capabilities of GPUs, supporting NVIDIA's Bluefield for high-performance networking, and all of the generative AI libraries that we've been working on, is now going to be offered as a special SKU by VMware's Salesforce, which is, as we all know, quite large because they reach some several hundred thousand VMware customers around the world. And this new SKU is going to be called VMware Private AI Foundation. And this will be a new SKU that makes it possible for enterprises. And in combination with HP Dell and Lenovo's new server offerings based on L40S, any enterprise could have a state-of-the-art AI data center and be able to engage generative AI. And so I think the answer to that question is hard to predict exactly what's gonna happen quarter to quarter, but I think the trend is very, very clear now that we're seeing a platform shift. Next, we'll go to Timothy Arcuri with UBS. Your line is now open. Thanks a lot. Can you talk about the attach rate of your networking solutions to the compute that you're shipping? In other words, is like half of your compute shipping with your networking solutions, you know, more than half, less than half, and is this something that maybe you can use to prioritize allocation of the GPUs? Thank you.

## ```網路解決方案與GPU的分配```

### Abstract

這段文字討論了該公司對於網路解決方案和GPU分配的方法，強調客戶選擇以及InfiniBand對大型基礎設施的效率。它還介紹了Spectrum X，這是一種對於依賴乙太網的客戶的解決方案。該公司的DGX雲端策略也被解釋了，強調其對AI合作夥伴、CSPs和內部工程師的好處。該公司的軟體業務也被討論了，注意到其成長以及其整合到各種產品中。

### Investment insight

該公司正在經歷高需求並正在擴大其生產能力。它也經常升級並新增產品，以應對不斷擴大的生成式AI領域。該公司的軟體業務正在成長，每年達到數千億美元。該公司也正在與Grace Hopper一同研究下一代加速運算和生成式AI。

### Keywords

網路解決方案，GPU的分配，InfiniBand，乙太網，Spectrum X，DGX雲端，軟體業務，人工智慧，雲服務提供商，生產能力，生成式人工智慧，Grace Hopper。

### 中文逐字稿

好的，從後往前看，我們並不會用這種方式來優先分配我們的GPU。我們讓客戶決定他們想要使用什麼網路。對於那些正在建設非常大型基礎設施的客戶來說，InfiniBand是，你知道，我不好意思說，但這實在是一個不二之選。原因是InfiniBand的效率非常高，比起一個價值十億美元的基礎設施，它的吞吐量高出10%、15%、甚至20%，這意味著巨大的節省。基本上，網路是免費的。所以，如果你有一個單一的應用，如果你願意，基礎設施，或者它主要用於大型語言模型或大型AI系統，InfiniBand實在是一個極好的選擇。然而，如果你正在為許多不同的用戶提供服務，並且以太網是你管理數據中心的核心，我們有一個我們剛剛宣布的優秀解決方案，它叫做Spectrum X。好的，我們將把InfiniBand的能力，如果你願意，不是全部，但部分，帶到以太網，這樣我們也可以在以太網的環境中，讓你獲得出色的生成AI能力。所以Spectrum X現在正在推廣。它需要Bluefield 3，並且支持我們的Spectrum 2和Spectrum 3以太網交換機。而額外的性能實在是驚人的。Bluefield 3使之成為可能，還有一大堆與之配套的軟體。Bluefield，如你們所知，是一個我非常關心的項目。它已經取得了一個極好的開始。我認為這是一個全壘打。這就是在網路計算中並將大量軟體放入計算網絡的概念，並且正在用Mood Field 3實現。這將是一個全壘打。我們最後的問題來自Milius的Ben Reitzis。你現在可以提問。嗨。下午好。晚上好。感謝你的問題並讓我參與這裡。我的問題與DGX Cloud有關。你能談談你看到的反響以及動力如何嗎？然後，Colette，你也能談談你的軟體業務嗎？現在的運行速度是多少，以及該業務的重要性？它似乎已經在一定程度上提高了利潤率。非常感謝。DJX Cloud的策略，讓我從這裡開始。DJX Cloud的策略是要實現幾個目標。第一，讓我們與全球CSP建立非常緊密的夥伴關係。我們認識到，我們的許多，嗯，我們與全球約30,000家公司合作，其中15,000家是初創公司，數千家是生成AI公司，而最快增長的部分，當然是生成AI。我們正在與全球所有的AI初創公司合作。最終，他們希望能夠在全球領先的雲中落地。所以我們在全球領先的雲中建立了DGX Cloud，這樣我們就可以同時與我們所有的AI夥伴合作，並幫助他們輕鬆地在我們的雲夥伴中落地。第二個好處是，它讓我們的CSP和我們自己能夠非常緊密地合作，以提高超大規模房屋的性能，這些房屋歷史上是為多租戶設計的，而不是為高性能分散式計算設計的，比如生成AI。所以能夠在架構上密切合作，讓我們的工程師攜手提高網路性能和計算性能，這實在是非常強大，非常出色的。然後，當然，NVIDIA自己也使用非常大的基礎設施，我們的自駕車團隊，我們的NVIDIA研究團隊，我們的生成AI團隊，我們的語言模型團隊，你知道，我們需要的基礎設施量相當大。我們的優化編譯器沒有我們的DGX系統是無法實現的。你知道，即使是編譯器這些天也需要AI，優化軟體和基礎設施軟體需要AI才能開發。我們的工程使用AI在我們的芯片上的事情已經被廣泛報導了。所以我們自己的AI消費，我們的機器人團隊，等等，omniverse團隊，等等，都需要AI。所以我們的內部消費也相當大，我們也把它放在DGX Cloud中。所以DGX Cloud有多種用途，多種驅動因素，並且已經取得了巨大的成功。我們的CSP喜歡它，開發者喜歡它，我們自己的內部工程師也在爭相要求更多。這對我們與所有AI用戶進行接觸和密切合作是一種很好的方式。讓我看看我能否回答你關於我們軟體收入的問題。在我們的開場白中，我們也提到了一部分，請記住，軟體是我們幾乎所有產品的一部分，無論是數據中心產品、GPU系統，還是我們在遊戲和未來的汽車產品中的任何產品。你是對的，我們也在單獨銷售它。而那個單獨的軟體業務持續增長，我們在那裡也提供軟體服務升級等。現在我們看到的是，我們的軟體業務每年可能有數百億美元的收入。我們正在考慮將NVIDIA AI Enterprise包含在我們銷售的許多產品中，比如我們的DGX，比如我們的H100的PCIe版本。我認為我們將在我們的CSE市場中看到更多的可用性。所以我們已經取得了一個很好的開始，我相信我們將看到這種增長在未來繼續。這就包括了今天的問答環節。我將會議交還給Jinsun Wong進行任何額外或結束的講話。一個新的計算時代已經開始。這個行業正在同時經歷兩個平台轉換，加速計算和生成AI。數據中心正在從通用計算平台轉向加速計算平台。全球價值一萬億美元的數據中心將轉向加速計算，以實現一個數量級更好的性能、能源效率和成本。加速計算使生成AI成為可能，現在正在推動軟體的平台轉換，並使得以前不可能的新應用成為可能。加速計算和生成AI一起推動著計算行業的廣泛平台轉換。我們的需求是巨大的。我們正在大幅擴大我們的產品生產能力。供應將在今年剩餘的時間和明年大幅增加。NVIDIA已經為此做了二十多年的準備，並創建了一個新的計算平台，全球的行業可以在此基礎上建立。使NVIDIA特殊的是，一，架構。NVIDIA加速了從數據處理、訓練、推理、每一種AI模型、實時語音到計算視覺，以及巨大的推薦器到向量數據庫的所有事物。我們架構的性能和多功能性轉化為最低的數據中心TCO和最佳的能源效率。二，安裝基礎。NVIDIA在全球擁有數以億計的CUDA兼容GPU。開發者需要一個大的安裝基礎來接觸終端用戶並發展他們的業務。NVIDIA是開發者首選的平台。更多的開發者創建更多的應用，使NVIDIA對客戶更有價值。三，覆蓋範圍。NVIDIA存在於雲端、企業數據中心、工業邊緣、PC、工作站、儀器和機器人中。每一個都有基本上獨特的計算模型和生態系統。系統供應商，如OEM，計算機OEM，可以放心地投資於NVIDIA，因為我們提供了顯著的市場需求和覆蓋範圍。規模和速度。NVIDIA已經達到了顯著的規模，並且100%投資於加速計算和生成AI。我們的生態系統夥伴可以相信，我們有專業知識、專注和規模，以提供強大的路線圖和覆蓋範圍，幫助他們成長。我們正在加速，因為這些能力的結果是相加的。我們大約每六個月就升級並添加新產品，而不是每兩年，以應對生成AI的擴展宇宙。當我們增加H100的輸出，用於訓練和推理大型語言模型時，我們正在推動我們的新L40S通用GPU，用於雲端擴展和企業服務器。Spectrum X，包括我們的以太網交換機、Bluefield 3 SuperNet和軟體，幫助希望在以太網基礎設施上獲得最佳可能AI性能的客戶。客戶已經在使用我們的Grace Hopper進行下一代加速計算和生成AI的工作。我們正在將NVIDIA AI擴展到全球的企業，這些企業需要生成AI，但需要模型的隱私、安全和主權。我們與全球領先的企業IT公司，Accenture、Adobe、Getty、Hugging Face、Snowflake、ServiceNow、VMware和WPP，以及我們的企業系統夥伴，Dell、HPE和Lenovo，一起將生成AI帶到全球的企業。我們正在建設NVIDIA Omniverse，以數字化並使全球數萬億美元的重工業能夠使用生成AI來自動化他們的建設和運營實體資產，並實現更大的生產力。生成AI從雲端開始，但最重要的機會在於全球最大的行業，這些公司可以實現數萬億美元的生產力增長。這對於NVIDIA、我們的客戶、夥伴和整個生態系統來說，都是一個令人興奮的時刻，推動這個計算的世代轉變。我們期待在下一季度更新我們的進展。這就結束了今天的電話會議。你現在可以斷開連接。

### 原文逐字稿

Well, working backwards, we don't use that to prioritize the allocation of our GPUs. We let customers decide what networking they would like to use. And for the customers that are building very large infrastructure, InfiniBand is, you know, I hate to say it, kind of a no-brainer. And the reason for that, Because the efficiency of InfiniBand is so significant, some 10, 15, 20% higher throughput for a billion dollars infrastructure translates to enormous savings. Basically, the networking is free. And so if you have a single application, if you will, infrastructure, or it's largely dedicated to large language models or large AI systems, InfiniBand is really a terrific choice. However, if you're hosting for a lot of different users and Ethernet is really core to the way you manage your data center, we have an excellent solution there that we had just recently announced and it's called Spectrum X. Well, we're gonna bring the capabilities, if you will, not all of it, but some of it, of the capabilities of InfiniBand to Ethernet so that we can also, within the environment of Ethernet, allow you to enable you to get excellent generative AI capability. So Spectrum X is just ramping now. It requires Bluefield 3, and it supports both our Spectrum 2 and Spectrum 3 ethernet switches. And the additional performance is really spectacular. And Bluefield 3 makes it possible, and a whole bunch of software that goes along with it. Bluefield, as all of you know, is a project really dear to my heart. And it's off to just a tremendous start. I think it's a home run. And this is the concept of in-network computing and putting a lot of software in the computing fabric is being realized with Mood Field 3. And it is going to be a home run. Our final question comes from the line of Ben Reitzis with Milius. Your line is now open. Hi. Good afternoon. Good evening. Thank you for the question and putting me in here. My question is with regard to DGX Cloud. Can you talk about the reception that you're seeing and how the momentum is going. And then, Colette, can you also talk about your software business? What is the run rate right now and the materiality of that business? And it does seem like it's already helping margins a bit. Thank you very much. DJX Cloud's strategy, let me start there. DJX Cloud's strategy is to achieve several things. Number one, to enable a really close partnership between us and the world CSPs. We recognize that many of our, well, we work with some 30,000 companies around the world, 15,000 of them are startups, thousands of them are generative AI companies, and the fastest growing segment, of course, is generative AI. We're working with all of the world's AI startups. And ultimately, they would like to be able to land in one of the world's leading clouds. And so we built DGX Cloud as a footprint inside the world's leading clouds. So that we could simultaneously work with all of our AI partners and help land them in easily in one of our cloud partners. The second benefit is that it allows our CSPs and ourselves to work really closely together to improve the performance of a hyperscale house which is historically designed for multi-tenancy and not designed for high performance distributed computing like generative AI. And so to be able to work closely architecturally to have our engineers work hand in hand to improve the networking performance and the computing performance has been really powerful, really terrific. And then thirdly, of course, NVIDIA uses very large infrastructures ourselves and our self-driving car team, our NVIDIA research team, our generative AI team, our language model team, you know, the amount of infrastructure that we need is quite significant. And none of our optimizing compilers are possible without our DGX system. You know, even compilers these days require AI and optimizing software and infrastructure software requires AI to even develop. It's been well publicized that our engineering uses AI on our chips. And so the internal, our own consumption of AI, our robotics team, so on and so forth, omniverse team, so on and so forth, all needs AI. And so our internal consumption is quite large as well we land that in DGX Cloud. And so DGX Cloud has multiple use cases, multiple drivers, and it's been off to just an enormous success. And our CSPs love it, the developers love it, and our own internal engineers are clamoring to have more of it. And it's a great way for us to engage and work closely with all of the AI users. And let's see if I can answer your question regarding our software revenue. In part of our opening remarks that we made as well, Remember, software is a part of almost all of our products, whether they are data center products, GPU system, or any of our products within gaming and our future automotive products. You're correct, we're also selling it in a standalone business. And that standalone software continues to grow, where we are providing both the software services upgrades across there as well. Now we're seeing at this point, probably hundreds of billions of dollars annually for our software business. And we are looking at NVIDIA AI Enterprise to be included with many of the products that we're selling, such as our DGX, such as our PCIe versions of our H100. And I think we're gonna see more availability with our CSE marketplaces. So we're off to a great start, and I do believe we'll see this continue to grow going forward. And that does include today's question and answer session. I'll turn the call back over to Jinsun Wong for any additional or closing remarks. A new computing era has begun. The industry is simultaneously going through two platform transitions, accelerated computing and generative AI. Data centers are making a platform shift from general purpose to accelerated computing. The trillion dollars of global data centers will transition to accelerated computing to achieve an order of magnitude better performance, energy efficiency, and cost. Accelerated computing enabled generative AI, which is now driving a platform shift in software and enabling new, never before possible applications. Together, accelerated computing and generative AI are driving a broad-based computer industry platform shift. Our demand is tremendous. We are significantly expanding our product production capacity. Supply will substantially increase for the rest of this year and next year. NVIDIA has been preparing for this for over two decades and has created a new computing platform that the world's industry, world's industries, can build upon. What makes NVIDIA special are, one, architecture. NVIDIA accelerates everything from data processing, training, inference, every AI model, real-time speech to computer vision, and giant recommenders to vector database. The performance and versatility of our architecture translates to the lowest data center TCO and best energy efficiency. Two, install base. NVIDIA has hundreds of millions of CUDA-compatible GPUs worldwide. Developers need a large install base to reach end users and grow their business. NVIDIA is the developer's preferred platform. More developers create more applications that make NVIDIA more valuable for customers. Three, reach. NVIDIA is in clouds, enterprise data centers, industrial edge, PCs, workstations, instruments, and robotics. Each has fundamentally unique computing models and ecosystems. System suppliers like OEMs, computer OEMs, can confidently invest in NVIDIA because we offer significant market demand and reach. Scale and velocity. NVIDIA has achieved significant scale and is 100% invested in accelerated computing and generative AI. Our ecosystem partners can trust that we have the expertise, focus, and scale to deliver a strong roadmap and reach to help them grow. We are accelerating because of the additive results of these capabilities. We're upgrading and adding new products about every six months versus every two years to address the expanding universe of generative AI. While we increase the output of H100 for training and inference of large language models, We're ramping up our new L40S universal GPU for scale for cloud scale-out and enterprise server. Spectrum X, which consists of our Ethernet switch, Bluefield 3 SuperNet and software, helps customers who want the best possible AI performance on Ethernet infrastructures. Customers are already working on next generation accelerated computing and generative AI with our Grace Hopper. We're extending NVIDIA AI to the world's enterprise that demands generative AI, but with the model privacy, security and sovereignty. Together with the world's leading enterprise IT companies, Accenture, Adobe, Getty, Hugging Face, Snowflake, ServiceNow, VMware, and WPP, and our enterprise system partners, Dell, HPE, and Lenovo, we are bringing generative AI to the world's enterprise. We're building NVIDIA Omniverse to digitalize and enable the world's multi-trillion dollar heavy industry to use generative AI to automate how they build and operate physical asset and achieve greater productivity. Generative AI starts in the cloud, But the most significant opportunities are in the world's largest industries, where companies can realize trillions of dollars of productivity gain. It is an exciting time for NVIDIA, our customers, partners, and the entire ecosystem to drive this generational shift in computing. We look forward to updating you on our progress next quarter. This concludes today's conference call. You may now disconnect.