LUISS

Dipartimento di Impresa e Management Corso di laurea in: Economia e Management

Cattedra di Informatica

Storia e Aziende che hanno portato allo sviluppo dell'Intelligenza Artificiale - il caso Nvidia

Prof. Luigi Laura

RELATORE

Stefano Brighenti Matr. 272361

CANDIDATO

Anno Accademico: 2023/2024

Indice:

Introduzione:

L'obbiettivo di ricerca di questa tesi è quello di approfondire le origini e il funzionamento dell'Intelligenza Artificiale analizzando le principali innovazioni introdotte e mettendo in evidenza il ruolo che alcune aziende, e in modo particolare NVIDIA hanno svolto nello sviluppo di questo settore.

Questa ricerca è svolta tramite l'utilizzo di fonti secondarie quali articoli scientifici, opere monografiche e pubblicazioni aziendali di NVIDIA stessa. La prima parte dell'elaborato esplora la storia dello sviluppo dell'IA e le aziende che maggiormente ne hanno contribuito. La seconda parte approfondisce, tramite la metodologia del caso studio, NVIDIA e le caratteristiche che l'hanno portata all'elevato successo finanziario e non solo.

NVIDIA è stata selezionata come oggetto di analisi di questa tesi per due motivazioni principali. La prima è il traguardo storico raggiunto il 18 giugno 2024, data nel quale è diventata l'azienda con la capitalizzazione di mercato più alta nel mondo. La seconda è la straordinaria lungimiranza dimostrata dall'azienda e dal suo Ceo (Jensen Huang) per quanto riguardo lo sviluppo di tecnologie come le GPU e l'intelligenza artificiale.

Capitolo 1: Introduzione al Funzionamento dell'Intelligenza Artificiale

Un primo Sguardo sull'Intelligenza Artificiale (IA)

Per introdurre questo argomento dobbiamo sicuramente prima dare una definizione di Intelligenza Artificiale[1] . Tale definizione la daremo semplicemente analizzando le parole delle quale si compone: Intelligenza in quanto l'obbiettivo di queste tecnologie è quello di emulare il ragionamento umano ed esso è Artificiale perché viene svolto da un macchinario[2].

Rispetto al normale utilizzo dei calcolatori, ovvero lo svolgimento di precise azioni da svolgere in un ordine predefinito, tramite le IA è possibile fornire un problema senza sapere, o aver già programmato tutti i passaggi necessari per ottenere il risultato. Questa caratteristica essenziale dell'IA è proprio dovuta all'emulazione dell'intelligenza umana alla quale si fornisce una domanda la risposta viene generata sul momento sulla base delle informazioni precedentemente acquisite. Le IA sono create tramite il pre-addestramento, che serve per la ricerca di pattern e rafforzamento degli stessi, tramite l'approvvigionamento di enormi quantità dati etichettati. Grazie all'apprendimento le IA sono in grado di affrontare problemi per i quali non sono state specificatamente preparate con una serie di minuscole azioni da eseguire pedissequamente, ma invece generare una soluzione al proplema fornito.

[1] D'ora in poi IA

[2] https://www.europarl.europa.eu/topics/it/article/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata#:~:text=L'intelligenza%20artificiale%20(IA),la%20pianificazione%20e%20la%20creativit%C3%A0

Essenzialmente il concetto di IA consiste nell'abilità di una macchina di mostrare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività.

Il concetto di IA viene spesso catalogato sulla base della finalità di utilizzo in quanto esistono IA deboli e IA forti. La differenza tra le due sta che la IA deboli hanno finalità di risolvere problemi specifici ed emulare il ragionamento umano in un ambito specifico, mentre le IA forti hanno come finalità unica di imitare il ragionamento umano nella sua totalità e per tanto possono essere utilizzate per svariate applicazioni

Prima di affrontare la storia dell'IA, è opportuno definire dei concetti contigui a quello dell'IA ovvero: Machine Learning (ML) e Deep learning (DL).

Partendo dal primo concetto del ML, questo è un sottocampo dell'IA, esso si interessa degli algoritmi in grado di apprendere compiti complessi e di elaborare previsioni tramite dati di esempio e può avere svariate applicazioni come: riconoscimento vocale, riconoscimento facciale, analisi delle sequenze genetiche e per la sicurezza informatica.

Per l'utilizzo predittivo del ML è necessario un processo detto *'feature engineering'* che comporta che degli esperti selezionino le caratteristiche informative dei dati del dataset necessario all'algoritmo di ML per la generazione della predizione in questione.

Mentre il DL è a sua volta un sottocampo del ML e attraverso esso si intende automatizzare il processo *'feature engineering'*, ciò apprendendo quali sono le caratteristiche ottimali dei dati campione. Inoltre, il DL riguarda le metodologie che fanno affidamento sulle reti neurali.

Le Reti Neurali sono sistemi di elaborazione dell'informazione i cui meccanismi di funzionamento si ispirano ai circuiti nervosi biologici[3]. Le reti neurali (e per tanto anche DL e ML) possono seguire uno o più dei seguenti tre protocolli di apprendimento[4]: con supervisione[5], senza supervisione[6] e con rinforzo[7].

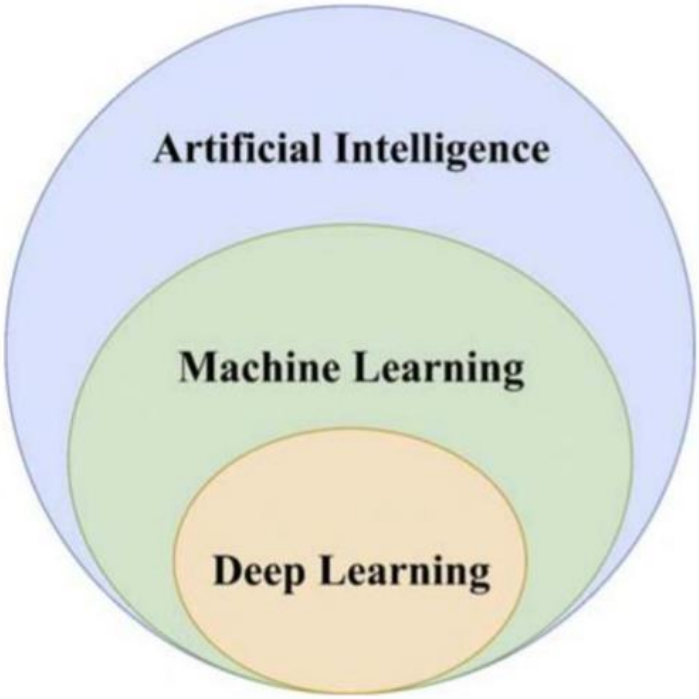


Figura 1: rappresentazione grafica del rapporto tra i concetti di AI, ML e DL.

Fonte: [8]

[3] "Manuale sulle Reti Neurali" – *di: Floreano Dario e Mattiussi Claudio* (Il mulino 2002) - passim
[4] INTRODUZIONE ALLE RETI NEURALI ARTIFICIALI - Marco Gori (MONDO DIGITALE • n.4 - dicembre 2003)
[5] Con supervisione: viene addestrata utilizzando un dataset etichettato, dove ogni input ha un'uscita associata nota.
[6] Senza supervisione: viene utilizzato quando i dati di input non sono etichettati. l'obiettivo della rete neurale è quello di scoprire pattern nascosti o strutture intrinseche nei dati.
[7] Con rinforzo: tramite l'interazione con un ambiente apprende una strategia (policy) per massimizzare una ricompensa cumulativa. A differenza dell'apprendimento supervisionato, non ci sono risposte giuste o sbagliate predefinite, ma solamente ricompense e penalità.
[8] "Machine Learning and Other Artificial Intelligence Applications, An Issue of Neuroimaging Clinics of North America" - Cap. "Brief history of artificial Intelligence" di Suresh K. Mukherji – (2020)

Storia IA:

Nel 1950 Alan Turing pubblicò su Mind l'articolo 'Computing Machinery and Intelligence' nel quale si chiedeva "Can machine think?" e introduceva quello che verrà chiamato Test di Turing[9]. Questo test vuole verificare se una macchina, creata appositamente per questo compito, possa convincere un esaminatore di essere un umano e che l'altro soggetto (un effettivo essere umano) invece non lo sia.

Con la conferenza di Dartmouth del 1956, nella quale venne coniato il termine IA da parte di John McCarthy[10]. Detto convegno fu essenziale per dare inizio alla ricerca di questo nuovo ambito. In tale conferenza venne anche avanzata la seguente, famosa. '*proposta di Dartmouth'*: "Lo studio procederà sulla base della congettura per cui, in linea di principio, ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli."; con questa premessa i ricercatori riuniti alla conferenza intendevano concentrare i loro sforzi per lavorare assieme e creare le basi per le future tecnologie hardware e software, per emulare l'intelligenza umana e dunque rispondere al quesito dell'articolo di Turing.

Dopo questo frangente di assoluto fermento della ricerca si andò incontro al periodo del cosiddetto 'AI winter' (tra anni Settanta e Ottanta del '900) nel quale l'assenza di fondi dedicati alla ricerca sul tema dell'IA comportò una stagnazione dell'avanzamento della conoscenza e della tecnologia. I mancati investimenti in questa materia furono dovuti agli altissimi costi che lo sviluppo dell'IA richiedeva e richiede. Infatti, per ogni progresso dell'IA sono imperative due risorse, che devono essere a disposizione in grande quantità: i dati e potenza di calcolo; queste due risorse erano estremamente costose in quel periodo, visto che anche la tecnologia dei computer era riservata quasi esclusivamente a centri di ricerca solo delle più prestigiose università. Inoltre, i possibili investimenti da enti privati erano ostacolati dal fatto che i ritorni economici di queste nuove tecnologie sembravano troppo lontani ed insicuri.

[9] "Computing machinery and intelligence" di Turing, A.M. - Mind, 59, 433-460 (1950)

[10] Successivamente proprio McCarthy sarà vincitore del premio Turing, nel 1971.

Ci fu una riscossa nel 1986 con Rumelhart e i suoi colleghi che introdussero la back-propagation in multi-layer neural networks[11] (ottimizzazione a gradiente discendente) capace di minimizzare l'errore del network e di aumentarne l'efficienza (riducendo il costo delle funzioni); queste due scoperte si rivelarono da subito rivoluzionarie inquanto aumentarono significativamente le capacità di apprendimento dei network neurali tramite un algoritmo di ottimizzazione.

Purtroppo, poco dopo questa ventata di aria fresca per la ricerca si andò incontro al secondo 'AI winter', questa volta dovuto sempre all'elevato prezzo della potenza di calcolo, ma anche alla assenza di scalabilità in queste grandi reti neurali.

La grande svolta che ridiede il via alla ricerca fu il lavoro di Boser[12] che nel 1992 riuscì ad ovviare al problema dell'insufficienza di potenza computazionale con un algoritmo di vettorizzazione. Questo piccolo cambiamento, ma dal grande impatto prese il nome di 'Kernel trick'. Anche oggi uno dei grandi problemi dell'aumento dell'efficienza dell'IA è dovuto all'enormità della potenza di calcolo necessaria e pertanto gli algoritmi in grado di ottimizzazione.

Le successive date di grande importanza fu verso la fine degli anni Novanta l'introduzione della prima rete neurale convoluzionale[13], grazie a essa è possibile risolvere problemi di classificazione, localizzazione, identificazione e region proposal[14].[15]

A inizio anni 2000 le due maggiori sfide, potenza degli hardware e costo dei dati, per il progresso delle tecnologie di IA, stavano per ricevere una grande scossa dovuta a due fenomeni: l'inizio della commercializzazione delle GPU (da parte di Nvidia nel 1999) e con la diffusione di internet e dei Personale Computer, tramite la quale il numero di dati

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Parallel Distributed Processing: Explorations in the Microstructures of Cognition, (Vol. I, pp. 318-362) by D. E. Rumelhart and J. L. McClelland - (Eds.) Cambridge, MA: MIT Press, (1986)

[12] "A Training Algorithm for Optimal Margin Classifiers." Di Bernhard E. Boser, Isabelle M. Guyon, Vladimir Vapnik. - Fifth Annual Workshop on Computational Learning Theory. ACM Press (Pittsburgh 1992)

[13] La convoluzione è un'operazione matriciale che viene svolta in abito grafico nel riconoscimento di immagini.

[14] Tramite un software che estrae zone di una immagine in cui è possibile che vi sia un oggetto.

in archiviazione salì a dismisura come mostrato dal grafico (si vuole far notare l'asse delle ordinate è espresso in scala logaritmica).

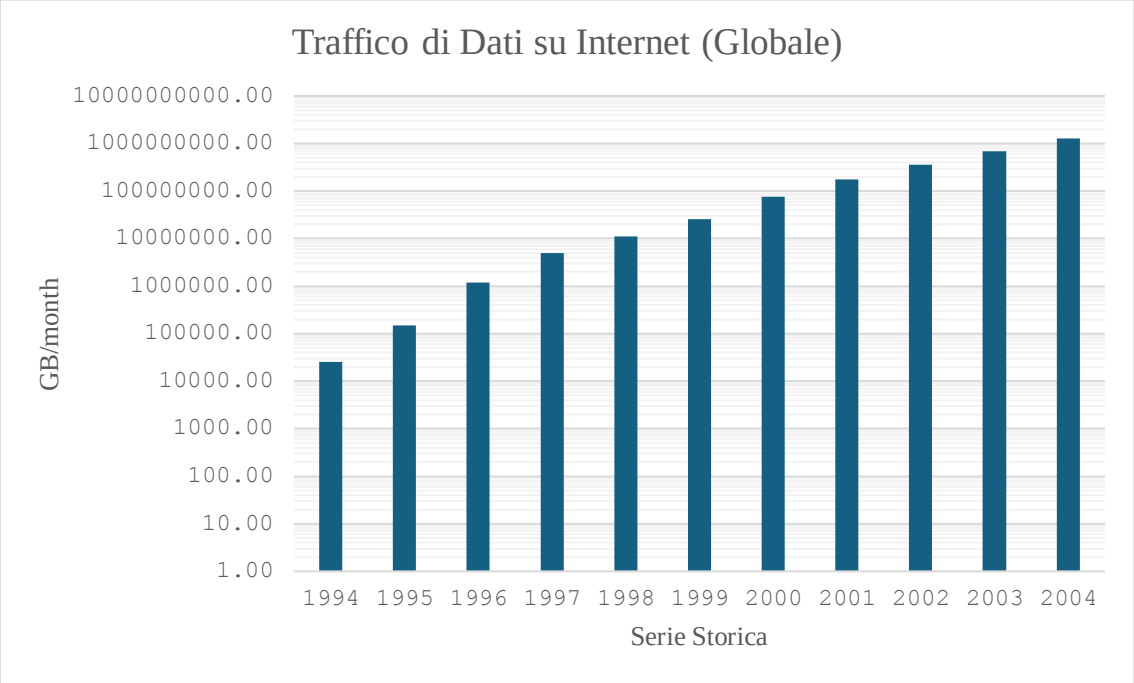


Grafico 1: Serie storica da 1994 a 2004 del traffico di internet globale

Fonte: "The History and Future of Internet Traffic" di Arielle Sumits, Cisco Blogs, (August 28, 2015)

Da questo momento in avanti la rivoluzione dell'IA incomincerà a raggiungere dimensioni di investimento compatibili con il mercato e pertanto l'innovazione sarà maggiormente trainata da singole aziende piuttosto che centri di ricerca specializzati.

Nei prossimi paragrafi analizzeremo come aziende alla stregua di IBM ed NVIDIA abbiano dato il loro contributo allo sviluppo dello stato dell'arte per quanto riguarda la tecnologia dell'IA.

Aziende che hanno Contribuito al Progresso dell'IA

IBM[16]:

Sin dagli anni Venti del Novecento, IBM era già presente settore dell'informatica nel quale occupava una posizione assieme ad altri pochi player. Già in quell'epoca, l'azienda

[16] https://www.ibm.com/history/advancing-humanitAy#Artificial+intelligence

si lancia nella traduzione automatica, con sperimentazioni destinate a mettere le basi per le tecnologie di NLP (Natural Language Processing), una delle prime attività per cui vengono utilizzate le IA. Nel giro di pochi decenni l'ambito della traduzione automatica si trasforma in un programma software sperimentale che traduce automaticamente dal russo all'inglese; il programma lavora sulla piattaforma 701 Electronic Data Processing Machine (che viene presentata per la prima volta nell'aprile del 1952), il primo computer scientifico commerciale al mondo. Lo sviluppo di questa tecnologia viene portato avanti fino ai giorni nostri, da parte di IBM con il lancio di Watson Language Translator (10 June 2023). Questo è uno strumento basato sulla tecnologia di traduzione automatica neurale, ed è in grado di tradurre istantaneamente documenti in decine di lingue.

Nel 1962[17] William C. Dersch ricercatore di IBM presentò la Shoebox, una macchina in grado di eseguire semplici calcoli matematici tramite comandi vocali. Shoebox venne presentata al pubblico all'Esposizione Universale del 1962 a Seattle (Washington). Il dispositivo riconosceva dieci cifre e sei parole di controllo, tra cui "più", "meno" e "totale", pronunciate attraverso un microfono. La creazione del Shoebox ha segnato un importante traguardo nella storia dell'informatica inquanto per la prima volta era possibile controllare un computer usando comandi vocali.

Numerosi progressi più recenti da parte di IBM nell'ambito dell'IA sono originati dal desiderio di esplorare il mondo ludico e della teoria dei giochi. I ricercatori di IBM notano come creare un ambiente con regole definite che riducono la complessità e il di scelte possibili, come avviene nel mondo dei giochi, fosse essenziale per lo sviluppo delle IA. L'azienda si imbarcherà in varie sfide per creare delle IA in grado di battere l'uomo in

[17] https://www.historyofinformation.com/detail.php?id=4989

vari giochi, partendo dai giochi da tavolo più diffusi. Nel 1992, IBM sviluppa la prima IA capace di giocare a Backgammon, un gioco relativamente semplice. Solo cinque anni dopo IBM sviluppo Deep blue, un IA per gli scacchi, che sarà in grado di battere Garry Kasparov, ancora oggi considerato da molti il più grande GM[18] di scacchi della storia.

Nel 2011 supercomputer IBM Watson fu in grado di battere Ken Jennings, il miglior giocatore umano di "Jeopardy!" di sempre, dimostrando la potenza dell'intelligenza artificiale. Negli anni successivi IBM ha investito molti milioni di dollari per promuovere Watson come assistente digitale che avrebbe aiutato ospedali e aziende agricole, ma anche uffici e fabbriche. IBM oltre al settore dell'IA è attiva anche nel settore dei supercomputer, strumenti fondamentali per la creazione dei modelli di IA impiegati in svariati settori. Secondo IBM, l'impatto più significativo del supercalcolo e dell'IA si manifesta nel miglioramento della vita quotidiana in modi che la maggior parte delle persone non percepisce.

NVIDIA:

Nel 1993 Jen-Hsun Huang (tuttora CEO dell'azienda), Chris Malachowsky e Curtis Priem fondano NVIDIA; con la visione che un giorno il PC sarebbe diventato un dispositivo di consumo per la fruizione di videogame e multimedia. Il settore nel quale si posizionano è quello della progettazione e produzione di chip grafici, all'epoca un mercato estremamente ridotto, che però incominciava a vedere l'inizio della fase di crescita del mercato.

[18] Gran Master: è il riconoscimento più alto che possa essere attribuito ad un giocatore di scacchi dalla Federazione Internazionale degli Scacchi.

Il lancio di NVIDIA NV1 nel 1995 ha segnato la prima incursione di NVIDIA nel mercato dell'elaborazione grafica. Questo processore multimediale era unico per l'epoca e integrava grafica 2D e 3D, capacità audio e persino una porta joypad compatibile con il Sega Saturn[19]. Il lancio del NV1 però non viene coronato da un completo successo: il mancato boom è dovuto principalmente al fatto che questa architettura utilizza un texture mapping quadratico, ossia questa architettura è compatibile solo con modelli 3D disegnati con rettangoli invece di quello che diventerà standard dell'industria ovvero l'utilizzo dei triangoli, che offrono una maggiore efficienza di calcolo[20], anche se proprio l'utilizzo del mapping quadratico, già all'epoca impopolare, è uno dei motivi per cui NVIDIA era stata scelta da Sega, in quanto anch'essa utilizza questa tecnologia.[21]

La collaborazione di NVIDIA con Sega è stata un aspetto significativo dello sviluppo dell'NV1. Sega era interessata a portare i suoi giochi per console Saturn sulla piattaforma PC, questa collaborazione ha portato anche al porting di alcuni titoli Sega sull'NV1. Tuttavia, nonostante le innovazioni tecnologiche, l'NV1 ha faticato a imporsi sul mercato a causa del suo approccio grafico non standard che invece privilegiava la grafica poligonale rispetto alla mappatura quadratica. Questa esperienza è stata fondamentale per NVIDIA, in quanto ha influenzato i progetti futuri dell'azienda e la sua decisione di allinearsi maggiormente agli standard emergenti del settore. Questa sfida iniziale non ha scoraggiato NVIDIA, che ha imparato lezioni preziose che hanno contribuito al successo dei suoi prodotti successivi, come la serie RIVA e le rivoluzionarie GPU GeForce. Questi dettagli evidenziano come le difficoltà iniziali di NVIDIA con l'NV1 e la sua partnership

[19] https://segaretro.org/NV1
[20] Il mapping 3D su base triangolare delle texture era già all'epoca quello utilizzato dalla maggioranza delle aziende del settore grazie alla sua minore complessità computazionale e in grado di evitare la deformazione delle texture.
[21] https://www.tomshardware.com/picturestory/715-history-of-nvidia-gpus.html

con Sega abbiano gettato le basi per il successivo successo dell'azienda nel settore della grafica. La scelta riguardo il mapping viene infatti cambiata nel 1997, con il lancio di Riva 128 (che questa volta adotta un texture mapping triangolare) e che è il primo processore 3D a 128 bit del mondo ad avere particolare successo, vendendo più di un milione di unità nei quattro mesi successivi al lancio.

Nel 1998 viene stretta una partnership strategica di durata pluriennale con Taiwan Semiconductor Manufacturing Company (TSMC), che inizia ad assistere NVIDIA nelle fasi di produzione. Questa partnership, tuttora duratura, è sicuramente uno dei passi più importanti nella storia dell'azienda, in quanto grazie a questo accordo di collaborazione strategica entrambe le parti sono potute crescere fino a diventare entrambe punti focali per i rispettivi settori. Inoltre, dal lato di NVIDIA avere un partner così importante dal lato della produzione ha fatto in modo che potesse avere un più o meno continuo approvvigionamento di chip anche nel periodo della crisi dei materiali semiconduttori (essenziali per la produzione di chip e di conseguenza per le schede video prodotte da NVIDIA).

Il lancio della GeForce 256 da parte di NVIDIA nel 1999 ha segnato una pietra miliare nella storia della grafica computerizzata, in quanto è stata la prima unità di elaborazione grafica (GPU) a integrare le capacità di trasformazione e illuminazione (T&L) direttamente sul chip. Questa innovazione ha cambiato radicalmente il panorama della grafica 3D, consentendo effetti visivi complessi e in tempo reale che in precedenza erano possibili solo con workstation di fascia alta. Questa svolta tecnologica ha posto le basi per il dominio di NVIDIA nel mercato della grafica e ha sottolineato la crescente importanza delle GPU nel computing al di là delle semplici operazioni di rendering, estendendosi ad aree quali l'elaborazione scientifica e l'IA. Questa GPU è un processore

a chip singolo con capacità di rendering in grado di elaborare un minimo di 10 milioni di poligoni al secondo (L'attuale stato dell'arte per le GPU è di oltre 7 miliardi di poligoni al secondo).

Sempre nel 1999, Nvidia procede alla IPO (offerta iniziale al pubblico) che avviene con una raccolta di 40 milioni di $, con azioni di Nvidia con un prezzo iniziale di 12$ e una valutazione di totale dell'azienda di 625 milioni di $[22].

Nel 2001 la società presenta la prima GPU programmabile del settore, NVIDIA GeForce3. Questo prodotto ha reso possibile agli sviluppatori di creare effetti visivi personalizzati, di ottimizzare le prestazioni e in generale ad avere una migliore grafica.

A segnare la crescita straordinaria del settore della grafica digitale a soli 2 anni dalla prima offerta di azioni al pubblico, NVIDIA viene inserita nel listino S&P 500 (è un indice che comprende i titoli di 500 società quotate in borsa negli Stati Uniti con i più alti valori di capitalizzazione di mercato[23]).

Il 2001 è anche l'anno nel lancio sul mercato da parte di Microsoft di Xbox, la sua prima console di gioco, per la quale sceglie NVIDIA come fornire di processori grafici. Questa console si rivelerà essere particolarmente di successo; essa permette a Microsoft un ingresso nel settore precedentemente occupato quasi esclusivamente da Sony (con cui nel 2005 collabora sviluppando il processore per Playstation 3) e Nintendo (con cui nel 2017 collabora sviluppando il processore per Nintendo Switch).

Grazie alla sua oramai stabile posizione di innovatore nel settore dei processori grafici, nel 2005 viene richiesto a NVIDIA da parte della NASA un aiuto nel ricostruire delle

[22] "I 24 anni delle azioni Nvidia in uno sguardo" di: Shanthi Rexaline – (22 maggio 2023) Yahoo finanza
[23] https://corporatefinanceinstitute.com/resources/equities/sp-500-index/

immagini del terreno di Marte. Utilizzando la tecnologia NVIDIA, i dati trasmessi dal Rover vengono renderizzati in una realtà virtuale fotorealistica, consentendo agli scienziati di esplorare Marte come se potessero davvero muoversi liberamente sulla superficie del pianeta rosso.

Poco più di dieci anni dopo la sua fondazione, nel 1993, NVIDIA strappa il posto di azienda leader per quanto riguarda la ricerca e sviluppo per le IA a IBM, grazie alla sua intuizione inerente al ruolo strategico dell'accelerated computing. Tale visione l'ha portata a diventare rapidamente uno dei colossi tech e dell'IA a livello globale.

La svolta cruciale portata da NVIDIA è la presentazione di CUDA, nel 2006, un insieme di strumenti rivoluzionari per il GPU Computing. CUDA è una libreria che rende possibile creare applicazioni ad alte performance grazie all'Accelerated Computing. L'Accelerated Computing consente di sfruttare le capacità di elaborazione in parallelo delle GPU per risolvere i problemi di calcolo più complessi assieme all'utilizzo della CPU.

Questa maggiore efficienza è dovuta alla differente metodologia di calcolo che avviene tra CPU e GPU; in quanto la CPU è progettata per elaborare dati sequenziali, mentre la GPU è progettata per elaborare dati paralleli. Inoltre, le GPU sono dotate di una grande quantità di memoria che consente di memorizzare temporaneamente i dati necessari all'elaborazione grafica. Questa memoria è molto più veloce rispetto alla memoria della CPU e permette di elaborare le immagini in tempo reale.

Questa essenziale differenza nello svolgimento dei calcoli tra CPU e GPU può essere sfruttata per velocizzare svariate applicazioni, grazie ad architetture come CUDA. Visto che esse rendono possibili la scrittura di programmi, normalmente svolti dalla CPU, di

modo che le operazioni possano essere svolti con il calcolo parallelo. Grazie alla trasposizione del programma in calcolo parallelo ed eseguito tramite GPU risulta generalmente in una sensibile diminuzione[24] del tempo di esecuzione. Questa pratica si chiama Accelerated Computing (AC) ed è ciò che ha reso possibile la creazione di Fundation Model da circa 1,76 trilioni di parametri[25] (ChatGPT-4).

La spinta di NVIDIA nello sviluppo delle tecnologie IA continua dopo CUDA (che viene continuamente aggiornato a sua volta con ulteriori librerie per ulteriori applicazioni dell'AC), tramite un continuo miglioramento dello stato dell'arte per quanto riguarda gli hardware. NVIDIA continua dai primi anni duemila ad ora a progettare, produrre e vendere nuove schede video basate su architetture sempre più potenti.

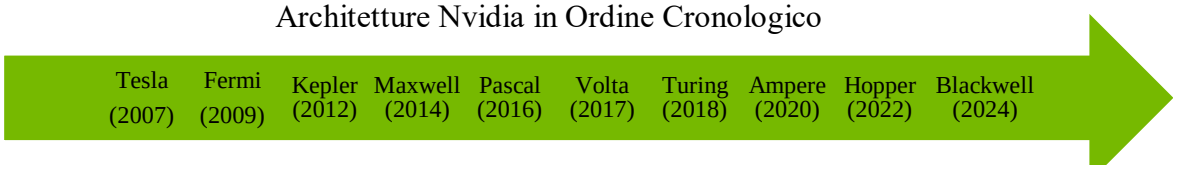


Grafico 2: Architetture Nvidia in Ordine Cronologico
Fonte:[26]

NVIDIA, inoltre, si stabilisce come standard per quanto riguarda i processori grafici per l'industria cinematografica e dei videogiochi.

A partire dal 2008 NVIDIA svolge diverse partnership, principalmente con centri di ricerca, con i quali sviluppa diversi supercomputer come: Tsubame[27] (170 TFLOPS[28] - 2008), Tianhe-1°[29] (4701 petaFLOPS - 2010), Titan[30] (27 petaFLOPS - 2012).

[24] Se eseguito con librerie ottimizzate si arriva ad incrementi anche ben superiori a 100x
[25] https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
[26] https://www.nvidia.com/it-it/technologies/
[27] Realizato in collaborazione con l'istituto Tokyo Tech
[28] FLOPS: sta per **FL**oating point **O**perations **P**er **S**econd è l'unità di misura utilizzata per misurare le performance di calcolo delle CPU e GPU
[29] Realizato in collaborazione con il National Supercomputing Center cinese
[30] Realizato in collaborazione con l'Oak Ridge National Laboratory

Nel 2016 NVIDIA lancia sul mercato DGX-1, il suo primo supercomputer di DL compatto. Uno degli aspetti più importanti del DGX-1 è la sua architettura, che include otto GPU Tesla P100 connesse tramite NVLink, un'interconnessione ad alte prestazioni sviluppata da NVIDIA. Questo setup consente al sistema di ottenere una velocità di elaborazione superiore rispetto ai sistemi tradizionali basati su PCIe[31], specialmente in applicazioni che richiedono un'intensa comunicazione tra le GPU, come il training di modelli di DL. L'importanza del lancio questo prodotto sta nella creazione di un mercato di "massa" in quello che prima avveniva solo su commessa è pertanto con elevati costi di progettazione e design. Con un prodotto come DGX NVIDIA dà la possibilità alle aziende di poter acquistare un modello standard di supercomputer con il quale possono creare la propria IA per incrementare la produttività aziendale.

La prima DGX-1 prodotta da NVIDIA viene regalata e consegnata di persona d Jensen Huang ad OpenAI. Grazie a questo supercomputer ad Open AI viene reso possibile l'allenamento su modelli più grandi e complessi, vista le incredibili prestazioni e efficienza energetica offerta. Fin da subito i supercomputer prodotti da NVIDIA, come DGX-1, vennero progettati per essere modulari e rendere possibili la costruzione di datacenter con centinaia di supercomputer, per i quali svilupperanno anche prodotti come NVLink; Un enorme trasmettitore di dati ad altissima velocità che rende possibile alle varie DGX-1 di lavorare come facenti parte dello stesso computer. DGX-1 rappresenta una soluzione potente e scalabile per creare modelli IA su larga scala, con applicazioni vastissime in quasi tutte le industrie.

[31] è uno standard di interfaccia di porte per il trasferimento dati

OpenAI:

L'11 dicembre 2015 viene fondata OpenAI, ad opera di un gruppo composto da Elon Musk, Sam Altman (presidente di Y Combinator), Reid Hoffman (co-fondatore di LinkedIn), Peter Thiel (Co-fondatore di PayPal) e Jessica Livingston (Co-fondatrice di Y Combinator). Si tratta di un gruppo di particolare spicco, composto da celebri founder di startup di incredibile successo: è proprio grazie ai guadagni di tali startup che i fondatori sono potuti entrare in questo progetto con investimenti iniziali importanti. Come già detto, nel 2016 i founder riceveranno da Jensen Huang in persona un supercomputer DGX-1 da parte di NVIDIA, fondamentale per l'allenamento di LLM.

Gli LLM (Large Language Model) sono un sottoinsieme dei fundation model, che vengono allenati su enormi ammontari di dati[32] non etichettati e tramite un apprendimento auto supervisionato. Il modello tramite l'analisi dei dati impara a riconoscere i pattern, in modo da produrre output generalizzabili. Gli LLM vengono definiti language model in quanti i dati che utilizzano sono testi o simili (es. righe di codice).

La mission di OpenAI consiste nella ricerca e sviluppo dell'IA di modo da garantire che essa porti benefici a tutta l'umanità[33]. Tale mission vuole essere attuata tramite tre pilastri di intenti che OpenAI stessa si impone: apertura, sicurezza e trasparenza.

Apertura in quanto il modello di innovazione di OpenAI si basa sull'open innovation, rendendo i loro modelli opensource e rendendo i loro tool a disposizione di tutti. La sicurezza è un altro pilastro di fondamentale importanza per OpenAI, ciò risulta di particolare rilevanza per quanto riguarda i dati che OpenAI ottiene tramite l'interazione degli utenti con LLM come ChatGPT 4. Il terzo ed ultimo pilastro riguarda la trasparenza

[32] Fino a petabyte (1.000.000.000.000.000 byte) di dati.
[33] https://openai.com/about/

della gestione e dell'operazione (almeno inizialmente, l'azienda ha adottato un atteggiamento proattivo nei confronti della condivisione delle informazioni e si presentava a quel tempo in opposizione a quelle che potrebbero essere considerate le pratiche più opache di altri giganti tecnologici).

Nel 2016 OpenAI rilascia due piattaforme di ricerca OpenAI, ovvero Gym e Universe. OpenAI Gym è un kit di strumenti per lo sviluppo e il confronto di algoritmi di reinforcement learning (RL). Consiste in una serie di ambienti simulativi (scritti in Python) come classici giochi per cabinati Atari, giochi da tavolo e simulazioni per il movimento di robot in uno spazio tridimensionale, per accelerazione della ricerca sulla RL.[34] Universe è una piattaforma software che rende possibile trasformare qualsiasi programma in un ambiente Gym per poterne misurare e addestrare l'intelligenza generale di un'IA35.

Nel 2019 si ha una grande novità per OpenAI in quanto la struttura dell'azienda passa[36] da no-profit a un modello di organizzazione "capped-profit"[37] con Sam Altman che ne diventa CEO.[38] La sua partecipazione, dal valore pari a un 1 miliardo di dollari, sempre nello stesso anno dall'ingresso da un investimento da parte di Microsoft. Questo

[34] https://openai.com/index/openai-gym-beta/

[35] https://openai.com/index/universe/

[36] Non vi è stato un completo cambio da no-profit a for-profit ma più che altro la creazione di una nuova for profit controllata e di proprietà dalla precedentemente fondata OpenAI non-profit.

[37] Secondo la quale gli investitori possono ottenere un massimo di 100 volte il loro investimento iniziale e qualsiasi profitto oltre tale soglia sarebbe reinvestito nell'organizzazione o ridistribuito.

[38] https://www.repubblica.it/tecnologia/2023/01/15/news/openai_chatgpt_storia_musk_altman-383409890/

passaggio porta il progetto OpenAI in diretta competizione con giganti tecnologici come Google e Facebook, che investono miliardi nello sviluppo dell'IA.

Nel 2018, avviene il rilascio di GPT-1[39] (Generative Pre-trained Transformer), un modello di GenAI sotto forma di LLM. Per il fundation model di GPT-1 OpenAI viene costruito su 117 milioni di parametri[40], destinato poi ad essere superato dal suo successore GPT-2 (rilasciato nel 2019).

GPT-2 è basato su 1.5 miliardi di parametri, migliorando la qualità generale dei token generati e riducendo le possibilità di allucinazione.

Nel 2020 OpenAI lancia GPT-3 (basato su 175 miliardi di parametri) con una scala di grandezza dei dati analizzati pari a 45 TB rispetto ai 40 GB del suo predecessore[41]. GPT-3 dimostra di essere uno strumento dalle straordinarie capacità di generazione di testo e di comprensione del linguaggio, essendo in grado di risolvere una vasta gamma di problemi di linguaggio naturale.

Nel Novembre del 2022 viene lanciato ChatGPT ovvero una versione particolarmente user frendly di GPT basata sul modello di GPT-3, con però degli ulteriori miglioramenti, viene pertanto denominato GPT-3.5 . Questo strumento rivoluzionario segna un totale cambio di marcia nel mondo dell'IA in quanto nel giro di soli cinque giorni dal lancio ha già superato il milione di utenti, che nei due mesi successivi diventano 100 milioni[42]. La facilità di utilizzo di ChatGPT rende disponibile alle masse uno strumento di enorme

[39] GPT è un modello di IA di LLM progettato per comprendere e generare testo in linguaggio naturale.
[40] "A commentary of GPT-3" di Min Zhang e Juntao Li - *MIT Technology Review* (2021)

[41] "A commentary of GPT-3" di Min Zhang e Juntao Li - *MIT Technology Review* (2021)
[42] https://tg24.sky.it/tecnologia/2023/12/04/chat-gpt-utenti-storia#05

potenza in grado di semplificare e aiutare lo svolgimento di un numero esorbitante di attività.

Il 2023 ha visto il rilascio di GPT-4, una versione ancora più avanzata del modello di linguaggio. GPT-4 è stato lodato per la sua maggiore accuratezza, comprensione contestuale e capacità di rispondere in maniera più complessa e articolata rispetto alle versioni precedenti. Questo modello ha ampliato ulteriormente le applicazioni dell'IA generativa in settori come la ricerca, la creatività e la produttività.

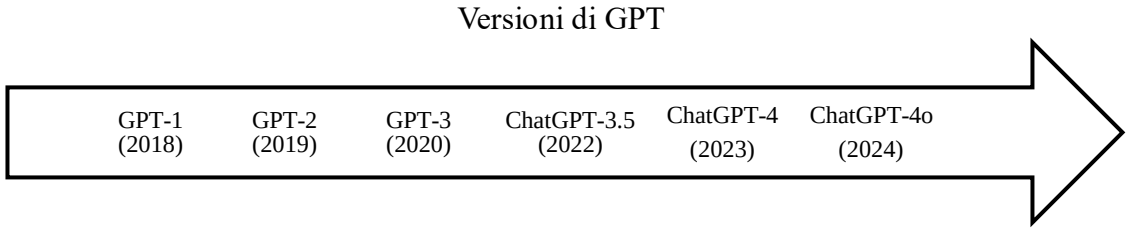


Grafico 3

Fonte: [43]

OpenAI continua ad essere un punto cardine nello sviluppo dell'IA grazie al suo impegno nella ricerca avanzata e nel lancio di prodotti innovativi facilmente utilizzabili; contribuendo a plasmare il panorama dell'IA su scala globale.

[43] https://medium.com/@dlaytonj2/chatgpt-how-we-got-to-where-we-are-today-a-timeline-of-gpt-development-f7a35dcc660e

Cap.2 CASO STUDIO: NVIDIA

I Fattori Esterni[44]:

Diversi eventi degli ultimi decenni si sono rivelati favorevoli all'ampiamento dei mercati in cui opera NVIDIA, ciò è solamente in parte dovuto alla sorte e maggiormente alle generali strategie di sviluppo tecnologico messe precedentemente in atto e alla reattività di risposta al cambiamento dell'ambiente.

Nel 2012 viene presentato AlexNet, un pionieristico algoritmo di DL sviluppato dai ricercatori dell'Università di Toronto utilizzando le GPU NVIDIA; tale algoritmo si è assicurato la vittoria nell'ImageNet Challenge, una prestigiosa competizione per il riconoscimento delle immagini. Questo riconoscimento ha scatenato un interesse diffuso per il DL e l'IA, innescando un'impennata nella domanda di soluzioni di calcolo ad alte prestazioni come le GPU NVIDIA.

Il 2016 ha visto la storica sconfitta del campione di Go Lee Sedol da parte di AlphaGo[45], questo risultato è stato ottenuto grazie alle GPU NVIDIA utilizzate per AlphaGo di DeepMind.

Durante la pandemia COVID-19 del 2020 si è verificata una rapida accelerazione nell'adozione dell'IA e dell'apprendimento automatico, in particolare in compiti come la diagnosi medica, lo sviluppo di vaccini e in generale in ambito medico. Le collaborazioni con istituzioni e aziende mediche per applicazioni sanitarie basate sull'IA hanno

[44] A Comprehensive Analysis of NVIDIA's Technological Innovations, Market Strategies, and Future Prospects - *John Wang, Jeffrey Hsu, Zhaoqiong Qin* - International Journal of Information Technologies and Systems Approach (Volume 17; Issue 1 2024)

[45] È un software per il gioco del go (sviluppato da Google DeepMind) che è stato il primo programma in grado di sconfiggere un maestro umano.

dimostrato come le tendenze del mercato in settori specifici abbiano contribuito in modo significativo al continuo successo di NVIDIA.

Un altro evento che si è rivelato estremamente favorevole a NVIDIA è stato il boom del mining di criptovalute, in particolare di Bitcoin. Il rapido sviluppo del settore delle criptovalute e della blockchain ha creato un'impennata senza precedenti nella domanda di GPU ad alte prestazioni. Le GPU di NVIDIA, in particolare quelle della serie GeForce, sono diventate ricercate per le loro eccezionali capacità di elaborazione in parallelo. L'inaspettata espansione nel mercato delle criptovalute ha messo in evidenza l'adattabilità dei prodotti NVIDIA, dimostrando la loro rilevanza al di là dei tradizionali ambiti di elaborazione grafica e gioco.

Trend Tecnologici Globali

Il panorama dell'informatica sta subendo cambiamenti significativi a partire dal 2010, anno da cui inizia l'adozione diffusa di servizi di cloud computing come: AWS[46] (Amazon Web Services) e Microsoft Azure[47]. Lo sviluppo di questi prodotti con tecnologia cloud porta a un'impennata della domanda di GPU ad alte prestazioni necessarie nei data center. A questo sviluppo NVIDIA risponde adattando strategicamente i propri prodotti e servizi a questo mercato in evoluzione, tramite l'inserimento di soluzioni GPU basate sul cloud e stringendo nuove partnership con i principali player del settore dei cloud provider.

[46] https://aws.amazon.com/it/
[47] https://azure.microsoft.com/it-it

Nel 2016 il numero di dispositivi Internet of Things (IoT)[48] ha visto un rapido aumento di assieme all'emergere dell'edge computing[49]. Queste tendenze hanno reso evidente l'importanza strategica di un computing efficiente, sempre più potente e maggiormente vicino alle fonti di dati; continuando così la crescita del settore dei data center e di conseguenza la domanda dei prodotti NVIDIA.

Nel 2020 con l'arrivo della quinta generazione della tecnologia cellulare (5G), portando un sensibile incremento della velocità di trasmissione dei dati wireless e consentendo applicazioni rivoluzionarie come la realtà aumentata (AR[50]).

L'introduzione di architetture GPU ad alta efficienza energetica, esemplificate dalle architetture NVIDIA Maxwell[51] e Pascal[52], si è sviluppata in perfetta concomitanza con la crescente consapevolezza e richiesta di soluzioni di un computing improntate alla riduzione del consumo elettrico e maggiormente sostenibili. L'impegno di NVIDIA per la sostenibilità ambientale si palesa nel suo approccio proattivo alla fornitura e sviluppo di prodotti che, oltre a soddisfare le richieste di prestazioni dei clienti, aderiscono ai principi del green computing.

L'adattabilità di NVIDIA a paradigmi mutevoli in diversi settori, che vanno dal cloud computing all'edge computing hanno garantito all'azienda flessibilità essenziale in un settore nel quale avvengono rapidi cambiamenti come quello della tecnologia. NVIDIA

[48] Per IoT si intende l'estensione di internet al mondo degli oggetti e dei luoghi concreti. Solitamente si tratta di oggetti comuni i quali vengono aggiunti apparati che gli danno la possibilità di poter comunicare con altri oggetti nella rete e poter fornire servizi agli utenti.
[49] Per Edge-computing si intende avvicinare il più possibile l'elaborazione dei dati a dove i dati vengono generati, migliorando i tempi di risposta e risparmiando sulla larghezza di banda.
[50] Per realtà aumentata si intende l'arricchimento della percezione sensoriale umana mediante informazioni, in genere manipolate e convogliate elettronicamente, che non sarebbero percepibili con i cinque sensi.
[51] https://www.nvidia.com/it-it/data-center/pascal-gpu-architecture/
[52] https://developer.nvidia.com/maxwell-compute-architecture

si è inoltre concentrata sull'efficienza energetica dei sui prodotti, mettendo una pronunciata enfasi sul green computing. Attenzioni di questo tipo dimostrano la lungimiranza di NVIDIA e la sua capacità di allineare il proprio portafoglio tecnologico alle esigenze in evoluzione del panorama globale.

La Competizione nel Settore Tecnologico

Nei settori altamente competitivi delle GPU per videogiochi e delle tecnologie per i data center[53], NVIDIA ha sempre dato evidenziato le sue virtù di resilienza e adattabilità. La continua competizione con concorrenti temibili, come AMD, hanno obbligato e reso necessaria la incessante innovazione, portando al continuo miglioramento di prodotti e servizi. La risposta strategica di NVIDIA alla crescente domanda di hardware incentrato sull'IA nei data center ha portato l'azienda a dover fronteggiare una concorrenza diretta con i produttori tradizionali di CPU come Intel.

Altrettanto, l'introduzione di NVIDIA DRIVE Hyperion 8 nel 2020[54] (una piattaforma per la guida autonoma) riflette l'ambizione dell'azienda di ergersi a leader nel mercato emergente della guida autonoma dei veicoli. Questa iniziativa rivela la volontà di NVIDIA di assumere rischi e investire in tecnologie dal grande potenziale, evidenziando il suo impegno per i progressi pionieristici nel settore dell'automotive. Il vantaggio dell'azienda nel settore dei chip per l'intelligenza artificiale, la posiziona in modo

[53] NVIDIA Investor Presentation - October 2023 disponibile al link: https://investor.nvidia.com/events-and-presentations/presentations/presentation-details/2023/NVIDIA-Investor-Presentation-October-2023/default.aspx

[54] NVIDIA DRIVE Hyperion Combines Orin With Best-in-Class Sensor Architecture for Production-Ready Platform - *by Gary Hicok* – Nvidia website (9 Novembre 2021)

strategico per ottenere conoscenze superiori per i prossimi progetti di chip, intensificando così la sfida per i concorrenti che cercano di colmare il divario tecnologico.

Le scelte strategiche di NVIDIA sia nel settore delle GPU che nelle tecnologie per i data center, assieme alla sua entrata nell'ambito nella guida autonoma e alle innovazioni rivoluzionarie nello sviluppo di chip AI, definiscono NVIDIA come un'azienda che non è solo in grado di risponde in maniera efficacemente alle dinamiche del mercato, ma capace di modellare in modo proattivo il panorama competitivo a suo favore.

Il Settore Governativo (B2G)[55]:

Nel 2013 con l'avvio della US BRAIN Initiative (un progetto di ricerca globale con partnership federali e non federali con l'obiettivo comune di accelerare lo sviluppo di neurotecnologie innovative) il governo USA ha stanziato fondi per la ricerca che utilizza l'IA e il HPC. Questo significativo investimento da parte del governo nordamericano ha agito da catalizzatore per l'aumento della domanda di GPU NVIDIA nelle applicazioni di ricerca scientifica.

Il 2017 ha segnato un momento cruciale con l'iniziativa strategica della Cina, che ha comportato ingenti investimenti del governo cinese adibiti allo sviluppo dell'IA e dei semiconduttori. Grazie a questa iniziativa a NVIDIA si sono aperte nuove strade per quanto riguarda il mercato cinese. L'importanza strategica dell'industria dei semiconduttori come essenziali per i prodotti NVIDIA ha comportato innumerevoli sfide dovute in particolare all'evoluzione delle politiche commerciali, delle tariffe e dei

[55] A Comprehensive Analysis of NVIDIA's Technological Innovations, Market Strategies, and Future Prospects - *John Wang, Jeffrey Hsu, Zhaoqiong Qin* - International Journal of Information Technologies and Systems Approach (Volume 17; Issue 1 - 2024)

controlli sulle esportazioni. L'abile navigazione e l'adattabilità di NVIDIA in risposta a queste dinamiche sono diventate cruciali per il sostentamento della sua catena di fornitura globale e preservare la sua leadership sul mercato.

Nel 2018, la Commissione europea ha adottato un piano coordinato con gli Stati membri, aggiornato nel 2021, per aumentare gli investimenti nell'intelligenza artificiale e adattare il quadro giuridico[56]. Questa mossa strategica dell'UE va a vantaggio di NVIDIA e altre aziende di tecnologia IA, promuovendo un ambiente favorevole all'innovazione e a spingere ulteriormente la crescita del mercato.

In definitiva, il percorso di NVIDIA attraverso queste importanti iniziative globali evidenzia la sua capacità di allinearsi ai principali progetti di ricerca e di affrontare le sfide geopolitiche. L'agilità dell'azienda nel rispondere alle dinamiche in evoluzione delle collaborazioni internazionali e degli scenari commerciali è stata fondamentale per mantenere e rafforzare la sua posizione di rilievo.

I Prodotti:

NVIDIA è un gigante high-tech. Analizzando in senso ampio il suo settore di riferimento è quello della tecnologia nel quale si confronta quasi esclusivamente con altri competitor di grandi dimensioni.

La produzione di NVIDIA consiste nel design di chip grafici ad alte prestazioni e nella realizzazione di software e piattaforme; i quali vengono utilizzati in svariati settori come: IA, ricerca scientifica, gaming, design creativo, veicoli autonomi, metaverso e robotica. I

[56] https://ec.europa.eu/commission/presscorner/detail/it/IP_18_4521

suoi chip grafici sono riconosciuti a livello globale per l'elevata potenza in rapporto all'energia utilizzata. Nello sviluppo di software per la creazione di IA, NVIDIA ricopre assieme ad altre aziende il ruolo di innovatore portando costantemente miglioramenti a livello tecnico al settore.

Pertanto, la produzione di NVIDIA è totalmente immateriale visto che la produzione di prodotti come schede video e chip (con marchio NVIDIA) viene effettuata da terze parti. Il principale produttore di chip progettati da NVIDIA e TSMC (Taiwan Semiconductor Manufacturing Company[57]). Il 60% dell'intera produzione[58] di TSMC sono chip di NVIDIA e le due aziende lavorano in stretto contatto sin dal 1998.[59]

I vari settori nei quali NVIDIA opera sono collegati strettamente dal know-how dell'azienda in fatto di funzionamento di microchip e come possano essere utilizzati al meglio tramite strumenti e software. Questi ultimi vengono principalmente utilizzati in materia di IA ed elaborazione grafica.

Il grafico sottostante illustra la percentuale delle entrate di NVIDIA per il 2023 suddivisa per i principali settori di attività dell'azienda. Osservando il diagramma a torta, si può notare come il settore dei Data-center rappresenti la quota più significativa delle entrate, seguito dal Gaming, che rimane un'area di grande importanza per l'azienda.

[57] TSMC è il più grande produttore indipendente di semiconduttori al mondo, con sede principale presso lo Hsinchu Science Park di Hsinchu, Taiwan.
[58] https://www.agendadigitale.eu/mercati-digitali/corsa-ai-chip-ai-tutti-contro-NVIDIA-chi-dominera-il-mercato/
[59] "CRONOLOGIA DI NVIDIA - Una storia di innovazione" on NVIDIA website – disponibile su: https://www.nvidia.com/it-it/about-nvidia/corporate-timeline/

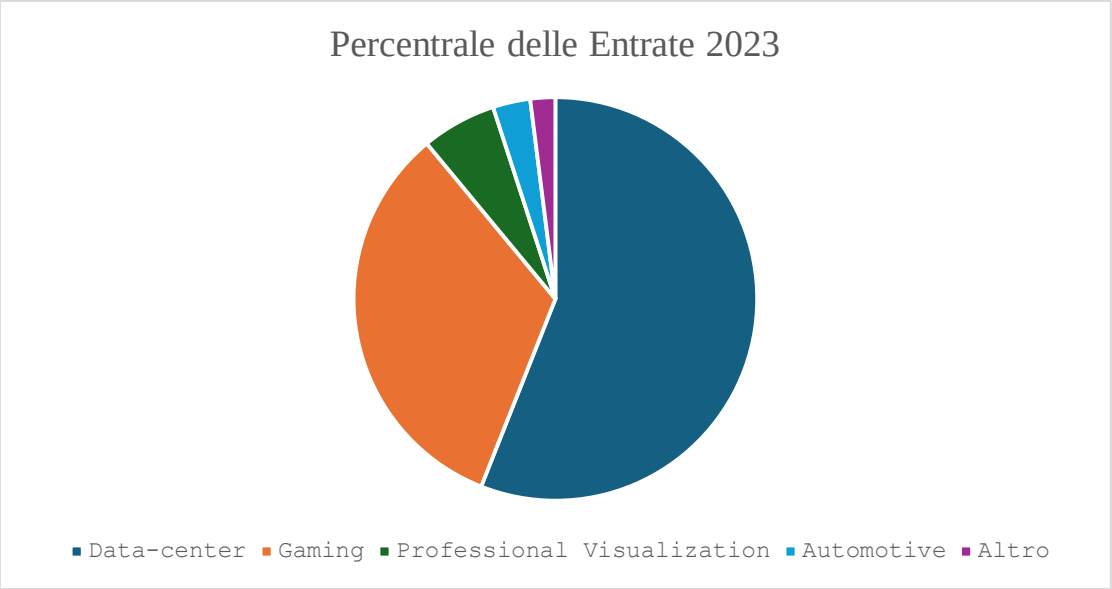


Grafico 4: Percentuale di contribuzione dei vari settori agli utili di Nvidia (2023)

Fonti:[60]

Il settore dei data-center oltre ad essere quello di maggior impatto sulle entrate di NVIDIA è anche quello con la crescita prevista per i prossimi 5 anni maggiore, inquanto secondo le stime di NVIDIA il CAGR (tasso di crescita annuale composto) sarà del 51%. I settori di gaming, visualizzazione professionale e automotive hanno invece un CAGR che si aggira attorno al 10%, ciò mette ancora maggiormente in risalto come il settore dei data-center sia in ascesa dell'IA. Traendo vantaggio dalla sua posizione di leader del mercato delle GPU ha continuato la sua espansione in svariati settori, sfruttando i cambiamenti di ambiente portandoli a suo favore.[61]

[60] NVIDIA Investor Presentation - October 2023 disponibile al link: https://investor.nvidia.com/events-and-presentations/presentations/presentation-details/2023/NVIDIA-Investor-Presentation-October-2023/default.aspx
[61] NVIDIA Investor Presentation - October 2023 disponibile al link: https://investor.nvidia.com/events-and-presentations/presentations/presentation-details/2023/NVIDIA-Investor-Presentation-October-2023/default.aspx

Come abbiamo già discusso principali strumenti software CUDA è sicuramente il progetto software più rilevante di NVIDIA in quanto rende possibile programmare codice per il calcolo simultaneo delle GPU e pertanto creare delle applicazioni accelerate.[62]

Gli altri importanti progetti software NVIDIA sono[63]: TensorRT (è un software di ottimizzazione dell'inferenza per massimizzare le prestazioni dei modelli di DL nelle applicazioni di IA), NeMo (è un toolkit open-source progettato per costruire, addestrare e mettere a punto modelli di IA per il riconoscimento vocale e l'elaborazione del NLP), Clara (è una piattaforma incentrata in campo medico con supporti per la diagnostica per immagini, studi del genoma e lo sviluppo di modelli di IA per applicazioni mediche), Triton Inference Server (è una piattaforma scalabile per la gestione dei modelli di inferenza dell'IA).

Tramite i suoi prodotti NVIDIA continua a ridefinire gli standard dell'innovazione tecnologica, mantenendo una posizione di leadership nella progettazione e sviluppo di strumenti avanzati che stanno trasformando il mondo digitale.

Il Mercato:

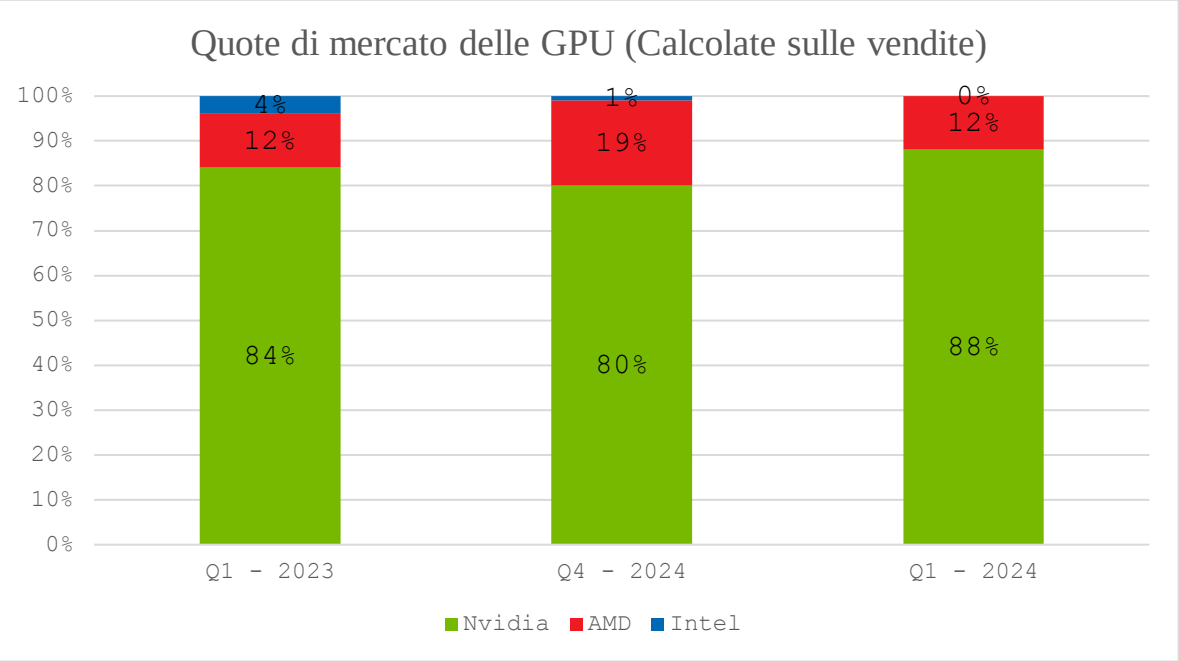
Per quanto riguarda il mercato di riferimento di NVIDIA, bisogna sempre fare la distinzione tra i prodotti hardware e quelli software prodotti da NVIDIA. Partendo dal settore del design delle GPU, NVIDIA si riferisce ad un mercato particolarmente ampio di clienti che vanno dai gamer ai graphic designer a miner di criptovalute. L'altra parte della domanda consiste invece nel B2B per il quale NVIDIA progetta anche delle schede-video apposite. Le aziende clienti di NVIDIA operano principalmente in settori come: la

[62] https://developer.nvidia.com/cuda-toolkit
[63] Catalogo software Nvidia: https://www.nvidia.com/it-it/software/

rivendita di hardware, produzione di computer e PC, data center e grandi aziende che utilizzano tecnologie di IA.

Nel 2024, il mercato globale delle GPU è stato valutato in 65,3 miliardi di dollari. In oltre le previsioni di crescita del settore per il 2029 che arriva a 274,2 miliardi di dollari[64]. Di questo mercato le vendite del primo trimestre del 2024 sono state per l'88% con marchio NVIDIA, 12%. Rispetto al trimestre precedente la quota di NVIDIA è aumentata del 8% mentre quella di AMD ha visto una riduzione del 7%.

Pertanto, la concentrazione del mercato delle GPU secondo l'indice HHI è di 7888 visto che significativamente maggiore di 1800[65] il mercato è definibile come molto concentrato.



[64] https://www.statista.com/statistics/1166028/gpu-market-size-worldwide/#:~:text=In%202024%2C%20the%20global%20graphics,percent%20from%202024%20to%202029

[65] https://www.justice.gov/atr/herfindahl-hirschman-index#:~:text=The%20agencies%20generally%20consider%20markets,Guidelines%20%C2%A7%202.1%20(2023).

Grafico 5: Quote di mercato delle GPU (Calcolate sulle vendite) – da Q1 2023 a Q1 2024

Fonti: [66]

Dal punto di vista territoriale, il mercato di riferimento di NVIDIA è esteso a livello globale. NVIDIA opera nelle varie nazioni solitamente con sezioni per singoli paesi o raggruppamenti regionali. Oltre alla sede centrale in California, negli ultimi anni i dipartimenti europei e asiatici stanno vivendo un periodo di particolare crescita dovuto all'aumento della domanda di GPU per data center e soluzioni AI.

Il mercato di riferimento per NVIDIA è estremamente vasto ma composto da settori chiave come: IA, ricerca scientifica, gaming, design creativo, veicoli autonomi, metaverso e robotica. Questi settori serviti sono frutto dell'utilizzo combinato o meno dei prodotti NVIDIA. Per rilevanza e importanza strategica dell'azienda, conviene partire dall'analisi del mercato delle GPU. Il mercato delle GPU è caratterizzato dalla presenza di importanti operatori dalle grandi dimensioni come: Intel Corporation, AMD (Advanced Micro Devices Inc) e NVIDIA Corporation. Gli operatori del mercato stanno adottando strategie quali partnership e acquisizioni per migliorare la propria offerta di prodotti e ottenere un vantaggio competitivo sostenibile.

Competitor:

Nei vari settori in cui NVIDIA opera si trova a competere con vari concorrenti. Rispetto ad aziende come Microsoft e AMD, NVIDIA si trova spesso in competizione in più settori.

[66] https://www.jonpeddie.com/news/shipments-of-graphics-add-in-boards-decline-in-q1-of-24-as-the-market-experiences-a-return-to-seasonality/

AMD:

Partendo da AMD, NVIDIA la ritrova come concorrente nel settore della progettazione di GPU e congruentemente anche in prodotti come le schede video. Negli ultimi anni questo mercato è sempre stato caratterizzato da una posizione dominante da parte di NVIDIA rispetto a AMD. Nonostante ciò, le ultime schede video lanciate da AMD sul mercato sono state riconosciute dagli esperti del settore come in grado di competere a livello di prestazione con la serie 4000 di NVIDIA[67].

Modello	Prezzo (MSRP[68])	Rasterizzazione (1440p)	Prestazioni Ray-Tracing (1080p)	VRAM	Efficienza energetica
NVIDIA RTX 4090	$1,699.99	Eccellente (oltre 120 FPS)	Eccellente (oltre 90 FPS)	24 GB	450 W
AMD Radeon RX 7900 XTX	$999.99	Ottima (120 FPS)	Buona (circa 70 FPS)	24 GB	355 W

Tabella 1: Confronto tra prestazioni e specifiche delle schede grafiche top di gamma di NVIDIA e AMD

Fonti : [69-70]

Ulteriormente AMD ha, proprio come NVIDIA, una presenza anche nel mercato dei software e strumenti per sviluppo delle IA. Il suo ecosistema di software per l'IA, in particolare attraverso ROCm. ROCm è uno stack software[71] AMD per la programmazione di unità di elaborazione grafica ed è la risposta di AMD al dominio di NVIDIA con

[67] Mentre negli anni precedenti NVIDIA aveva una posizione di leader non solo a livello di vendite, ma anche a livello tecnologico, questo divario si è ridotto significativamente nell'ultimo periodo.
[68] manufacturer's suggested retail price
[69] https://www.computerbase.de/2023-09/graphics-card-ranking-gpu-comparison-geforce-rtx-radeon-arc/
[70] https://www.techradar.com/news/computing-components/graphics-cards/amd-vs-NVIDIA-who-makes-the-best-graphics-cards-699480
[71] L'insieme di componenti che lavorano insieme per supportare l'esecuzione dell'applicazione.

CUDA. AMD sta lavorando per espandere la propria influenza nel settore dell'IA offrendo alternative competitive agli sviluppatori che necessitano di soluzioni di elaborazione ad alte prestazioni ed economicamente vantaggiose. Tuttavia, l'adozione complessiva del software AI di AMD è ancora in crescita rispetto a NVIDIA.

Microsoft:

NVIDIA si trova a competere con Microsoft in settori come lo sviluppo di strumenti per la creazione di applicazioni di IA e nel settore del metaverso. Microsoft offre come prodotto software per la creazione di strumenti IA Azure[72]: una piattaforma cloud che offre una vasta gamma di servizi con una forte enfasi sull'IA e il ML. Azure[73] compete con NVIDIA inquanto offre ai clienti la possibilità di sfruttare risorse di calcolo potenti e scalabili, compresi cluster di GPU. Tramite Microsoft Azure è possibile accedere a infrastrutture hardware all'avanguardia come le GPU NVIDIA. Microsoft Azure offre in oltre servizi software di alto livello come Azure Machine Learning, che permettono agli sviluppatori di poter addestrare i modelli di IA su larga scala senza dover gestire l'infrastruttura fisica. Azure si trova in competizione con NVIDIA CUDA soprattutto nel settore delle piattaforme di IA basate su cloud.

Riguardo il metaverso, Microsoft è investita in questo settore in particolare con applicazioni di supporto a Copilot[74] tramite realtà aumentata e realtà virtuale. Ad esempio, Mootup è una applicazione per la creazione di ambienti tridimensionali per svolgere riunioni. Microsoft sempre tramite l'utilizzo di Azure con la tecnologia di Azure Digital Twins, consente la creazione di rappresentazioni virtuali di oggetti fisici e ambienti reali.

[72] https://learn.microsoft.com/it-it/azure/cloud-adoption-framework/get-started/what-is-azure

[73] Microsoft Azure Handbook - passim

[74] È un chat bot basato su un modello linguistico di grandi dimensioni che integra le funzionalità IA con ad esempio gli strumenti di Microsoft office per aumentarne la produttività e facilità d'uso.

Le applicazioni di questo metaverso sono prettamente industriali, dando la possibilità alle aziende di modellare e simulare sistemi complessi, come fabbriche o città intelligenti, nel mondo digitale. Tramite Azure Digital Twins[75], Microsoft supporta lo sviluppo di metaversi per il settore manifatturiero, nel quale gli utenti possono interagire con versioni digitali di impianti, macchinari o prodotti, migliorando la gestione delle risorse.

E' possibile riassumere la situazione dei concorrenti di NVIDIA tramite la tabella sottostante, che comunque considera solamente i player maggiormente rilevanti dei rispettivi settori:

Settore:	Concorrenza:
Progettazione chip grafici e schede video:	AMD
Metaverso:	Meta, Alphabet, Microsoft
Piattaforme, software e strumenti per sviluppo IA:	HPE, IBM, Intel, AWS, Microsoft, AMD

Table 1: Concorrenti di NVIDIA nei principali settori in cui opera

Le Dimensioni dell'Azienda:

Dal punto di vista della struttura aziendale, NVIDIA impiega 13.800 addetti (2020)[76] a livello globale, dimostrando la sua capacità di gestire operazioni complesse su larga scala. L'azienda adotta una modalità produttiva industriale caratterizzata da processi ripetitivi e standardizzati, focalizzati su un output altamente omogeneo, supportato da un uso

[75] https://azure.microsoft.com/it-it/products/digital-twins
[76] https://it.tradingeconomics.com/nvda:us:employees

intensivo di capitale e tecnologie avanzate. Questa modalità operativa sottolinea la natura capital intensive dell'azienda, la quale richiede ingenti investimenti in infrastrutture tecnologiche per mantenere la sua posizione di leadership nel mercato.

La struttura organizzativa di NVIDIA è altrettanto complessa, come indicato dal modello di business adottato, che le consente di essere altamente flessibile e reattiva rispetto alle sfide del mercato globale. Questo approccio garantisce all'azienda un notevole potere di mercato, supportato da un altrettanto forte potere finanziario e da un'influenza globale significativa, che si estende anche a livello politico. NVIDIA domina il settore delle GPU, un comparto cruciale per numerosi altri settori industriali, poiché i chip prodotti sono essenziali per un'ampia gamma di applicazioni tecnologiche.

La Struttura Organizzativa Aziendale:

NVIDIA è la settima azienda al mondo per dimensioni, con una capitalizzazione di mercato di 1.000 miliardi di dollari (2023)[77]. A causa delle sue grandi dimensioni e della portata delle sue operazioni, l'azienda non ricade esattamente in una singola struttura organizzativa; è evidente che NVIDIA presenti una struttura organizzativa ibrida tra quella funzionale e matriciale. La multinazionale tecnologica divide le sue pratiche commerciali a seconda della funzione come: ingegneria e sviluppo del prodotto, catena di fornitura, operazioni, risorse umane, finanza e contabilità e legale.

La struttura organizzativa funzionale offre una serie di vantaggi a NVIDIA. In particolare, i dipendenti vengono assegnati a gruppi di lavoro sulla base delle loro competenze

[77] "NVIDIA Organizational Structure: functional and hybrid" By John Dudovskiy – Business Research Methodology (17 Giugno 2023)

specifiche. Inoltre, la struttura organizzativa funzionale di NVIDIA viene integrata con linee di reporting chiare, questo migliora sensibilmente la velocità e la qualità del processo decisionale. La struttura di NVIDIA ha però anche alcuni componenti della struttura organizzativa a matrice. In particolare, utilizza gruppi di progetto interfunzionali per affrontare compiti specifici, come lo sviluppo di nuovi prodotti o l'espansione in nuovi mercati. Questi team coinvolgono membri provenienti da diversi dipartimenti, permettendo una collaborazione trasversale e una maggiore agilità operativa e tecnologica. Si formano spesso gruppi temporanei di prodotto o di progetto e i membri del gruppo riferiscono sia al capogruppo sia ai loro diretti superiori all'interno della struttura organizzativa.

Figure 2: Organigramma NVIDIA

Fonte:[78]

In questo contesto, l'azienda si posiziona non solo come un player chiave nel mercato dei semiconduttori, ma anche come un attore globale capace di condizionare l'evoluzione di settori adiacenti, grazie alla sua capacità di innovare e di anticipare le tendenze tecnologiche future.

[78] https://research-methodology.net/NVIDIA-organizational-structure-functional-and-hybrid/

Strategie (di Innovazione) e Politiche:

Come già riscontrato dall'analisi della storia di NVIDIA, si è notata la predisposizione all'innovazione dell'azienda e lo sviluppo di prodotti rivoluzionari come: GPU (1999), CUDA (2006) e Ray Tracing in tempo reale (2018)[79]. Queste tappe fondamentali del percorso di NVIDIA non hanno solamente plasmato la traiettoria dell'azienda, ma hanno anche influenzato in modo significativo il panorama delle unità di elaborazione grafica, dell'IA e del DL.

Diversificazione di Portafoglio e Open Innovation[80]:

Le strategie che NVIDIA ha messo in atto si sono col tempo rilevate di grande successo ed in particolare la diversificazione del proprio portafoglio e l'adozione di tattiche di open innovation, come la creazione di ambienti aperti facendo così in modo di accelerare lo sviluppo di tecnologie ancora in fase embrionale. Le decisioni di avventurandosi nell'IA e nel calcolo ad alte prestazioni con anticipo rispetto al boom globale, e al contempo la promozione di un ecosistema collaborativo per facilitare il progresso del settore ha fatto sì che NVIDIA si stabilisse come leader dell'innovazione del settore.

L'impegno dell'azienda verso l'innovazione, che abbraccia sia l'hardware che il software, fa sì che NVIDIA rimanga un precursore nel dinamico panorama delle tecnologie emergenti. Unendo a ciò un approccio proattivo all'innovazione in grado di reagire o addirittura anticipare l'andamento tecnologico, NVIDIA è riuscita a posizionarsi come un'azienda chiave in grado di plasmare il futuro dell'IA e di altri progressi tecnologici.

---

[79] "CRONOLOGIA DI NVIDIA - Una storia di innovazione" on NVIDIA website – disponibile su: https://www.nvidia.com/it-it/about-nvidia/corporate-timeline/

[80] A Comprehensive Analysis of NVIDIA's Technological Innovations, Market Strategies, and Future Prospects - *John Wang, Jeffrey Hsu, Zhaoqiong Qin* - International Journal of Information Technologies and Systems Approach (Volume 17; Issue 1 2024)

Rimanendo costantemente all'avanguardia della tecnologia, NVIDIA ha mantenuto un vantaggio competitivo e ha ulteriormente dimostrato la capacità di anticipare le tendenze del mercato.

Partnership Strategiche

NVIDIA ha stretto partnership strategiche sia con leader del settore della ricerca ed altrettanto con player di rilievo nello sviluppo di prodotti sia fisici che digitali. Ciò ha ampliato la sua portata sul mercato e facilitando l'integrazione delle sue tecnologie in diverse applicazioni. Ha svolto collaborazioni in settori come: il gaming, i data center e l'automotive. Tali collaborazioni dimostrano l'impegno di NVIDIA nell'allineare le proprie capacità alle esigenze in continua evoluzione dei vari settori.

Di grande importanza strategica è stata l'integrazione delle GPU NVIDIA in piattaforme cloud (con Azure di Microsoft e AWS di Amazon), portando l'azienda ad estendere la propria influenza ad un pubblico più vasto. Impegnandosi attivamente con i leader dell'industria e gli attori chiave di vari settori, NVIDIA ha ampliato la propria portata sul mercato. Inoltre, l'utilizzo del proprio know-how in vari settori ha contribuito in modo significativo a far progredire la tecnologia e a promuovere l'innovazione in aree critiche come l'IA, il cloud computing e i veicoli autonomi.

Cultura aziendale e Gestione delle Risorse Umane:

Sin dal 2000 NVIDIA intraprende i suoi sforzi di scouting e acquisizione di talenti (provenienti da 3dfx Interactive) ai quali offre grandi possibilità di carriera. La scelta si rivela subito fondamentale per consolidare la sua posizione nel settore della grafica.

Questa acquisizione ha portato ingegneri esperti e preziosa proprietà intellettuale, catalizzando lo sviluppo delle successive architetture di GPU.

Nel 2006, NVIDIA ha effettuato un'altra acquisizione strategica portando al suo interno Mellanox (un'azienda di spicco nelle soluzioni di rete ad alte prestazioni). Questa mossa ha migliorato strategicamente l'offerta di NVIDIA per quanto riguarda il settore dei data center, consentendo la fornitura di soluzioni integrate per l'IA e il calcolo ad alte prestazioni.

L'enfasi di NVIDIA nel coltivare i talenti interni è evidente attraverso il suo impegno nello sviluppo della leadership. L'azienda ha assistito all'ascesa di dirigenti chiave, tra cui lo stesso CEO Jensen Huang, attraverso i suoi ranghi. Questa politica aziendale che favorisce la promozione di dipendenti interni riflette l'impegno di NVIDIA nel coltivare la leadership dall'interno. Questo genere di politica è essenziale nella creazione di una cultura aziendale che valorizza e investe nella crescita della propria forza lavoro.

Inoltre, NVIDIA promuove una cultura dell'innovazione e dell'assunzione di rischi, dando ai propri dipendenti la possibilità di esplorare nuove idee. Questa cultura ha portato a progressi rivoluzionari, come gli acceleratori per il DL e le piattaforme di guida autonoma (come NVIDIA DRIVE Hyperion 8). Incoraggiando un ambiente di lavoro dinamico e inventivo, NVIDIA ha costantemente spinto i confini di ciò che è possibile fare nel campo dell'intelligenza artificiale, dell'elaborazione grafica e delle tecnologie emergenti.

La gestione accurata dei fattori interni, le acquisizioni strategiche, la gestione dei talenti e la cultura dell'innovazione, hanno giocato un ruolo fondamentale nel successo e nella leadership di mercato di NVIDIA. La capacità dell'azienda di adattarsi ai cambiamenti

del panorama e il suo impegno nella crescita interna e nelle partnership esterne l'hanno posizionata come un attore chiave nel settore tecnologico in continua evoluzione.

Politiche[81]:

Con l'intento di rimanere un leader innovatore del settore delle GPU e dell'IA Nvidia sta attuando numerose politiche in particolarmente con l'intento di aumentare la penetrazione dell'azienda nei vari settori nei quali opera.

Per quanto riguarda il settore dei data center e degli assistenti IA NVIDIA ha sviluppato AI Enterprise un sistema operativo completo per la scienza dei dati e l'IA, integra una vasta gamma di librerie e strumenti volti a semplificare l'implementazione dell'IA in ambito aziendale. Questo sistema è ottimizzato per l'integrazione nel cloud e consente alle aziende di tutte le dimensioni di accedere a strumenti avanzati di IA senza dover ricorrere a costosi hardware. Nella missione di NVIDIA di democratizzare l'AI NVIDIA AI Enterprise svolge un ruolo fondamentale dando acceso a numerose aziende al mondo dell'IA portando grandi ventate di open-innovation al settore.

NVIDIA AI Workbench[82] aggiunge un ulteriore livello di accessibilità allo sviluppo dell'intelligenza artificiale. Utilizzabile nel cloud, offre strumenti come ChatUSD[83], un assistente di codice USD (Universal Scene Description) che semplifica il processo di generazione di codice Python in risposta alle domande degli utenti. L'integrazione dell'elaborazione del linguaggio naturale (NLP) con gli strumenti di sviluppo dell'IA

[81] Conferenza: "GTC March 2024 Keynote with NVIDIA CEO Jensen Huang" disponibile su: https://www.youtube.com/watch?v=Y2F8yisiS6E

[82] https://www.nvidia.com/it-it/deep-learning-ai/solutions/data-science/workbench/

[83] Nvidia press release: "NVIDIA Omniverse Opens Portals to Vast Worlds of OpenUSD New Omniverse Cloud APIs Help Developers Adopt OpenUSD; Generative AI Model ChatUSD LLM Converses in USD; RunUSD Translates USD to Interactive Graphics, DeepSearch LLM Enables Semantic 3D Search" *by Kasia Johnston* - (8 Agosto 2023)

basati sul cloud rende i complessi processi di IA più accessibili sia agli sviluppatori che alle aziende.

Tramite l'alleanza con altri giganti del settore della grafica digitale come: Pixar, Adobe, Apple e Autodesk, NVIDIA ha co-fondato il progetto Alliance for OpenUSD (AOUSD)[84]. OpenUSD consiste nell'ambizioso progetto di creare uno standard globale per quanto riguarda la grafica digitale tramite il quale e possibile caricare i progetti su Omniverse. Il grande rivoluzione che questo progetto porta è quella di fornire uno spazio di lavoro collaborativo in cloud in grado di supportare progetti di grandissime dimensioni.

Un'altra innovazione legata a questo progetto consiste nella possibilità del Text to 3D in tempo reale che grazie a strumenti NLP e GenIA all'interno di Omniverse rende possibile generare ambienti tridimensionali tramite comandi vocali.

La continua innovazione di NVIDIA nella tecnologia hardware è stata la forza trainante del suo successo. L'ultimo prodotto hardware presentato da Jensen Huang al GTC 2024 (GPU Technology Conference) è l'architettura NVIDIA Blackwell[85]. Il Superchip GB200 Grace Blackwell vanta 40 petaFLOPS di prestazioni AI, una cifra che evidenzia la crescita esponenziale della potenza di calcolo di NVIDIA nel corso degli anni. Negli ultimi otto anni, infatti, NVIDIA ha aumentato la potenza di calcolo di un fattore 1000, a testimonianza della sua incessante ricerca di efficienza e innovazione.

Uno dei componenti di spicco dell'ecosistema hardware di NVIDIA è il chip di commutazione NVLink, che contiene 50 miliardi di transistor[86]. Questa tecnologia consente una comunicazione ultraveloce fra le GPU, trasformando di fatto più GPU in

[84] https://aousd.org/
[85] Conferenza: "GTC March 2024 Keynote with NVIDIA CEO Jensen Huang" disponibile su: https://www.youtube.com/watch?v=Y2F8yisiS6E
[86] https://www.nvidia.com/it-it/data-center/gb200-nvl72/

un'unica unità coesa. Il DGX GB200 NVL72[87], che è composto da questa tecnologia, può funzionare come una "GPU gigante" con un'incredibile prestazione di 720 petaFLOPS di training FP8 e 1,44 hexaFLOPS di prestazioni di inferenza FP4. Nonostante le sue immense capacità di calcolo, questo sistema opera con un costo energetico di 120 kW per rack, dimostrando l'attenzione di NVIDIA per l'efficienza energetica anche ai massimi livelli di potenza di calcolo.

Quando si tratta di addestrare modelli di IA su larga scala, il contrasto fra la vecchia architettura Hopper di NVIDIA e la nuova architettura Blackwell è sorprendente. L'addestramento di un modello GPT da 1,8 trilioni di parametri con Hopper richiederebbe 15 megawatt di potenza, 8.000 GPU e tre mesi. In confronto, l'architettura Blackwell riduce questo requisito a soli 4 megawatt e 2.000 GPU, mantenendo la stessa durata di addestramento. Questa riduzione dei requisiti energetici e hardware rappresenta una svolta, soprattutto per le aziende che vogliono ridurre l'impatto ambientale della formazione dell'intelligenza artificiale su larga scala.[88]

Inoltre, l'architettura Blackwell migliora significativamente le capacità di inferenza dell'IA, fondamentali per le applicazioni in tempo reale. Con un throughput 30 volte superiore per GPU e una maggiore interattività per utente, Blackwell riduce drasticamente il costo dell'inferenza, migliorando al contempo le prestazioni e consentendo una generazione più rapida ed efficiente di token nei modelli linguistici.

[87] NVIDIA GB200 NVL72 connette 36 superchip GB200 Grace Blackwell con 36 CPU Grace e GPU 72 Blackwell in un design rack.

[88] Conferenza: "GTC March 2024 Keynote with NVIDIA CEO Jensen Huang" disponibile su: https://www.youtube.com/watch?v=Y2F8yisiS6E

L'approccio completo di NVIDIA allo sviluppo dell'IA, che comprende piattaforme basate su cloud, soluzioni aziendali e hardware all'avanguardia, sta rivoluzionando il settore. Democratizzando l'accesso all'IA attraverso piattaforme come NVIDIA AI Enterprise e DGX Cloud, le aziende di tutte le dimensioni possono sfruttare potenti strumenti di IA senza significativi investimenti iniziali. Nel frattempo, le innovazioni hardware di NVIDIA, fra cui l'architettura Blackwell e le workstation RTX, stanno superando i limiti delle prestazioni di calcolo, consentendo una formazione e un'inferenza dei modelli di IA più rapide ed efficienti.

Assetto Proprietario:

NVIDIA ha un assetto di public company (equivalente a una SPA italiana quotata in borsa), le sue azioni vengono scambiate sul NASDAQ. La struttura di NVIDIA è molto simile a quella di altre grandi aziende americane e per tanto il capitale di rischio è principalmente in mano grandi istituzioni finanziarie, che ne detengono il 66,88%[89]. Di queste Vanguard Group e BlackRock insieme possiedono all'incirca 400 milioni di azioni, che equivalgono a quasi il 15% dell'intero capitale sociale dell'azienda. Il singolo individuo con il maggior numero di azioni NVIDIA è il suo CEO Jensen Huang, il quale ad ora (24 maggio 2024) possiede 86,827,600 azioni dal valore di 82.8 miliardi di dollari. Altre figure del top management possiedono grandi numeri di azioni come la CFO Colette Kress, con 643,148 azioni, e Jay Puri (executive vice president of Worldwide Field Operations) con 532,401 azioni[90].

[89] https://finance.yahoo.com/quote/NVDA/key-statistics/?guccounter=1
[90] "Who Owns the Most NVIDIA Stock Besides CEO Jensen Huang?" di Ryan Vanzo - The Motley Fool (24 maggio 2024)

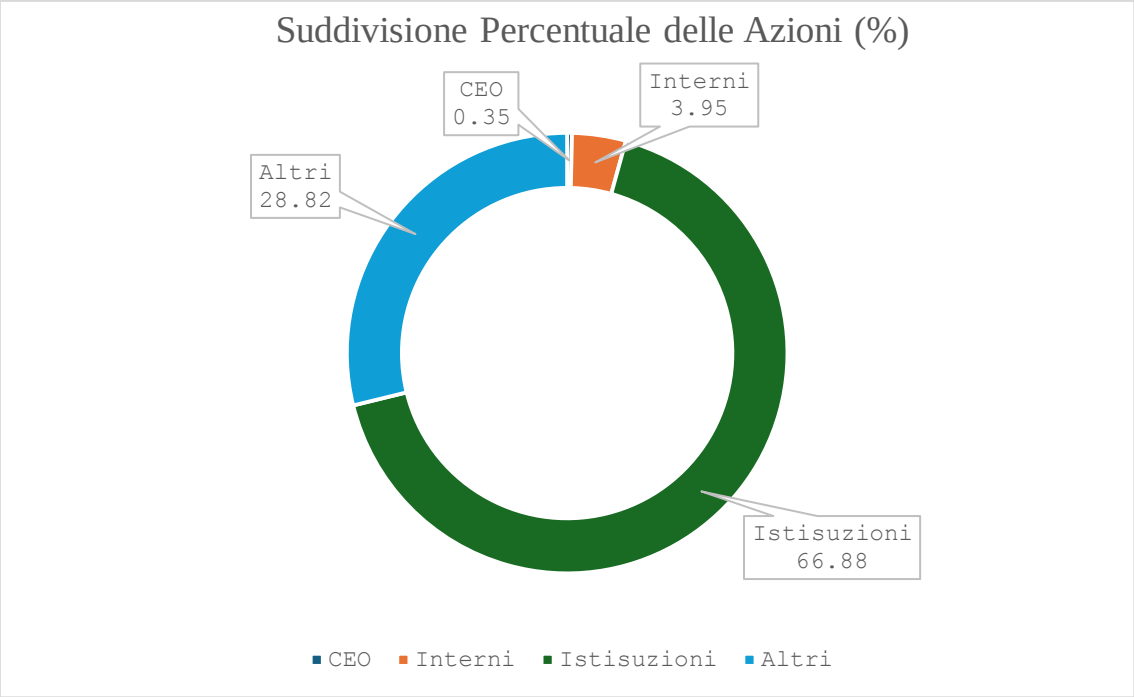


Grafico 6: Suddivisione Percentuale delle Azioni (%)

Fonte:[91]

Prestazioni (Finanziarie e Sociali):

Performance Finanziarie

NVIDIA è leader indiscusso nel mercato dei chip grafici utilizzati per lo sviluppo della GenIA, ha ottenuto significativi vantaggi dalla nuova era della rivoluzione tecnologica. Nel primo trimestre dell'anno fiscale 2024-2025, l'azienda ha registrato ricavi pari a 26,04 miliardi di dollari, superando le aspettative del consensus, che prevedeva circa 24 miliardi di dollari, con una variazione del +/- 2%. Questo risultato riflette una crescita del 9% rispetto al trimestre precedente e un impressionante aumento del 234% su base annua. In termini di utili per azione (EPS), le stime prevedono un valore pari a 5,52 dollari, in

[91] https://finance.yahoo.com/quote/NVDA/key-statistics/?guccounter=1

crescita rispetto ai 4,55 dollari del trimestre precedente, con un aumento anno su anno di circa il 400%.

Sociali[92]:

NVIDIA è attivamente investita nella sostenibilità in molti settori che vanno dall'efficienza energetica dei prodotti, valorizzazione dei dipendenti, gestione dei rifiuti e riduzione dei consumi d'acqua.
Per NVIDIA l'efficienza energetica dei prodotti è uno degli obiettivi principali in ogni fase dei processi di ricerca, sviluppo e progettazione. Per quanto riguarda lo sviluppo di hardware, il miglioramento delle prestazioni e l'efficienza energetica vanno di pari passo, progettando prodotti in grado di ridurre l'intensità delle emissioni.
I modelli di IA stanno costantemente aumentando in complessità e dimensioni, per tanto NVIDIA punta a migliorare l'IA generativa che consentono un'accelerazione delle ricerche scientifiche. I moderni data center data center richiedono piattaforme di calcolo accelerate per eseguire efficacemente questi carichi di lavoro. Le GPU NVIDIA Blackwell (l'ultima architettura annunciata da NVIDIA) sono 20 volte più efficienti dal punto di vista energetico rispetto alle CPU tradizionali per carichi di lavoro di AI e HPC. Se questi carichi di lavoro HPC e AI passassero dall'infrastruttura di CPU alle operazioni accelerate da GPU, si stima che il mondo potrebbe risparmiare quasi 30.000 miliardi di wattora di energia all'anno, equivalenti al fabbisogno di elettricità di quasi 4 milioni di case.

[92] NVIDIA Sustainability Report Fiscal Year 2024 – disponible a: https://images.nvidia.com/aem-dam/Solutions/documents/FY2024-NVIDIA-Corporate-Sustainability-Report.pdf

Per quanto riguarda la valorizzazione dei dipendenti, NVIDIA fornisce la possibilità ai propri dipendenti di presentare, visualizzare e votare i suggerimenti direttamente al CEO. Inoltre, NVIDIA presenta un tasso di turnover dei dipendenti complessivo pari al 2,7%, che è sensibilmente inferiore alla media del settore dei semiconduttori (pari al 17,7%). L'azienda segue una politica aziendale che favorisce la promozione di dipendenti interni riflette l'impegno di NVIDIA nel coltivare la leadership dall'interno. Questa politica è essenziale nella creazione di una cultura aziendale che valorizza e investe nella crescita della propria forza lavoro. Promuove una cultura dell'innovazione e dell'assunzione di rischi, dando ai propri dipendenti la possibilità di esplorare e proporre nuove idee.

Conclusioni:

NVIDIA, leader nell'elaborazione grafica e nell'hardware per l'IA, ha trasformato il panorama tecnologico globale attraverso una serie di progetti e innovazioni all'avanguardia. In questa tesi, tramite l'esplorazione delle radici della storia dell'IA, è stata evidenziata l'importanza di NVIDIA nello sviluppo di questa tecnologia. Nel settore dell'IA, NVIDIA ha sicuramente contribuito allo sviluppo dello stesso, tramite la sua posizione di leader innovatore. Grazie proprio alla estremamente rapida crescita del settore IA, NVIDIA è diventata la società con il valore di più alto al mondo (18/06/2024) con una capitalizzazione di mercato di 3.350.000.000.000$. Nonostante l'incredibile ascesa nell'ultimo decennio dell'azienda amministrata da Jensen Huang, il futuro riserva sicuramente nuove ed importanti sfide, sia dovute all'evoluzione dei mercati sia al mantenimento della sua posizione dominante strettamente legata all'innovazione. Nei prossimi anni il settore dell'IA, al momento in fase di sviluppo, andrà inevitabilmente incontro a un rallentamento del ritmo di crescita portando a un consolidamento di pochi grandi player e dunque un aumento delle forze competitive all'interno del mercato. Da non sottovalutare sono inoltre i cambiamenti a livello legislativo del settore dell'IA, che già ora sta sollevando questioni riguardo l'utilizzo morale di questi strumenti e di diritto d'autore sui dati utilizzati per l'allenamento dei modelli.

Bibliografia:

“A commentary of GPT-3” di Min Zhang e Juntao Li - MIT Technology Review (2021)

“A Comprehensive Analysis of NVIDIA’s Technological Innovations, Market Strategies, and Future Prospects” di John Wang, Jeffrey Hsu, Zhaoqiong Qin - International Journal of Information Technologies and Systems Approach (Volume 17; Issue 1 2024)

“A Training Algorithm for Optimal Margin Classifiers.” Di Bernhard E. Boser, Isabelle M. Guyon, Vladimir Vapnik. - Fifth Annual Workshop on Computational Learning Theory. ACM Press (Pittsburgh 1992)

“Che cos’è l’intelligenza artificiale?” - Parlamento Europeo (03-09-2020)

“Computing machinery and intelligence” di Turing, A.M. - Mind, 59, 433-460 (1950)

“CRONOLOGIA DI NVIDIA - Una storia di innovazione” on NVIDIA website

“From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference” di S. S. Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, Vijay Gadepally- 2023 IEEE HPEC (2023)

“INTRODUZIONE ALLE RETI NEURALI ARTIFICIALI” - di Marco Gori (MONDO DIGITALE • n.4 - dicembre 2003)

“Learning internal representations by error propagation, Parallel Distributed Processing: Explorations in the Microstructures of Cognition” (Vol. I, pp. 318-362) di D. E. Rumelhart, G. E. Hinton, and R. J. Williams, - by D. E. Rumelhart and J. L. McClelland - (Eds.) Cambridge, MA: MIT Press, (1986)

“Machine Learning and Other Artificial Intelligence Applications, An Issue of Neuroimaging Clinics of North America” - Cap. “Brief history of artificial Intelligence” di Suresh K. Mukherji – (2020)

“Manuale sulle Reti Neurali” – di: Floreano Dario e Mattiussi Claudio (Il mulino 2002)

“Manuale sulle Reti Neurali” – di: Floreano Dario e Mattiussi Claudio (Il mulino 2002) – passim

“NVIDIA DRIVE Hyperion Combines Orin With Best-in-Class Sensor Architecture for Production-Ready Platform” - by Gary Hicok – Nvidia website (9 Novembre 2021)

“NVIDIA Omniverse Opens Portals to Vast Worlds of OpenUSD - New Omniverse Cloud APIs Help Developers Adopt OpenUSD; Generative AI Model ChatUSD LLM Converses in USD; RunUSD Translates USD to Interactive Graphics, DeepSearch LLM Enables Semantic 3D Search” by Kasia Johnston - Nvidia press release (8 Agosto 2023)

“NVIDIA Organizational Structure: functional and hybrid” By John Dudovskiy – Business Research Methodology (17 Giugno 2023)

“The 2018 NVIDIA AI City Challenge” di Milind Naphade, Ming-Ching Chang, Anuj Sharma, David C. Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming-Yu Liu, Rama Chellappa, Jenq-Neng Hwang, Siwei Lyu

“The History and Future of Internet Traffic" di Arielle Sumits, Cisco Blogs, (August 28, 2015)

“Towards Digital Twins for NVIDIA's Earth-2 Initiative: Pushing the Limits of Deep Auto-regressive Fourier Neural Operator and Transformer models for Earth System Emulation” di Kashinath Karthik, Pritchard Mike S., Anandkumar Anima, Pathak Jaideep, Mardani Morteza, Kurth Thorsten, Hall David, Messmer Peter, Posey Stan, Subramanian Shashank, Harrington Peter -

“Who Owns the Most NVIDIA Stock Besides CEO Jensen Huang?” di Ryan Vanzo - The Motley Fool (24 maggio 2024)

AGU Fall Meeting 2022, held in Chicago, IL, 12-16 December 2022 (December 2022)

Conferenza: “GTC March 2024 Keynote with NVIDIA CEO Jensen Huang”

Microsoft Azure Handbook – passim

NVIDIA Investor Presentation - October 2023

NVIDIA Investor Presentation - October 2023

NVIDIA Sustainability Report Fiscal Year 2024

Sitografia:

https://aousd.org/

https://aws.amazon.com/it/

https://azure.microsoft.com/it-it

https://azure.microsoft.com/it-it/products/digital-twins

https://corporatefinanceinstitute.com/resources/equities/sp-500-index/

https://developer.nvidia.com/cuda-toolkit

https://developer.nvidia.com/maxwell-compute-architecture

https://ec.europa.eu/commission/presscorner/detail/it/IP_18_4521

https://finance.yahoo.com/quote/NVDA/key-statistics/?guccounter=1

https://it.tradingeconomics.com/nvda:us:employees

https://learn.microsoft.com/it-it/azure/cloud-adoption-framework/get-started/what-is-azure

https://medium.com/@dlaytonj2/chatgpt-how-we-got-to-where-we-are-today-a-timeline-of-gpt-development-f7a35dcc660e

https://openai.com/about/

https://openai.com/index/openai-gym-beta/

https://openai.com/index/universe/

https://research-methodology.net/NVIDIA-organizational-structure-functional-and-hybrid/

https://segaretro.org/NV1

https://tg24.sky.it/tecnologia/2023/12/04/chat-gpt-utenti-storia#05

https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

https://www.agendadigitale.eu/mercati-digitali/corsa-ai-chip-ai-tutti-contro-NVIDIA-chi-dominera-il-mercato/

https://www.computerbase.de/2023-09/graphics-card-ranking-gpu-comparison-geforce-rtx-radeon-arc/

https://www.europarl.europa.eu/topics/it/article/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata#:~:text=L'intelligenza%20artificiale%20(IA),la%20pianificazione%20e%20la%20creativit%C3%A0

https://www.historyofinformation.com/detail.php?id=4989

https://www.ibm.com/history/advancing-humanity#Artificial+intelligence

https://www.jonpeddie.com/news/shipments-of-graphics-add-in-boards-decline-in-q1-of-24-as-the-market-experiences-a-return-to-seasonality/

https://www.justice.gov/atr/herfindahl-hirschman-index#:~:text=The%20agencies%20generally%20consider%20markets,Guidelines%20%C2%A7%202.1%20(2023)

https://www.nvidia.com/it-it/about-nvidia/corporate-timeline/

https://www.nvidia.com/it-it/data-center/gb200-nvl72/

https://www.nvidia.com/it-it/data-center/pascal-gpu-architecture/

https://www.nvidia.com/it-it/deep-learning-ai/solutions/data-science/workbench/

https://www.nvidia.com/it-it/software/

https://www.nvidia.com/it-it/technologies/

https://www.repubblica.it/tecnologia/2023/01/15/news/openai_chatgpt_storia_musk_altman-383409890/

https://www.statista.com/statistics/1166028/gpu-market-size-worldwide/#:~:text=In%202024%2C%20the%20global%20graphics,percent%20from%202024%20to%202029

https://www.techradar.com/news/computing-components/graphics-cards/amd-vs-NVIDIA-who-makes-the-best-graphics-cards-699480

https://www.tomshardware.com/picturestory/715-history-of-nvidia-gpus.html