

Il Gtc è l'evento che ogni anno fa il punto sull'evoluzione delle tecnologie che usano le Gpu per compiti diversi dalla grafica visuale, come accelerare il calcolo scientifico riducendo i tempi di risoluzione di algoritmi complessi

La rivoluzione scientifica spinta dalle Gpu

Il Gtc (Gpu Technology Conference) è l'evento con il quale da qualche anno Nvidia raccoglie, nella Silicon Valley, addetti della stampa specializzata, analisti, aziende, ricercatori, sviluppatori e tutti gli appassionati dei settori direttamente o indirettamente influenzati dalle tecnologie legate alle Gpu, ovvero ai processori grafici, ormai utilizzati in campi che esulano dalla grafica vera e propria. Sono state più di 3.000 le persone che si sono registrate e che nei giorni dell'evento hanno affollato le stanze del San Jose Convention Center per seguire i moltissimi seminari in calendario. Per l'Italia hanno partecipato circa 30 ricercatori provenienti da diverse Università (nazionali, ma non solo) con più di 20 progetti di ricerca. L'evento più importante della manifestazione è, ovviamente, il keynote di apertura tenuto da Jen-Hsun Huang, Ceo e cofondatore di Nvidia, ma soprattutto una delle personalità visionarie della Silicon Valley che a distanza di anni dal suo debutto ha mantenuto inalterato il coraggio di sfidare il mercato con le proprie idee. Cominciamo questo reportage del Gtc 2013 proprio dal keynote che Jen-Hsun Huang ha articolato in cinque punti: l'incipit è stato un ritorno alle origini per mostrare qual è stato l'impatto

dei processori grafici non solo nel mondo della grafica e dei videogiochi, ma in tutto l'universo informatico. Jen-Hsun ha quindi proiettato la platea nei progetti futuri di Nvidia, per tornare poi al presente.

Le rivelazioni sul presente e sul lancio di un prodotto sono uno dei momenti più interessanti perché permettono di capire a che punto della propria strategia si trova l'azienda.

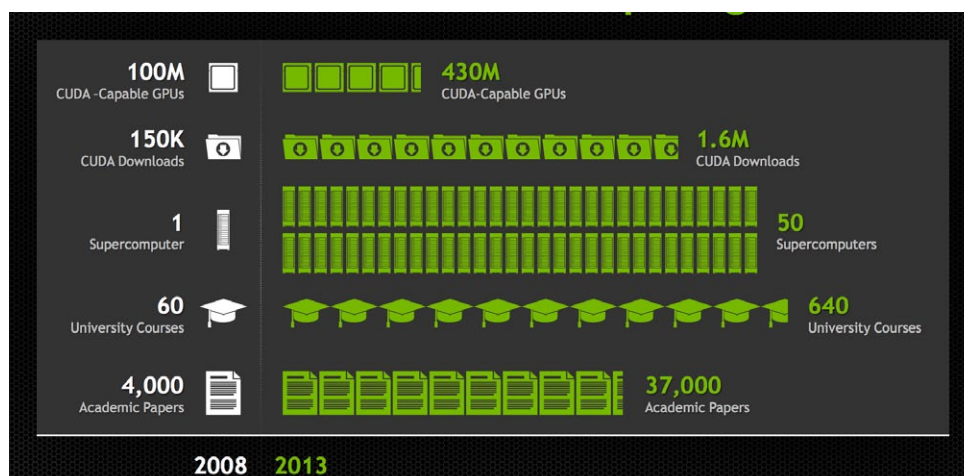
Jen-Hsun è partito dall'obiettivo principale che ha guidato lo sviluppo delle Gpu, cioè raggiungere l'eccellenza nell'elaborazione dei pixel che compongono le

immagini, per allargare l'orizzonte verso i traguardi futuri perché la maggior parte dei dispositivi con i quali interagiamo utilizzano un display e un'interfaccia visuale. L'industria cinematografica e quella dei videogiochi hanno sfruttato la crescente potenza di calcolo delle Gpu nel tentativo di ricreare volti, personaggi e ambienti sintetici, ma non distinguibili da quelli reali.

Un obiettivo ambizioso e molto complesso da raggiungere, ma sempre più vicino proprio grazie alla potenza di calcolo che ogni generazione di Gpu riesce a fornire. A trainare lo sviluppo delle

tecnologie e architetture legate alle Gpu non c'è solo il mondo della grafica, ma anche quello più ampio del calcolo scientifico perché questi processori sono estremamente efficienti e adatti a svolgere calcoli di tipo parallelo. L'impatto delle Gpu sulle strutture di calcolo e supercalcolo è evidente così come Jen-Hsun Huang sottolinea con un grafico che riporta l'incremento di prestazioni. La strategia di Nvidia sarà di continuare a sviluppare Gpu che possano trovare il loro spazio sia nel settore professionale sia in quello consumer, perché solo in questo modo è possibile rientrare degli

I NUMERI DELLA DIFFUSIONE DI CUDA



Nei cinque anni trascorsi dalla sua prima apparizione, era il 2008, la tecnologia Cuda si è diffusa non solo in ambito professionale, ma anche consumer fino a superare i 430 milioni di Gpu vendute e capaci di sfruttare la tecnologia Nvidia.

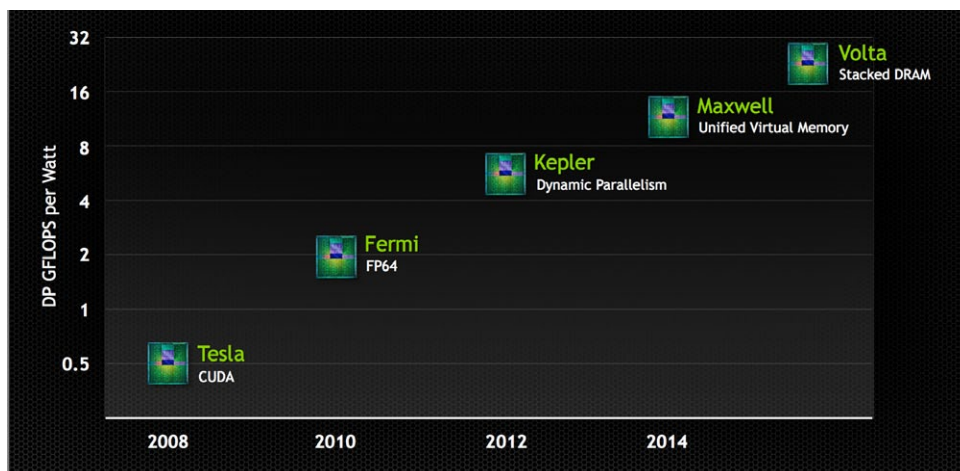
altissimi investimenti in ricerca e sviluppo: ogni generazione di Gpu richiede più di 1 miliardo di dollari prima di arrivare al processo di produzione e commercializzazione.

Il futuro delle Gpu

Dopo l'architettura Kepler, che tra poche settimane raggiungerà il primo anno di età, Nvidia rilascerà l'architettura Maxwell. La caratteristica principale di Maxwell sarà il supporto alla tecnologia *Unified Virtual Memory* che permetterà di indirizzare in modo omogeneo la memoria direttamente connessa al processore grafico insieme a quella di sistema.

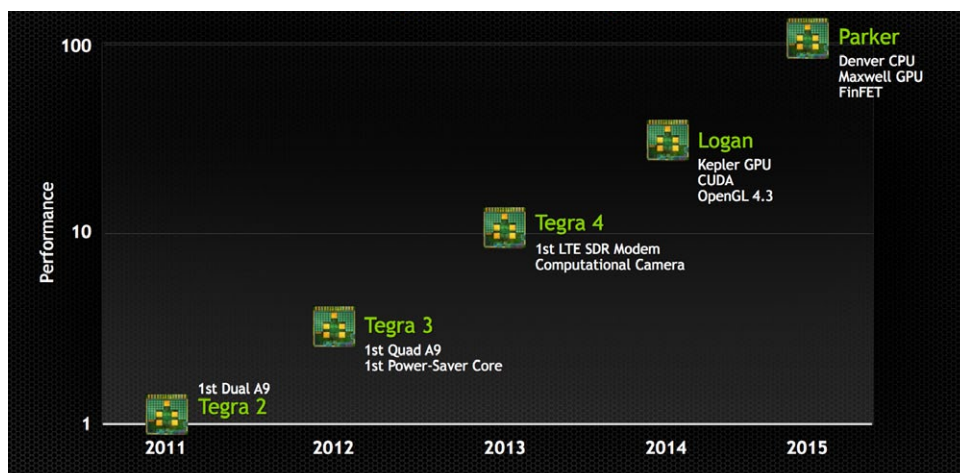
I potenziali vantaggi offerti da questa tecnologia, che peraltro è tutt'altro che nuova, sono molteplici e favoriscono in modo particolare gli sviluppatori. Questo perché al momento per utilizzare la Gpu come acceleratore di calcolo è necessario esplicitare nel codice Cuda le operazioni di copia dei dati da elaborare, sia per quanto riguarda l'operazione verso la memoria grafica sia per quanto riguarda quella inversa. Permettendo l'allocazione omogenea di tutta la memoria presente nel sistema o nel nodo di calcolo, chi sviluppa il

LA ROADMAP DELLE GPU



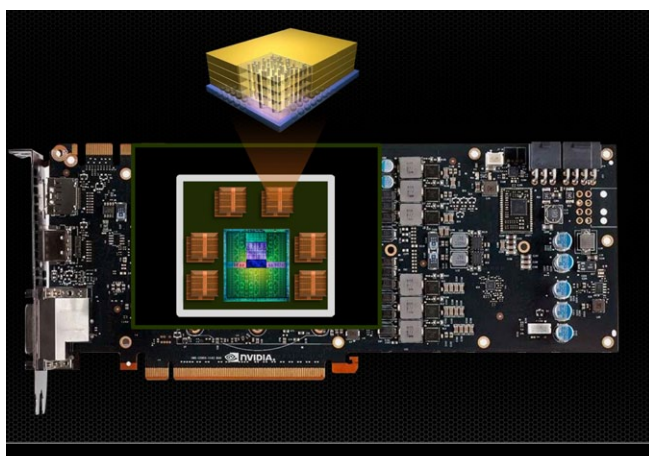
Il prossimo anno Nvidia lancerà l'architettura Maxwell che introdurrà la tecnologia Unified Virtual Memory.

LA ROADMAP DI TEGRA



All'inizio del 2014 arriverà Tegra 5 (nome in codice Logan) che sarà il primo processore mobile con supporto Cuda.

NVIDIA VOLTA E LA MEMORIA IMPILATA



Entro il 2017 Nvidia prevede d'impacchettare nella Gpu memoria di tipo stacked, cioè a celle sovrapposte. Questo garantirà trasferimenti fino a 1 Tbyte/s.

software Cuda dovrebbe poter saltare questi passaggi.

Jen-Hsun si è però spinto oltre, presentando quella che sarà l'architettura successiva a Maxwell: Volta, questo è al momento il nome in codice, non arriverà prima del 2016, ma potrebbe rappresentare il prossimo salto di qualità nel campo delle Gpu; la tecnologia che Nvidia ha in previsione di adottare è denominata *Stacked Dram*. Il processore Volta integrerà all'interno del suo package non solo la Gpu vera e propria, ma anche i chip di memoria. Lo spostamento dei chip di memoria dal Pcb

all'interno del processore grafico porterà vantaggi importanti per quanto riguarda i tempi di latenza e di banda di trasmissione dati; per quest'ultima si arriverà, infatti, a toccare la soglia di 1 Tbyte al secondo, contro un valore medio odierno di 300 Gbyte al secondo sul prodotto di fascia più alta.

Maxwell e Volta saranno le Gpu a guidare lo sviluppo nelle soluzioni a elevata potenza di calcolo, ma Jen-Hsun Huang, tornando a sottolineare come ogni dispositivo con un display ha bisogno di un

processore capace di elaborare immagini. Il progetto della casa californiana è quindi di riuscire a realizzare un processore adatto a ogni tipo di dispositivo, qualunque esso sia. Nvidia ha cominciato a percorrere questa strada con il progetto Tegra che nel corso degli anni è maturato in Tegra 2 e poi Tegra 3 (il primo quad core della famiglia) e che oggi è giunto alla versione 4.

La prossima evoluzione di Tegra, in arrivo all'inizio del prossimo 2014, sarà Logan e sarà il primo processore di classe mobile a integrare un'architettura Gpu derivata dal progetto Kepler e a supportare Cuda 5 e OpenGL 4.3. La roadmap Nvidia si spinge però oltre e mostra anche il progetto Parker che sarà realizzato utilizzando una componente Cpu di tipo Denver, con un'architettura Arm a 64 bit, e una Gpu derivata da Maxwell.

L'obiettivo di Jen-Hsun Huang è di offrire con Parker una potenza di calcolo 100 volte superiore a quella di cui era capace il Tegra 2 introdotto nel 2011.

Nvidia Grid Vca

Il momento più atteso del keynote di apertura del Gtc è quello conclusivo: Jen-Hsun Huang si è soffermato sul computing remoto e di come trasferire l'elaborazione dalle postazioni personali fisse a workstation delocalizzate per migliorare l'efficienza del flusso di lavoro e l'accessibilità ad applicazioni professionali attraverso terminali con ridotta potenza di calcolo. Grid Vca (*Visual Computing Appliance*), disponibile in due varianti, è il prodotto annunciato durante il Gtc 2013 e che va a completare la famiglia delle soluzioni Grid finora composta da quelle dedicate al gaming e al settore enterprise.

Grid Vca è un rack in formato 4U all'interno del quale possono trovare posto un massimo



Le unità Nvidia Grid Vca possono essere acquistate solo dai Var certificati che gestiscono la vendita e il supporto tecnico di applicazioni professionali.



Il rack 4U del Grid Vca può contenere un massimo di due Cpu Intel Xeon a otto core e fino a di otto acceleratori Nvidia Grid per gestire 16 utenti concorrenti.

di due processori Intel Xeon con architettura a 8 core e un massimo di otto acceleratori della famiglia Nvidia Grid, ciascuno dei quali è dotato di due Gpu con architettura Kepler. La gestione del Grid Vca è affidata a un componente software di virtualizzazione (Vgx) sviluppata da Nvidia; è la prima volta che la casa californiana chiede il pagamento di una licenza di utilizzo per il proprio software: si tratta di circa 300 dollari per ciascuna Gpu presente all'interno del rack Grid Vca.

La soluzione Nvidia permette di creare ambienti virtualizzati Microsoft Windows 7 su all'interno dei quali possono essere installate applicazioni professionali. La prima versione del software Vgx installato sulle unità Grid Vca permette di costruire una relazione uno a uno tra Gpu e macchina virtuale, per un massimo di 16 utenti concorrenti e altrettante applicazioni virtualizzate.

Il Grid Vca mostra all'utente che si collega solo l'applicazione e i progetti ad essa collegati e non l'intera macchina virtuale. Durante gli incontri con Greg Estes (Industry Executives Media & Entertainment), Will Wade (Director, Product Management Professional Solutions Business) e Michael DeNeffe (Director of Marketing Grid Gaming, Grid Vca, Grid Enterprise) abbiamo approfondito quale sia il target della soluzione Grid Vca e quali i suoi futuri sviluppi. Grid Vca è indirizzato ai piccole realtà lavorative – 15 o 20 persone – che non dispongono di un reparto IT dedicato alla gestione e manutenzione dell'infrastruttura informatica. In questi ambienti Nvidia ritiene che gli appliance Grid Vca possano fornire grandi vantaggi in termini di costo, ma soprattutto di efficienza nel flusso di lavoro.

Il Grid Vca è infatti un dispositivo pronto all'uso dopo che sono state installate le

applicazioni. Attraverso uno qualunque dei dispositivi per cui sia già disponibile il client Vgx di accesso (workstation, desktop, notebook, tablet) è possibile lavorare con le applicazioni installate sul Grid Vca; tutto questo senza che sia necessario configurare le singole macchine (pensiamo ad esempio a workstation portatili di collaboratori esterni o a contratto) e senza che i dati sensibili lascino l'ambiente di lavoro in cui sono installati il Grid Vca e l'archivio dati.

La prossima versione del software Vgx permetterà di assegnare anche più di una Gpu virtuale a una singola applicazione così da avere accesso in modo dinamico e a specifiche esigenze di tutta la potenza di calcolo disponibile. In futuro Nvidia prevede inoltre di permettere la gestione di un numero di utenti concorrenti superiore al numero delle Gpu presenti nel sistema attraverso una gestione dinamica delle risorse. I sistemi Grid Vca saranno distribuiti solo attraverso il canale dei Var (*value-add reseller*) che vendono le applicazioni supportate dall'appliance. In questo modo Nvidia è in grado di fornire garanzie sulla professionalità del punto vendita e assicurare il supporto tecnico necessario.

Gpu Conference Technology

Il Gtc è molto di più del keynote di apertura e questa edizione si è articolata in quattro giorni densi di seminari: da quelli propedeutici per a chi si avvicina al mondo dell'accelerazione Gpu, fino a quelli più tecnici dove gli sviluppatori hanno potuto confrontarsi in modo diretto con chi ogni giorno lavora nei laboratori di Nvidia e con chi utilizza queste soluzioni per risolvere problemi nel campo della ricerca scientifica. Erano tanti i progetti universitari in mostra (trovate tutti i poster all'indirizzo www.gputechconf.com/page/posters.html).