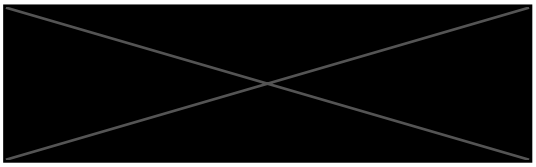
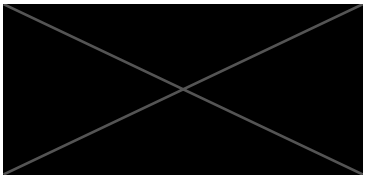


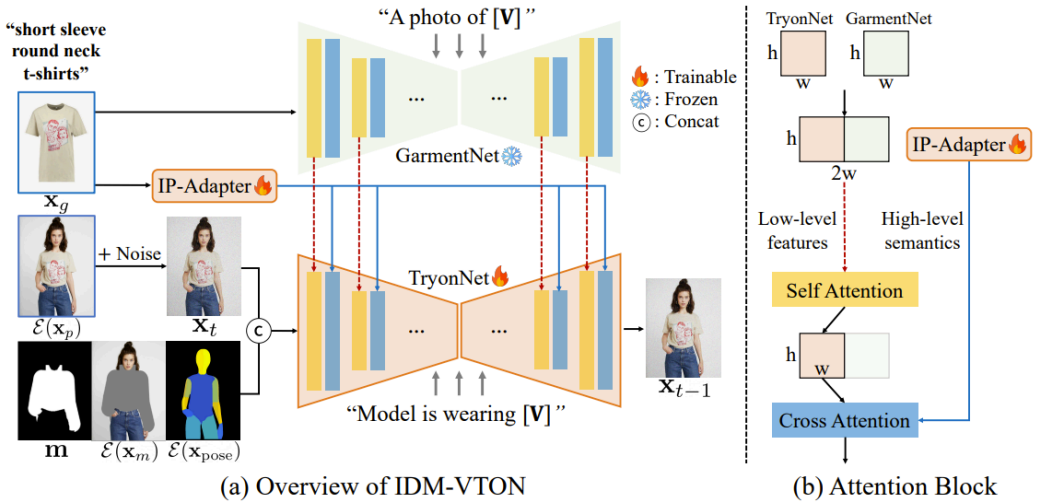


Report - Studio costi per singolo try-on

Introduzione:

L'obiettivo di questo studio è stimare il costo per singolo try-on utilizzando il modello di intelligenza artificiale "IDM-VTON: Improving Diffusion Models for Authentic Virtual Try-on in the Wild". Il try-on è definito come l'inferenza del modello AI che, data un'immagine di una persona e un'immagine di un capo di abbigliamento, genera un'immagine della persona che indossa quel capo.

Descrizione del Processo di Try-On Virtuale



(a) Overview of IDM-VTON (b) Attention Block

Introduzione all'Architettura IDM VTON

L'architettura IDM VTON (Image-based Virtual Try-On Network) è progettata per permettere agli utenti di provare virtualmente abiti su immagini di persone. Questa tecnologia utilizza avanzati algoritmi di visione artificiale e reti neurali convoluzionali (CNN) per adattare in modo realistico i capi d'abbigliamento all'immagine del corpo dell'utente. IDM VTON si basa su un processo di inpainting basato su esempio, che mira a riempire l'immagine mascherata con un'immagine di riferimento.

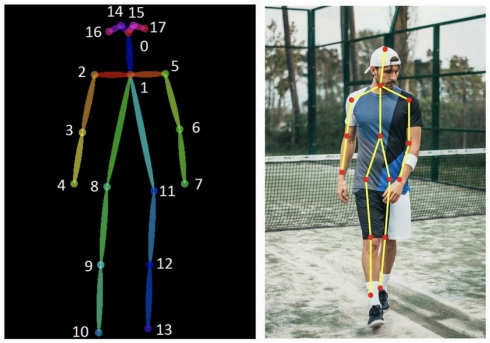
Componenti dell'Architettura

IDM VTON richiede diversi input:

- Immagine della persona
- Immagine del capo d'abbigliamento
- Maschera del capo dall'immagine della persona
- Embedding DensePose dell'immagine della persona

Per ottenere questi input, è necessario utilizzare diversi modelli di preprocessing:

OpenPose

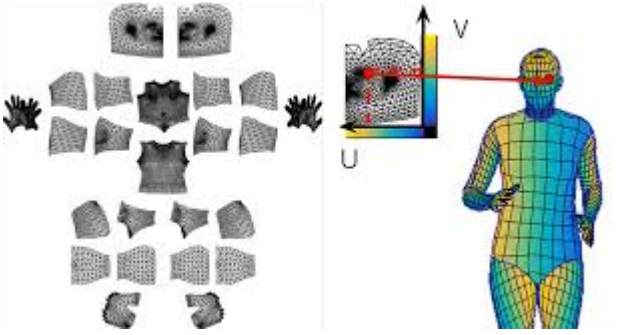


- **Funzione**: Estrarre i punti chiave della posa della persona, identificando le coordinate delle principali articolazioni del corpo.
- **Descrizione**: Questo modello rileva i keypoints che descrivono la postura e la posizione del corpo dell'utente. Questi punti sono fondamentali per comprendere come il corpo è orientato nello spazio.

HumanParsing

- **Funzione**: Calcolare la maschera del capo d'abbigliamento, utilizzando le informazioni sui keypoints estratti da OpenPose.
- **Descrizione**: Questo modello segmenta l'immagine della persona per identificare l'area specifica dove il capo d'abbigliamento sarà applicato. La maschera risultante permette di delimitare con precisione l'area da modificare.

DensePose



- **Funzione**: Mappare i pixel dell'immagine della persona su una superficie 3D del corpo umano, fornendo una rappresentazione dettagliata della geometria del corpo.
- **Descrizione**: Questo modello fornisce un embedding dettagliato che rappresenta la forma tridimensionale del corpo della persona. Questo embedding è cruciale per adattare correttamente il capo d'abbigliamento all'anatomia dell'utente.

Oltre ai dati di input ottenuti dai modelli di preprocessing, IDM VTON richiede una descrizione del capo d'abbigliamento e una del modello. Nei nostri test, queste descrizioni sono state lasciate in bianco.

Generazione dell'immagine
Una volta processati i dati di input con i vari modelli di pre-processing, si può passare alla parte centrale dell'intera soluzione, ovvero il modello che genera l'immagine finale del modello che indossa il capo selezionato. Questa fase è nota come "image_generation" ed utilizza un approccio di generazione autoregressiva. In questo processo, l'immagine generata viene riutilizzata come input per un certo numero di iterazioni, comunemente chiamate "step". Nei nostri test, abbiamo utilizzato 30 step, come indicato nel valore di default nella documentazione.

L'architettura di IDM VTON rappresenta una soluzione avanzata nel campo del virtual try-on, permettendo agli utenti di visualizzare in modo realistico come appariranno indossando diversi capi d'abbigliamento.

Piattaforme Cloud per AI:

L'IA richiede enormi risorse computazionali, soprattutto per l'addestramento di modelli complessi come le reti neurali profonde. Le schede video, o GPU (Graphics Processing Unit), giocano un ruolo cruciale in questo contesto. Grazie alla loro architettura parallela, le GPU sono particolarmente adatte per eseguire i calcoli massicci richiesti dall'IA.

Le GPU sono costruite con migliaia di core che possono eseguire operazioni in parallelo, rendendole ideali per l'elaborazione di grandi quantità di dati simultaneamente. Questo le rende essenziali per l'addestramento di modelli di deep learning, che necessitano di elevate capacità di calcolo per l'elaborazione dei dati e l'ottimizzazione dei pesi del modello.

Per soddisfare questa esigenza, sono nati i servizi cloud di GPU on demand, dove gli utenti possono affittare GPU per il tempo necessario, pagando solo per l'uso effettivo. Questi servizi permettono agli sviluppatori e ai ricercatori di accedere a potenti risorse computazionali senza dover investire in costose infrastrutture hardware.

Funzionamento dei Servizi Cloud di GPU on Demand

I servizi cloud di GPU on demand funzionano attraverso piattaforme che offrono l'accesso a macchine virtuali dotate di GPU potenti. Gli utenti possono configurare queste macchine con le specifiche desiderate e utilizzarle per il tempo necessario, pagando un costo orario o su

base mensile. Questo modello offre flessibilità e scalabilità, permettendo di gestire facilmente carichi di lavoro variabili.

Di seguito, confronteremo tre delle principali piattaforme che offrono GPU as a Service: Paperspace, RunPod e Lambda Labs.

GPU Popolari e Performanti

Le GPU più popolari e performanti nel campo dell'intelligenza artificiale includono:

- **NVIDIA 4090**: Una delle GPU più potenti sul mercato, particolarmente apprezzata per le sue eccellenti prestazioni nei compiti di AI e rendering. La 4090 offre un ottimo rapporto costo-prestazioni, rendendola una scelta ideale per applicazioni che richiedono alta capacità di elaborazione e scalabilità.
- **NVIDIA A100**: La GPU con architettura Ampere, offre prestazioni straordinarie per applicazioni AI e HPC.
- **NVIDIA H100**: La GPU di ultima generazione con architettura Hopper, che rappresenta il massimo in termini di prestazioni per applicazioni di deep learning e calcolo ad alte prestazioni, offrendo innovazioni significative rispetto alle generazioni precedenti.

Confronto tra Paperspace, RunPod e Lambda Labs

Paperspace

Paperspace è una piattaforma cloud che offre potenti GPU per applicazioni di intelligenza artificiale, machine learning e altre attività ad alta intensità computazionale. Fornisce ambienti preconfigurati e personalizzabili attraverso Gradient, il loro ecosistema di sviluppo ML.

GPU offerte e prezzi

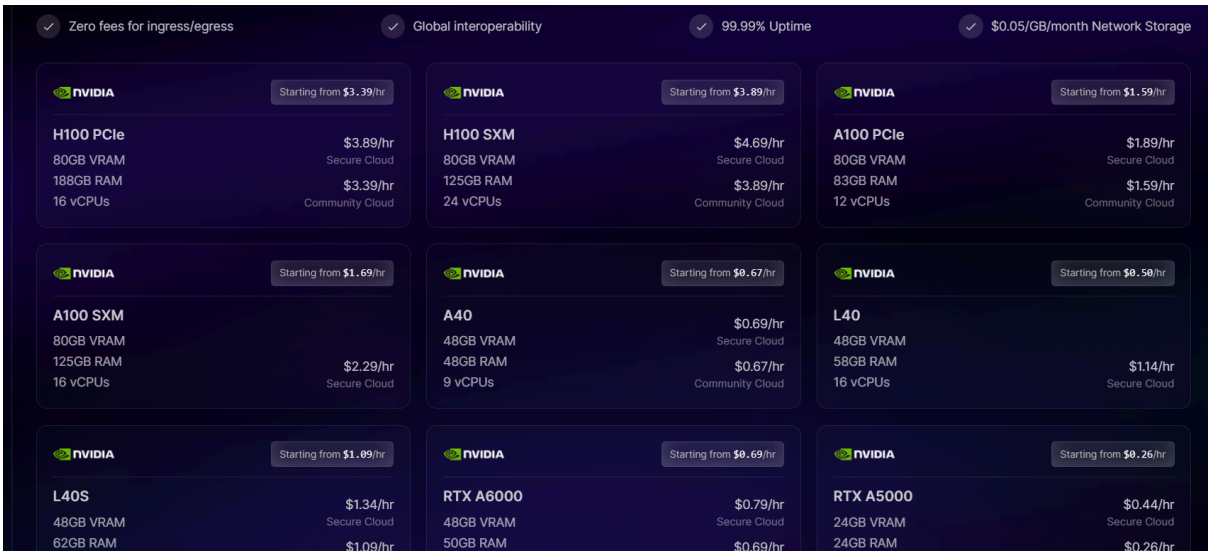
Pros

- Ampia gamma di GPU tra cui scegliere.
- Gradient offre strumenti per tutto il ciclo di vita del machine learning.
- Prezzi competitivi e flessibili.
- Interfaccia user-friendly e supporto per Jupyter Notebooks.

RunPod

Descrizione: RunPod è una piattaforma che si concentra sulla fornitura di GPU on demand per sviluppatori e data scientists. Offre un'interfaccia semplice e accesso a una varietà di GPU con un modello di prezzo pay-as-you-go.

GPU Offerte e pricing:



Pro:

- Prezzi competitivi, specialmente per GPU di fascia media.
- Facile da usare con una semplice interfaccia web.
- Buona selezione di GPU adatte a vari carichi di lavoro.

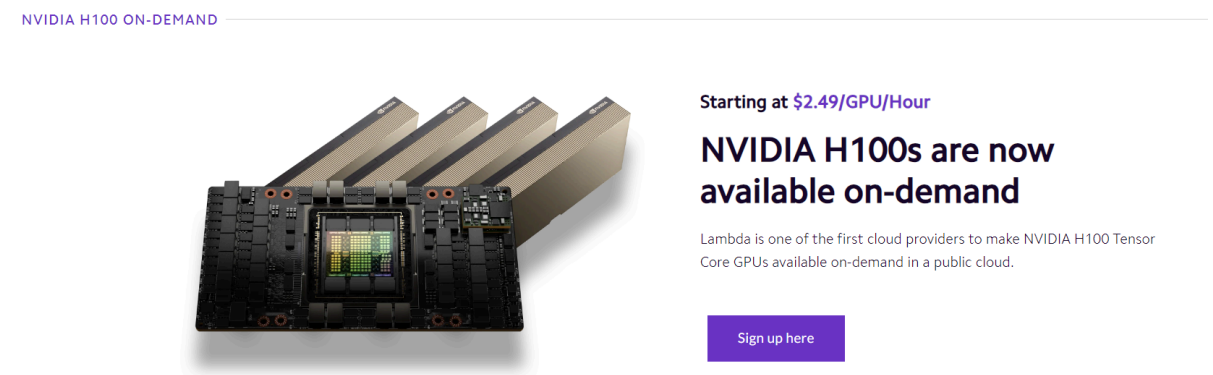
Contro:

- Meno opzioni di GPU rispetto ad altre piattaforme.
- Mancanza di strumenti integrati per lo sviluppo ML.

Lambda Labs

Descrizione: Lambda Labs offre istanze GPU per il deep learning con un focus particolare su hardware potente e configurazioni ottimizzate per il machine learning. È conosciuta per la sua specializzazione nel supporto di progetti IA di alto livello.

GPU Offerte e pricing:



Pro:

- Specializzazione nel deep learning e configurazioni ottimizzate.
- Prezzi competitivi per GPU di fascia media e alta.
- Supporto tecnico eccellente.

Contro:

- Interfaccia meno intuitiva rispetto a Paperspace.
- Meno opzioni di personalizzazione dell'ambiente di sviluppo.

Dopo aver esaminato le caratteristiche, i prezzi e le GPU offerte da Paperspace, RunPod e Lambda Labs, scegliamo **RunPod** per i costi competitivi e la versatilità del servizio. RunPod offre una vasta gamma di GPU, strumenti integrati per lo sviluppo di machine learning con Gradient, e una piattaforma facile da usare, rendendola la scelta ideale per sviluppatori e ricercatori in cerca di una soluzione completa per l'intelligenza artificiale.

Metodologia:

Per stimare il costo per singolo try-on e determinare la convenienza delle diverse GPU, abbiamo seguito questi passaggi:

1. **Raccolta Dati**: Raccolta dei tempi di esecuzione per ogni fase del processo di try-on utilizzando tre diverse GPU: NVIDIA A100, H100 PCIe, e RTX 4090.
2. **Calcolo dei Tempi Medi**: Calcolo dei tempi medi di esecuzione per ciascuna fase del processo.
3. **Analisi dei Costi**: Calcolo del costo per singolo try-on basato sui tassi orari di noleggio delle GPU.
4. **Analisi della Capacità**: Calcolo del numero di richieste per ora che ogni GPU può gestire.
5. **Analisi della Scalabilità**: Valutazione della scalabilità delle GPU in base al numero di richieste per ora con diverse quantità di GPU.

Risultati:

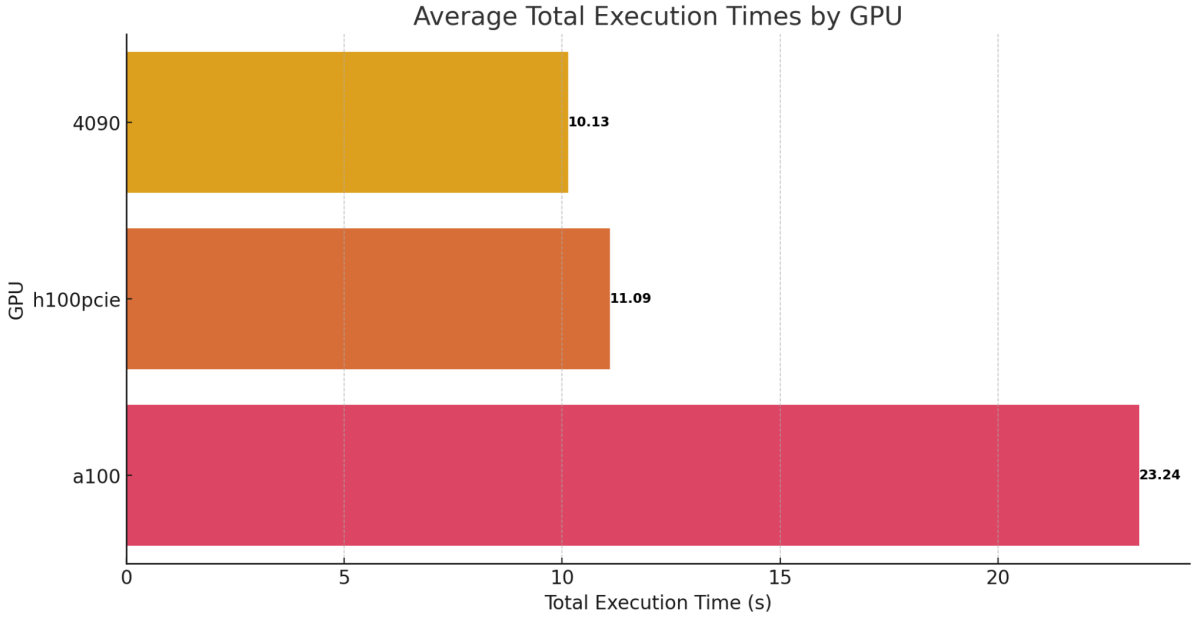
In questo capitolo, esamineremo i risultati ottenuti dai test di performance eseguiti per valutare l'efficienza del modello IDM VTON su diverse GPU. L'analisi si concentra su vari aspetti chiave, inclusi i tempi di esecuzione, i costi per inferenza e il throughput per diverse quantità di GPU. Le GPU testate includono la NVIDIA A100, la H100 PCIE e la Nvidia 4090, ognuna delle quali offre differenti prestazioni e capacità computazionali.

I risultati presentati in questo capitolo sono fondamentali per comprendere quale GPU offre il miglior compromesso tra tempo di esecuzione, costo e capacità di scalabilità. Questa analisi fornisce indicazioni preziose per ottimizzare il processo di try-on virtuale, permettendo di scegliere l'hardware più adatto alle esigenze specifiche di performance e costo. I dati ottenuti ci permettono di identificare i punti di forza e le debolezze di ciascuna GPU, aiutando a prendere decisioni informate per implementazioni future.

Nelle sezioni seguenti, analizzeremo in dettaglio:

- **Tempi di Esecuzione Medi per GPU**: Valutazione dei tempi necessari per completare un try-on virtuale utilizzando diverse GPU.
- **Tempi di Esecuzione per le Diverse Fasi del Processo**: Analisi dei tempi di esecuzione suddivisi per le diverse fasi del processo di try-on, considerando ciascuna GPU.
- **Costi per Inferenza**: Calcolo del costo medio per inferenza per ogni GPU, assumendo che le GPU siano già pronte e caricate in memoria.
- **Throughput per Diverse Quantità di GPU**: Valutazione del throughput in termini di richieste per ora in funzione del numero di GPU utilizzate.

Tempi di Esecuzione Medi



La figura sopra mostra i tempi di esecuzione medi per generare un try-on virtuale utilizzando tre diverse GPU: NVIDIA A100, H100 PCIE e 4090. I tempi sono espressi in secondi (s) e riflettono la durata necessaria per completare l'operazione con la GPU già pronta e caricata in memoria.

NVIDIA 4090
Tempo di Esecuzione Medio: 10.13 secondi
Analisi: La GPU NVIDIA 4090 ha registrato il tempo di esecuzione medio più basso tra le GPU testate. Questo risultato suggerisce che la 4090 è estremamente efficiente nel gestire i processi di generazione del try-on, rendendola una scelta ottimale per applicazioni che richiedono tempi di risposta rapidi.

H100 PCIE
Tempo di Esecuzione Medio: 11.09 secondi

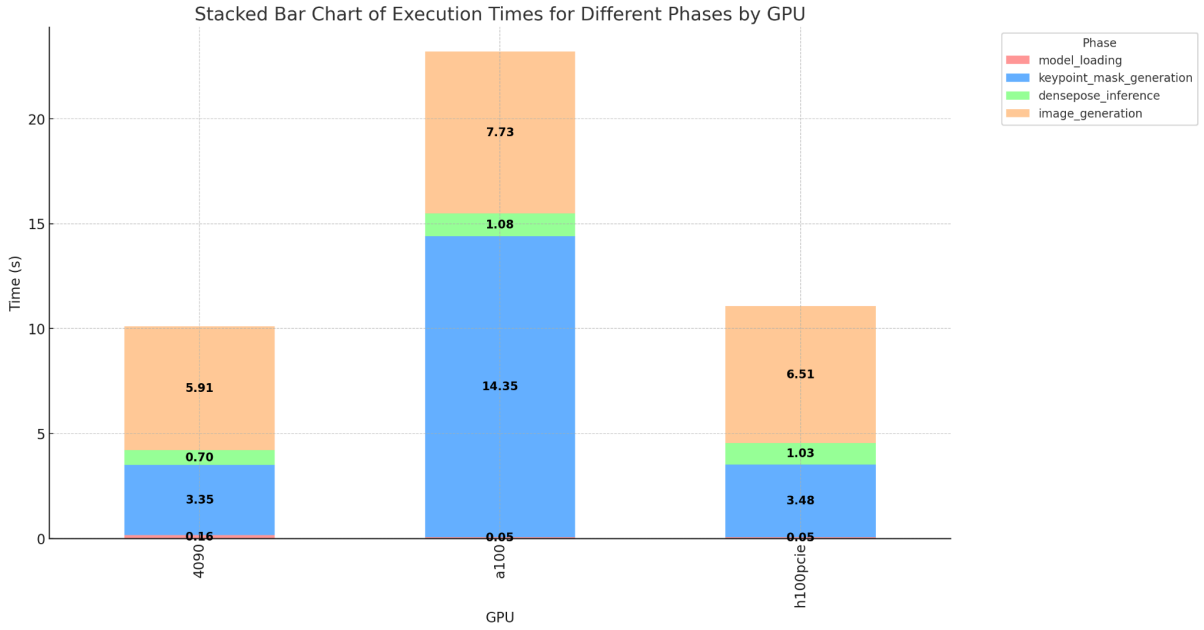
Analisi: La H100 PCIE ha mostrato un tempo di esecuzione leggermente superiore rispetto alla 4090, con una differenza di circa 0.96 secondi. Sebbene sia marginalmente più lenta, la H100 PCIE rappresenta comunque una valida alternativa, offrendo prestazioni competitive.

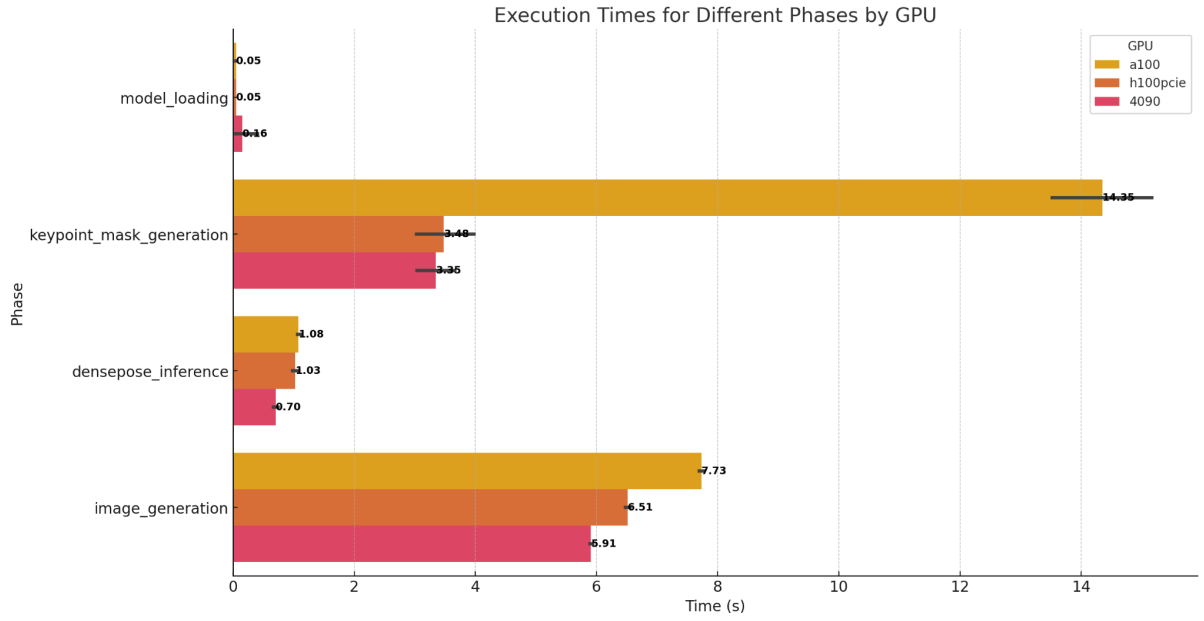
NVIDIA A100

Tempo di Esecuzione Medio: 23.24 secondi

Analisi: La GPU NVIDIA A100 ha registrato il tempo di esecuzione medio più alto, pari a 23.24 secondi. Questo risultato indica che, rispetto alle altre GPU testate, la A100 è meno efficiente nel contesto specifico della generazione del try-on virtuale. Tuttavia, è importante considerare che la A100 potrebbe eccellere in altri tipi di carichi di lavoro, data la sua architettura orientata al deep learning e all'intelligenza artificiale.

Tempi di Esecuzione per le Diverse Fasi su GPU Differenti





La figura sopra mostra i tempi di esecuzione per ciascuna fase del processo di generazione del try-on virtuale suddivisi per le diverse GPU utilizzate nei test: NVIDIA A100, H100 PCIE e 4090. I tempi sono espressi in secondi (s) e riflettono la durata media delle fasi di caricamento del modello (model_loading), generazione della maschera dei punti chiave (keypoint_mask_generation), inferenza DensePose (densepose_inference) e generazione dell'immagine (image_generation).

Model Loading
Tempi di Esecuzione Medi:

- A100: 0.05 secondi
- H100 PCIE: 0.05 secondi
- 4090: 0.16 secondi

Analisi: La fase di caricamento del modello è molto rapida per tutte le GPU testate, con i tempi di esecuzione che variano leggermente. La GPU 4090 mostra un tempo di caricamento leggermente superiore rispetto alle altre due GPU, ma la differenza è marginale.

Keypoint Mask Generation
Tempi di Esecuzione Medi:

- A100: 14.35 secondi
- H100 PCIE: 3.48 secondi
- 4090: 3.35 secondi

Analisi: La generazione della maschera dei punti chiave è significativamente più lunga sulla GPU A100 rispetto alle altre due GPU. La H100 PCIE e la 4090 mostrano tempi di esecuzione simili, con la 4090 leggermente più veloce. Questo suggerisce che la A100 potrebbe non essere ottimale per questa specifica fase del processo.

DensePose Inference
Tempi di Esecuzione Medi:

- A100: 1.08 secondi
- H100 PCIE: 1.03 secondi
- 4090: 0.70 secondi

Analisi: La fase di inferenza DensePose mostra tempi di esecuzione comparabili tra A100 e H100 PCIE, mentre la GPU 4090 è significativamente più veloce. Questo indica che la 4090 ha un vantaggio in termini di rapidità per questa fase specifica.

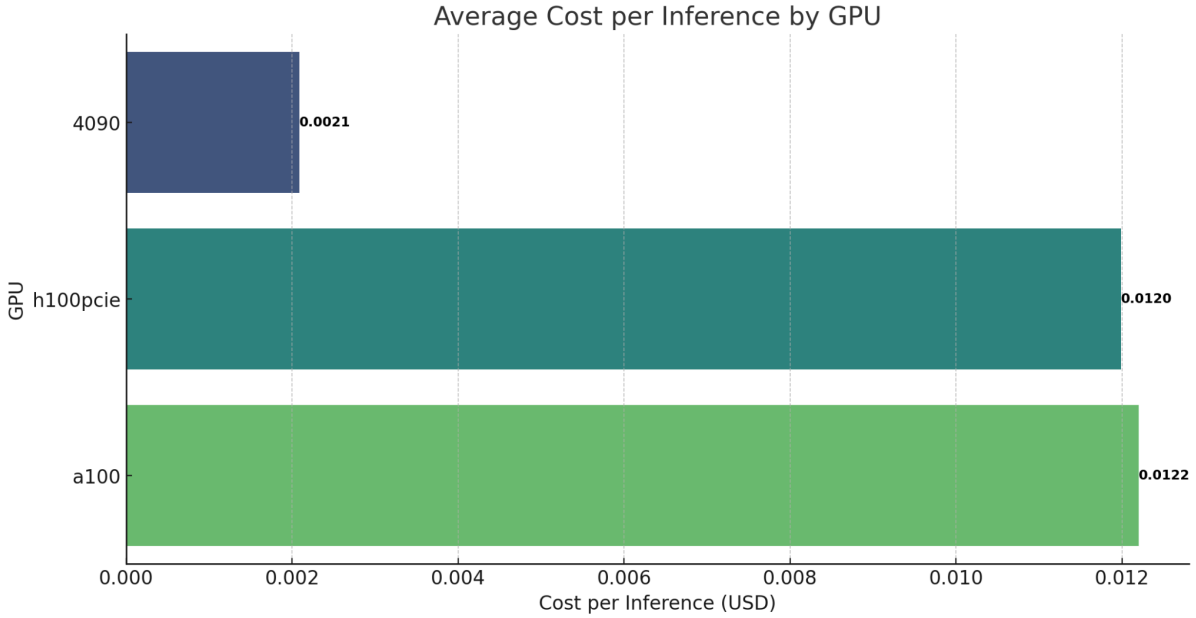
Image Generation
Tempi di Esecuzione Medi:

- A100: 7.73 secondi
- H100 PCIE: 6.51 secondi
- 4090: 5.91 secondi

Analisi: La generazione dell'immagine è la fase più dispendiosa in termini di tempo per tutte le GPU. La GPU 4090 è la più veloce, seguita dalla H100 PCIE e infine dalla A100. Questo

pattern conferma che la 4090 offre le migliori prestazioni anche per la fase più complessa del processo.

Costo per Singolo Try-on (runtime)



La figura sopra mostra il costo medio per inferenza (in USD) per ciascuna delle GPU utilizzate nei test: NVIDIA A100, H100 PCIE e 4090. È importante notare che i costi sono calcolati assumendo che le GPU siano già pronte e caricate in memoria, eliminando quindi i tempi e i costi associati al cold boot.

NVIDIA 4090

- Costo Medio per Inferenza: $0.0021
- Analisi: La GPU NVIDIA 4090 ha il costo per inferenza più basso tra le GPU testate, rendendola la scelta più economica per eseguire inferenze con IDM VTON. Questo basso costo, combinato con i suoi tempi di esecuzione rapidi, fa della 4090 una soluzione altamente efficiente sia in termini di costi che di prestazioni.

H100 PCIE

- Costo Medio per Inferenza: $0.0120
- Analisi: La GPU H100 PCIE presenta un costo per inferenza intermedio. Anche se più costosa rispetto alla 4090, offre comunque un buon compromesso tra costo e prestazioni. Questo la rende una valida alternativa per applicazioni che richiedono una combinazione di efficienza e costi moderati.

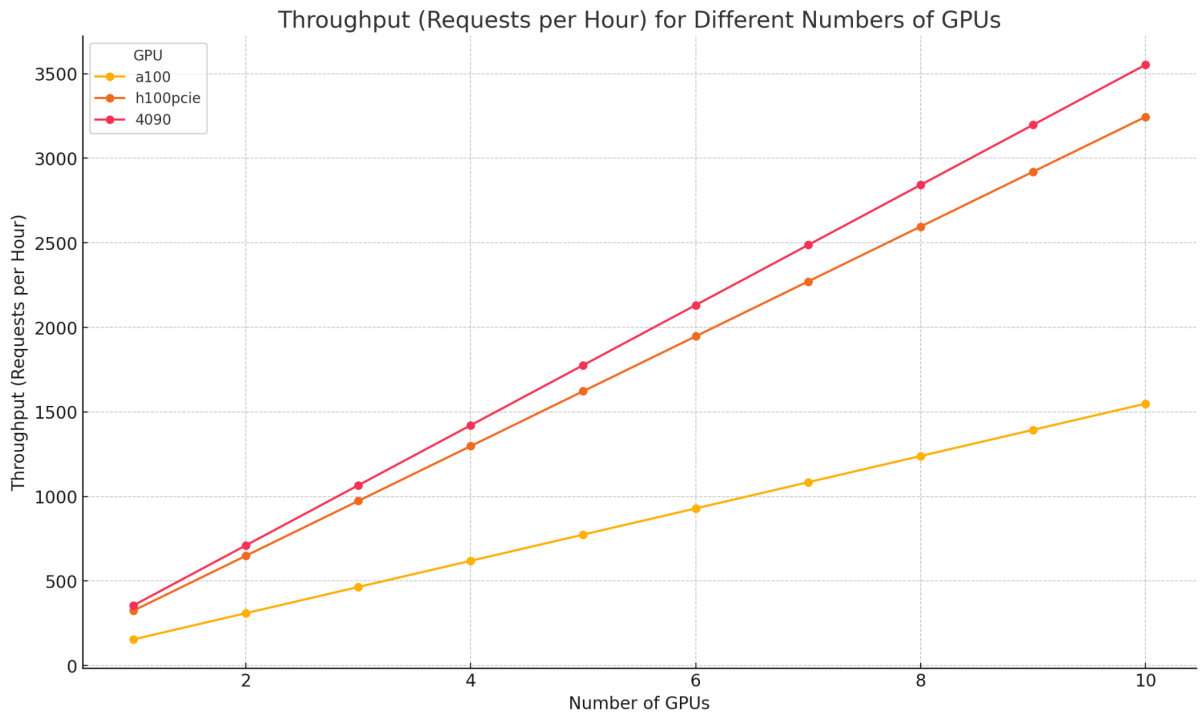
NVIDIA A100

- Costo Medio per Inferenza: $0.0122
- Analisi: La GPU NVIDIA A100 ha il costo per inferenza più alto tra le GPU testate. Questo costo elevato può essere giustificato solo se la A100 offre benefici significativi in altri aspetti del carico di lavoro che non sono riflessi direttamente nei tempi di esecuzione o nei costi per inferenza.

	A100	H100 PCIe	RTX 4090
Cost per **Try-on** (**USD**)	0.0122	0.011987	0.002083

L'analisi dei costi medi per inferenza rivela che la GPU NVIDIA 4090 è la soluzione più economica per eseguire inferenze con IDM VTON, seguita dalla H100 PCIE e infine dalla A100. Considerando che le GPU sono già pronte e caricate in memoria, la 4090 non solo offre i tempi di esecuzione più rapidi ma anche il costo per inferenza più basso, rendendola la scelta ideale per applicazioni di try-on virtuale che necessitano di un'alta efficienza a un costo contenuto. Questi risultati sono fondamentali per ottimizzare le risorse e ridurre i costi operativi nelle implementazioni pratiche di IDM VTON.

Scalabilità



Analisi del Throughput (Richieste per Ora) per Diverse Quantità di GPU
La figura sopra mostra il throughput, misurato in richieste per ora, per diverse quantità di GPU (da 1 a 10), utilizzando tre modelli di GPU: NVIDIA A100, H100 PCIE e 4090. L'obiettivo di questa analisi è determinare come il numero di GPU influisce sulla capacità di elaborare richieste per il modello IDM VTON.

NVIDIA 4090

- Andamento: Il throughput per la GPU 4090 aumenta linearmente con l'aumento del numero di GPU. Con una GPU, il throughput è di circa 500 richieste per ora, mentre con 10 GPU, il throughput raggiunge circa 3500 richieste per ora.

- Analisi: La GPU 4090 mostra la scalabilità migliore tra le GPU testate, con un aumento costante delle prestazioni man mano che vengono aggiunte più GPU. Questo indica che la 4090 è altamente efficiente e adatta per applicazioni che richiedono alta scalabilità.

H100 PCIE

- Andamento: Anche il throughput per la GPU H100 PCIE aumenta linearmente con il numero di GPU. Con una GPU, il throughput è di circa 500 richieste per ora, e raggiunge circa 3000 richieste per ora con 10 GPU.
- Analisi: La H100 PCIE mostra una buona scalabilità, anche se leggermente inferiore rispetto alla 4090. Questa GPU rappresenta comunque una scelta valida per applicazioni che richiedono un'elevata capacità di elaborazione e scalabilità.

NVIDIA A100

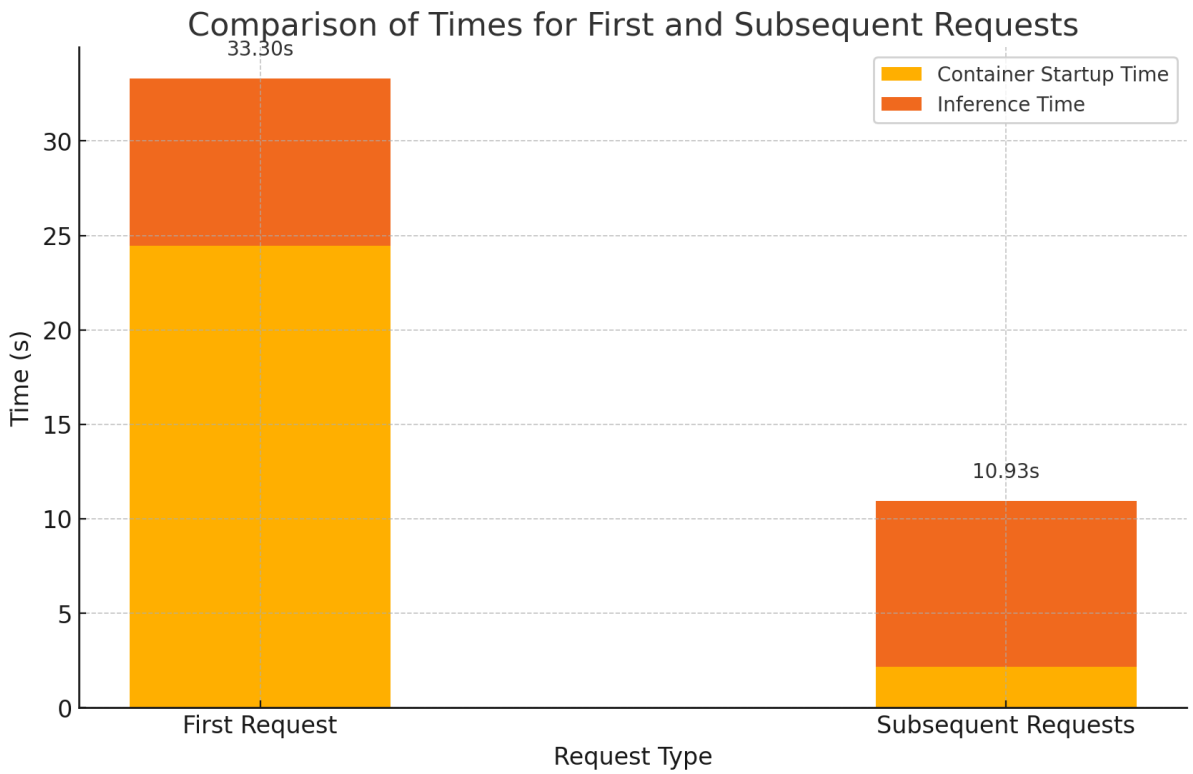
- Andamento: Il throughput per la GPU A100 segue un aumento lineare con l'aumento del numero di GPU, ma con valori inferiori rispetto alle altre due GPU. Con una GPU, il throughput è di circa 250 richieste per ora, mentre con 10 GPU, il throughput raggiunge circa 1500 richieste per ora.
- Analisi: La GPU A100 ha un throughput significativamente inferiore rispetto alle GPU 4090 e H100 PCIE. Sebbene mostri una scalabilità lineare, l'incremento delle prestazioni è meno pronunciato, indicando che l'A100 potrebbe non essere l'opzione migliore per applicazioni che richiedono alta scalabilità e throughput elevato.

	A100	H100 PCIe	RTX 4090
Capacità di richieste per ora a pieno carico.	154.924001	324.508505	355.295883

Costo singolo Try-on serverless

In questa sezione, analizziamo il costo del singolo try-on utilizzando un'architettura serverless, implementata tramite Runpod. Questo approccio è comunemente utilizzato nei servizi web in produzione poiché offre una scalabilità dinamica e consente di gestire efficacemente carichi di lavoro variabili. Descriveremo il funzionamento del modello come servizio serverless, il processo di gestione delle richieste e i tempi di esecuzione associati, evidenziando le differenze tra i tempi di avvio iniziali e quelli delle richieste successive. Per effettuare i test è stata utilizzata la scheda video Nvidia 4090 poichè la migliore in inferenza nei test precedenti.

Quando il modello IDM VTON è implementato come servizio serverless su Runpod, esso viene eseguito all'interno di un container Docker. Il modello viene avviato automaticamente in risposta alle richieste API, consentendo di scalare dinamicamente in base al volume delle richieste.



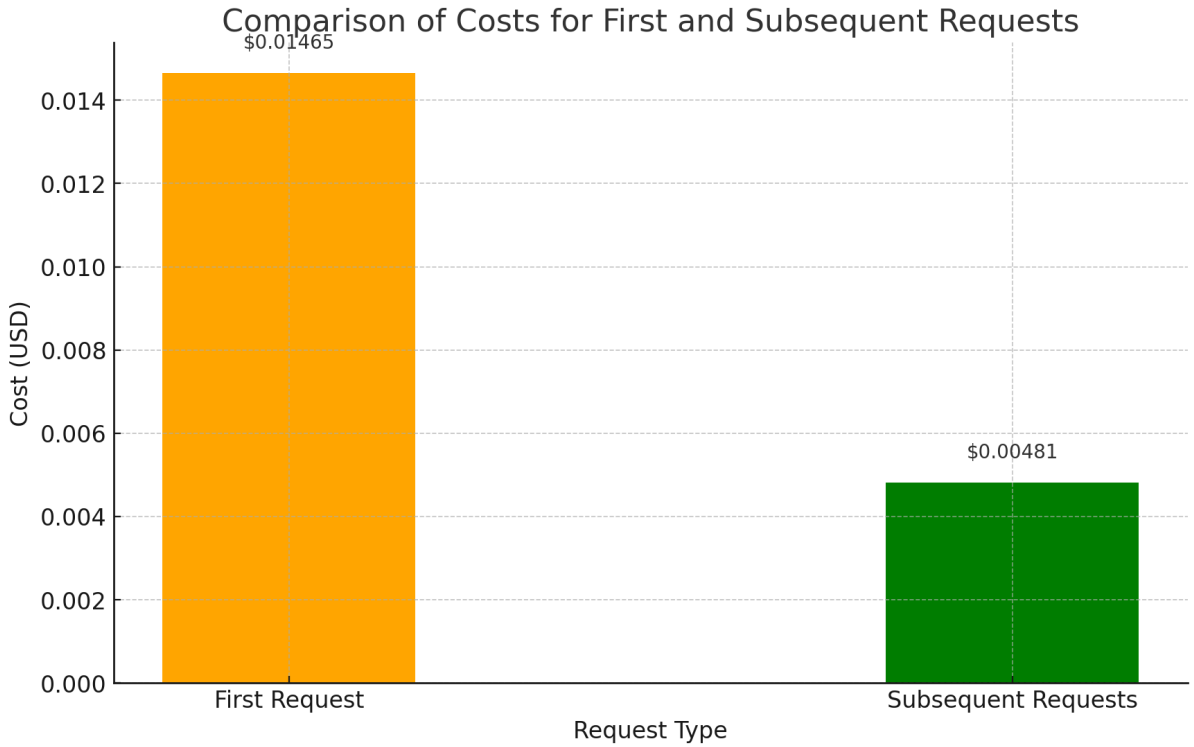
Avvio del Container e Istanziazione dei Worker

- Processo: Alla prima richiesta, Runpod avvia un container Docker che contiene l'immagine del modello IDM VTON. Questo processo di avvio include l'istanziazione di tre worker principali più due worker extra per gestire le richieste.
- Tempo di Avvio: Il tempo necessario per avviare il container è di 24.45 secondi.

Gestione delle Richieste

- Prima Richiesta: Poiché il container deve essere avviato, la prima richiesta ha un tempo di esecuzione maggiore:
    - Media dei Tempi di Avvio del Container: **24.45 secondi**
    - Media dei Tempi di Inferenza: **8.85 secondi**
    - Media totale di inferenza: **133.30 secondi**
- Richieste Successive: Una volta che il container è avviato, i worker possono gestire le richieste con tempi di esecuzione significativamente inferiori. I tempi rilevati per le richieste successive sono i seguenti:
    - Media dei Tempi di Avvio del Container: **2.18 secondi**
    - Media dei Tempi di Inferenza: **8.76 secondi**
    - Media totale di inferenza: **10.93 secondi**

Calcolo del costo



In questo sottocapitolo, analizziamo i costi associati alle richieste serverless per il try-on virtuale, confrontando i costi della prima richiesta con quelli delle richieste successive. Il costo del servizio serverless per una Nvidia RTX 4090 è di $0.00044 per secondo. I calcoli si basano sui tempi medi di esecuzione precedentemente discussi.

- **Prima Richiesta**: La prima richiesta ha un costo maggiore a causa del tempo di avvio del container, con un costo totale di **$0.014652.**
- **Richieste Successive**: Una volta avviato il container, le richieste successive hanno un costo significativamente inferiore, con un costo totale di **$0.0048092.**

	Prima richiesta	RIchieste successive
Costo (**USD**)	**$0.014652**	**$0.0048092**

Il confronto dei costi evidenzia l'efficienza economica dell'architettura serverless una volta superato il costo iniziale di avvio del container. Mentre la prima richiesta comporta un costo maggiore ($0.014652), le richieste successive hanno un costo ridotto ($0.0048092), rendendo questo approccio particolarmente vantaggioso per applicazioni con un elevato volume di richieste. Questo calcolo dimostra l'importanza di ottimizzare l'avvio dei container per minimizzare i costi operativi complessivi.

Conclusione:

In base all'analisi dei costi e delle prestazioni, possiamo concludere quanto segue:

RTX 4090: La GPU più conveniente e performante. Offre il costo per singolo try-on più basso e la capacità di gestire il maggior numero di richieste per ora.

H100 PCIe: È più costosa della RTX 4090 ma offre comunque una capacità elevata di gestione delle richieste.

A100: La GPU meno conveniente in termini di costo per singolo try-on e capacità di gestione delle richieste.

Considerando i vantaggi del modello serverless, la GPU RTX 4090 è particolarmente adatta per questo tipo di implementazione. Utilizzare la 4090 in un'architettura serverless offre i seguenti benefici:

- **Scalabilità Dinamica**: La capacità di scalare con le richieste consente di gestire carichi di lavoro variabili in modo efficiente.
- **Costo Efficiente**: Pagare per secondo per ogni richiesta ricevuta piuttosto che su base oraria permette di ottimizzare i costi operativi. Il costo per singola richiesta serverless con la RTX 4090 è significativamente inferiore, con un costo per richiesta calcolato a **$0.00481**.

In sintesi, implementare il servizio IDM VTON in modalità serverless utilizzando la GPU RTX 4090 non solo offre prestazioni elevate, ma consente anche una gestione dei costi molto più efficiente, rendendo questa soluzione ideale per applicazioni web in produzione che richiedono alta flessibilità e scalabilità.