PADRONEGGIARE L'HARDWARE PER L’ AI LOCALE: LA GUIDA DEFINITIVA

Versione: [1.0] | Data: [01.10.2025] | Stato: [Approvato]

Via Maestri Comacini 7
6830 Chiasso
Svizzera
prointelligentiaartificiali@protonmail.com

Sommario

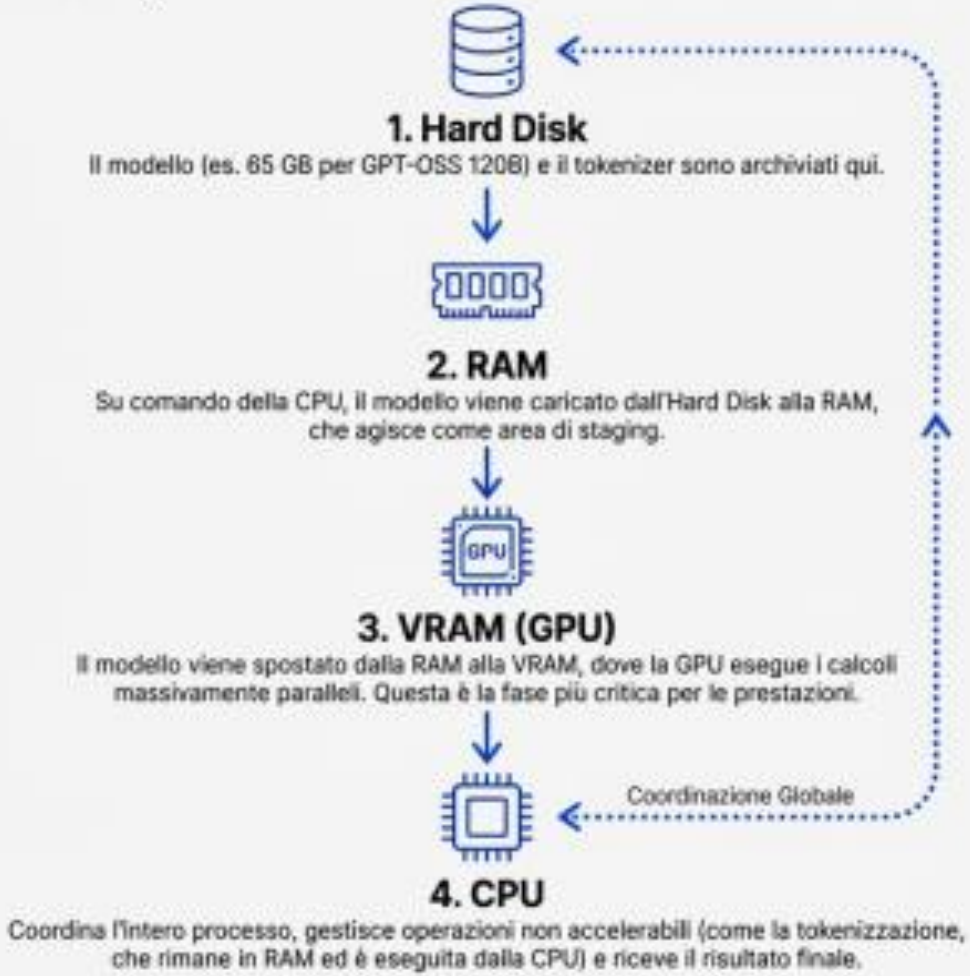
Il Dilemma dell'Ottimizzazione: Trovare il Proprio Equilibrio

Scegliere l'hardware per l'AI non è una questione di 'meglio' in assoluto, ma di trovare il giusto compromesso tra tre fattori chiave. Il tuo obiettivo è posizionarti nel punto ideale di questo triangolo.

L'obiettivo di questa guida è darti gli strumenti per navigare questo dilemma in modo informato.

Come Funziona il Flusso Dati nell'Inferenza AI

Quando esegui un modello, i dati seguono un percorso preciso attraverso il tuo sistema. Capire questo flusso è essenziale per identificare i colli di bottiglia.

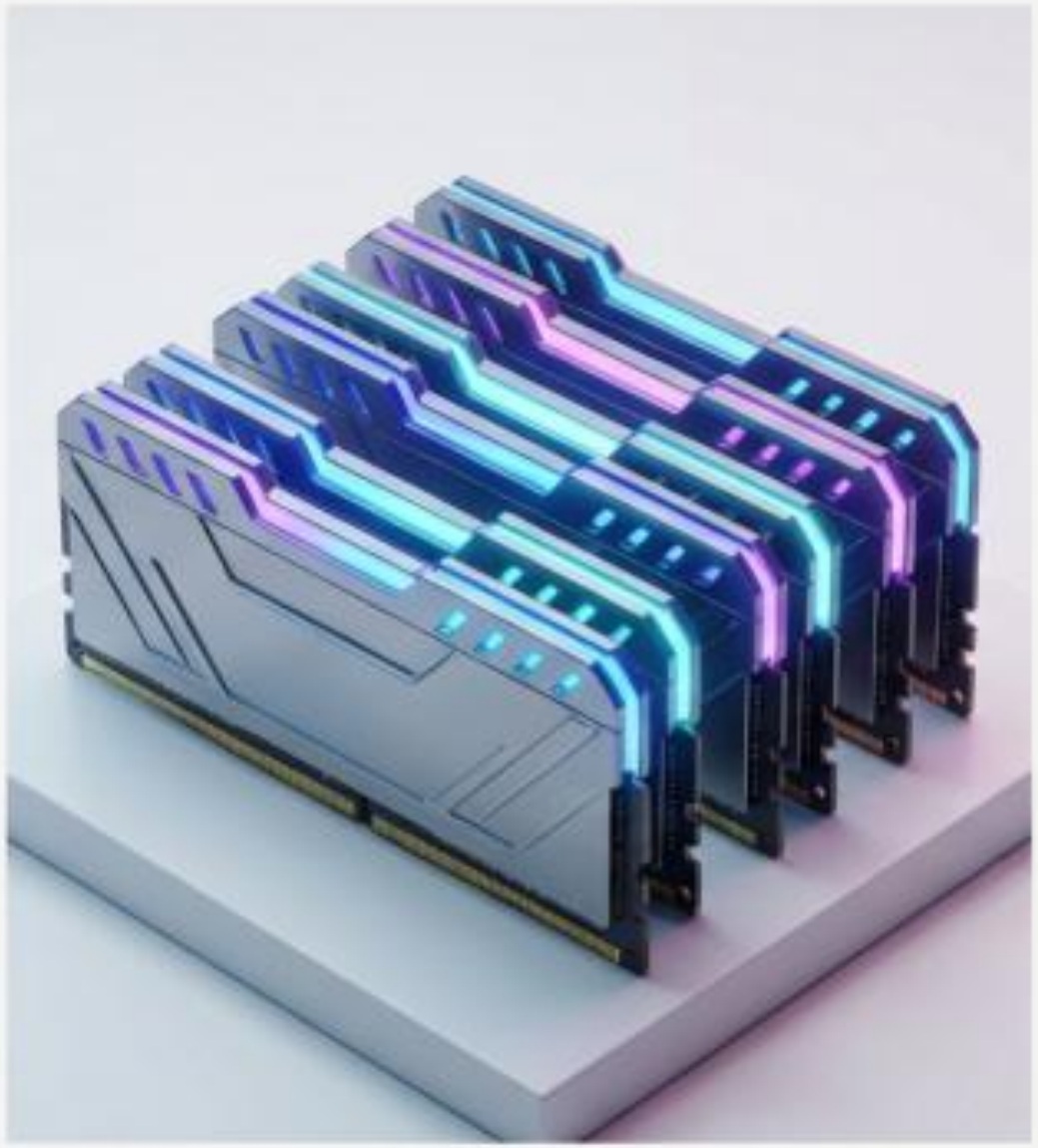


Il Protagonista: La GPU (★★★★★)

Il motore del calcolo, dove la VRAM è la valuta principale.

Punti Chiave

- **Funzione:** Esegue in parallelo le operazioni su matrici e tensori, fondamentali per LLM, modelli di visione e diffusione.
- **La VRAM è Tutto:** Serve a caricare il modello e il suo contesto (KV-cache). Se non basta, si ricorre all' "offloading" su RAM, ma le prestazioni crollano drasticamente.
- **L'Ecosistema:**
  - **NVIDIA (CUDA):** L'ecosistema più maturo e supportato.
  - **AMD** (ROCm): In netto miglioramento e oggi un'alternativa valida.
  - **Apple Silicon**: L'architettura a memoria unificata (Unified Memory) fonde RAM e VRAM, cambiando le regole del gioco.



Comprimario Essenziale: La RAM (⭐⭐⭐⭐)

Il buffer di sicurezza per l'offloading e i contesti lunghi.

Punti Chiave

- **Funzione**: Mantiene in memoria parti del modello non caricate in VRAM, gestisce il contesto (KV-cache), i buffer di sistema e l'OS.
- **Regola Fondamentale**: Avere almeno **2 volte la VRAM** come RAM per operare comodamente.
- **Esempi Pratici**:
  - GPU con 8 GB VRAM → **16-32 GB RAM** consigliati.
  - GPU con 24 GB VRAM → **32-64 GB RAM** consigliati.
- **Nota**: Per inferenza solo su CPU o per modelli con contesti molto lunghi (es. 128k token), 32-64 GB di RAM diventano cruciali.

I Ruoli di Supporto: CPU (★★★) e Hard Disk (★★)

CPU

- **Ruolo:** Coordina il carico, esegue task non accelerati (tokenizzazione, I/O).
- **Con GPU:** Non è il collo di bottiglia principale, ma un modello moderno evita stalli.
- **Senza GPU (CPU-only):** L'inferenza è possibile ma molto più lenta. Si raccomandano CPU con più core, supporto AVX2/AVX-512 e l'uso di build ottimizzate (llama.cpp, GGML).

Hard Disk

- **Ruolo:** Archivia modelli, checkpoint e dataset.
- **Impatto sulle Prestazioni:** La velocità (lettura/scrittura) incide solo sul tempo di caricamento del modello, non sui token/s durante l'inferenza.
- **Consiglio:** Un NVMe (con velocità di 3+ GB/s) è fondamentale per ridurre i tempi di attesa. Lo spazio è critico: un modello da 70B in formato FP16 richiede oltre 140 GB.

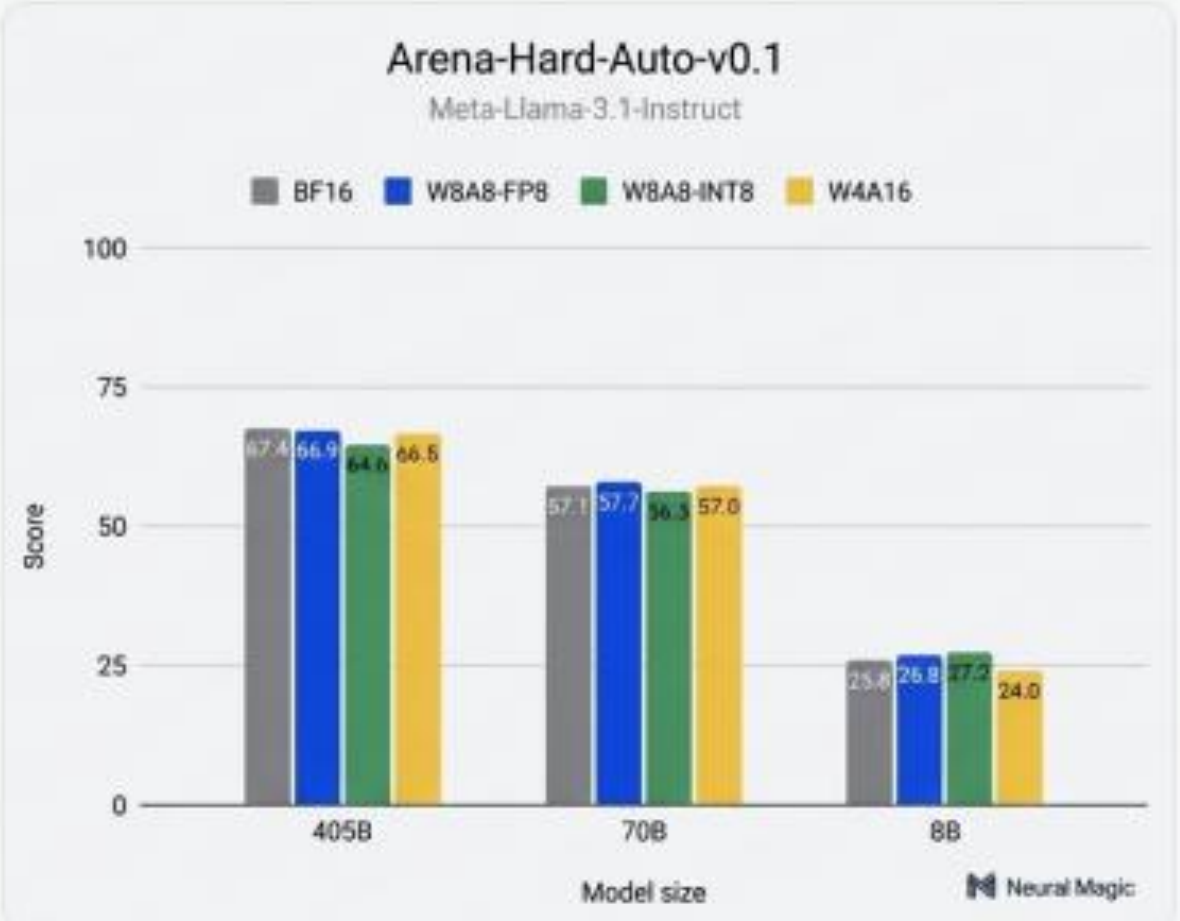
L'Arma Segreta: La Quantizzazione per Piegare le Regole

La quantizzazione è una tecnica che riduce la precisione numerica dei pesi di un modello (es. da 16-bit a 4-bit). Questo comprime drasticamente le dimensioni del modello e ne accelera l'esecuzione.

Il Trade-off

Vantaggi: Meno VRAM richiesta, meno spazio su disco, inferenza più veloce.

Svantaggio: Una perdita di accuratezza, generalmente trascurabile fino a 4-bit, ma che aumenta con compressioni più aggressive.



Calcola la Tua VRAM: La Formula per Non Sbagliare

Usa questa formula base per stimare la memoria necessaria a caricare un modello, in base ai suoi parametri e al livello di quantizzazione.

FP16/BF16 (Full Precision)

Parametri (B) × **2** = GB di VRAM

Esempio: Modello da 8B → 16 GB VRAM

INT8 (8-bit Quantized)

Parametri (B) × **1** = GB di VRAM

Esempio: Modello da 8B → 8 GB VRAM

INT4 (4-bit Quantized)

Parametri (B) × **0.5** = GB di VRAM

Esempio: Modello da 8B → 4 GB VRAM

ℹ Aggiungere sempre un **~15-20% di VRAM extra** per il contesto (KV-cache) e l'overhead del sistema.

Scenario #1: Eseguire GPT-OSS 120B a Massima Precisione

I requisiti hardware per la 'massima potenza', senza compromessi.

Requisiti Hardware (Full Precision - FP16)

VRAM: 120B parametri × 2 = **240 GB di VRAM**.

- Equivalente a **3 GPU NVIDIA H100** (80 GB ciascuna).
- Costo GPU: ~$24,500 × 3 = **~$73,500**.

RAM: Almeno il doppio della VRAM → **~480 GB di RAM**.

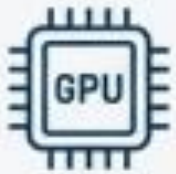
- Costo RAM (stimato): Un modulo da 256 GB costa ~€3,400. Costo totale **~€7,000**.

Spazio Disco: Almeno 2 TB NVMe.

Un costo totale di **oltre €80.000** per un singolo sistema in grado di gestire il modello a piena velocità e per più utenti.

Scenario #2: Il Punto di Equilibrio "Smart"

La configurazione consigliata per la maggior parte degli utenti, un ottimo bilanciamento tra costo e prestazioni.



GPU

Componente Cruciale

- VRAM: min. **8-12 GB** (ideale **16+ GB**)
- Modello: NVIDIA **RTX 3060 (12GB)** o superiore
- Supporto: CUDA/ROCm

ⓘ Nota: Richiesta per inferenza efficiente

CPU

- Core: **6+ core** (8+ consigliati)
- Frequenza: **3.0+ GHz**

RAM

- Minimo: **16 GB**
- Consigliati: **32+ GB**
- Regola: **Almeno 1.5x-2x** la VRAM della GPU

Scenario #3: La Soluzione Disruptive dei Mini PC

Massima efficienza di costo per l'inferenza a singolo utente, senza GPU dedicata.

Questi sistemi sfruttano CPU potenti (con acceleratori AI) e grandi quantità di RAM (fino a 128 GB) per eseguire modelli di grandi dimensioni quantizzati.

Esempio di Prestazioni: Un Mini PC con **AMD Ryzen AI 395** e **128 GB di RAM** può eseguire **GPT-OSS 120B (quantizzato a 4-bit)** a circa **16 token/s**.

Caso d'Uso Ideale: Un professionista (es. avvocato,
Caso d'Uso Ideale: Un professionista (es. avvocato, commercialista) che necessita di un sistema RAG locale per i propri documenti. Gestisce un singolo utente in modo efficace.

Costo: Tipicamente tra **€1.200 e €1.500**.

Vademecum dell'Hardware AI: La Sintesi Strategica

Gerarchia dei Componenti

- **GPU:** ⭐⭐⭐⭐⭐ (Il motore)
- **RAM:** ⭐⭐⭐⭐ (Il supporto)
- **CPU:** ⭐⭐⭐ (Il coordinatore)
- **Hard Disk:** ⭐⭐ (L'archivio)

Raccomandazioni Chiave

VRAM: È il fattore limitante. Parti da 12 GB.

RAM: Almeno 2x la VRAM.

Storage: Solo NVMe per caricamenti veloci.

Ecosistema: NVIDIA/CUDA offre la massima compatibilità.

La Formula della VRAM

FP16: `Parametri(B) * 2`

INT8: `Parametri(B) * 1`

INT4: `Parametri(B) * 0.5`

(+15-20% di overhead)

Grazie

Via Maestri Comacini 7, 6830 Chiasso, Svizzera • prointelligentiaartificiali@protonmail.com

UID: [inserire UID]