SCHEDA TECNICA · PIATTAFORMA NVIDIA · REV. 1.0

Tutto il sistema. In una scatola da un litro.

Hardware, modelli e budget di memoria di Nova. Ogni componente della catena — ragionamento, lettura documenti, ricerca semantica, generazione — viene eseguito sulla macchina. Nessun account, nessuna chiave API, nessun dato in uscita.

128 GB
Memoria unificata coerente CPU e GPU, stesso indirizzo

1 PFLOP
FP4 con sparsità GPU Blackwell

1,13 litri
240 W una scatola da un litro, alla lettera

0 byte
Trasmessi a terzi durante l'inferenza

01 Piattaforma hardware

Una scatola da 1,13 litri. Nessun rack, nessuna sala server, nessun sistemista.

Sotto inferenza continua la ventola è udibile, come in un portatile che sta lavorando davvero. Non richiede però né un locale tecnico né raffreddamento dedicato.

COMPONENTE	SPECIFICA
Superchip	NVIDIA GB10 Grace Blackwell — piattaforma *DGX Spark* (ASUS Ascent GX10)
CPU	20 core Arm — 10× Cortex-X925 + 10× Cortex-A725
GPU	Blackwell con Tensor Core di 5ª generazione — **1 PFLOP in FP4** (con sparsità)
Memoria	128 GB LPDDR5X unificata e **coerente** — CPU e GPU condividono lo stesso spazio di indirizzamento
Banda di memoria	273 GB/s — è il fattore che determina la velocità di generazione
Storage	NVMe PCIe Gen4 — 1 TB (Desktop) · 4 TB (Server)
Rete	Wi-Fi 7, Bluetooth 5.4, NVIDIA ConnectX integrato
Sistema operativo	DGX OS (Ubuntu, ARM64). **Nessun parametro kernel da configurare:** i 128 GB sono nativamente accessibili alla GPU
Formato e consumo	1,13 litri · 240 W
Espandibilità	Due unità si collegano via ConnectX per reggere modelli più grandi

PERCHÉ LA MEMORIA UNIFICATA COERENTE È IL PUNTO

Su una scheda video la memoria è un limite invalicabile: una GPU da 24 GB non può caricare un modello da 60 GB, punto. Qui CPU e GPU non condividono soltanto la capacità — condividono **lo stesso spazio di indirizzamento**. Nessuna copia, nessun travaso tra due memorie separate: il modello sta dove serve, e ci resta.

IL VANTAGGIO CHE NON SI VEDE NEI BENCHMARK

La lettura dei documenti, non la scrittura delle risposte.

La velocità con cui Nova *scrive* è limitata dalla banda di memoria (273 GB/s): siamo sui ~55–60 token al secondo, in linea con qualunque altra macchina da 128 GB. Ma la velocità con cui Nova **legge** — ingerire un contratto di ottanta pagine prima di ragionarci — è un compito di *calcolo*, e lì la GPU Blackwell con 1 PFLOP in FP4 opera su un altro ordine di grandezza rispetto a una grafica integrata.

È esattamente il carico di lavoro di Nova: assorbire documenti aziendali, poi ragionarci sopra. Il tempo che l'utente aspetta prima di vedere la prima parola della risposta è quello che si accorcia — ed è il tempo che percepisce come lentezza.

Due configurazioni: Desktop e Server

Lo stesso hardware in due configurazioni. Cambia come le intelligenze vengono allocate — dedicate a un singolo utente per la massima precisione, o ottimizzate per servire un intero team.

	DESKTOP	SERVER
Utenti	Un utente alla volta	Più utenti in contemporanea
Intelligenze	Un modello specializzato per ogni agente	Modelli ottimizzati per servire tutti gli agenti
Risorse per agente	Dedicate	Condivise fra gli agenti
Precisione	Estrema — ogni agente lavora al massimo	Ottimizzata per il carico condiviso
Storage	1 TB NVMe	4 TB NVMe
Per chi	Il singolo professionista	Team e reparti con più utenti

02 Le intelligenze di Nova

Quattro intelligenze specializzate, ognuna per un compito. Tutte a pesi aperti, tutte in locale, tutte caricate in memoria durante il lavoro.

Cervello SEMPRE CALDO

`Tipo di intelligenza: MoE — molti parametri complessivi, pochi attivi per ogni parola generata`

Il motore di ragionamento di Nova. Interpreta la richiesta, decide quali dati leggere e quali strumenti usare, incrocia le informazioni dei gestionali e produce il risultato finale — un report, un'email, un documento pronto. L'architettura MoE attiva solo una frazione dei suoi parametri a ogni passo: è ciò che la rende veloce pur restando capace sui compiti complessi. Resta sempre in memoria, così tra una richiesta e l'altra non c'è attesa di caricamento.

Occhi SEMPRE CALDO

`Tipo di intelligenza: Visione (VLM)`

La comprensione dei documenti. Legge fatture, contratti, moduli e scansioni, e non si limita a estrarne il testo: ne capisce la struttura — distingue l'imponibile dalla scadenza, la controparte dall'oggetto. È ciò che permette a Nova di lavorare sui documenti che l'azienda già riceve, senza doverli reinserire a mano.

Memoria SEMPRE CALDO

`Tipo di intelligenza: Embedding semantico`

L'archivio che ritrova per significato. Indicizza in continuazione i documenti trasformandoli in una rappresentazione che ne coglie il senso, non le parole esatte. Così una richiesta come «i contratti con rinnovo automatico» trova i documenti giusti anche quando quelle parole precise non compaiono. È la base della ricerca interna e del contesto su cui il Cervello ragiona.

Immagini A RICHIESTA

`Tipo di intelligenza: Diffusion`

La generazione visiva. Produce immagini a partire da una descrizione, quando servono — un mockup, un'illustrazione per un documento. È l'unica intelligenza non sempre attiva: si carica al bisogno, perché serve solo occasionalmente e in quei momenti può prendere spazio in memoria senza sottrarlo alle altre.

Tutte le intelligenze sono a pesi aperti, con licenze permissive per l'uso commerciale, senza telemetria e interamente eseguibili offline. Girano tutte sull'ecosistema software maturo della piattaforma — la differenza tra un sistema che si aggiorna e uno che ogni sei mesi va rimesso in piedi.

FP4 NATIVO

Il modello gira nel formato per cui è stato addestrato.

Il motore di ragionamento è distribuito nativamente in **FP4**, e i Tensor Core Blackwell lo eseguono senza conversione. Nessuna riquantizzazione, nessuna perdita di qualità introdotta per farlo entrare in memoria: i pesi che arrivano sono i pesi che girano.

03 La regola che governa la scelta

La velocità dipende dai parametri **attivi**, non da quelli totali.

Architettura MoE

~55 t/s

molti parametri totali · pochi attivi per token

Architettura densa

~18 t/s

stessi parametri totali · tutti attivi per token

Stessa dimensione dichiarata, **tre volte la differenza di velocità.** Il criterio di selezione è quindi uno solo: molti parametri totali, pochi attivi. Architetture MoE (Mixture of Experts), mai modelli densi — un denso da 70B striscia a 5–15 t/s, perché ogni token deve attraversare tutti i pesi e la banda di memoria diventa il collo di bottiglia.

04 Budget di memoria

128 GB coerenti, di cui ~118 GB disponibili a intelligenze e contesto. L'intera catena operativa avviene **senza mai scaricare un'intelligenza dalla memoria.**

VOCE	MEMORIA	STATO	NOTE
Sistema operativo	~10 GB	fisso	Ubuntu ARM64, runtime, applicazione Nova
Cervello	~63 GB	SEMPRE CALDO	Ragionamento, orchestrazione, produzione documenti
Occhi	~7 GB	SEMPRE CALDO	OCR e comprensione documenti
Memoria	~1,2 GB	SEMPRE CALDO	Indicizzazione continua
Contesto (cache KV)	~20 GB	variabile	Più generoso: il prefill veloce rende il contesto lungo davvero usabile
Totale residente	**~101 GB**	—	**Margine libero: ~27 GB**
Immagini	~12 GB	A RICHIESTA	Entra nel margine senza sfrattare nulla

IL MARGINE IN PIÙ

Restano circa **27 GB liberi** — più che sulle piattaforme concorrenti, perché non serve riservare una quota fissa di memoria alla GPU: è tutta coerente e disponibile. È spazio per crescere: un contesto ancora più ampio, un'intelligenza aggiuntiva tenuta sempre calda, o i carichi di picco — senza dover sfrattare nulla dalla memoria.

ESPANDIBILITÀ

Due unità si collegano via ConnectX e mettono in comune la memoria: 256 GB coerenti, abbastanza per modelli che oggi non entrerebbero. È l'unico percorso di crescita disponibile in questa classe di macchine — e significa che l'acquisto non è una scommessa irreversibile.

05 Sovranità del dato

PROPRIETÀ	GARANZIA
Inferenza	Interamente locale. Nessuna chiamata a provider esterni.
Dati aziendali	Non lasciano mai la macchina. Documenti, gestionali e conversazioni restano su disco locale.
Apprendimento	Ciò che Nova impara — regole, correzioni, tono — resta sulla macchina. Non addestra modelli di terzi.
Funzionamento offline	Sì. Il sistema opera con il cavo di rete scollegato da internet.
Telemetria	Nessuna richiesta dai modelli.
Verificabilità	Osservabile sulla rete del cliente: assenza di traffico in uscita durante l'inferenza.

LA PROVA

Stacca il cavo di rete.
Nova continua a lavorare.

È la sola garanzia che conti davvero sui dati riservati: non una clausola contrattuale, ma un fatto che si verifica sulla propria rete in trenta secondi.

Nova — un prodotto Ncode Studio

team@ncodestudio.it