CAPITOLATO DI GARA

Fornitura infrastruttura AI/HPC basata su GPU – Compute, Storage, Networking e Software di orchestrazione

Premessa e Obiettivi

La presente gara ha per oggetto la fornitura, l'installazione, la configurazione e il collaudo di un'infrastruttura AI/HPC on-prem basata su GPU, comprensiva di nodi di calcolo basato su sistema NVIDIA DGX H200, storage NFS, head node e nodi di control plane, rete 25/100 GbE e software di gestione/orchestrazione. L'infrastruttura dovrà supportare **sviluppo, training, fine-tuning e inferenza** su modelli LLM e workload AI/ML, con requisiti di **scalabilità, affidabilità, osservabilità** e **sicurezza** a livello enterprise. Sarà oggetto della fornitura anche lo sviluppo di use case di modelli LLM da eseguire sulla infrastruttura di calcolo oggetto della fornitura.

1. Specifiche Tecniche Della Fornitura

I prodotti, le prestazioni e le relative specifiche tecniche elencate nel presente capitolato si intendono quali prescrizioni tecniche inderogabili, il cui mancato rispetto comporterà l'esclusione dalla presente procedura, fermo restando il rispetto del principio di equivalenza funzionale comprovato secondo le modalità di cui al successivo articolo del presente capitolato.

PRINCIPIO DI EQUIVALENZA FUNZIONALE

Si precisa che le marche e modelli degli elementi specificati di seguito sono da intendersi con prescrizioni legate a prodotti che soddisfano pienamente le esigenze della stazione appaltante.

Eventuali prodotti equivalenti proposti dovranno avere prestazioni pari o superiori sia singolarmente (singolo componente) che nell'intero complesso (sistema) sotto tutti punti di vista (es. potenza di calcolo, velocità di calcolo, velocità di elaborazione, velocità di trasferimento dati, semplicità di utilizzo, consumi energetici, scalabilità del sistema, strumenti e risorse disponibili on line, contenitori e framework proprietari e open source disponibili, ecc.). La non equivalenza anche per un solo punto di vista di un solo componente comporterà l'esclusione dell'offerta.

La proposta di prodotti equivalenti andrà evidenziata nell'offerta tecnica e accompagnata da relativa scheda tecnica. Nel caso di mancata specificazione si intende che i prodotti saranno della marca e modello richiesti.

2. Oggetto della Fornitura

La fornitura dovrà comprendere obbligatoriamente, a pena di esclusione dell'offerta:

- **Sistema di calcolo NVIDIA DGX H200**
- **Nodi per control plane**, **Nvidia Base Command Manager** e **Nvidia Run:ai**
- **PDU e tranciver SFP**
- **Software gestione cluster:**
  - **NVIDIA Base Command Manager**
  - **Kubernetes**
  - **NVIDIA Run:ai**
  - **NVIDIA AI Enterprise**
- **Servizi**
  - **Installazione e configurazione cluster**
  - **Integrazione con infrastruttura di calcolo esistente**: montaggio in armadio rack marca Lenovo 42U 1100mm contenente uno switch NVIDIA SN3420 25GbE, un server ThinkSystem SR680a, due gruppi di continuità RT11kVA 6U Rack mount;
  - **Installazione e configurazione sistema di gestione cluster e provisioning delle risorse (GPU, CPU, Storage)**
  - **Sviluppo di use case su modelli LLM**
  - **Collaudo**
  - **Formazione**

3. Specifiche Tecniche Minime

3.1 Sistema NVIDIA DGX H200

Fornitura, consegna, installazione, messa in opera e collaudo di un sistema NVIDIA DGX™ H200 di ultima generazione per sviluppo, training, fine-tuning e inferenza su modelli AI/LLM e carichi HPC, includendo 3 anni di supporto enterprise (Hardware + Software). Il sistema deve includere la suite software NVIDIA DGX (DGX OS, Nvidia Base Command) e NVIDIA AI Enterprise.

Requisiti funzionali e di piattaforma

Il sistema Nvidia richiesto dovrà includere/garantire:

- 8 GPU NVIDIA H200 Tensor Core connesse tramite NVLink/NVSwitch per dominio GPU a bassa latenza intra-nodo.
- NVIDIA Base Command™ (orchestrazione/gestione) e suite NVIDIA AI Enterprise (framework, runtime, microservizi e supporto).
- Prestazioni di sistema: fino a 32 PFLOPS in FP8 e memoria GPU aggregata pari a 1.128 GB.

Specifiche tecniche della fornitura

I prodotti, le prestazioni e le relative specifiche tecniche elencate nel presente capitolato si intendono quali prescrizioni tecniche inderogabili, il cui mancato rispetto comporterà l'esclusione dalla presente procedura, fermo restando il rispetto del principio di equivalenza funzionale comprovato secondo le modalità di cui al successivo articolo del presente capitolato. La fornitura dovrà comprendere obbligatoriamente, a pena di esclusione dell'offerta:

- nr. 1 Server NVIDIA DGX H200 come da specifiche del produttore e da caratteristiche di dettaglio fornite nei successivi paragrafi(https://resources.nvidia.com/en-us-dgx-systems/dgx-h200-datasheet?ncid=no-ncid)
- DGX OS (base software della piattaforma DGX).
- NVIDIA Base Command software stack
- Supporto nativo da parte del produttore NVIDIA

Acceleratori GPU

- Quantità e modello: 8× NVIDIA H200 in formato SXM.
- Memoria per GPU: 141 GB HBM3e e banda memoria 4,8 TB/s per GPU.
- Interconnessione intra-nodo: NVLink (4ª gen) con 18 connessioni per GPU e fino a 900 GB/s di banda bidirezionale per GPU; NVSwitch (4 unità) per 7,2 TB/s bidirezionali totali.

CPU, memoria di sistema e storage

- CPU: 2× Intel® Xeon® Platinum 8480C (totale 112 core).
- RAM di sistema: 2 TB minimo.
- Storage locale: 30 TB NVMe per dati (8× 3,84 TB Gen4 NVMe) + unità NVMe per OS (2× 1,92 TB).

Networking e gestione

- Connettività ad alte prestazioni: 10× NVIDIA ConnectX-7 400 Gb/s (InfiniBand/Ethernet), fino a 1 TB/s di banda di rete bidirezionale di picco.
- Sfp Tranciver compatibili con lo switch NVIDIA SN3420 25GbE;
- Rete di management: interfacce dedicate e BMC/gestione out-of-band; DGX OS con strumenti di gestione integrati.

Software incluso

- DGX OS (base software della piattaforma DGX).
- NVIDIA Base Command™.
- NVIDIA AI Enterprise (tool, framework e supporto).

Riepilogo caratteristiche minime

Component	Description
GPU	For H200: 8 x NVIDIA H200 GPUs that provide 1,128 GB total GPU memory
CPU	2 x Intel Xeon 8480C PCIe Gen5 CPUs with 56 cores each 2.0/2.9/3.8 GHz (base/all core turbo/Max turbo)

Component	Description
NVSwitch	4 x 4th generation NVLinks that provide 900 GB/s GPU-to-GPU bandwidth
Storage (OS)	2 x 1.92 TB NVMe M.2 SSD (ea) in RAID 1 array
Storage (Data Cache)	8 x 3.84 TB NVMe U.2 SED (ea) in RAID 0 array
Network (Cluster) card	4 x OSFP ports for 8 x NVIDIA® ConnectX®-7 Single Port InfiniBand Cards Each card provides the following speeds: • InfiniBand (default): Up to 400Gbps • Ethernet: 400GbE, 200GbE, 100GbE, 50GbE, 40GbE, 25GbE, and 10GbE
Network (storage and in-band management) card	2 x NVIDIA® ConnectX®-7 Dual Port Ethernet Cards Each card provides the following speeds: • Ethernet (default): 400GbE, 200GbE, 100GbE, 50GbE, 40GbE, 25GbE, and 10GbE • InfiniBand: Up to 400Gbps
System memory (DIMM)	2 TB using 32 x DIMMs
BMC (out-of-band system management)	1 GbE RJ45 interface Supports Redfish, IPMI, SNMP, KVM, and Web user interface
System management interfaces (optional)	Dual port 100GbE in slot 3 and 10 GbE RJ45 interface
Power supply	6 x 3.3 kW

Specifiche Fisiche

Feature	Description
Form Factor	8U Rackmount
Height	14” (356 mm)
Width	19” (482.3 mm) max
Depth	35.3” (897.1 mm) max
System Weight	287.6 lbs (130.45 kg) max

Prestazioni attese e capacità

- Performance AI: fino a 32 PFLOPS FP8 per nodo DGX H200.

- Memoria GPU aggregata: 1.128 GB (8×141 GB).
- Banda memoria HBM: 4,8 TB/s per GPU (HBM3e).

Supporto, licenze e durata (3 anni)

- Inclusione di 3 anni di supporto enterprise NVIDIA per il sistema DGX H200 (Hardware + Software), con accesso al supporto tecnico, aggiornamenti/patch DGX OS/firmware e entitlement a NVIDIA AI Enterprise.
- Sono ammessi profili Business Standard. La registrazione del seriale del sistema è obbligatoria per l'attivazione.

Sono richiesti i seguenti servizi, descritti in modo esteso nel paragrafo "servizi".

- Consegna, installazione e messa in opera on-site, comprensive di racking, cablaggi e integrazione di rete.
- Configurazione di DGX OS, NVIDIA Base Command e attivazione NVIDIA AI Enterprise.
- Verifica funzionale e collaudo (job AI di test, validhe GPU/NVLink/NVSwitch, throughput di rete).
- Documentazione (runbook, piani di aggiornamento, report collaudo) e formazione operativa minima sullo stack DGX (amministratori e utenti).

Collaudo e prove di accettazione

- Riconoscimento hardware di 8× H200 con 141 GB ciascuna e NVSwitch operativo (topologia all-to-all) – test nvidia-smi topo e strumenti NVIDIA.
- Prestazioni baseline: benchmark AI (FP8/FP16) su più GPU fino a saturare NVLink/NVSwitch; verifica memoria aggregata 1.128 GB.
- Rete: validazione 10× ConnectX-7 400 Gb/s, misure banda fino a 1 TB/s bidirezionale aggregata (test multistream).
- Software: avvio e accesso a Base Command, deploy container NVIDIA AI Enterprise (es. Triton/NeMo) e job DGX OS previsti.
- Supporto: consegna prove di attivazione supporto 3 anni (seriale registrato, livello di servizio, portale).

Conformità, sicurezza e gestione

- Conformità alle specifiche NVIDIA DGX e normative di sicurezza elettrica/termica per datacenter.
- Gestione out-of-band via BMC; aggiornabilità DGX OS e firmware tramite canali NVIDIA.

Documentazione di offerta (obbligatoria)

- Scheda tecnica ufficiale DGX H200 (NVIDIA Datasheet).
- Descrizione software incluso (Base Command, NVIDIA AI Enterprise) e politiche di supporto.
- Dichiarazione requisiti elettrici/termici e piano di installazione.

3.2 NODI CONTROL PLANE / BCM / RUN:AI

Caratteristiche minime (per nodo):

- **Chassis**: **1U**, **single-socket**; supporto **Intel Xeon 6300 series** o **Xeon E-2400**; **DDR5**
- **Rete**: **2× 1 GbE** onboard (minimo) + **NIC 25** dedicata per control plane/etcd e gestione out-of-band
- **Almeno 32 gb di memoria ram**
- **Espandibilità**: bay frontali **8× 2.5"** / **4× 3.5"**, controller **PERC/HBA**;
- **Sicurezza e gestione**: architettura **Cyber Resilient**, **Secure Boot**, **TPM 2.0**, **OpenManage**.
- **3 anni supporto On Site NBD erogato dal vendor**

3.3 NETWORKING

Si richiede la fornitura, installazione

- N. sufficiente di SFP trancevier almeno da 10Gb compatibili con switch NVIDIA SN3420

3.4 PDU

- N.4 Vertiv PowerIT Rack PDU, Switched (Unit Level), EC, 1U,input IEC 60309 230V 32A, combi outlets (12)C13 or C19

3.5 SOFTWARE DI GESTIONE/ORCHESTRAZIONE

Si chiede la fornitura, installazione, configurazione dei seguenti software di gestione, orchestrazione, provisioning di risorse di calcolo.

- **NVIDIA Base Command Manager (BCM)**
    - Funzioni: **provisioning cluster**, **workload management**, **monitoring**, integrazioni container/K8s e mission control; manuali e guide ufficiali disponibili.
- **Kubernetes**
    - Distribuzione open o enterprise (es. **OpenShift** o K8s upstream) con **operator NVIDIA** per GPU/network; integrazione con **Run:ai** e **BCM**
- **NVIDIA Run:ai**
    - Piattaforma Kubernetes-native per orchestrazione AI con pooling dinamico GPU, fractional GPU, preemption, quote/policy e multi-tenant;
- **NVIDIA AI Enterprise**
    - Suite **cloud-native** con **NIM/NeMo microservices**, **Triton**, driver/operator K8s; licensing **per-GPU** e supporto **Business Standard/Critical**; branch **Production/LTS** per stabilità API.

4. SERVIZI RICHIESTI

Si richiedono i seguenti servizi professionali.

4.1 Installazione e Configurazione

- **Racking, cablaggi e alimentazione** di tutti gli apparati oggetto di fornitura
- **Installazione** di DGX H200 e nodi storage e di gestione, configurazione **NVLink/NVSwitch** e fabric Ethernet.
- **Deploy** di **Nvidia Base Command Manager**, **Kubernetes** (control plane/worker), **NVIDIA Run:ai** e componenti **NVIDIA AI Enterprise** (operator, runtime, servizi NIM).
- **Configurazione NFS** su server Storage NFS (RAID, export, sicurezza, performance, mount ottimizzati per AI pipeline).

4.2 Sviluppo Use Case su LLM

Dovranno essere sviluppati i seguenti use case:

Obiettivi

Realizzare una pipeline LLM end-to-end per:

- ingest di documenti scientifici, brevetti, paper, report tecnici
- indicizzazione semantica con vector database
- generazione risposte citate (Grounded Answer)
- interfaccia demo web (dashboard o micro-app)
- capacità di esecuzione on-prem su DGX H200

Componenti obbligatori

- Triton Inference Server
- modello LLM open-weight ottimizzato per H200
- Vector DB GPU-ready
- Demo finale eseguibile dall’utente

Deliverable

- codice di pipeline RAG

- configurazione Triton e scheduler Run:AI
- dashboard dimostrativa
- documentazione tecnica
- **Porting** e **fine-tuning** di modelli (es. Llama/Nemotron) con pipeline **RAG/inference** su H200/DGX.
- **Ottimizzazione** di throughput e **schedulazione GPU** con Run:ai (batch, preemption, fractional GPU).

Criterio di valutazione sviluppo Use case su LLM

Sarà oggetto di valutazione tecnica la descrizione dettagliata nel documento di offerta delle modalità di sviluppo degli use case.

La valutazione dello sviluppo del modello LLM avverrà sulla base dei seguenti criteri tecnici, indipendentemente dall'effort dichiarato.

Architettura: Completezza pipeline RAG. Punteggio variabile da 0–10
Stack NVIDIA: Coerenza con DGX / Run:ai / Triton. Punteggio variabile da 0–10
Performance: Approccio a tuning e scaling. Punteggio variabile da 0–5
Demo: Usabilità e completezza demo. Punteggio variabile da 0–10

Lo sviluppo LLM si intende accettato a seguito della dimostrazione di:

- esecuzione della pipeline RAG su DGX H200;
- ingest di un set documentale fornito dal committente;
- risposta a query con citazione delle fonti;
- utilizzo delle GPU tramite NVIDIA Run:AI visibile da dashboard.

Lo sviluppo dei casi d'uso LLM è da intendersi a responsabilità di risultato rispetto ai deliverable richiesti.
Non sono ammesse offerte basate esclusivamente su giornate uomo o effort senza descrizione architetturale e funzionale.

4.3 Formazione

Dovrà essere fornita formazione agli utenti per l'utilizzo dei software di orchestrazione e provisioning:

- **NVIDIA Base Command Manager** (admin e utenti),
- **Run:ai**, **NVIDIA AI Enterprise** (operatori, NIM).

Nell’offerta tecnica dovrà essere presentato il programma di formazione che sarà oggetto di valutazione. Dovrà essere dimostrata la partecipazione da parte di chi eseguirà la formazione a corsi di formazione/certificazione sugli specifici software sopra indicati con data della partecipazione.

Il programma di formazione sui software NVDIA BCM e NVIDIA Run AI sarà oggetto di valutazione premiale secondo i parametri di seguito indicati:

- Attinenza del programma rispetto agli scopi del progetto
- Esperienza del docente sulle materie oggetto di formazione
- Accesso a materiali del produttore dei software (vendor).

4.4 Collaudo

L’Accettazione della fornitura sarà soggetta all’esecuzione dei seguenti test di collaudo che dovranno essere effettuati dall’operatore economico in presenza della stazione appaltante.

- **Test funzionali**: creazione cluster, provisioning job, logging/telemetry, resilienza control plane.
- **Benchmark AI**: inferenza/training LLM su DGX H200 (test di latenza/throughput) e nodi GPU.
- **Networking**: throughput 25/100 GbE, RoCE/NVMe-oF, telemetria WJH/INT.
- **Storage NFS**: IOPS/throughput, failover (se previsto), integrazione con K8s CSI.

Checklist di Collaudo Infrastruttura AI/HPC

1. Collaudo Hardware

- Verifica installazione fisica e cablaggio di tutti i server (DGX H200, compute node, Nodo Storage NFS, Nodi di gestione (head node, control plane), Netoworking 100 GbE /100 GbE)
- Accensione e POST senza errori di tutti i nodi
- Rilevamento e funzionamento di tutte le GPU NVIDIA H200
- Verifica memoria installata (RAM, HBM, storage NVMe)
- Test ridondanza alimentazione e ventole
- Verifica connessioni di rete 25/100 GbE (link up, velocità, bonding/LACP)

2. Collaudo Storage

- Verifica configurazione RAID
- Test export NFS e montaggio da compute node e Nvidia DGX H200
- Benchmark throughput/IOPS NFS (rispetto ai requisiti minimi)
- Verifica snapshot/backup

3. Collaudo Networking

- Verifica configurazione e reachability switch (routing, VLAN, VXLAN)
- Test throughput 25/100 GbE tra nodi (iperf3 o equivalente)
- Verifica RoCEv2 e GPUDirect RDMA
- Test telemetria WJH/INT e logging eventi critici

4. Collaudo Software e Orchestrazione

- Deploy e avvio cluster Kubernetes (control plane e worker)
- Installazione e configurazione NVIDIA Base Command Manager
- Deploy e test NVIDIA Run:ai (allocazione job, pooling GPU, preemption)
- Installazione NVIDIA AI Enterprise (operator, runtime, NIM, Triton)
- Verifica accesso e funzionalità dashboard BCM, Run:ai, K8s

5. Collaudo Funzionale AI/ML

- Esecuzione job di training LLM su server NVIDIA DGX H200 (benchmark throughput/token/s)
- Esecuzione job di inferenza su compute node GPU e DGX
- Test fine-tuning modello open (es. Llama, Nemotron)
- Test multi-tenant e scheduling GPU (Run:ai)
- Test pipeline RAG e inferenza distribuita

6. Collaudo Sicurezza e Gestione

- Verifica hardening OS e K8s (RBAC, auditing, segregazione reti)
- Test accesso bastion e logging centralizzato
- Verifica funzionalità iDRAC/OpenManage su tutti i server
- Test Secure Boot, TPM 2.0, Silicon Root of Trust

7. Collaudo Documentazione

- Consegna runbook operativo e diagrammi di rete/cablaggio

- Inventario seriali e BOM aggiornato
- Report collaudo con risultati test e parametri accettazione

Requisiti di Consegna e Documentazione

- **Piano di progetto** con Gantt: fasi di consegna, installazione, configurazione, collaudo, hand-over.
- **Runbook** operativo (BCM/K8s/Run:ai/NFS/monitoring).
- **Diagrammi** di rete e cablaggio; indirizzamenti; inventario seriali.
- **Report collaudo** con risultati dei test e parametri accettazione.

5. Garanzia, Supporto e SLA

- **Hardware**: garanzia minima **36 mesi** on-site per server/switch/DGX, con parti e manodopera. Si richiede garanzia erogata direttamente dal produttore degli apparati.
- **Software**:
    - Garanzia sul sistema operativo Nvidia DGX OS erogata dal vendor
    - licenze **NVIDIA AI Enterprise per-GPU** con supporto **Business Standard** incluso per 3 anni (subscription);
    - licenze **NVIDIA Run AI** con supporto **Business Standard** incluso per 3 anni (subscription);

6. Criteri di Valutazione delle Offerte

- **Aderenza tecnica** ai requisiti minimi (compute, DGX, storage, networking, software/licenze).
- **Prestazioni e scalabilità** dimostrabili (benchmark/documentazione).
- **Servizi e tempi** (installazione, configurazione, collaudo, formazione).

7. Requisiti di presentazione offerta

Dovrà essere predisposta un’offerta tecnica dettagliata contenente la descrizione della soluzione offerta. Oltre alla descrizione e dimostrazione del rispetto di tutti i criteri minimi indicati dal capitolato, dovranno essere illustrati i seguenti macro punti, oggetto di valutazione premiale:

1.CARATTERISTICHE DELLA SOLUZIONE OFFERTA

1.1 Aderenza alla tipologia di soluzione indicata in capitolato;

La proposta tecnica dovrà contenere

- **schema architetturale** con disegni e distinta base (BOM) dettagliata.
- **Cronoprogramma** lavori e **piano di test**.
- **Elenco licenze** con durata e SLA di supporto.
- **Capitolato tecnico** risposta punto-punto (matrix di conformità).

1.2 Unicità di produttore per Server GPU, Networking, Software di Orchestrazione e provisioning, erogazione supporto su Hardware e Software;

Criterio di valutazione: tabellare

1.3 Qualità del progetto complessiva;

Criterio di valutazione: saranno oggetto di valutazione i seguenti parametri:

- Affidabilità delle soluzioni proposte;
- Semplicità di utilizzo delle tecnologie;
- Armonicità delle tecnologie impiegate;
- Modalità di esecuzione dei servizi di integrazione con l’infrastruttura esistente;

2. Progetto creazione Use Case LLM

Dovrà essere illustrata nell’offerta tecnica la proposta relativa al progetto di creazione di Use Case LLM. La chiarezza, completezza ed efficacia del progetto saranno oggetto di valutazione secondo i seguenti parametri:

- 2.1 Qualità complessiva delle use case LLM realizzato secondo i criteri indicati
  - Architettura: Completezza pipeline RAG. Punteggio variabile da 0–10
  - Stack NVIDIA: Coerenza con piattaforma NVIDIA DGX /NVIDIA Run:AI / Triton. Punteggio variabile da 0–10

- Performance: Approccio a tuning e scaling. Punteggio variabile da 0–5
- Demo: Usabilità e completezza demo. Punteggio variabile da 0–10

Modalità installazione, configurazione e formazione

3.1 Modalità di installazione e configurazione dell'infrastruttura in risposta ai requisiti di capitolato

L'offerta tecnica dovrà riportare una descrizione delle modalità di installazione, configurazione ed integrazione adottate. Saranno oggetto di valutazione premiale il livello di dettaglio fornito e l'efficacia delle soluzioni adottate nonché le tempistiche proposte.

3.2 Programma di formazione su infrastruttura AI e moduli

Nell'offerta tecnica dovrà essere presentato il programma di formazione che sarà oggetto di valutazione. Dovrà essere dimostrata la partecipazione da parte di chi eseguirà la formazione a corsi di formazione/certificazione sugli specifici software sopra indicati con data della partecipazione.

Il programma di formazione sui software NVDIA BCM e NVIDIA Run AI sarà oggetto di valutazione premiale secondo i parametri di seguito indicati:

Attinenza del programma rispetto agli scopi del progetto

Esperienza del docente sulle materie oggetto di formazione

Accesso a materiali del produttore dei software (vendor).

CERTIFICAZIONI AZIENDALI ATTINENTI ALL’EROGAZIONE DEI SERVIZI

4.1 Certificazioni aziendali attinenti all’erogazione dei servizi

Saranno oggetto di valutazione il possesso delle seguenti certificazioni aziendali, a comprova della qualità della modalità dei servizi offerti.

- Certificazione ISO 9001. Possesso certificazione: 1 punti; mancato possesso: 0 punti
- Certificazione ISO 27001. Possesso certificazione: 1 punti; mancato possesso: 0 punti"

Tabella punteggi tecnici

	ELEMENTI E SUB-ELEMENTI DI VALUTAZIONE	D/Q		PESO PONDERALE
1,	**CARATTERISTICHE DELLA SOLUZIONE OFFERTA**		Sub-peso ponderale	35
	1.1 Aderenza alla tipologia di soluzione indicata in capitolato	D	18	
	1.2 Unicità di produttore per Server GPU, Networking, Software di Orchestrazione e Provisioning, erpgazione supporto su Hardware e Software	T	7	
	1.3 Qualità del progetto complessiva	D	10	
2,	**Progetto creazione Use Case LLM**		Sub-peso ponderale	35
	2.1 Qualità complessiva dello use case LLM realizzato secondo i criteri indicati - Architettura: Completezza pipeline RAG. Punteggio variabile da 0–10 - Stack NVIDIA: Coerenza con piattaforma NVIDIA DGX /NVIDIA Run:AI / Triton. Punteggio variabile da 0–10 - Performance: Approccio a tuning e scaling. Punteggio variabile da 0–5 - Demo: Usabilità e completezza demo. Punteggio variabile da 0–10	D	35	
3,	**Modalità installazione, configurazione e formazione**		Sub-peso ponderale	8
	3.1 Modalità di installazione e configurazione dell'infrastruttura in risposta ai requisiti di capitolato	D	4	
	3.2 Programma di formazione su infrastruttura AI e moduli	D	4	
4,	**CERTIFICAZIONI AZIENDALI ATTINENTI ALL'EROGAZIONE DEI SERVIZI**		Sub-peso ponderale	2
	4.1 Certificazioni aziendali attinenti all'erogazione dei servizi -Certificazione ISO 9001. Possesso certificazione: 1 punti; mancato possesso: 0 punti -Certificazione ISO 27001. Possesso certificazione: 1 punti; mancato possesso: 0 punti	T	2	
			TOTALE PUNTEGGIO	80