CAPITOLATO DI GARA

Fornitura infrastruttura AI/HPC basata su GPU – Compute, Storage, Networking e Software di orchestrazione **da impiegare nell'ambito del progetto denominato "Trasferimento tecnologico Intelligente per il Digital hEalth" - acronimo TIDE – PIANO NAZIONALE PER GLI INVESTIMENTI COMPLEMENTARI AL PIANO NAZIONALE DI RIPRESA E RESILIENZA, LINEA DI INVESTIMENTO PNC-I.1 - PROGRAMMA PNC-DM MEF 15 LUGLIO 2021.**

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.

Saranno valutate anche configurazioni complessive equivalenti o migliorative.

2. Oggetto della Fornitura

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

- **Sistema di calcolo NVIDIA DGX H200**
- **Nodo storage NFS**
- **Nodi per control plane**, **Nvidia Base Command Manager** e **Nvidia Run:ai**
- **Networking 25 GbE e 100 GbE** per integrazione storage/compute fabric.
- **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** (Sistema di calcolo basato su piattaforma Nvidia DGX/HGX H200).
    - **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.
- 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)
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

Component	Description
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 STORAGE NFS

Sarà richiesta la fornitura, installazione e configurazione di un sistema che svolgerà la funzione di storage NFS.

Caratteristiche minime Storage NFS:

Fornitura, installazione e messa in esercizio di un sistema di **storage NFS** ad alte prestazioni, idoneo a esportare volumi NFS v4/v4.1/v4.2 verso nodi di calcolo, con adeguati livelli di **affidabilità, scalabilità e gestibilità** enterprise.

Requisiti di capacità e prestazioni (minimi)

- **Spazio RAW**: **≥ 23 TB RAW** su supporti **NVMe Gen5** destinati ai dati applicativi.
- **CPU / Core**: **≥ 2 CPU x86-64** di classe datacenter con **≥ 16 core fisici ciascuna** (**≥ 32 core** complessivi; **≥ 64 thread**).
- **Memoria di sistema**: **≥ 256 GB** di RAM **DDR5** (profilo prestazionale; frequenza non inferiore a **6400 MT/s**).
- **Banda di rete aggregata** lato dati: **≥ 250 Gb/s** lordi complessivi, ottenibili tramite porte 10/25/100 GbE
- **Sottosistema boot**: sottosistema dedicato al sistema operativo **separato dai dischi dati**, con **2× M.2 NVMe** in **RAID 1** (**≥ 480 GB** ciascuno) gestiti da controller hardware dedicato, in modo da liberare completamente gli slot dati frontali.
- **RAID**: non è richiesto RAID hardware sui dischi dati NVMe “direct”; eventuali politiche di protezione potranno essere implementate a livello software/file system secondo best practice del fornitore.

Requisiti di rete (data & management)

- **Dati (front-end/back-end)**:
  - **≥ 1 NIC OCP 3.0** con **≥ 4 porte 10/25 GbE SFP28**.
  - **≥ 1 NIC PCIe** con **≥ 2 porte 100 GbE QSFP56**.
- **Gestione**: interfaccia di **management out-of-band** con **API Redfish**, **KVM-over-IP**, **media virtuali**, logging/telemetria e gestione allarmi; supporto cifratura TLS e autenticazione.

Requisiti di affidabilità e sicurezza

- **Alimentazione**: **≥ 2 PSU hot-plug** in **ridondanza (1+1)**, **potenza ≥ 1500 W** ciascuna, efficienza **80Plus Titanium** o superiore.
- **Raffreddamento**: ventilazione ad **alte prestazioni** idonea a carichi continuativi su NVMe Gen5 (“high performance” fans o equivalente).

- **Sicurezza di piattaforma**: **Secure Boot UEFI** abilitabile, catena di avvio verificata, protezioni a livello firmware e supporto a **Root-of-Trust** hardware o equivalenti.
- **Configurazione BIOS**: modalità **UEFI/GPT**; profilo energetico configurabile (es. “power saving”) senza impatto sulle soglie minime di prestazione richieste.

Requisiti di espandibilità e I/O

- **Chassis**: rack **2U** con **backplane NVMe Gen5** per **≥ 16** slot **E3.S** complessivi (occupati inizialmente da **≥ 6** SSD come al § 3).
- **PCIe**: disponibilità di **≥ 4 slot PCIe Gen5 x16** full-height per future espansioni; **≥ 2 slot OCP 3.0** (almeno uno x16) per networking modulare.
- **Cavi/alimentazione rack**: fornitura di **≥ 2 cavi** C13/C14, **2 m**, **10 A**.

Requisiti software e protocollo

- **Protocollo**: **NFS v4/v4.1/v4.2** con autenticazione e sicurezza conformi alle policy del committente.
- **OS**: sistema operativo **Linux** enterprise (o altro equivalente proposto dall’offerente) supportato dal fornitore del server e delle NIC; stack NFS ottimizzato per NVMe Gen5.

Requisiti di gestione e manutenzione

- **BMC/Controller di gestione**: come al § 4 (API Redfish, KVM, virtual media, telemetria).
- **Inventario/Asset**: etichettatura e **asset tag** con seriali e MAC address riportati su chassis e documentazione.
- **Documentazione**: runbook operativo (configurazione NFS, rete, tuning, procedure di aggiornamento firmware) e diagramma porte/slot.

Servizi e garanzia

- **Garanzia**: **≥ 36 mesi** con **assistenza on-site NBD**; inclusa opzione di **supporto avanzato** on-site **36 mesi** erogata dal produttore degli apparati (vendor).
- **Installazione e collaudo**: racking, cablaggio, configurazione rete e NFS, test funzionali (montaggi NFS, throughput, failover alimentazione), consegna report di collaudo.

3.3 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**

3.4 NETWORKING

Si richiede la fornitura, installazione e configurazione di un sistema di networking atto ad esporre la componente Storage Fabric.

Caratteristiche minime Sistema di networking:

- **Porte**: **48× SFP28 10/25 GbE** + **12× QSFP28 40/100 GbE** in **1U** ToR.
- **Capacità di switching**: fino a **~2,4 Tb/s** con **3,58 Bpps**; **latenza costante** cut-through; **RoCEv2** e **VXLAN routing**.
- **Telemetry/OS**: compatibile **Cumulus Linux/SONiC**; **What-Just-Happened (WJH)**, **INT** e **streaming telemetry**.

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 Modalità di pagamento e fatturazione del compenso

I pagamenti verranno effettuati nel termine di 30 giorni dal ricevimento della fattura. La fatturazione avverrà a seguito di esito positivo del collaudo dell'apparecchiatura effettuato

dall’appaltatore. La fattura dovrà essere intestata all’Università degli Studi di Foggia _____ _____ ed inoltrata mediante il sistema di fatturazione elettronica il cui CUU (Codice Univoco Ufficio) è _____.

8 **Controlli sull’esecuzione del contratto**

La Stazione appaltante ha diritto di verificare in ogni momento l’esecuzione delle prestazioni contrattuali. A tal fine, l’appaltatore si impegna a prestare piena collaborazione per rendere possibili dette attività di verifica. La Stazione appaltante evidenzia le eventuali “non conformità” riscontrate rispetto agli obblighi contrattuali in forma scritta e l’appaltatore è chiamato a rispondere alla stazione appaltante, evidenziando le azioni correttive che intende porre in essere per garantire il regolare adempimento delle condizioni contrattuali, ferma restando l’applicazione delle penali previste nel presente capitolato. La Stazione appaltante, ove le “non conformità” evidenziassero oggettivamente i presupposti di gravi inadempienze contrattuali, ha la facoltà di risolvere il contratto.

9 **Subappalto**

In conformità a quanto previsto dall’art. 119 D.lgs. 36/2023 è ammesso il subappalto, alle condizioni, limiti e modalità ivi riportati.

10 **Penali**

In caso di ritardo nell’espletamento delle prestazioni e/o nella consegna dei beni oggetto di fornitura rispetto alle tempistiche del contratto e del capitolato è applicata una penale calcolata in misura giornaliera nella misura del**l'1,5 per mille** dell’ammontare netto contrattuale, senza superare il 10% di quest’ultimo, fermo restando il risarcimento dell’eventuale maggior danno.

11 **Garanzia definitiva**

Ai sensi dell’art. 117 D.lgs. 36/2023, in vista della sottoscrizione del contratto l'appaltatore costituisce una garanzia, a sua scelta sotto forma di cauzione o fideiussione con le caratteristiche e modalità previste dall’art. 106 D.lgs. 36/2023, in misura pari al 10% dell’importo contrattuale. Si applicano le riduzioni previste dall’art. 117 co. 3 D.lgs. 36/2023. La garanzia è prestata per l'adempimento di tutte le obbligazioni del contratto e per il risarcimento dei danni derivanti dall'eventuale inadempimento delle obbligazioni stesse, nonché per il rimborso delle somme pagate in più all'esecutore rispetto alle risultanze della liquidazione finale, salva comunque la risarcibilità del maggior danno verso l'appaltatore. La garanzia cessa di avere effetto solo alla data di emissione del certificato di regolare esecuzione e secondo le modalità previste dall’art. 117 co 8. D.lgs. 36/2023. La Stazione appaltante ha il diritto di valersi della garanzia, nei limiti dell'importo

massimo garantito, per l'eventuale maggiore spesa sostenuta per il completamento delle forniture o nel caso di risoluzione del contratto disposta in danno dell'esecutore.

11 Risoluzione del contratto

La Stazione appaltante può risolvere il contratto ai sensi dell’art. 122 D.lgs. 36/2023, qualora si verifichino le ipotesi, i presupposti e le condizioni ivi riportati. In particolare, il contratto di appalto può inoltre essere risolto per grave inadempimento delle obbligazioni contrattuali da parte dell'appaltatore, tale da compromettere la buona riuscita delle prestazioni. Il direttore dell'esecuzione quando accerta un grave inadempimento ai sensi del primo periodo avvia in contraddittorio con l’appaltatore il procedimento disciplinato dall’articolo 10 dell’allegato II.14 al D.lgs. 36/2023. All’esito del procedimento, la Stazione appaltante, su proposta del RUP, dichiara risolto il contratto con atto scritto comunicato all’appaltatore. Qualora, al di fuori di quanto previsto dal paragrafo precedente, l'esecuzione delle prestazioni sia ritardata per negligenza dell'appaltatore rispetto alle previsioni del contratto, il direttore dell'esecuzione gli assegna un termine che, salvo i casi d'urgenza, non può essere inferiore a 10 (dieci) giorni, entro i quali deve eseguire le prestazioni. Scaduto tale termine, e redatto il processo verbale in contraddittorio, qualora l'inadempimento permanga, la Stazione appaltante risolve il contratto, con atto scritto comunicato all’appaltatore, fermo restando il pagamento delle penali irrogate. In tutti i casi di risoluzione del contratto l'appaltatore ha diritto soltanto al pagamento delle prestazioni relative alle forniture regolarmente eseguite.

12 Recesso dal contratto

Fermo restando quanto previsto dagli artt. 88 co. 4-ter e 92 co. 4 D.lgs. 159/2011, la Stazione appaltante può recedere dal contratto in qualunque momento tenendo indenne l'appaltatore mediante pagamento delle forniture regolarmente eseguite nonché del valore dei materiali utili esistenti in magazzino, oltre al decimo dell'importo delle forniture non eseguite, calcolato secondo quanto previsto dell'allegato II.14 del D.lgs. 36/2023.

13 Certificato di verifica di conformità/regolare esecuzione

Entro 30 giorni dall'ultimazione della fornitura il RUP o, se nominato, il DEC, procederà alla verifica di conformità per certificare il rispetto delle caratteristiche tecniche, economiche e qualitative delle prestazioni di fornitura, nonché degli obiettivi e dei tempi, in conformità delle previsioni e pattuizioni contrattuali. 7 Resta salva la facoltà per la Stazione appaltante, per prestazioni contrattuali di importo inferiore alle soglie comunitarie, di sostituire l'emissione del certificato di verifica di conformità con il certificato di regolare esecuzione emesso, dal DEC e confermato dal RUP.

14 Definizione delle controversie

Per le eventuali controversie che dovessero insorgere sia durante l'esecuzione del contratto che al termine del contratto stesso tra la Stazione appaltante e l'appaltatore, che non si siano potute definire con l'accordo bonario ai sensi dell'art. 211 del D.lgs. 36/2023, è competente in via esclusiva il Foro di Foggia.