



PART C

FORMS TO BE SUBMITTED WITH PROPOSAL



Technical Proposal - Standard Forms

- **Form B1. Technical Proposal Submission Form**
- **Form B2. Profile of lead bidder**
- **Form B3: Co-location & Hosting Facility (Primary +DR)**
- **Form B4: Proven Capability & Delivery Track Record of the Firm/Bidder**
- **Form B5: Proposed AI Compute Infrastructure**
- **Form B6: Sovereign AI Language Model Capability, Data Pipelines & Tiered Inference Services**
- **Form B7: Proposed AI Stack- Sovereignty & Architecture**
- **Form B8:- NAIH User-facing Access Portal and AI Innovation Marketplace**
 - B8-1: -: NAIH User-facing Access Portal
 - B8-2: AI Innovation Marketplace
- **Form B9: Proposed Methodology & Project Management Plan**
 - B9-1: Deployment plan and timelines
 - B9-2: Commissioning & Testing Strategy
 - B9-3: Operations & SLA Framework
 - B9-4: Stakeholder Engagement & Workload Onboarding
 - B9-5: Security & Compliance Approach
 - B9-6: Platform Sustainability & Resource Monetisation Model
 - B9-7: Exit, Migration & Handover Strategy
- **Form B10: Capacity Building & Knowledge Transfer**
 - B10-1: Capacity Building Methodology Approach
 - B10-2: Operational Enablement, Shadowing and Transition Strategy
 - B10-3: Training Infrastructure, Certification & Output Metrics
 - B10-4: Capacity Building & Output Metrics
- **Form B11: Proposed AI Infrastructure Project Team**
 - B11-1: Team Structure & Deployment Model
 - B11-2: Key Personnel Profiles
 - B11-3: Team Capability & Experience
 - B11-4: Consortium / Partner Roles
- **Form B12: Sovereignty & Supply-Chain Assurance**



B1. TECHNICAL PROPOSAL SUBMISSION FORM (TO BE SUBMITTED ON BIDDER'S LETTER HEAD)

[Location, Date]

To:

Manager Procurement

Ignite – National Technology Fund

3rd Floor, TF Complex, 7 Mauve Area, G-9/4

Islamabad, Pakistan

Tel: +92-51- 910 7441 - 46

Cell: +92-306-199-1234

Fax: +92-51- 910 7447

Email: procurement@ignite.org.pk

Sir,

We, the undersigned, offer to provide the services for the **Design, Deployment, Operations, and Maintenance of Sovereign National AI Infrastructure under the National AI Innovation Hub (NAIIH)** in accordance with your Request for Proposal dated [ADVERTISEMENT DATE]. We are hereby submitting our Proposal, which includes this Technical Proposal and Financial Proposal sealed under a separate envelope.

Our Technical Proposal shall be binding upon us subject to the modifications resulting from Contract negotiations, up to expiration of the validity period of the Proposal, which is 180 calendar days from the date of submission.

We understand you are not bound to accept any Proposal you receive.

We remain,

Yours sincerely,

Authorized Signature:

Name and Title of Signatory:

Name of Firm:

Address:



B2. PROFILE OF LEAD BIDDER

S #	Criteria	Bidder Response
1	Profile of the lead bidder: i. Registered age of Company (in years) - Duly verified from certificate of incorporation or equivalent. ii. Legal status iii. Ownership structure iv. Names of Managers/ Owners/ CEO/ Directors/ Partners	
2	Operational Capability i. Location of agency office/sub office ii. Number of branches and their contact numbers. iii. Total number of employees at different branches.	
3	Financial Position i. Name of Banks ii. Certificate of Financial position (to be Issued by relevant Bank) iii. Copy of audited Annual Accounts (of last 3 years) iv. Tax Registration (NTN/STN/FTN)	



B3: COLOCATION & HOSTING FACILITY (PRIMARY + DR)

Total: 90 Marks

Instructions to Bidders:

- Complete all sub-sections (A through E) in full. Incomplete responses will be scored zero for affected criteria.
- Every scored criterion must be substantiated by the specific evidence type listed. Narrative descriptions without supporting documents will not be accepted.
- Both Primary and DR facility details must be provided where sub-sections apply to both sites.
- Power and cooling thresholds are outcome-matched to the hardware declared in Form B5 Section A and Section E.

Section A — Facility Resiliency & Certification

25 marks

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Resiliency tier — Primary	Minimum Uptime Institute Tier III, or TIA-942 Rated-3, or independently-audited equivalent — certified for the constructed/operational facility, not design-only; concurrently maintainable (N+1 power and cooling)	Certificate for the built facility; auditor identity	6
2	Committed availability SLA	≥ 99.982% annual facility availability (Tier III-class), with defined service-credit remedies for breach	Binding SLA with remedy schedule	4
3	Concurrent maintainability	Any single power-path, cooling unit, UPS, or generator maintainable/replaceable with zero IT-load downtime	Maintainability design; single-path-isolation analysis	3
4	Backup power autonomy	N+1 generators with ≥ 48 hours on-site fuel at full IT load, plus a fuel-replenishment contract	Generator/fuel design; replenishment SLA	3



5	UPS ride-through	UPS sized to carry full IT load through generator start/transfer, $\geq$ N+1 redundancy	UPS design and redundancy statement	3
6	Operational track record	Evidence the facility has operated to its tier in production (uptime history / references); facilities in active certification may qualify conditionally with certification as a contract milestone	Operational history; references; certification roadmap if applicable	3
7	Cloud Office Accreditation	The co-location partner must have initiated the Cloud office accreditations under Pakistans Cloud First Policy	Evidence of accreditation process initiation/ documents submission.	3

Section B — Power & Cooling — Outcome-Matched to Delivered Hardware

20 marks

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Power matched to load + headroom	Facility guarantees power capacity $\geq$ the bidder's delivered hardware peak draw + 25% growth headroom, at the per-rack/per-cabinet density the delivered systems require	Power-commitment letter reconciled against Form B5 Section E declared power envelope	6
2	High-density cooling matched to load	Facility cools the delivered hardware at full sustained load with no throttling, supporting the bidder's cooling method — direct-liquid (CDU/facility water loop), rear-door heat exchanger, or high-density air — as the hardware requires	Cooling design; capacity calculation at full delivered load	6
3	Floor loading & form-factor accommodation	Facility accommodates the physical dimensions, weight/floor-loading (including heavy integrated super-node cabinets, often 1,500+ kg), busbar, and cooling-interface of the delivered systems	Floor-loading and form-factor compatibility statement reconciled with Form B5 Section E	3
4	Power usage effectiveness	Design PUE $\leq$ 1.5 (climate-adjusted for Pakistan; $\leq$ 1.4 Preferred), reported monthly	PUE figure and measurement methodology	3



5	Metered power billing	Transparent per-circuit metered power billing; no flat-rate over-provisioning charges	Metering and billing methodology	2
---	-----------------------	---	----------------------------------	---

All thresholds are matched to the bidder's declared hardware in Form B5 — not prescriptive fixed standards.

Section C — Security & Sovereignty

15 marks

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	In-country location (sovereignty gate)	Both Primary and DR facilities physically located within Pakistan; all hosted data and equipment subject solely to Pakistani jurisdiction; foreign-only siting fails this criterion	Site addresses; jurisdiction attestation	4
2	Physical access control	Multi-factor access, 24×7 manned security, mantrap/airlock entry, CCTV with ≥ 90-day retention, full visitor logging	Security design; access-control and CCTV documentation	3
3	Dedicated caged/suite space	Dedicated locked cage or private suite for the deployment, physically segregated from other tenants, at both sites	Floor plan; segregation statement	3
4	Facility certifications	ISO 27001 (or recognised equivalent) for information security; compliance with applicable Pakistani data-protection regulation	Certification documents	3
5	Diverse network entry	≥ 2 physically diverse fibre/carrier entry paths per facility	Connectivity/diversity design	2

Section D — Disaster Recovery / Secondary Site

20 marks

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	DR site tier	Secondary site at minimum Tier III-equivalent (same standard basis as B3 Section A #1), in-country	DR facility certificate	6



2	Geographic separation	DR site separated from Primary by sufficient distance to survive a regional event (flood/seismic/grid), while supporting the required replication — bidder states distance and justifies	Site separation rationale; risk-zone analysis	4
3	Replication & data protection	Inter-site replication with stated RPO; bidder declares replication architecture and bandwidth between sites	Replication design; committed RPO	4
4	Failover capability & RTO	Defined failover process with committed RTO for resuming critical services at the DR site	DR runbook; committed RTO; test methodology	3
5	DR readiness testing	Scheduled DR failover testing ($\geq$ annually) with documented results	DR test plan and reporting commitment	3

Section E — Service & Commercial Terms

10 marks

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Remote-hands / smart-hands	24×7×365 on-site technical support at both sites; Severity-1 response $\leq$ 1 hour with defined escalation	Service catalogue and SLA	3
2	Lease term & escalation caps	Committed hosting term matching the equipment lifecycle, with defined renewal and price-escalation caps	Lease/contract terms	3
3	Onboarding & commissioning	Installation/commissioning plan and timeline for racking, powering, and bringing the infrastructure online at both sites	Onboarding plan with milestones	2
4	Exit & decommissioning	At term end: secure equipment removal and certified data destruction (recognised sanitisation standard), no hostage-data or punitive exit fees	Exit plan; data-destruction standard	2



B4: PROVEN CAPABILITY & DELIVERY TRACK RECORD OF THE FIRM/BIDDER

Total: 120 Marks

Instructions to Bidders:

- Provide a separate reference entry for each scored criterion where a qualifying reference exists.
- All references must be verifiable: provide a named client contact with current email/phone. Unverified entries score zero.
- References may be drawn from the lead bidder or named consortium partners.

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Large-scale AI/HPC infrastructure delivery	≥ 2 reference deployments where the bidder (or named partner) delivered and commissioned accelerator/HPC clusters of comparable scale (state accelerator count and scale of each)	Reference list with scope, scale, dates, client, contactable referee	20
2	Production LLM training / fine-tuning delivery	≥ 1 reference where the bidder (or named partner/ OEM) has successfully trained and deployed a production LLM of 100B+ parameters using the proposed AI platform, ideally the reference should include a low-resource or non-English language	Project description; model/scale; model company certification/ letter; outcome; contactable referee	20
3	Multi-tenant AI platform / AlaaS operation	≥ 1 reference operating a multi-tenant AI platform or AI-as-a-Service offering with tiered users at scale (state user/tenant count)	Reference with tenancy model, scale, uptime achieved, referee	15
4	Marketplace / developer-ecosystem delivery	≥ 1 reference building and operating an AI/software marketplace, API marketplace, or developer ecosystem (catalogue, onboarding, monetisation, SDKs)	Reference with ecosystem scope, adoption metrics, referee	15



#	Requirement	Threshold / Outcome	Substantiation Required	Marks
5	Data-pipeline & governance delivery	≥ 1 reference establishing large-scale data pipelines with multi-organisation data sourcing under a governance/consent framework	Reference with data scope, governance model, referee	15
6	Sovereign / government-scale programme	≥ 1 reference delivering a sovereign, government, or regulated-sector platform with data-residency and sovereignty obligations	Reference with the scope of sovereignty, jurisdiction, and referee	15
7	Local presence & delivery footprint	Demonstrated in-country (Pakistan) delivery, support, or partner presence capable of sustaining the programme; named local entities/partners	Local-presence evidence; partner agreements/LOIs	10
8	Operational maturity (certifications)	Organisational delivery/operations certifications (e.g. ISO 9001, ISO 20000, ISO 27001, or recognised equivalents) for bidder and key partners	Certificates	10

Note: All marks are awarded on a per-criterion binary basis (0 or full marks) except where the criterion specifies a tiered approach. Evidence must be submitted; narrative assertions without documentary support score zero.



Reference Submission Templates

Complete a separate table below for each reference being submitted against criteria 1–7

Reference number (e.g. Ref 9.1-A)	
Criterion reference(s) claimed (e.g. 9.1, 9.3)	
Client / organisation name	
Project / programme title	
Project scope and scale (accelerator count, user count, data volume, etc.)	
Your organisation's specific role	
Project duration (start and end dates)	
Project status (Completed / Ongoing)	
Client contact name, title, email, telephone	
Any consortium partner involved in this reference	

Duplicate the above table as needed for additional references.



B5: PROPOSED AI COMPUTE INFRASTRUCTURE

Total: 220 Marks

Instructions to Bidders:

- Sections A-E must each be completed in full.
- All specifications must be substantiated by the evidence type listed. OEM datasheets, architecture diagrams, and calculation sheets are mandatory for threshold claims.
- Scoring prioritises sustained performance, scaling efficiency, and sovereignty compliance over peak theoretical specifications.
- Individual item scores are tiered (partial marks awarded for meeting lower thresholds) unless otherwise stated.

Section A — AI Accelerators

60 marks

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Accelerator class, quantity & form factor	Brand-new, training-class AI accelerator suitable for both large-scale distributed training and high-throughput inference. ≥ 64 accelerators. Must be a baseboard/module-integrated form factor (SXM, OAM, or proprietary integrated module). Standalone PCIe add-in-card-only configurations not meeting interconnect/ form-factor are not acceptable. Rack-scale / super-node architectures are explicitly permitted.	OEM datasheets confirming training-class designation; architecture documentation; cluster interconnect topology diagram	8
2	Low-precision tensor throughput (per accelerator)	≥ 1.5 PFLOPS low-precision tensor throughput per accelerator at a vendor-declared datatype (FP8, FP6, FP4, INT8, or equivalent). Vendor must state the datatype and whether dense or sparse; sparse claims must also disclose the equivalent dense figure.	OEM datasheet stating throughput, datatype, and dense/sparse basis	4



3	16-bit training compute (per accelerator)	≥ 0.4 PFLOPS dense BF16/FP16 (or vendor-equivalent 16-bit floating-point) per accelerator, ensuring compatibility with established training frameworks dependent on 16-bit precision.	OEM datasheet	4
3.1	Enhanced 16-bit training compute (per accelerator)	≥ 0.8 PFLOPS dense BF16/FP16 (or vendor-equivalent 16-bit floating-point) per accelerator, ensuring compatibility with established training frameworks dependent on 16-bit precision. Scored only if mandatory threshold (3) is met.	OEM datasheet	2
3.2	Premium 16-bit training compute (per accelerator)	≥ 1.2 PFLOPS dense BF16/FP16 (or vendor-equivalent 16-bit floating-point) per accelerator, ensuring compatibility with established training frameworks dependent on 16-bit precision. Scored only if mandatory threshold (3) is met.	OEM datasheet	4
4	Per-accelerator high-bandwidth memory	≥ 96 GB on-package high-bandwidth memory (HBM or vendor-equivalent on-package high-bandwidth memory) per accelerator. Capacity and memory type must be stated; every accelerator in the configuration must meet this threshold.	OEM accelerator datasheets and memory specifications.	4
4.1	Enhanced Per-accelerator high-bandwidth memory	≥ 128 GB on-package high-bandwidth memory (HBM or vendor-equivalent on-package high-bandwidth memory) per accelerator. Capacity and memory type must be stated; every accelerator in the configuration must meet this threshold. Scored only if mandatory threshold (4) is met	OEM accelerator datasheets and memory specifications.	2
4.2	Premium Per-accelerator high-bandwidth memory	≥ 144 GB on-package high-bandwidth memory (HBM or vendor-equivalent on-package high-bandwidth memory) per accelerator. Capacity and memory type must be stated; every accelerator in the configuration must meet this threshold. Scored only if mandatory threshold (4) is met	OEM accelerator datasheets and memory specifications.	4



5	Per-accelerator memory bandwidth	≥ 3.5 TB/s sustained on-package memory bandwidth per accelerator	OEM datasheet	4
6	Aggregate cluster high-bandwidth memory	≥ 6 TB aggregate on-package high-bandwidth memory across the proposed cluster (= per-device capacity $\times$ device count). Host DRAM is explicitly excluded; the calculation must be shown.	Cluster sizing calculation; architecture design; OEM datasheets.	4
6.1	Enhanced Aggregate cluster high-bandwidth memory	≥ 12 TB aggregate on-package high-bandwidth memory across the proposed cluster (= per-device capacity $\times$ device count). Host DRAM is explicitly excluded; the calculation must be shown. Scored only if mandatory threshold (6) is met	Cluster sizing calculation; architecture design; OEM datasheets.	2
6.2	Premium Aggregate cluster high-bandwidth memory	≥ 18 TB aggregate on-package high-bandwidth memory across the proposed cluster (= per-device capacity $\times$ device count). Host DRAM is explicitly excluded; the calculation must be shown. Scored only if mandatory threshold (6) is met	Cluster sizing calculation; architecture design; OEM datasheets.	4
7	Scale-up (intra-node) accelerator interconnect	All accelerators within a node/super-node shall be directly interconnected by a high-bandwidth, low-latency fabric providing ≥ 700 GB/s accelerator-to-accelerator bandwidth per accelerator (any topology: all-to-all, switched mesh, or fabric, or equivalent), supporting direct peer-to-peer memory access between accelerators without host CPU/DRAM staging. This minimum reflects the tight intra-domain coupling required for efficient large-model (tensor- and pipeline-parallel) training at national scale; architectures whose intra-node accelerator-to-accelerator bandwidth falls below this threshold cannot sustain the required collective-communication efficiency and are not acceptable.	OEM/architecture documentation stating per-accelerator aggregate interconnect bandwidth and topology	4



8	Scale-out (inter-node) connectivity	High-speed, RDMA-capable inter-node connectivity (any standards-based or proprietary RDMA transport — e.g., RoCEv2, InfiniBand, UBoE, or vendor-equivalent) sufficient for distributed multi-node training, with direct accelerator-to-network data paths that bypass host CPU staging. Per-node injection bandwidth must be declared.	Architecture documentation; NIC/fabric datasheet; declared per-node injection bandwidth	4
9	Hardware partitioning / multi-tenant isolation	Hardware-enforced partitioning into ≥ 2 isolated instances per accelerator, each with dedicated compute and memory. Minimum partitioning granularity declared.	OEM documentation describing partitioning mechanism and isolation guarantees	2
10	Confidential computing / trusted execution	Accelerator supports hardware-isolated trusted execution environment protecting model weights and data in use from privileged host administrators.	OEM security documentation	2
11	Power & thermal envelope	Per-accelerator and per-node maximum power draw and supported cooling method (air and/or direct liquid) declared.	OEM thermal/power datasheets; rack integration documentation	2
12	Software ecosystem & toolchain	Complete supported software toolchain: compiler, drivers, optimised libraries, and framework integration for PyTorch plus ≥ 1 other major framework. Migration/portability tooling documented.	Toolchain documentation; supported framework matrix; support/lifecycle statement	2
13	Product availability & authorised supply channel	Latest-generation hardware in current production (not end-of-life/end-of-sale), supplied through authorised official OEM channels.	OEM letter of authorisation; product lifecycle/availability statement	4



Section B — AI Interconnect & Networking Fabric

45 marks

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Fabric oversubscription ratio	1:1 (zero oversubscription) bisectional, non-blocking across all switch tiers at full cluster scale	Formal non-blocking / bisectional-bandwidth analysis signed by OEM; topology diagram showing every tier with port and link counts	6
2	Path diversity (resilience)	≥ 4 equal-cost independent (disjoint) paths between any two accelerator endpoints	Routing/topology analysis demonstrating disjoint-path count for worst-case endpoint pair	4
3	Single-switch fault domain	Failure of any single switch isolates ≤ 12.5% of accelerators (≥ 7/8 remain connected)	Fault-domain analysis mapping each switch to its dependent accelerators	4
4	Topology scale ceiling	Proposed topology supports ≥ 100% accelerator growth beyond delivered count without adding switch tiers or re-architecting. Accelerator bandwidth between two nodes should be ≥ 580GB/s.	Scalability analysis stating maximum supported endpoints at the delivered tier count	4
5	Per-node fabric injection bandwidth	≥ 3.2 Tb/s aggregate per accelerator node into the scale-out fabric	NIC/port OEM datasheets; per-node bandwidth calculation	6
6	RDMA losslessness	Zero packet loss for RDMA traffic under full congestion	Protocol documentation; hardware flow-control and congestion-control mechanism description; lossless validation/test method	4
7	Collective-operation efficiency	Sustained ≥ 90% of theoretical bisectional bandwidth on all-reduce at full cluster scale	All-reduce benchmark results at stated scale, OR OEM-validated performance model with all assumptions disclosed	5
8	Inter-node small-message latency	≤ 3 μs end-to-end accelerator-to-accelerator for messages < 4 KB	Latency benchmark report or OEM-validated figure with test conditions stated	4



9	Storage-plane aggregate reach	All accelerator nodes simultaneously reach 100% of aggregate storage bandwidth (non-blocking storage plane)	Storage-fabric design with bandwidth calculation; non-blocking analysis for the storage plane	4
10	Direct accelerator-to-network DMA	Host-CPU/DRAM-bypass DMA path present on inter-node accelerator-to-accelerator transfers	OEM architecture documentation describing the direct-DMA data path	4

Section C — Compute Infrastructure: Host, Control-Plane & Data-Pipeline Tiers

45 marks

a — Accelerator Compute Host Tier (15 marks)

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Host CPU provision	Dual-socket (or vendor-equivalent multi-socket) hosts; ≥ 128 physical cores per accelerator host, aggregate across sockets (x86-64 or Arm)	OEM CPU datasheet; aggregate per-host core calculation	3
2	Host system memory	≥ 1.5 TB DDR5/equivalent per accelerator host; ≥ 8 -channel memory subsystem; $\geq 4,800$ MT/s or equivalent effective bandwidth	OEM memory configuration and population diagram	2
3	Host-to-accelerator I/O	$\geq$ PCIe Gen5 (or vendor-equivalent coherent link) to every accelerator; full host-to-device bandwidth stated	OEM block diagram showing host-to-accelerator lanes/links	2
4	Local NVMe data cache	Aggregate ≥ 240 TB raw NVMe across the accelerator-compute tier, distributed so every accelerator has a local high-speed staging path supporting direct-to-accelerator DMA	OEM drive datasheets; aggregate capacity calculation; DMA-path documentation	2



#	Requirement	Quantified Threshold	Substantiation Required	Marks
5	Boot/OS storage redundancy	Dedicated mirrored (RAID 1) OS media per host, separate from data NVMe	OEM storage configuration	2
6	Host network (data plane)	$\geq 2 \times 100$ GbE (or equivalent) host-plane ports per host, separate from accelerator-fabric NICs	OEM NIC datasheet; port allocation diagram	2
7	Power redundancy	Titanium-grade PSUs, N+N or A/B-grid redundant, sized to peak node draw	OEM PSU datasheet; node power-budget calculation	2

b — Management & Control-Plane Node Tier (15 marks)

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Cluster lifecycle nodes	≥ 2 physical nodes in HA pair; automatic failover ≤ 60 s	OEM HA documentation; failover validation method	3
2	Orchestration control-plane nodes	≥ 3 physical nodes forming a quorum (tolerates ≥ 1 node failure with no loss of control-plane availability)	Control-plane HA architecture; quorum design	3
3	Control-plane CPU & memory	Each node: ≥ 96 physical cores per node; ≥ 512 GB DDR5/equivalent	OEM CPU/memory datasheets; aggregate per-node core calculation	3
4	Control-plane local storage	Each node: mirrored (RAID 1) OS media + ≥ 15 TB raw NVMe data, SED-capable	OEM storage datasheet; encryption-capability statement	2
5	Control-plane network	Each node $\geq 4 \times 25$ GbE (or equivalent) + $\geq 2 \times 1$ GbE dedicated OOB management ports	OEM NIC datasheet; port allocation	2
6	Physical fault separation	Control-plane nodes distributed across ≥ 2 racks and separate PDU circuits	Rack elevation and power-distribution diagram	2



c — Data-Pipeline / General-Purpose Compute Tier (15 marks)

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Dedicated data-pipeline tier	A dedicated, physically separate compute tier (no accelerators required) that tolerates the failure of any single node without halting pipeline processing	OEM server datasheets; tier role statement; single-node-failure tolerance description	3
2	Aggregate pipeline CPU cores	≥ 1,500 physical cores aggregate across the data-pipeline tier (x86-64 or Arm)	Per-node × node-count core calculation; OEM CPU datasheets	3
3	Pipeline memory	≥ 1 TB DDR5/equivalent per data-pipeline node for in-memory dataset transforms	OEM memory configuration	3
4	Working / scratch storage	≥ 500 TB raw flash-class (NVMe or vendor-equivalent) storage for staging and shuffling datasets; sustained aggregate ≥ 100 GB/s read and ≥ 100 GB/s write	OEM storage datasheet; aggregate throughput figure with test/measurement basis	2
5	Pipeline-to-fabric connectivity	Each node ≥ 2× 25 GbE (or equivalent) into the storage/data plane	OEM NIC datasheet; port allocation	2
6	Pipeline-to-storage path	Demonstrated continuous data path from the pipeline tier to the parallel storage system and onward to accelerators (no architectural dead-ends)	End-to-end data-flow diagram	2



Section D — Parallel Storage System

40 marks

#	Requirement	Quantified Threshold	Substantiation Required	Marks
1	Total raw capacity	≥ 1 PB raw across the external parallel storage system	OEM capacity configuration; drive count × capacity calculation	4
2	Hot-tier performance capacity	≥ 200 TB usable (post-data-protection) for the active training working set (shards, checkpoints, KV cache);non-disruptively expandable to ≥ 1 PB	Tier design; usable-capacity calculation explicitly showing erasure-coding/data-protection overhead	4
3glob	Bulk-tier capacity	A capacity tier providing the remainder of the ≥ 1 PB pool for datasets and archived checkpoints, within a tiered namespace shared with the hot tier	Namespace/architecture documentation; capacity breakdown	4
4	Aggregate read throughput	≥ 380 GB/s sustained aggregate read to the accelerator fleet at full cluster scale	Benchmark or OEM-validated aggregate-throughput figure with test conditions	4
5	Aggregate write throughput	≥ 190 GB/s sustained aggregate write (checkpoint-class workload) at full cluster scale	Benchmark or OEM-validated figure with test conditions	3
6	Simultaneous per-node delivery	≥ 6 GB/s sustained storage bandwidth deliverable to every accelerator node concurrently under full all-node load, with no node falling below this floor	Per-node bandwidth analysis under simultaneous all-node read load	3
7	Direct storage-to-accelerator DMA	A direct DMA data path from storage to accelerator memory that bypasses host CPU and system DRAM; any implementation accepted	Architecture documentation describing the direct-DMA path and accelerator-memory target	3
8	Parallel tiered namespace	A single parallel/distributed filesystem presenting tiered namespaces across all tiers and nodes, concurrently mountable by all accelerator and pipeline nodes	Filesystem architecture; concurrent-mount and namespace documentation	3



9	Linear scalability	Capacity and aggregate throughput scale together at $\geq 80\%$ of linear efficiency up to $\geq 2\times$ delivered capacity, with no single-namespace re-architecting required	Scalability analysis; throughput-vs-nodes data or OEM-validated statement	3
10	Multi-tenancy & QoS	Enforcement of per-tenant capacity quotas and bandwidth/IOPS QoS limits, plus RBAC, within the shared namespace	Multi-tenancy and QoS feature documentation	3
11	Data protection	Erasure coding or equivalent, tolerating ≥ 2 concurrent device failures with no data loss and no namespace outage during rebuild	Data-protection scheme; failure-tolerance and rebuild-behaviour statement	3
12	Storage-fabric independence	Storage connects over a standards-based RDMA-capable fabric; storage-fabric switches NOT required to share a vendor with the compute fabric, provided an interoperable solution is demonstrated	Storage-network design; interoperability and single-point-of-support statement	3

Section E — Infrastructure & Equipment Management

30 marks

a — Out-of-Band Management (12 marks)

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Dedicated OOB network	Physically isolated management network (dedicated switches, separate cabling) carrying no AI-compute, storage, or tenant data-plane traffic; redundant (≥ 2 switches, hot-standby)	Network isolation design; switch redundancy statement	2



#	Requirement	Threshold / Outcome	Substantiation Required	Marks
2	Per-node BMC access	Every node (compute, control-plane, storage, pipeline) reachable via dedicated BMC ports (primary + failover) on the OOB network only	Port allocation; OOB connectivity diagram	2
3	Open management API	Open, standards-based RESTful management API (DMTF Redfish or equivalent) on all BMCs for provisioning, power control, inventory, sensor telemetry, and firmware updates	BMC documentation; management-API version statement	2
4	Remote console	HTML5 KVM-over-IP with session recording and audit logging	BMC/KVM documentation	2
5	Hardware telemetry & alerting	Per-node and per-accelerator telemetry (temperature, power, fan, ECC/error counters, utilisation) with threshold-based alerting forwarded to NOC/SOC within ≤ 30 s of breach	Telemetry/alerting documentation; forwarding-latency statement	2
6	Smart PDU management	Per-outlet metered PDUs with remote power cycling, manageable over the OOB network via open protocol (Redfish/SNMP)	PDU datasheet; management-protocol statement	2
7	DCIM / inventory	Automated hardware inventory/CMDB (serials, accelerator/NIC IDs, firmware versions, rack location) updated on hardware-change events	DCIM/CMDB documentation	2

b — Equipment Power (6 marks)



#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Redundant power delivery	Titanium-grade PSUs; N+N or A/B-grid redundancy across all node classes; sized to peak draw with no single PSU/feed as a single point of failure	Power architecture; redundancy and peak-draw calculation	2
2	OOB power independence	All OOB switches and remote-console devices on UPS-protected, electrically separate circuits, maintaining management access during a primary-power event	Power-distribution diagram for the management plane	2
3	Declared power envelope	Per-rack/per-cabinet and total-solution peak power declared and reconciled against the capacity guaranteed in Form B3 (Colocation & Hosting Facility)	Power-budget document	2

c — Thermal & Cooling (Equipment) (6 marks)

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Thermal dissipation	Cooling solution dissipates the full declared per-rack/per-cabinet thermal load with no accelerator/CPU throttling at sustained 100% load; any method (air, direct-liquid, hybrid) accepted	Thermal design; heat-load and cooling-capacity calculation	2
2	Cooling redundancy	Redundant, hot-swappable cooling components such that a single fan/pump/CDU failure does not cause throttling or shutdown	Cooling redundancy statement; failure-mode analysis	2



#	Requirement	Threshold / Outcome	Substantiation Required	Marks
3	Facility integration	Equipment physically compatible with, and deployable within, the floor-loading, dimensional, power, and cooling envelope guaranteed in Form B3. Any non-standard rack/cabinet form factor, busbar, or liquid-cooling interface must be fully specified.	Form-factor, weight/floor-loading, and cooling-interface specification; facility-compatibility statement	2

d — Physical & Hardware Security (6 marks)

#	Requirement	Threshold / Outcome	Substantiation Required	Marks
1	Hardware root of trust	Silicon root of trust + UEFI Secure Boot + TPM 2.0 (or recognised equivalent) on all node classes	Security architecture; component statement	2
2	Signed firmware & lockdown	Cryptographically signed firmware verified at every boot; system lockdown preventing unauthorised configuration drift; SOC alert on attestation failure	Firmware-security documentation	2
3	Physical tamper detection	Chassis intrusion detection logged and alerted via the OOB management plane	Intrusion-detection documentation	2



B6: SOVEREIGN AI LANGUAGE MODEL CAPABILITY, DATA PIPELINES & TIERED INFERENCE SERVICES

Total: 110 Marks

This form evaluates the bidder's capability to provide, deploy, and operate **sovereign AI Language model(s)** under full ownership and control of the Government of Pakistan, suitable for national-scale training, fine-tuning, and inference within NAIH.

Key Definitions

Term	Definition
Sovereign AI Model	A large language model for which the Government of Pakistan holds full intellectual property ownership of the delivered weights, owns the national training corpus assembled under the contract, controls all deployment infrastructure, and retains unrestricted rights to maintain, extend, redistribute, and rebrand the model independently of the original developer — in perpetuity.
Large Language Model (LLM)	For the purposes of this RFP, an LLM is a transformer-based language model with a minimum of 30 billion parameters (dense equivalent or MoE active parameters per forward pass). The target model scale for this engagement is 70 billion parameters.
Open-Weight Model	A model whose trained weight parameters are publicly downloadable. Apache 2.0 and MIT or equivalent are the only base licences acceptable under this RFP for sovereign redistribution and national branding. Licences with volume caps, revenue triggers, mandatory name attribution, or redistribution restrictions do not qualify.
Dense vs. MoE Parameters	Dense models activate all parameters for every token processed. Mixture-of-Experts (MoE) models activate only a subset of parameters per token. For this RFP, parameters per forward pass is the relevant measure for compute and capability comparison. Vendors must declare both figures if proposing an MoE architecture.
IP Assignment	Full legal transfer of intellectual property ownership of model weights to the Government of Pakistan — not a licence. Only an assignment gives the government unrestricted rights to use, modify, redistribute, and rebrand the model without reference to the original developer.
Technology Transfer	The structured transfer of knowledge, skills, training code, data pipelines, and institutional capability to Pakistani engineers and national institutions, enabling Pakistan to develop, maintain, and extend AI models independently after the contract period.



Section A — Model Training & Fine-Tuning Platform (25 marks)

#	Requirement	Threshold / outcome	Substantiation required	Marks
1	Operational training platform	A fully configured, end-to-end training and fine-tuning platform deployed on the procured infrastructure, supporting distributed pre-training, continued pre-training, and fine-tuning (full and parameter-efficient)	Platform architecture; deployment/acceptance plan	8
2	Experiment & lifecycle management	Experiment tracking, dataset/version lineage, model registry, checkpoint management, and reproducible training runs	Tooling documentation	6
3	Evaluation harness	Automated evaluation framework for benchmarking models, including Urdu/local-language and domain-specific evaluation sets	Evaluation methodology; sample eval sets	5
4	Platform handover & independence	Platform fully operable by the in-country team without bidder dependency at contract end; no licensing or technical lock preventing continued independent fine-tuning	Handover plan; dependency/lock-in disclosure	6

Section B — Governed Data Pipelines & Local-Context Datasets (25 Marks)

#	Requirement	Threshold / outcome	Substantiation required	Marks
1	Data-partnership framework	A working framework and signed initial/ data-sharing arrangements with willing Pakistani organisations, operating under MoITT/Ignite governance and consent/legal compliance	Partnership engagement plan; governance/consent model	5



#	Requirement	Threshold / outcome	Substantiation required	Marks
2	Urdu & local-language corpora	Assembly, cleansing, and curation of Urdu language corpora and other local-context datasets (regional languages, domain data) to a stated minimum scale and quality	Dataset specification; scale (tokens/documents) and quality methodology	10
3	End-to-end pipeline operation	Production data pipelines (ingestion → cleansing → de-duplication → PII handling → curation → training-ready) running on the platform, integrated with the B5 Section C compute tier and Section D storage	Pipeline architecture; data-flow documentation	5
4	Data sovereignty & governance	All collected/curated data resident in-country, owned by GoP, with full provenance, audit trail, and consent/withdrawal handling under MoITT/Ignite governance	Data-governance and residency documentation	5

Section C — Sovereign Fine-Tuned LLM (the national model) (35 Marks)

#	Requirement	Threshold / outcome	Substantiation required	Marks
1	Base-model license (sovereignty foundation)	Base model under Apache 2.0, or an equivalent OSI-approved permissive license providing an explicit patent grant, with an unbroken permissive license chain from base weights through every intermediate artefact to the delivered model — such that the resulting model is freely modifiable, redistributable, and commercialisable by GoP with no foreign veto, MAU cap, geographic, field-of-use, or output restriction	Full license chain documentation for base + all derivative artefacts; legal attestation of permissive continuity	10



#	Requirement	Threshold / outcome	Substantiation required	Marks
2	Fine-tuning for Urdu & local context	Model fine-tuned/aligned on the Urdu corpora and local-context datasets of Section B, with demonstrated capability uplift on Urdu/local evaluation sets versus the untuned base	Before/after evaluation results on Urdu/local benchmarks	8
3	Full sovereignty transfer	Complete transfer to GoP of: model weights, training/fine-tuning code, datasets, configs, and documentation — sufficient for GoP to independently retrain, re-fine-tune, and redeploy	Sovereignty transfer package manifest; rights-assignment instrument	7
4	Alignment, safety & local values	Safety alignment, content controls, and bias/appropriateness handling tuned to the Pakistani legal and cultural context	Alignment methodology; safety evaluation	5
5	Model documentation & card	Full model card: architecture, training data summary, capabilities, limitations, eval results, intended-use and licensing statement	Model card document	5

Section D — Tiered Public Inference Service (≥ 20 open-weight models) (25 Marks)

#	Requirement	Threshold / outcome	Substantiation required	Marks
1	Model catalogue (≥ 20)	≥ 20 open-weight models in production inference, stood up in parallel with (not after) the sovereign-model training; catalogue spans sizes/capabilities (general, coding, reasoning, multilingual). These need not be permissively licensed, provided each model's license permits the intended hosted-inference service	Model list with per-model license and intended-tier mapping; deployment evidence	6



#	Requirement	Threshold / outcome	Substantiation required	Marks
2	Four-tier access model	Distinct access tiers — Academia/Researchers, Startups, Public Sector, Private Sector — each with defined eligibility, quota/rate limits, data-handling/residency terms, and SLA. Tier definitions enforced via the multi-tenancy/RBAC of Section 5	Tier specification matrix; enforcement mechanism	6
3	Inference availability & launch	Inference service live and publicly accessible within a stated timeline from contract start (parallel track), with committed uptime SLA	Launch plan with milestone date; uptime SLA	6
4	API & access management	Standards-based API (OpenAI-compatible or documented equivalent per 5.6), self-service onboarding, authentication, per-tier metering and quota enforcement	API and access-management documentation	4
5	Usage metering & reporting	Per-tier, per-model usage metering and reporting to MoITT/Ignite (utilisation, adoption, cost-allocation), aligned with the FinOps capability of 5.16	Metering and reporting specification	3



B7: PROPOSED AI STACK & ARCHITECTURE

Form B7 constitutes the **core technical submission** of the bidder and shall form the primary basis for evaluation of the proposed AI stack, its sovereignty, functional completeness, and readiness for production deployment under the National AI Innovation Hub (NAIHH).

Section A - AI Software Stack Requirements (110 Marks)

#	Requirement	Capability threshold (outcome)	Substantiation required	Marks
1	Accelerator runtime & operator library	A supported low-level runtime and optimised operator/kernel library for the proposed accelerator, exposing a documented programming interface	Runtime/library documentation; supported-version statement	8
2	Training framework	A distributed training framework supporting multi-node data/tensor/pipeline parallelism, with native support for — or a documented supported import path from — PyTorch, and at least one of TensorFlow/ONNX	Framework documentation; parallelism and framework-compatibility matrix	8
3	Model alignment & customisation	Tooling for fine-tuning and alignment (e.g. SFT, RLHF, DPO, or equivalent) and dataset curation/preparation	Tooling documentation	6
4	Inference serving engine	A production serving engine supporting dynamic/continuous batching, concurrent multi-model execution, multi-accelerator/multi-node parallelism, exposing gRPC and HTTP/REST endpoints with a metrics interface	Serving-engine documentation; endpoint and metrics specification	8



#	Requirement	Capability threshold (outcome)	Substantiation required	Marks
5	Inference optimisation	Quantisation (vendor-declared low-precision formats), KV-cache management, and throughput/latency optimisation (e.g. in-flight batching, speculative decoding, or equivalent)	Optimisation feature documentation	6
6	OpenAI-compatible API surface	Inference services exposable via an OpenAI-compatible (or documented equivalent open) API for client portability	API documentation	4
7	Accelerated data-science libraries	Accelerator-native libraries for dataframe/ETL operations, classical ML, and/or graph analytics for pipeline acceleration	Library documentation	4
8	Kubernetes-native (or equivalent) orchestration	The full stack runs on a conformant container-orchestration platform (CNCF-conformant Kubernetes distribution or equivalent); container runtime and accelerator device-plugin integration provided	Orchestration architecture; device-plugin documentation	8
9	Accelerator lifecycle operator	Automated deployment/management of accelerator drivers, runtime, and device plugins across the cluster (driver/firmware lifecycle)	Operator/automation documentation	5
10	High-performance network operator	Automated deployment of RDMA drivers and fabric-management components for the proposed interconnect	Network-automation documentation	6
11	Serverless inference, autoscaling & model versioning	Concurrency/throughput-based autoscaling, scale-to-zero, and versioned model	Autoscaling, versioning, and rollout documentation	6



#	Requirement	Capability threshold (outcome)	Substantiation required	Marks
		management with traffic-splitting/canary rollout and instant rollback		
12	Accelerator fractioning & isolation	Fractional accelerator allocation with hardware- or software-enforced memory isolation between co-located workloads, supporting ≥ 4 isolated workloads per accelerator; mechanism is vendor's choice	Partitioning/QoS documentation; isolation-guarantee statement	5
13	Multi-tenancy & RBAC	Isolated tenant namespaces/virtual clusters with resource quotas, network policy, and role-based access control	Multi-tenancy architecture	5
14	Zero-trust control plane	Control plane operable without inbound open ports to the cluster API (outbound-only / mutually-authenticated), API not exposed to public internet; full user-level audit trail	Security architecture; audit-logging documentation	5
15	Air-gapped operation	Full core operation in an air-gapped/on-premises environment: offline container registry and model/artifact distribution, no dependency on external public-cloud services	Air-gap deployment documentation	8
16	FinOps metering & chargeback	Real-time per-tenant/project utilisation metering (accelerator/CPU/memory) with token-level inference metering and chargeback/cost-allocation APIs	Metering and billing-API documentation	4



#	Requirement	Capability threshold (outcome)	Substantiation required	Marks
17	Support SLA & long-term support	24×7×365 enterprise support with defined Severity-1 response time; a long-term-support branch with stated API-stability duration	SLA terms; LTS/lifecycle policy	6
18	Distributed-training fault tolerance	Automatic checkpointing and resumption of distributed training after a worker/node failure without loss of completed progress; maximum lost-work window stated ($\leq$ last checkpoint interval)	Fault-tolerance/elastic-training documentation; stated lost-work window	4
19	Unified observability	Cluster-wide observability providing metrics, logging, and accelerator/job telemetry through standards-based interfaces (Prometheus-compatible, OpenTelemetry, or equivalent)	Observability-stack documentation; interface specification	4

Section B - Ecosystem Adoption & Software Maturity (30 Marks)

This section evaluates the maturity, breadth, and real-world adoption of the accelerator's software ecosystem and developer base — the factors that most determine whether the platform can be productively used, maintained, and built upon. Each criterion is scored on three levels (Poor / Moderate / Good) against the evidence stated. Marks per level: Poor = 0, Moderate = half, Good = full.

#	Criterion	Poor (0 marks)	Moderate (50% marks)	Good (full marks)	Evidence required	Max
1	Framework & toolchain maturity	Limited/experimental support for major frameworks; frequent porting needed	Supports major frameworks with some gaps; periodic porting/tuning required	Native, day-zero support across major frameworks (training & inference) with maintained, production-grade toolchain	Supported-framework matrix; release/maintenance history; third-party validation	5



#	Criterion	Poor (0 marks)	Moderate (50% marks)	Good (full marks)	Evidence required	Max
2	Developer adoption & community	Small/niche developer base; sparse documentation and community support	Growing developer base; reasonable documentation; some community activity	Large, active global developer base; extensive documentation, forums, and trained talent pool	Evidence of community size, repositories, contributors, available skilled-talent indicators	5
3	Pre-optimised models, libraries & kernels	Few optimised models/libraries; most work must be hand-tuned	Moderate library of optimised models, kernels, and reference implementations	Comprehensive, continuously updated library of optimised models, kernels, and reference solutions available out-of-the-box	Catalogue of optimised models/libraries; update cadence	5
4	Third-party & ISV ecosystem	Minimal third-party/ISV integration or partner tooling	Some third-party tools, integrations, and partner solutions	Broad third-party/ISV ecosystem with mature integrations across the stack	List of integrations, ISV partners, marketplace presence	5
5	Proven production deployments at scale	Few or no comparable large-scale production references	Some production deployments at moderate scale	Multiple, verifiable large-scale production deployments comparable to this programme	Reference deployments with scale, dates, contactable referees	5
6	Migration & interoperability tooling	Little or no tooling to migrate from common ecosystems; high lock-in	Partial migration/portability tooling; some friction	Mature migration and interoperability tooling enabling low-friction portability to/from common ecosystems	Migration tooling documentation; portability evidence	5



B8 – NAIH USER-FACING ACCESS PORTAL AND AI INNOVATION MARKETPLACE

Bidders must provide detailed proposals for both platforms in the tables below. Both must be fully sovereign, deployed within Pakistan, with complete source code ownership vested in the Government of Pakistan and no external dependencies.

B8.1: NAIH User-Facing Access Portal & Developer Platform (20 marks)

Provide details of the proposed user-facing portal supporting onboarding, access management, workload submission and monitoring, and interaction with NAIH services.

#	Requirement	Description	Supporting Evidence / Documents	Marks	Bidder Response
1	Proposed Architecture & Technology approach	Proposed platform architecture, modularity, scalability, and technology stack	Architecture diagram, technology list	4	
2	Key Features & Functionality	User onboarding, authentication, workload submission, dashboards, quota management, experiment tracking, managed developer environments, self-service training/ fine-tuning workspace provisioning and monitoring.	Detailed description + wireframes/screenshots	6	
3	Sovereignty, Security & Compliance	Data residency, IAM, audit logging, deployment sovereignty, security approach	Security architecture, compliance approach description	4	
4	Integration with AI Infrastructure & Software Stack	Approach for integration with compute, orchestration, APIs and AI stack components	Integration methodology and diagrams	3	
5	Deployment Maturity & Handover Readiness	Production readiness, deployment evidence, documentation, source code transfer approach	Client references/screenshots/document list	3	



B8.2: AI Innovation Marketplace

(20 marks)

Provide details of the proposed marketplace platform supporting AI ecosystem participation, innovation programs, and solution catalog management.

#	Requirement	Description	Supporting Evidence / Documents	Marks	Bidder Response
1	Proposed Architecture & Technology approach	Platform architecture, extensibility, modularity, and technology approach	Architecture diagram, technology list	4	
2	Key Features & catalogue management	Searchable catalog, solution showcasing, program/hackathon management, demo days, investor matching, recommendation engine.	Detailed description, Feature list, screenshots, wireframes	6	
3	Governance, Transparency & Compliance	Workflow approvals, audit capability, transparency mechanisms	Governance model and compliance approach	4	
4	Integration with NAIH Access Portal & AI Infrastructure	Integration with portal and infrastructure services	Integration methodology and diagrams	3	
5	Deployment Maturity & Handover Readiness	Deployment Maturity & Handover Readiness	Client references/ screenshots/deployment proof	3	

- All claims regarding deployment status must be supported by verifiable evidence.
- Both platforms must be delivered with **full source code**, configuration files, APIs, and complete documentation.
- No proprietary or externally hosted components that compromise sovereignty will be accepted.

B9 – PROPOSED METHODOLOGY & PROJECT MANAGEMENT PLAN

Bidders shall provide a complete, practical, and realistic lifecycle methodology covering deployment, commissioning, operations, governance, security, onboarding, sustainability, and exit strategy for the Sovereign AI Infrastructure.

Responses must be **implementation-oriented**, with strong technical narrative as the primary evaluation basis. Diagrams and supporting documents are important evidence but shall not override the quality of the technical response.

B9-1: Deployment Plan & Timelines

(15 Marks)

Describe the phased deployment strategy for the 6-month implementation period, including sequencing, key milestones, integration approach, dependencies, and risk management.

Bidder Response

- End-to-end phased deployment including OEM backed LLD (design → Responsibility matrix → procurement → installation → integration → testing → go-live)
- Key milestones and delivery roadmap
- Hardware, software, AI stack, and Colocation deployment (and DR)
- Dependencies, assumptions, and constraints
- Risk identification and mitigation strategy

Supporting Evidence (mandatory)

- Detailed **Gantt Chart**
- High-level deployment architecture diagram

B9-2: Commissioning & Testing Strategy

(10 Marks)

Describe the approach to validation, benchmarking, scalability, disaster recovery, security validation, and acceptance procedures. Acceptance criteria must reference and meet the minimum standards declared in Form B5.

Bidder Response

- Performance benchmarking (training & inference workloads)
- Scalability and stress testing methodology (multi-node / cluster scale)
- Disaster recovery and failover testing (RPO/RTO defined)
- Security validation approach
- Go-live acceptance criteria and readiness process
- Core Component Implementation and Testing (should be done by OEM)

Supporting Evidence (mandatory)

- Testing & Commissioning Plan
- Benchmarking methodology summary



B9-3: Operations & SLA Framework

(15 Marks)

Describe the operational approach ensuring stable, secure, and scalable production operations. KPIs mentioned in TORs section 33 must also be addressed.

Bidder Response

1. Monitoring, Observability & Alerting

- Metrics collection framework (compute, storage, accelerator utilization)
- Centralized logging and observability platform
- Real-time alerting and dashboarding approach

2. Incident Management

- L1/L2/L3 escalation model
- Incident classification and response workflow
- Root cause analysis and resolution tracking

3. Patch, Upgrade & Change Management

- OS, firmware, accelerator driver, and AI stack update process
- Hardware replacement
- Maintenance windows and rollback strategy
- Change approval workflow

4. Multi-Tenant Governance & Resource Isolation

- Tenant separation model (logical / physical where applicable)
- Resource quota management
- Scheduling and allocation policy for AI workloads
- O&M administration capability to manage and maintain resources across assigned tenants

5. SLA Framework & Enforcement

- Availability and uptime commitments
- SLA monitoring methodology
- Breach handling and enforcement mechanism

6. Reporting & Operational Governance

- Monthly operational reports (performance, utilization, incidents)
- Quarterly service reviews with government stakeholders
- Continuous improvement mechanism
- Real-time dashboards for usage analytics and performance monitoring, as well as documented methodology for impact measurement and governance reporting.



Supporting Evidence (mandatory)

- SLA Framework (including targets and enforcement approach)
- Administrator role hierarchy and delegated permission model
- List of key operational policies and procedures
 - Access control usage policies;
 - Security, incident response & risk management;
 - Data handling & compliance;
 - Change/upgrade/config management;
 - Ops & maintenance procedures.
- Sample dashboard architecture/mock-up and impact-measurement methodology note

B9-4: Stakeholder Engagement & Workload Onboarding

(10 Marks)

Describe the onboarding approach for government, academia, startups, and industry users.

Bidder Response

- Stakeholder engagement and workload onboarding model (including community forums/ knowledge base, outreach, helpdesk with defined SLA response time and workshop, data pipelines)
- User onboarding and access provisioning workflow
- Identity integration (SSO / IAM / federation)
- Workload lifecycle support (data ingestion → training → deployment → inference)
- Usage and allocation (including quotas, resource tracking and billing/ chargeback support)
- Feedback collection systems (dashboard)

Supporting Evidence (mandatory)

- Onboarding workflow diagram
- Sample user journey
- Dashboard architecture

B9-5: Security & Compliance Approach

(10 Marks)

The bidder shall describe their approach to securing the AI infrastructure platform across all layers. Responses must be supported by the evidence listed under each area.

Bidder Response

- Data protection (encryption at rest/in transit, Pakistan data residency enforcement, data classification and handling policy)



- Logging, audit, and monitoring framework (including threat detection capability, real-time alerting on anomalous access or activity)
- AI security controls, including model theft/extraction protection, dataset poisoning mitigation, adversarial attack defenses, and model misuse/API exploitation controls

Supporting Evidence (mandatory)

- Security architecture diagram
- Summary of security controls (IAM, data protection, logging, AI risks)

B9-6: Platform Sustainability & Resource Monetization Model

(10 Marks)

Describe the proposed model for platform cost recovery, resource pricing, and sustainable access across user categories.

Bidder Response

- User tiering model (government, academic, startup, commercial etc.)
- Quota allocation and usage governance
- Pricing and cost recovery approach (subsidised + commercial mix)
- Capacity utilization, optimization, and demand balancing strategy
- Governance model for pricing and concessional access approvals
- Revenue reinvestment strategy

Supporting Evidence (mandatory)

- User tier & pricing summary
- Indicative 3-year cost recovery projection (assumption-based)
- Governance framework description

B9-7: Exit, Migration & Handover Strategy

(10 Marks)

Describe how full operational independence will be achieved at the end of contract

Bidder Response

- Knowledge transfer milestones
- Reduction of operational dependency strategy
- Final handover criteria
- Migration readiness to alternate infrastructure
- Infrastructure-as-Code (IaC) / automation frameworks — all configs delivered as reproducible code

Supporting Evidence (mandatory)

- Handover roadmap and checklist



- Knowledge transfer plan
- Infrastructure-as-Code (IaC) Delivery Plan

B10 – CAPACITY BUILDING & KNOWLEDGE TRANSFER

B10-1: Capacity Building Methodology Approach

(15 Marks)

The bidder shall propose a comprehensive capacity building and knowledge transfer methodology for NAIH, covering the full lifecycle of training, operational enablement, and transition to independent operations.

#	Requirements	Evidence Required	Description
1	Overall training strategy	Detailed written strategy document (max 5 pages) including sample curriculum, & high-level roadmap diagram,	
2	Target audience (TOT, Relevant Ignite staff, capability + external ecosystem)	Clear breakdown table showing number of participants per category and role	
3	Skill levels to be developed	Skill matrix / competency framework table (beginner, intermediate, advanced)	
4	Training delivery approach (hands-on, labs, etc.)	Training delivery methodology with mode-wise breakdown	
5	Tools, platforms & environments to be used	List of tools/platforms with justification and integration with proposed AI stack	
6	Approach for real-world operational exposure	Plan showing how trainees will access live NAIH infrastructure for practical exposure	
7	Approach for continuous learning & upskilling	Post-training upskilling and refresher program plan	

B10-2: Operational Enablement, Shadowing and Transition Strategy

(15 Marks)

The bidder shall describe their approach for enabling trained resource/ teams to progressively take over operations, including knowledge transfer, joint operations, and full independence.

#	Requirements	Evidence Required	Description
1	Approach for knowledge transfer	Knowledge transfer methodology document + topics matrix	
2	Shadowing / joint operations model	Detailed shadowing plan with roles, duration, and weekly schedule	
3	Transition phases & timelines	Phased transition timeline (Gantt chart or table with clear start/end dates)	

4	Reduction of dependency on bidder	Declining dependency schedule (table or graph showing % responsibility shift over months)	
5	Criteria for operational handover	Measurable handover readiness criteria checklist (with success indicators)	
6	Risk mitigation during transition	Transition risk register with mitigation measures and responsible parties	

B10-3 Training Infrastructure, Certification & Output Metrics

(15 Marks)

The bidder shall define the training ecosystem, certification approach, and measurable outcomes of the capacity-building program.

#	Requirements	Evidence Required	Bidder Response
1	Training infrastructure (labs, sandbox, etc.)	Description + architecture diagram of proposed training labs / sandbox environment	
2	Access to real compute resources	Access policy and schedule showing how trainees will use the actual AI compute cluster	
3	Practical training approach (labs, simulations, etc.)	List of hands-on labs / practical modules with objectives	
4	Certification methodology	Certification framework document (including issuing authority and validity)	
5	Assessment & evaluation mechanisms	Assessment methodology + sample evaluation rubrics / test formats	
6	Documentation & knowledge assets to be delivered	List of all documents/runbooks/manuals to be delivered with delivery timeline	
7	National talent development approach (beyond NAIH)	Strategy for external ecosystem (academia, startups, industry) with outreach plan	

B10-4 Capacity Building Output Metrics

(15 Marks)

The bidder must commit to the following:

#	Metric	Evidence Required	Target to be filled by the bidder
1	No. of staff trained	Supporting justification	
2	Training of Trainers (ToT)	Strategy for ToT and pathway to further trainings for capability development	
3	No. of certified professionals	Certification pathway explanation	

RFP Ignite National AI Innovation Hub – Sovereign AI Infrastructure



4	No. of training hours	Breakdown by module / phase	
5	No. of lab sessions	List of proposed lab sessions	
6	% staff ready for independent ops	Definition + measurement methodology for “ready”	
7	No. of external ecosystem participants trained	Outreach and selection criteria/ justification for total trained	



B11 – PROPOSED AI INFRASTRUCTURE PROJECT TEAM

The bidder shall propose a **complete team structure** for deployment, operations, security, and management of the AI infrastructure

B11-1: Team Structure & Deployment Model

(10 Marks)

Role / Function	No. of Resources	Full-Time / Part-Time	Key Responsibilities	Location
A. Core Management & Architecture				
AI/HPC Architect		FT		
AI Platform Lead		FT		
LLM / Foundation Model Lead		FT		
B. Engineering & Technical Staff				
Infrastructure Engineer- Compute		FT		
Infrastructure Engineer- Storage & Network		FT		
Infrastructure Engineer- MLOps & Monitoring		FT		
C. Capacity Building & Enablement				
Capacity Building & Training Lead		FT		
D. Part-time Staff				
(Specify Position)		PT		
(Specify Position)		PT		
E. On-call / Specialist Consultants				
DR/ BCP Specialist		On-call		
AI Safety & Alignment Specialist		On-call		
Legal & Regulatory Consultant		On-call		
(Add rows where required)				
TOTAL TEAM (all categories)				

Mandatory Attachments: Organizational Chart showing all roles and reporting lines. Role mapping to project phases (deployment vs operations)



B11-2: Key Personnel Profiles

(20 Marks)

Name	Role	Years of Experience	Relevant Projects	Certifications	Availability (%)
A. Regular Full-time Staff					
	AI/HPC Architect				100%
	AI Platform Lead				100%
	LLM / Foundation Model Lead				100%
	Infrastructure Engineer- Compute				100%
	Infrastructure Engineer- Storage & Network				100%
	Infrastructure Engineer- MLOps & Monitoring				100%
	Capacity Building & Training Lead				100%
B. Part-time Staff					
	(PT Position 1)				
	(PT Position 2)				
C. On-call / Specialist Consultants					
	DR/ BCP Specialist				
	AI Safety & Alignment Specialist				
	Legal & Regulatory Consultant				

(Add rows in form where necessary.)

The bidder shall ensure continuity of key personnel throughout the contract period and propose a mitigation plan for resource unavailability.

Mandatory Attachments:

- Signed CVs (template provided)
- Certifications
- Experience references (with project names)



B11-3: Team Capability & Experience

(10 Marks)

Capability Area	Minimum Experience	Experience Description
AI / HPC Architecture	≥ 5–7 years	
LLM / Foundation Model	≥ 5–7 years	
Distributed training systems	≥ 5–7 years	
AI platform / MLOps	≥ 3–5 years	
Infrastructure operations (24/7)	≥ 5–7 years	
Capacity Building	≥ 3 years	
Systems Engineers	≥ 3 years	

The bidder shall demonstrate that the proposed team possesses the required expertise in AI infrastructure design, deployment, and operations



B11-4: Consortium / Partner Structure

(10 Marks)

Partner	Role	Scope of Work	Key Responsibilities	Accountability
Lead Bidder				
Co-Location Partner				
Any other partner				

Please add rows where applicable.

The bidder shall share a mandatory signed declaration on the letterhead from all partners confirming:

- Lead bidder is **fully responsible for delivery**
- Clear separation between:
 - Infrastructure hosting (data centres)
 - AI platform control
 - Any other consortium partner's responsibilities
- AI platform remains under **Ignite / Government control**



B12 – SOVEREIGNTY & SUPPLY-CHAIN ASSURANCE

This section evaluates the degree to which the proposed solution delivers genuine national sovereignty: independence from foreign export-control or supply interruption, the strength of binding high-level continuity commitments, technology and capability transfer, and freedom from foreign operational control. Each criterion is scored on three levels (Poor / Moderate / Good) against the evidence stated. Marks per level: Poor = 0, Moderate = half, Good = full.. Commitments scored here shall be at the highest available level (e.g. binding OEM and, where applicable, government-backed undertakings).

#	Criterion	Poor (0 marks)	Moderate (50% marks)	Good (full marks)	Evidence required	Max
1	Freedom from foreign export-control / supply interruption	Supply subject to foreign export licensing that can restrict or halt delivery; no mitigation	Some exposure to foreign licensing, with partial mitigation or alternative paths	Supply demonstrably free of any single foreign government's export-control veto; no foreign kill-switch over supply	Supply-chain/jurisdiction analysis; export-control exposure statement; mitigation evidence	5
2	Binding continuity-of-supply commitment (highest level)	No binding commitment; best-efforts only	Binding OEM commitment of limited scope/duration	Binding OEM and, where applicable, government-backed continuity-of-supply commitment covering the full programme lifecycle, incl. spares and replacements	Signed/committed undertakings; duration and scope; backing level	5
3	Domestic / sovereign supply & assembly	Fully foreign-manufactured with no local content or assembly	Partial local assembly, integration, or in-country stocking	Substantial in-country assembly/integration and/or committed local supply, reducing external dependency	Local-content/assembly plan; in-country stocking commitments	5
4	Technology & capability transfer	No meaningful transfer; full dependence on vendor	Partial transfer (operations/training) with residual dependence	Comprehensive transfer enabling independent national operation, including documentation, training, and (where applicable) source/IP access for sovereign components	Transfer plan; scope of IP/source/documentation; independence acceptance	5
5	Freedom from foreign operational control	Operation depends on foreign-controlled services, licences, or remote dependencies that could be withdrawn	Some foreign operational dependencies with mitigation	Fully operable under national control with no foreign-controlled runtime dependency, licence check-in, or remote kill-switch; air-gap capable	Operational-dependency analysis; air-gap capability evidence	5
6	Data & model sovereignty alignment	Architecture impedes in-country data/model ownership or residency	Supports residency with some external dependencies	Fully supports in-country data and model residency and Government of Pakistan ownership of sovereign artefacts, with no external dependency	Residency/ownership architecture; alignment with sovereign-model requirements	5