File No.: INDAI/14/2025-INDAI
Ministry of Electronics and Information Technology
Digital India Corporation
IndiaAI IBD

Electronics Niketan, 6, CGO Complex,
New Delhi - 110003
Dated: 13.03.2026

Office Memorandum

Subject: Issuance of the document titled "Ready Reckoner for Compute Users" on the IndiaAI Compute Portal.

The undersigned is directed to convey that the document titled "Ready Reckoner for Compute Users" is enclosed for reference of users of the IndiaAI Compute Portal.

2. The document provides indicative guidance to assist users in mapping AI use-cases to appropriate GPU configurations available through the IndiaAI Compute Portal. The guidance is illustrative and non-exhaustive and is intended solely to support informed decision-making.

3. The ready reckoner shall also be made available on the Compute Portal.

This issues with the approval of Competent Authority.

13/3/2026

(Sushil Kumar Jangid)
Scientist 'B'
Tel. No.: 011-24301784

Encl.: As above.

To:
All the members of the PMEC
All users of the IndiaAI Compute Portal.

Copy to:

i. PS to Secretary, MeitY
ii. PS to Chief Executive Officer, IndiaAI
iii. PS to Chief Operating Officer, IndiaAI
iv. PS to Joint Secretary, MeitY
v. PS to Director, IndiaAI
vi. General Manager (Finance), IndiaAI

Ready Reckoner for Compute Users

This document provides indicative guidance to assist users in mapping AI use-cases to appropriate GPU configurations available through the IndiaAI Compute Portal. The guidance is illustrative and non-exhaustive and is intended solely to support informed decision-making. Users are advised to undertake their own due diligence and technical assessment while determining their compute and GPU requirements, prior to submitting requests on the Compute Portal.

NVIDIA A100 Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
A100 40GB (8X)	• **Large Scale OCR**: NVOCR (150 pages/doc). • **ASR Training**: Conformer models (Long duration). • **LLM Inference**: Llama 3.3 70B (Low concurrency).	₹3175.66
A100 80GB (1X)	• **Specialized CV**: Satellite & Crowd Analytics. • **Healthcare**: X-ray Inferencing. • **Light GenAI**: Llama 8B, Nemotron 12B.	₹135.9
A100 80GB (4X)	• **VLM Training**: Handwritten Text Recognition. • **High-Load ASR**: 90 concurrent users. • **MoE Inference**: Mixtral 8x7B.	₹543.6
A100 80GB (8X)	• **GenAI Production**: 70B LLM Inferencing. • **Heavy Finetuning**: CV/NLP models.	₹1087.2

	• **Complex OCR**: 200+ page documents.	

NVIDIA H100 & H100 NVL Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
H100 NVL (4X)	• **Memory-Intensive LLM**: RAG, Agents. • Medium LLM: 56B-70B models.	₹674.96
H100 SXM (8X)	• **Foundation Model Training**: 70B-100B. • **Video Search**: Summarization at scale.	₹1224.0
H100 PCIe (8X)	• **Small Model Inference Clusters**: ASR training. • **Multiple Instances**: 40B LLM (1 Card/Instance).	₹2008.0

AMD MI300X Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
MI300X (Single OAM Module)	• Memory-Intensive LLM: RAG, Agents. • Medium LLM: 56B–70B models.	MI300X.1X - ₹168.22
MI300X (UBB Multi-GPU System)	• Foundation Model Training: 70B–100B. • ASR & Vision Transformers at scale.	MI300X.8X - ₹1416.56

MI300X (Cluster Deployment)	• Enterprise-scale LLM serving. • Multi-agent workloads.	MI300X.8X - ₹1416.56

NVIDIA H200 Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
H200 SXM (1X)	• **Online Document Summarization**: 8B - 30B • **ASR Finetuning**.	₹140.0
H200 SXM (2X)	• **LLM Inference**: Mixtral 8x7B. • **Long Context**: 128k token processing.	₹510.0
H200 SXM (4X)	**Medium LLM:** LLM Finetuning, RAG, Agents, Inferencing 56B - 70B	₹1020.0
H200 SXM (8X)	• **Video Search**: Massive index search. • **Large LLM**: Nemotron 100B. • **GenAI**: Mixtral 8x7B (70-80 concurrent users).	₹1125.0

AMD MI325X Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 27th Feb' 2026**

MI325X (1 GPU)	• Online Document Summarization: 8B–30B models. • ASR Finetuning.	MI325X.1X - ₹169.2
MI325X (8 GPUs in UBB)	• Medium LLM Inference: 56B–70B models. • RAG, Agents, Multi-user concurrency.	MI325X.8X - ₹1351.8
MI325X (Cluster Deployment)	• Large LLMs: 100B+ parameter models. • Enterprise-scale GenAI workloads.	MI325X.8X - ₹1351.8

NVIDIA B200 Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
B200 SXM (1X)	• **VLM Training**: Qwen 3 VL 8B. • **High Concurrency Inference**.	₹323.0
B200 SXM (4X/8X)	• **Multi-Agent Systems**: Nemotron 100B. • **Video Summarization**.	**B200 SXM 4X** - ₹1292.0 **B200 SXM 8X** – ₹2584.0

NVIDIA L4 & L40S Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
L4 (1X - 8X)	• **Signature Detection**. • **Translation Transformers**.	L4.1x - ₹44.86 L4.2x - ₹98.84 L4.4x - ₹196.68 L4.8x - ₹507.94

	• **Small LLM Inference**.	
L40S (1X - 8X)	• **Face Verification**. • **Medium LLM Inference**. • **Image Quality Assessment**.	L40S.1x - ₹67.5 L40S.2x - ₹135.0 L40S.4x - ₹306.0 L40S.8x - ₹540.0

Intel Gaudi3 Usage Guide

Instance Configuration	**Primary Use-Cases**	**Pricing (On-Demand/hour) as per 13th March 2026**
Gaudi2 (1X – 8X) **Gaudi3 (1X - 8X)**	• Enterprise RAG • Enterprise Dataware house Integration • Conversational AI • Human-In-The-Loop business process optimizations.	Gaudi2.1x - ₹57.6 Gaudi2.2x - ₹115.2 Gaudi2.4x - ₹230.4 Gaudi2.8x - ₹460.8 Gaudi3.1x - ₹153.0 Gaudi3.2x - ₹306.0 Gaudi3.4x - ₹612.0 Gaudi3.8x - ₹1224.0