LICDIA
Laboratorio de Investigación en Ciencia de Datos e Inteligencia Artificial

JORNADAS 2026 DE CIENCIA, TECNOLOGÍA Y EXTENSIÓN
DEPARTAMENTO DE CIENCIAS BÁSICAS

Línea de I+D - Modelos de lenguaje abiertos

InferHive: préstamo colaborativo de GPU de consumo para cargas de LLM con planificación por reservas y créditos de uso verificable

Petrocelli, D.[1,2]; Hernando, A.[1,3]; Fernández, J. M.[1]

[1] Laboratorio de Investigación en Ciencia de Datos e Inteligencia Artificial (LICDIA) - [2] Departamento de Tecnología, Universidad Nacional de Luján -[3] Facultad de Ciencias Exactas, Físicas y Naturales (FCEFyN), Universidad Nacional de Córdoba

dpetrocelli@unlu.edu.ar - alex.hernando@mi.unc.edu.ar - jmfernandez@unlu.edu.ar

InferHive ya está en uso: alumnos de grado y posgrado de la UNLu trabajan con modelos de lenguaje sobre cinco GPU de consumo compartidas, sin dinero de por medio.

1 Problema

Ejecutar y ajustar modelos de lenguaje de gran escala (LLM) requiere GPU, cuya oferta se concentra en nubes y marketplaces, mientras en hogares y laboratorios hay GPU de consumo ociosas.

2 Plataforma

InferHive presta GPU de consumo ociosas mediante un plano de control en la nube. El uso se registra como créditos sin valor monetario en un libro contable verificable.

Usuarios
interfaz compatible con OpenAI · reserva de GPU · kubectl

↓

Plano de control en la nube
clúster k3s híbrido

↓

GPU de consumo
sin abrir puertos en la red de sus dueños

Créditos sin valor monetario
libro contable encadenado y firmado

3 Planificador con tres niveles de servicio

Considera

- **Memoria de video de cada GPU**
- **Franjas reservadas**
- **Regreso del dueño**

→ **Planificador**
línea de investigación →

Niveles de servicio

- **Reserva exclusiva**
- **Inferencia compartida**
- **Carga interrumpible**

4 Medido y por validar

MEDIDO EN LA PLATAFORMA

5 GPU de consumo en operación	**~41** tokens por segundo

METAS DEL PLANIFICADOR, POR VALIDAR

1,5 veces la utilización	**Ninguna** reserva violada
≥ 0,8 índice de Jain	**< 5 %** de trabajo perdido

Con cuatro solicitudes en paralelo: 116 a 133 tokens por segundo. Las metas se comparan con un préstamo exclusivo.

5 En operación

Uso medido, no declarado
La plataforma mide GPU y tokens sin confiar en el nodo.

Aislamiento entre usuarios
Cada alumno en su espacio, con cuotas y sin privilegios.

Tarifa por potencia medida
Cada GPU cobra según su prueba; referencia: RTX 4060.

Franjas de clase
Las GPU quedan bloqueadas durante la cursada.

Acceso desde código y agentes
API compatible con OpenAI; SDK, CLI y servidor MCP.

Cómputo desde dispositivos comunes
Aportan sin unirse al clúster, con tareas de control.

6 Conclusiones

APORTE
Tres niveles en producción
Reserva exclusiva, inferencia compartida y carga interrumpible, con créditos verificables.

POR VALIDAR
Cuatro metas del planificador
Utilización, reservas, equidad y trabajo perdido; faltan trazas reales.

PLATAFORMA
Operativa con cinco GPU
RTX 3090, 4060, 5060, 3060 y GTX 1650 SUPER, en un clúster k3s híbrido.

USO EN DOCENCIA
Grado y posgrado
Sistemas Distribuidos en grado; diplomaturas de IA generativa y Blockchain en posgrado.

Sigue: validar las cuatro metas con trazas reales, que se publicarán anonimizadas.

Referencias

Petrocelli, D. Plataforma colaborativa, distribuida, escalable y de bajo costo para tareas de cómputo intensivo. Tesis doctoral, UNLP, 2021.

Jain et al. A Quantitative Measure of Fairness and Discrimination. DEC-TR-301, 1984.

AI Horde. Software y servicio público. github.com/Haidra-Org/AI-Horde.

Andrade et al. Discouraging Free Riding in a Peer-to-Peer CPU-Sharing Grid. HPDC, 2004.

Dettmers et al. QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS, 2023.

Borzunov et al. Petals: Collaborative Inference and Fine-tuning of Large Models. ACL, 2023.

Anderson, D. P. BOINC: A Platform for Volunteer Computing. J. Grid Computing, 2020.

Portal

Detalle técnico

Líneas de I+D - Investigación aplicada en ciencia de datos e IA, con fuerte impronta en los problemas regionales

Modelos de lenguaje abiertos · IA en educación · Diálogo y RAG · Gobernanza y transformación digital · Blockchain y credenciales · Imágenes y datos geoespaciales

licdia.unlu.edu.ar