UM-CLOUD: INFRAESTRUCTURA DE IA ON-PREMISE PARA EDUCACIÓN SUPERIOR

Diego Navarro[1], Juan José Ciarlante[1], Ignacio Bosh[1], Mariela Asensio[1]
[1]UM-Cloud, Facultad de Ingeniería, Universidad de Mendoza
diego.navarro@um.edu.ar

RESUMEN
Las universidades enfrentan una brecha significativa en recursos de infraestructura de IA comparadas con la industria. Este trabajo presenta UM-Cloud, infraestructura on-premise de IA de la Universidad de Mendoza que implementa arquitectura híbrida sobre Kubernetes utilizando KubeAI con motores duales Ollama y vLLM. La solución sirve 6 modelos LLM en producción sobre 5 nodos GPU heterogéneos (RTX 3090, RTX 5060 Ti, RTX 2080 Super), con contexto extendido de 32K tokens que habilita casos de uso educativos avanzados. El enfoque prioriza soberanía tecnológica e independencia operativa.
Área: Arquitectura, Redes y Sistemas Operativos

CONTEXTO
UM-Cloud es una línea de investigación consolidada de la Facultad de Ingeniería iniciada en 2015, enfocada en arquitecturas de servicios cloud. El proyecto responde al principio institucional de entender la tecnología construyéndola para luego enseñarla.

Evolución histórica:
• 2015: OpenStack
• 2019: Kubernetes + Rook + OpenStack
• 2020-2021: Laboratorio Remoto COVID-19
• 2022: Ampliación x4
• 2023: kros'k + cátedra Cloud Computing
• 2024: Extensión a servicios IA con KubeAI

Equipo: Mg. Ing. Diego Navarro (Director Ingeniería en Informática, líder UM-Cloud desde 2015), Mg. Ing. Juan José Ciarlante (Arquitecto UM-Cloud, ex-Google SRE, CNCF CKA), Ing. Ignacio Bosch (Titular Cátedra Inteligencia Artificial), Esp Lic. Mariela Asensio (Directora Instituto Informática – Sede San Rafael)

Infraestructura: https://cloud.um.edu.ar
Repositorio: https://github.com/umcloud

ARQUITECTURA TÉCNICA
Stack sobre Kubernetes:
• KubeAI Operator (orquestación)
• Motores duales: Ollama (GGUF) + vLLM (AWQ 4-bit)
• Open WebUI (OAuth @um.edu.ar)
• CephFS (200Gi persistente)
• NVIDIA GPU Operator v25.3.4 con time-slicing

Infraestructura GPU:
• 2× RTX 3090 24GB (Ampere) - modelos grandes
• 2× RTX 5060 Ti 16GB (Blackwell) - medianos + HA
• 1× RTX 2080S 8GB (Turing) - ligeros

POSICIONAMIENTO REGIONAL
Primera universidad privada latinoamericana con infraestructura Kubernetes de IA documentada.

Diferenciación en 4 dimensiones:
1. Institución privada (vs. centros públicos/nacionales)
2. Orquestación Kubernetes (vs. SLURM regional)
3. GPUs consumer-grade (vs. datacenter)
4. Orientación educativa directa

Contexto regional:
• Argentina: Clementina XXI (296 Ponte Vecchio), UNC-CCAD (44 A30)
• Brasil: Santos Dumont (248 H100 + 144 GH200)
• Chile, México, Colombia: SLURM sobre datacenter GPUs

MODELOS DESPLEGADOS

Modelo	GPU	VRAM	Caso de uso
gemma3-4b	RTX 2080S	~6GB	Chat ligero
gemma3n-e2b	RTX 2080S	~2GB	Embeddings
nemotron-3-nano-30b	RTX 3090	~20GB	Razonamiento MoE
glm47-flash	RTX 3090	~20GB	Texto/Código especializado
nomic-embed-text	CPU	~0.3GB	Embeddings
gpt-oss-20b	**2× RTX 5060Ti**	**~14GB**	**32K ctx (HA)**

DESGLOSE TÉCNICO CONTEXTO 32 k
Contexto extendido de 32K tokens en RTX 5060 Ti (16GB) mediante cuantización dual:

• Modelo: GGUF Q4_K_M (40GB FP16 → 11.5GB)
• KV cache: Q4_0/Q8_0 (12-14GB → 3-4GB)
• Overhead: 0.5GB para estabilidad

Cálculo VRAM:
Pesos (11.5GB) + KV cache (3-4GB) + Overhead (0.5GB) = 15.5GB / 16GB

Esto maximiza el contexto para análisis de documentos extensos, agentes de código (tipo OpenCode) y código completo en aplicaciones educativas.

RESULTADOS OBTENIDOS / ESPERADOS

Infraestructura validada:
• 6 modelos LLM en producción / OpenWebUI + OAuth institucional (@um.edu.ar) / API OpenAI-compatible / Storage CephFS persistente

Capacidad operativa estimada:
• Uso ligero (1-2 req/min): 40-60 usuarios
• Uso medio (5-10 req/min): 25-35 usuarios
• Uso intenso (15+ req/min): 12-20 usuarios

Validación tecnológica:
• Stack alineado con industria 2026 / KubeAI / Arquitectura híbrida maximiza hardware heterogéneo

Próximos pasos:
• Métricas de uso real / Evaluación expansión GPU / Documentación casos educativos

BIBLIOGRAFÍA
[1] Stanford IT. "Unlock Marlowe's Potential for Breakthrough Research". 2025.
[2] Kempner Institute. "~400 H100 GPUs Computing Cluster". 2023.
[3] Princeton AI. "300 GPU cluster for academic AI research". 2024.
[4] Kantrowitz. "Universities Under-Resourced For AI". 2024.
[5] Argentina.gob.ar. "Clementina XXI". 2023.
[6] UNC-CCAD. "Cluster Mendieta Fase 2". 2024.
[7] Eviden. "Santos Dumont upgrade". 2025.
[8] National Research Platform. nrp.ai, 2025.
[9] Weitzel et al. "NRP: Stretched Kubernetes Cluster". PEARC '25.
[10] KubeAI Documentation. kubeai.org, 2025.
[11] Kwon et al. "PagedAttention". SOSP '23.
[12] Song et al. "PowerInfer". arXiv:2312.12456, 2024.
[13] NVIDIA. "GPU Operator Documentation". 2025.