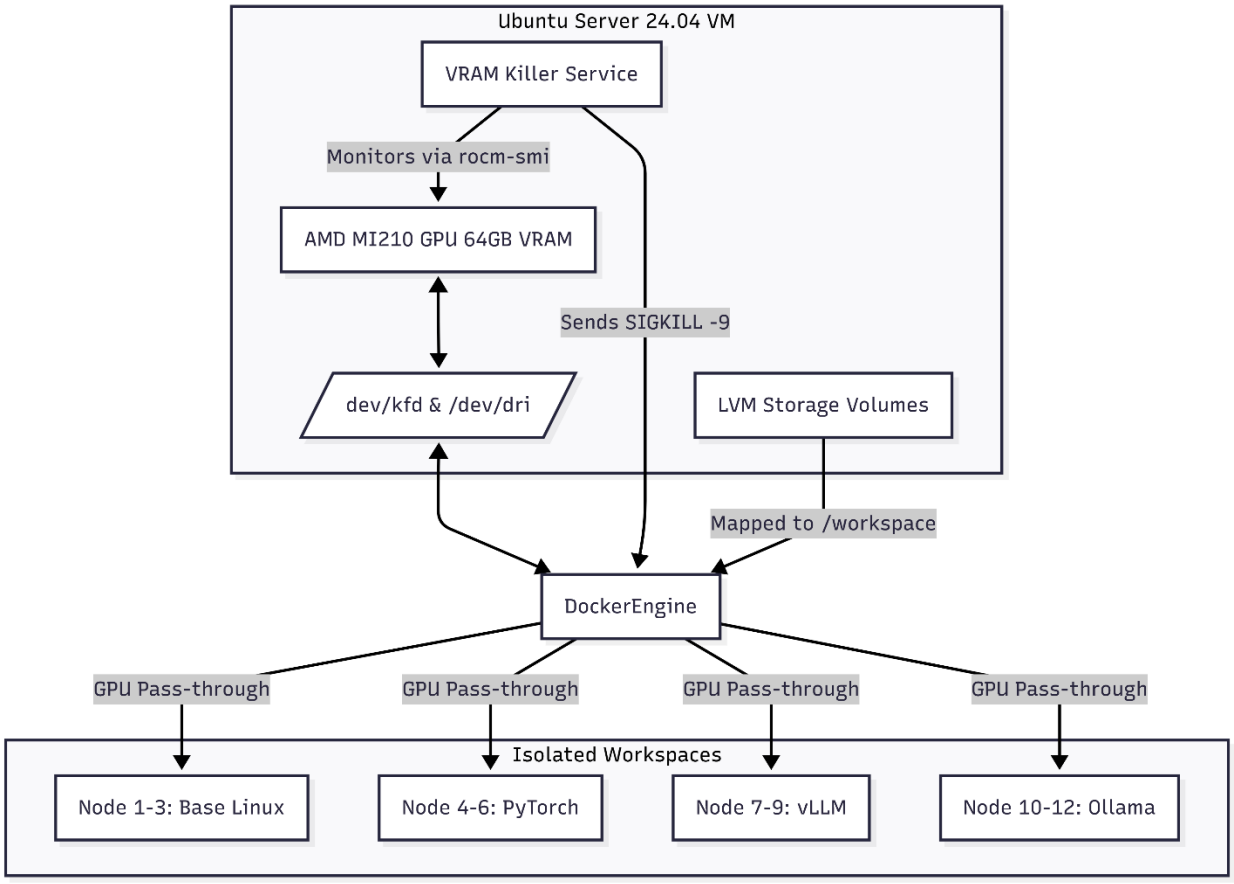
MCP Server Services by Docker Environment

❖ System Architecture

The MCP server supports 12 concurrent users. Resources are completely isolated using Docker containers, with hardware-level access management.



The service configuration of each node can be adjusted according to user requirements without affecting other active nodes. Four service images are currently available:

- Base Linux is based on rocm/pytorch, allowing users to install any additional software they require.
- PyTorch is based on rocm/pytorch.
- vLLM is based on rocm/vLLM.

- Ollama is based on rocm/Ollama.

❖ System Environment

- Each user (container) has a dedicated 100 GB volume mounted at /workspace for persistent data storage across container restarts.
- Each user accesses their own container via SSH.
- Memory is limited to 16 GB per container.
- GPU VRAM usage limit: 5 GB per user.
  - VRAM usage cannot be limited for individual users through standard mechanisms because the hardware does not support this capability. Technically, every user can access the full VRAM capacity.
  - To address this limitation, a monitoring service checks VRAM usage every 5 seconds. When a user's VRAM usage exceeds the 5 GB limit, the relevant process is terminated immediately. The user who started the process receives a termination notice and/or a process error message in their SSH terminal, and can continue using the SSH terminal normally.
- Users log in with the assigned username, password, and port using a command such as: ssh user01@ai-g3 -p 2201