Infrastructure Design and Deployment: GPU-Accelerated Virtual Machines for the Faculty of Computer Science

Client & Team

Client: Faculty of Computer Science, Dalhousie University

- Jeremy Dempsey (Tech Lead)
- Jeff Allen (FCS Infrastructure Manager)
- Devi

Assignee: Dalhousie ITS

- Edward White (Supervisor)
- Kevin Nguyen (Technician & Infrastructure Designer)

Project Overview

In response to the Faculty of Computer Science's need for a secure, high-performance virtual environment tailored to research and computational workloads, I was assigned to design and implement a virtual infrastructure capable of supporting GPU-accelerated Linux machines with deep learning capabilities, Horizon remote access, and role-based administrative control.

The target users included:

- Undergraduate Students (e.g., lab environments)

- Graduate Researchers (GPU workloads)

- Faculty Members (general-purpose Linux VM usage)

Proposed Infrastructure Design

1. **Operating System & Environment**

- Base OS: Ubuntu 22.04 LTS chosen for its lightweight architecture, wide compatibility with research libraries (CUDA, PyTorch, TensorFlow), and intuitive CLI support.

- Display Manager: Transitioned from GDM to LightDM for improved compatibility with the Horizon Agent, resolving Blast protocol integration issues and ensuring seamless remote rendering.

2. Virtualization & Access Control

- Platform: VMware Horizon (latest stable release)

- Entitlements: Integrated through Horizon Connection Server, with pooled desktop provisioning and floating VM assignment for scalability.

- Security Model: Active Directory (AD)-based access using:

- Group Policy Objects (GPOs) to restrict elevated permissions.

- Role Separation through domain groups:

    - CS Lab: No sudo

    - CS Researchers: Limited sudo access

    - Admins: Full access

Security Architecture

Active Directory Authentication

Initial challenges arose due to blocked Kerberos requests from our VM subnet. Domain join attempts silently failed. I conducted firewall packet inspections using tcpdump and monitored AD logs with Donald (Windows Server Admin). Together, we identified a subnet registration issue on the domain controller, where the subnet lacked the appropriate GPO inheritance.

Resolution:

- Proposed temporary access workaround (local account provisioning).

- Long-term fix: Coordinated with domain admin to register the subnet and validate AD policies for Linux clients using SSSD.

Sudo Access Management

Concerns over sudo privilege abuse led to a multi-path proposal:

1. Minimal sudo elevation to trusted research groups only (chosen solution).

2. Implement RSA-logged root sessions (auditing via /var/log/auth.log).

3. Develop a custom webhook monitoring script that logs and notifies sudo attempts via REST API to a logging server (not implementing).

SSSD & PAM Integration

Configured sssd.conf with:

- [domain/dal.ca] section including AD realm, LDAP fallback

- pam_sss.so for session authentication

- Auto-creation of home directories with secure permissions

GPU Integration & Optimization

NVIDIA Licensing & Horizon Conflicts

After installing the NVIDIA vGPU driver stack and CUDA toolkit, Horizon Agent disabled GPU acceleration by overriding display sessions.

Upon further investigation:

- BlastServer handshake was failing due to TLS certificate errors and improper GPU device recognition.

- Manually issued and applied a self-signed certificate (testing phase).

Reverted to hardware-level troubleshooting:

- Identified Horizon Agent compatibility mismatch with vGPU guest OS kernel modules.

- Upgraded Horizon Agent and Display Manager, which resolved rendering failures.

Driver Consistency

Driver versions on the ESXi host and Ubuntu guest OS were misaligned:

- Contacted Glen (ESXi Host Admin) to provision and synchronize vGPU-compatible drivers.

Infrastructure Design and Deployment: GPU-Accelerated Virtual Machines for the Faculty of Computer Science

- Ensured that nvidia-smi and nvidia-settings returned proper device context.

Policy-Driven Access Flow

A structured group-based access framework was designed:

[VMware Horizon]

[Group Policies from AD Domain]

> CS_Students_Lab > No sudo, locked-down environment

> CS_Research_Users > Limited sudo, GPU access

> CS_Admin > Extended access, minimal restrictions

All user access was authenticated via Kerberos + LDAP, with sudo rights determined by their OU membership. Group membership was enforced through GPO and verified via id command on login.

Monitoring, Feedback & Support Loop

To ensure the infrastructure met academic and operational requirements, I:

- Maintained direct communication with all stakeholders.

- Collected continuous feedback from test users (students, researchers, faculty).

- Iteratively improved performance, access policies, and documentation.

- Documented all configurations and proposed maintenance policies for future ITS technicians.

Key Takeaways & Outcomes

- Successfully delivered a scalable, GPU-enabled VM solution for the Faculty of Computer Science.

- Mitigated security risks by implementing policy-driven sudo control and hardened Linux AD integration.

Designed a solution that aligns with cloud infrastructure principles, emphasizing:

- Centralized access control (AD)

- Role-based entitlement (GPO)

- Virtualization at scale (VMware Horizon)

- Hardware acceleration (NVIDIA vGPU stack)

This project emphasized the importance of problem solving under pressure, cybersecurity-aware design, and collaborative troubleshooting with system administrators, laying the groundwork for secure academic computing at scale.