Document Version: 1.0
February 26, 2025
By Jim Venuto

The Phase 0 User Guide is designed for Cleveland Clinic's Lerner Research Institute (LRI) to self-manage cloud GPU workloads on IBM Cloud. This guide is sourced from official IBM Cloud documentation and includes references (URLs) so LRI teams can find the underlying details. It's written from the perspective of providing *guidance only*—all configuration, governance, and actual deployment steps remain the responsibility of the LRI technical team.

Table of Contents

1. Overview of Phase 0

Objective: Provide individual labs at LRI with on-demand GPU capacity in IBM Cloud, **without** any centralized HPC scheduler or unified namespace. Each lab manages its own environment, from data upload and resource provisioning to cost tracking and tear-down.

- **Key Characteristics**:

  - Ephemeral GPU instances, spun up and down by the lab.
  - Labs manually transfer data using standard tools (e.g., SFTP, Aspera).
  - No automation of job scheduling or data tiering—entirely self-managed.
- **Risks/Responsibilities**:
  - Ensuring data security (e.g., no inadvertent PHI exposure if relevant).
  - Watching for cost overruns (e.g., forgetting to shut off GPU nodes).
  - Manually monitoring usage, usage logs, or cost analytics in IBM Cloud.

Who Should Use This Guide?

- LRI researchers or HPC support staff who need additional GPU resources beyond the on-prem cluster.
- Users comfortable administering their own cloud environments with minimal guidance from IBM.

2. IBM Cloud Account Setup & Billing

1. **Sign Up for IBM Cloud**
   - Visit https://cloud.ibm.com/registration to create an IBM Cloud account.
   - If your lab already has an account, ensure you have sufficient permission (Org Manager, Billing Manager, or higher) to create new resources.
2. **Billing & Subscription Model**
   - Phase 0 is typically "pay-as-you-go" (PAYG). You pay only for the GPU and storage hours used.
   - For more info on billing plans: https://cloud.ibm.com/docs/billing-usage?topic=billing-usage-billing-overview
3. **Resource Groups and Organization**
   - You might create a dedicated **Resource Group** for HPC or GPU workloads, so usage is clearly separated from other projects.
   - **IBM Cloud Docs**: https://cloud.ibm.com/docs/account?topic=account-rgs
4. **Recommended Action**:
   - Each lab or HPC sub-group might set up its own resource group or have separate accounts to track usage.
   - Decide early on who holds the billing responsibility so that cost ownership is transparent.

3. Basic Identity and Access Management (IAM)

- **IBM Cloud IAM**
  - Controls who can create VMs, provision GPUs, or access storage.
  - For labs that share a single account, the account owner can grant specific roles (Viewer, Operator, Editor, Admin) to each user.
- **Key IAM Tasks**

1. **Create/Invite Users**: Each researcher or HPC admin at LRI gets an IBMid or is invited to the existing account.

2. **Assign Roles**: For GPU provisioning, you typically need the **Editor** or **Administrator** role on the resource group.

3. **Policies**: If certain labs must not see each other's resources, consider using custom access policies.

- **Docs**: https://cloud.ibm.com/docs/account?topic=account-userroles

4. Choosing and Provisioning GPU Resources

4.1 Virtual Private Cloud (VPC) Overview

Most HPC users at LRI will use IBM's **Virtual Private Cloud** (VPC) service to create and manage GPU-enabled Virtual Server Instances (VSIs). VPC provides a secure, isolated network environment.

- **Docs**: https://cloud.ibm.com/docs/vpc?topic=vpc-getting-started

4.2 GPU Profiles

IBM Cloud offers various **GPU profiles** (e.g., NVIDIA Tesla, NVIDIA A100) with different memory and GPU core configurations. Labs should pick a profile matching their software's GPU requirements (CUDA, ML frameworks, etc.).

- **Docs**: https://cloud.ibm.com/docs/vpc?topic=vpc-gpu-profiles

4.3 Creating a GPU-Enabled VSI

1. **Navigate to VPC**
   - In the IBM Cloud console, go to **Menu** → **VPC Infrastructure** → **Instances** → **Create**.
2. **Specify Compute & GPU Profile**
   - Under **Profile**, select a GPU-enabled option (e.g., mx2-2x16-gpu2).
3. **Configure Networking**

   - Choose a **VPC** (create one if needed), a **Subnet**, and a **Public or Private** IP address. For external access (SFTP, etc.), a public IP or a VPN is required.

4. **Attach SSH Keys**
   - SSH keys allow secure access to your instance. If labs need Windows GPU VMs, they can use RDP or a bastion approach.

5. **Provision**
   - After verifying your selections, click **Create**. The VSI typically spins up in a few minutes.

- **Docs**:
   - Creating a Virtual Server: https://cloud.ibm.com/docs/vpc?topic=vpc-creating-virtual-servers
   - Best Practices for GPU usage: https://cloud.ibm.com/docs/vpc?topic=vpc-gpu-vms

5. Storage Options

5.1 Block Storage for VM Instances

A GPU VSI typically uses **Block Storage** for the OS disk. You can attach one or more Block Storage volumes to store data temporarily during processing. If your labs are dealing with large datasets, consider:

- **Performance tiers** (IOPS per GB) based on the workload.
- Automatic encryption at rest is included by default.
- **Docs**: https://cloud.ibm.com/docs/vpc-block-storage?topic=vpc-block-storage-getting-started

5.2 IBM Cloud Object Storage (COS) for Bulk Data

For large datasets, labs typically use **IBM Cloud Object Storage**. They can then download relevant data to the GPU instance's block storage only when needed.

- **Docs**: https://cloud.ibm.com/docs/cloud-object-storage?topic=cloud-object-storage-getting-started

Common Phase 0 Workflow:

1. Upload data to an Object Storage bucket.
2. Spin up the GPU VSI.
3. Download the needed data from COS into the VSI's block storage.
4. Run the HPC/ML workload.

5. Upload final outputs back to COS.
6. Terminate the VSI.

5.3 Data Transfer Tools (Aspera, SFTP)

- **Aspera**: Good for large data sets, high-speed transfers.
  - Docs: https://cloud.ibm.com/docs/aspera?topic=aspera-getting-started
- **SFTP**: Straightforward but typically slower.
  - Labs can set up an SFTP server inside the GPU instance or use an SFTP client to connect if the GPU VSI has a public IP or is behind a VPN.

6. Security & Compliance Considerations

1. **Data Classification**
   - Validate that data is approved for external cloud use. If you handle sensitive data (PHI), confirm your compliance needs with local IT or cybersecurity teams.
2. **IAM & Access Control**
   - Carefully assign roles so that only intended researchers can provision and destroy GPU resources.
3. **Network Security**
   - If labs open SSH or RDP ports publicly, ensure strong passwords, limited access via IP whitelisting, or a VPN.
   - Some labs prefer a "private-only" VSI + a bastion host, isolating the GPU instance from direct internet access.
4. **Encryption**
   - By default, Block Storage volumes are encrypted at rest. For data in COS, consider SSE (server-side encryption) if needed, or set up your own encryption keys using IBM Key Protect.
5. **Official IBM Cloud Security Overview**:
   - https://cloud.ibm.com/docs/security?topic=security-get-started

7. Monitoring Usage & Controlling Costs

1. **Cost & Usage Dashboard**
   - In IBM Cloud console, go to **Manage** → **Billing and usage** → **Usage** to see spending trends.

  - You can set up budget alerts to notify labs if monthly spending surpasses a threshold.

2. **Resource Tags**
   - Tag each GPU instance with the lab name or project code (e.g., lab=smith), so cost shows up per lab.
3. **Check Idle Resources**
   - Some HPC tasks might not run 24/7. Encourage labs to shut down or delete GPU VMs once they're done.
4. **Reference**:
   - https://cloud.ibm.com/docs/billing-usage?topic=billing-usage-cost

8. Shutting Down and Cleaning Up

1. **Delete or Stop the VSI**
   - If the lab expects to run the workload again soon, they might just **stop** the instance (to remove CPU/GPU costs) but keep paying for the block storage.
   - Otherwise, **delete** the VSI to avoid further charges.
2. **Remove Block Volumes** (If no longer needed)
   - Each volume continues to incur charges until deleted. Make sure the lab's data is backed up before removal.
3. **Object Storage Cleanup**
   - Stored data also incurs a monthly cost. Labs should decide how long they need to keep results. Consider life-cycle policies (e.g., auto-delete after 30 days).
4. **IAM Access**
   - Revoke unneeded permissions if a researcher no longer needs GPU privileges. This mitigates accidental resource sprawl.

9. Additional IBM Cloud Documentation References

Below are select IBM Cloud Docs pages for deeper detail:

- **IBM Cloud Getting Started**:
  https://cloud.ibm.com/docs?tab=publiccloud
- **IBM VPC**:
  https://cloud.ibm.com/docs/vpc

- **GPU Profiles**: https://cloud.ibm.com/docs/vpc?topic=vpc-gpu-profiles
- **VPC Block Storage**: https://cloud.ibm.com/docs/vpc-block-storage
- **IBM Cloud Object Storage**: https://cloud.ibm.com/docs/cloud-object-storage
- **Aspera**: https://cloud.ibm.com/docs/aspera?topic=aspera-getting-started
- **Billing & Usage**: https://cloud.ibm.com/docs/billing-usage

Conclusion

Phase 0 is designed for **immediate** GPU cloud bursting with minimal overhead from LRI's side. Each lab takes full ownership—**from provisioning and configuration** to data transfer, compliance, and tear-down. This approach delivers rapid GPU access but also demands vigilance with security, cost oversight, and manual operations.

Should Cleveland Clinic decide to reduce the manual load or ensure compliance on a scale, the advanced features proposed in a **Phase 1 pilot** (e.g., automated job scheduling, integrated tiered storage, unified identity) would offer central governance and streamlined HPC operations. Until then, this guide should enable labs to stand up ephemeral GPU resources, run HPC workloads, and carefully manage associated costs.