Cloud GPU Workloads for LRI: Phase 0 and Beyond

Cleveland Clinic QBR

Jim Venuto,

jvenuto@us.ibm.com

February 26, 2025

Introduction

Purpose:

- Explain Phase 0 for self-managed cloud GPU usage at LRI

Context:

- Supplements on-premises GPU cluster with additional cloud resources

What is Phase 0?

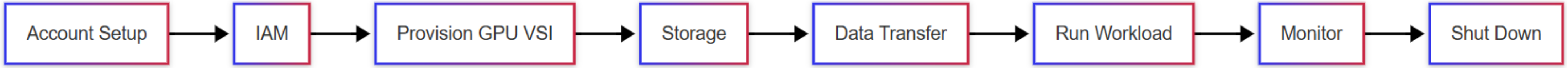
Definition:

- On-demand, self-managed GPU capacity on IBM Cloud

Key Features:

- Each lab independently manages its GPU instances and data
- Ephemeral GPU resources, spun up and down as needed
- No centralized scheduling or automation

Phase 0 Workflow



Account Setup → IAM → Provision GPU VSI → Storage → Data Transfer → Run Workload → Monitor → Shut Down

Steps: The Phase 0 User Guide.docx [Full Phase 0 User Guide]

- Set up IBM Cloud account and billing
- Manage access with Identity and Access Management (IAM)
- Provision GPU-enabled Virtual Server Instances (VSIs) in Virtual Private Cloud (VPC)
- Use Block Storage and Object Storage for data
- Transfer data with tools like SFTP or Aspera
- Monitor usage and costs, then clean up resources

Responsibilities in Phase 0

Data Security:

- Ensure compliance (e.g., PHI if applicable)

Cost Management:

- Avoid overruns through vigilant monitoring

Resource Handling:

- Manually provision and tear down GPUs

Monitoring:

- Track usage and performance independently

Introducing Phase 1

Overview:

- Future phase with advanced features

Key Enhancements:

- Automated job scheduling
- Integrated tiered storage
- Unified identity and central governance

Goal:

- Reduce manual effort and improve efficiency

Comparison: Phase 0 vs. Phase 1

Aspect	Phase 0	Phase 1
Management	High manual effort	Automated scheduling and management
Cost Control	Risk of overruns if not monitored	Centralized monitoring and control
Security	Lab-specific, harder to enforce	Unified identity and governance
Scalability	Limited by lab capacity	Better resource utilization
Initial Setup	Quick to start	More time and resources required

Trade-offs

Phase 0:

- *Pros*: Immediate access, full lab control, low initial setup
- *Cons*: Manual effort, cost risks, security challenges

Phase 1:

- *Pros*: Automation, efficiency, enhanced security
- *Cons*: Higher setup time and cost, less lab autonomy

Considerations:

- Urgency, expertise, budget, compliance needs

Key Considerations for Phase 0

Cost Management:

- Set budget alerts
- Use resource tags
- Regularly check for idle resources

Security:

- Validate data compliance
- Leverage IAM and encryption
- Consider VPN for network security

Data Transfer:

- Use Aspera for large datasets, SFTP for smaller ones

Conclusion

Phase 0:

- Offers quick GPU access with self-management

Phase 1:

- Automated, centralized job scheduling, storage, cost, controls

Decision Point:

- Balance immediate needs with long-term efficiency

Next Steps:

- Questions and discussion