Document Version: 1.0
September 20, 2025
Jim Venuto

Objective

Stand up a **manual, self-managed GPU burst** capability on IBM Cloud for up to **5 labs** (min 4), proving feasibility and establishing artifacts (runbooks, templates) to hand over and to inform Phase 1 automation.

Outcomes

1. Researchers can provision a GPU VSI, move synthetic data **via SFTP**, run a workload, persist outputs to COS, and clean up—**without IBM hands-on** beyond documentation.
2. Tagging/budgets applied for visibility and guardrails.
3. Artifacts complete: **UI → CLI → Schematics** runbooks and BOM.

1) Scope

In Scope

- Foundation: **VPC**, subnets, security groups, SSH keys, **COS instance**, budget alerts/tags.
- **Per-lab stack (×4–5)**: 1 GPU VSI; boot + data volumes; **SFTP** (SSH); COS bucket or prefix.
- **Dev/MVP sandbox** (for Jim): 1 GPU VSI + storage to build UI → CLI → Schematics paths.
- Runbooks: **Web UI**, **ibmcloud CLI**, **Schematics**; start/stop/terminate; SFTP; COS sync; cleanup.
- Evidence capture screenshots/logs; acceptance checklist.

Out of Scope (Phase 0)

- No Aspera, no VPN requirement, no centralized scheduler, no PHI/PII, no Watsonx services.

2) Constraints & Assumptions

- **Synthetic data only** (no PHI/PII) for all labs.
- Public IP allowed for Phase 0 minimal path; **bastion** is optional.
- IAM federation preferred if available; otherwise **IBMid** with MFA.
- Region(s) chosen must have available GPU capacity.
- IBM Cloud credits/service funding available via CC IT to maintain “No-Cost PoC”.

3) Functional Requirements

3.1 Provisioning & Access

- Create **VPC** with at least **2 subnets** and **Public Gateways**.
- Upload **SSH key(s)**; disable password SSH; restrict inbound to approved CIDRs.
- Provision **GPU-enabled VSI** per lab; attach boot + data volume(s).
- Allocate **public IP** (or use bastion pattern, optional).
- Install GPU drivers/tooling (per chosen profile); verify with accelerator utility (e.g., nvidia-smi when NVIDIA).

3.2 Data Services & Movement

- Create **COS instance** and **bucket(s)** (one per lab or shared with prefixes).
- **SFTP** enabled on the GPU VSI for upload/download of synthetic data.
- Provide COS CLI commands for **put/get**, including checksum validation.

3.3 Identity & Authorization

- Resource Group: LRI-Phase0.
- Access Groups: LRI-Phase0-Admins (Jim + designated CC/IBM), LRI-Phase0-Researchers.
- Roles: Admin (RG-scoped) for Jim; Editor/Reader for researchers; no broad account admin.

3.4 Governance & Cost Hygiene

- **Tags required** on all resources: org=LRI, phase=0, lab=<name>, owner=<PI>, env=lab, data=synthetic, cost-center=<id>.
- **Budget alerts** configured; email recipients: Jim + CC IT distro.
- Idle control: document **stop/terminate** steps; validate no idle GPU > 8h.

3.5 Operations

- Start/stop/terminate GPU VSIs using UI and CLI commands.
- Persist results to **COS**; document cleanup (detach/delete volumes; delete VSI/public IP).
- Maintain evidence: provisioning logs, SFTP logs, COS listings, budget alert screenshot.

4) Non-Functional Requirements

- **Security**: encryption at rest for Block and COS; least-privilege IAM; SSH from allow-listed CIDRs; root login disabled; MFA enforced on admin identities.
- **Reliability**: each lab completes **two** full provision→run→cleanup cycles.
- **Usability**: time-to-GPU (UI) **≤ 30 min**; CLI flow **≤ 45 min**; Schematics plan/apply succeeds.
- **Observability**: retain console/CLI outputs; resource inventories before/after cleanup.
- **Portability**: runbooks usable by LRI without IBM console access beyond Admin approval.
- **Compliance**: synthetic data only; no PHI/PII; no Aspera; no VPN requirement.

5) Architecture Requirements

5.1 Network

- **VPC**: lri-phase0-vpc.
- **Subnets**: at least 2 AZs.
- **Public Gateways**: 1 per subnet (Phase 0 minimal).

- **Security Groups**:
    - sg-default-outbound (egress allow).
    - sg-ssh-inbound (TCP/22 from CC CIDRs + Jim's IP).
- **Optional**: bastion host (CPU VSI) for private-only pattern (1 lab demonstration).

5.2 Compute

- **Per lab**: 1 GPU VSI (profile to be selected per capacity—A100/L40s/Gaudi 3 family acceptable), Ubuntu LTS/RHEL baseline.
- **Dev**: 1 GPU VSI (same family), smaller data volume acceptable.

5.3 Storage

- **Block**: boot (auto) + **data volume** per lab (size t-shirt default 256–1024 GB).
- **COS**: 1 instance; bucket per lab **or** shared bucket with per-lab prefixes.

6) Software & Configuration Baseline

- OS hardening basics: SSH key-based auth; disable root SSH; package updates.
- GPU drivers/toolkit compatible with selected profile; CUDA/cuDNN (if NVIDIA) or equivalent stack per accelerator.
- CLI tools: ibmcloud with VPC & COS plugins; openssl or sha256sum for checksum.
- SFTP service via OpenSSH; document home/working directories and permissions.

7) Deployment Paths (all three required)

1. **Web UI**
    - Foundation (VPC, subnets, SGs, key, COS), then per-lab stack.
2. **ibmcloud CLI**
    - Scripts: login/profile selection; VPC/SG; VSI create; volume attach; public IP; COS put/get; stop/terminate; cleanup.
3. **IBM Cloud Schematics (Terraform)**
    - **Workspace A (foundation)**: VPC, subnets, SGs, key, COS.

    - **Workspace B (lab)**: variables for lab, gpu_profile, data_volume_gb, ssh_key_name, allowed_cidrs, cos_bucket/prefix.
    - Expectation: plan clean; apply succeeds; destroys all lab resources.

8) Evidence & Deliverables

- **BOM.md** (no prices), **Foundation-UI-Runbook.md**, **CLI-Runbook.md**, **Schematics-README.md**.
- **Acceptance Checklist** with screenshots/logs: provisioning, nvidia-smi (or equivalent), SFTP checksums, COS list/get, budget alert, inventory before/after cleanup.
- **Demo deck** for CC IT + LRI leadership (summary of outcomes and Phase 1 deltas).

9) Resource Counts (parameterized)

Item	Shared	Per-lab	Total (4 labs)	Total (5 labs)
VPC	1	–	1	1
Subnets	2	–	2	2
Public Gateways	2	–	2	2
Security groups (baseline)	2	–	2	2
Security group (lab)	–	1	4	5
SSH key	1	–	1	1
COS instance	1	–	1	1
COS buckets (or prefixes)	0/1	1	4	5
GPU VSI	–	1	4	5
Boot volumes	–	1	4	5
Data volumes	–	1	4	5

Item	Shared	Per-lab	Total (4 labs)	Total (5 labs)
Public IPs (min path)	–	1	4	5
Schematics workspaces	1	1	5	6

(If using a single shared COS bucket with prefixes, set Shared=1 and Per-lab=0.)

10) Acceptance & Success (traceability)

- **Map to Must-Pass criteria**:
  - Time-to-GPU (UI) ≤ 30 min; CLI ≤ 45 min; Schematics plan/apply clean.
  - SFTP ingress (≥ 5 GB) with checksum; COS egress verified with checksum.
  - Tagging 100%; budget alert triggered; no idle GPU > 8 h; cleanups leave **0 orphans**.
  - Two full cycles per lab completed: acceptance signed by CC IT + LRI sponsor.

11) Risks & Mitigations

- **GPU capacity constraints** → pre-approve two regions/profiles; keep fallback profile.
- **Cost drift from idle** → budget alerts; explicit stop/terminate steps; audit idle windows.
- **Operator error** → golden runbooks; minimal variables; examples provided.
- **Security exposure** → SSH CIDR allow-lists; option to demonstrate bastion for one lab.

12) Change Control

- Any addition (e.g., VPN, Aspera, scheduler, managed services, PHI) is **Phase 1+** material and not permitted in Phase 0 PoC without written sponsor approval and scope update.

13) Variables (Schematics / CLI)

Variable	**Example**
region	us-south
vpc_name	lri-phase0-vpc
subnet_cidrs	10.10.1.0/24, 10.10.2.0/24
ssh_key_name	lri-phase0-key
lab	lab-smith
gpu_profile	a100.* / l40s.* / gaudi3.*
image	ubuntu-22.04
data_volume_gb	256–1024
allowed_cidrs	x.x.x.x/yy (CC ranges + Jim)
cos_bucket / cos_prefix	lri-phase0-lab-smith / lab-smith/
tags	org=LRI,phase=0,lab=lab-smith,owner=…,env=lab,data=synthetic

14) Definition of Done

- All **Must-Pass** success criteria met for **≥ 4 labs** (target **5**).
- All artifacts delivered in Box.
- Acceptance checklist signed by CC IT + LRI sponsor.
- Phase 1 automation/gov backlog captured from Phase 0 lessons.