**A-SWARM Pilot Beachhead & KPIs (Final)**

**Beachhead**

**AI datacenters (GPU clusters / colos).**  
Scope includes GPU/CPU hosts, hypervisors, storage nodes, ToR/leaf switches, and PLC-adjacent gateways for facility controls (read-only). Safety PLCs/SIS are excluded from actuation.

**Why this first:** homogeneous stacks, 24/7 high-value workloads, strong observability, clear incident economics, fast decision cycles.

**KPI #1 — MTTD for Coordinated Anomalies < 200 ms**

**What it means:** Time to *detect and elevate* a coordinated anomaly from its ground-truth onset to A-SWARM’s quorum-elevated alert.

* **Formula:** MTTD = t\_detect\_elevated − t\_anomaly\_start
* **Targets:** P95 **≤ 200 ms**, P99 **≤ 300 ms** (per scenario)
* **In-scope scenarios (min 5):**
  1. East-west lateral surge across 50+ hosts
  2. Malicious RDMA/NVLink pattern during GPU training
  3. Compromised K8s node spawning privileged pods
  4. Data exfil over atypical control plane (mTLS but novel SNI)
  5. Supply-chain backdoor process beaconing
* **Measurement method:**
  1. Ground truth t\_anomaly\_start stamped by the injector/twin replay.
  2. “Elevated” = anomaly crosses policy threshold and is quorum-confirmed on the pheromone mesh.
  3. Time sync via PTP/NTP; all events sealed in Action Certificates.

**KPI #2 — MTTR to Micro-Containment < 5 s**

**What it means:** Time from elevated detection to **effective** Ring-1 containment (rate-limit/quarantine/token revoke/process kill) observed on the affected asset/segment.

* **Formula:** MTTR = t\_containment\_effective − t\_detect\_elevated
* **Targets:** P95 **≤ 5 s**, P99 **≤ 7 s**
* **Effectiveness proof:**
  + Network: malicious flow drops to ≤1 pkt/s (or ≤1% baseline) on the target interface/segment.
  + Host: offending process terminated and IOC blocked; no re-spawn within TTL.
* **Constraints:** All Ring-1 acts are pre-authorized, reversible, TTL-bounded, and auto-revert on no-harm within **≤60 s** unless extended.

**KPI #3 — False-Positive Rate at Ring-1 < 0.5%**

**What it means:** Share of Ring-1 micro-actuations later adjudicated as unnecessary given available context.

* **Formula:** FPR = FP / (TP + FP) for Ring-1 actions over the pilot window
  + **FP (false positive):** post-mortem (within 7 days) classifies the actuation as unwarranted (no threat, unjustified disruption).
  + **TP (true positive):** actuation addressed a confirmed threat or policy-violating behavior.
* **Targets:** **< 0.5%** FPR; **0** Ring-2 false positives (none permitted).
* **Quality gates:** Each Ring-1 act has an Action Certificate (context → decision → actuation → outcome) to support adjudication.

**Pilot Acceptance Criteria (Pass/Go)**

* Achieve KPI targets across **all** five scenarios in twin and **≥3** live/synthetic drills in prod.
* **Blast-radius reduction ≥ 80%** vs. customer baseline in at least two drills.
* **No** changes to safety PLC/SIS; zero Ring-2 acts without quorum/human approval.

**Instrumentation & Evidence (Built-in)**

* **Time sync:** PTP/NTP normalized stamps across injectors, sentinels, switches, and SIEM.
* **Evidence:** Action Certificates exported to SIEM/SOAR; twin replays reproducible; blue/green canaries for all policy pushes.
* **Reporting:** Daily run chart for MTTD/MTTR (P50/P95/P99) and weekly FPR audit; end-of-pilot signed report.