BUSINESS CONTINUITY AND DISASTER RECOVERY POLICY

Dedicated NVIDIA GB300 NVL72 AI Infrastructure

CambridgeNexus, Inc.
Revision 1.3
Effective: August 2026
Policy Owner: VP of Operations & Delivery

1. Purpose and Control Status

This document defines CambridgeNexus, Inc. (“CNEX”)'s approach to maintaining and restoring operations on dedicated, bare-metal NVIDIA GB300 NVL72 infrastructure following a disruptive event. It consolidates our operational continuity policy, the recovery mechanics of our infrastructure platform, and the architectural approach behind our multi-site design into a single reference for prospective customers evaluating CNEX for regulated or business-critical AI workloads.

The Plan is designed so that:

- Recovery objectives are defined per asset class, measurable, and validated through exercises.
- Roles, decision authority, escalation paths, and recovery procedures are established before an event.
- Customer communications occur within contractual and operational timeframes.
- Recovery is validated against a defined technical gate before infrastructure is returned to customer service.
- The posture we report is never stronger than the evidence behind it (see Section 12).

2. Scope

This Plan covers dedicated customer infrastructure (GB300 NVL72 compute, NVMe storage, NVLink/NVSwitch fabric, ConnectX networking), the management plane (NVIDIA Mission Control, ProphetStor Federator.ai Cortex), supporting cloud and identity services, and CNEX corporate systems that support incident response and customer communication.

Customer workload continuity inside a customer-dedicated environment remains a customer responsibility under the Shared Responsibility Model described in Section 14. CNEX is responsible for restoring the infrastructure layer to a condition in which the customer can resume workloads.

3. Business Continuity vs. Disaster Recovery

Business Continuity and Disaster Recovery are related but distinct disciplines, and CNEX designs and reports on them separately.

	Business Continuity (BC)	Disaster Recovery (DR)
Goal	Keep essential services operating during disruption.	Restore IT services, data, and workloads after a major disruption.
Representative controls	Power / cooling / network redundancy; 24×7 monitoring and escalation; staff coverage and vendor support; capacity management and customer communications.	Backup, replication and checkpoint protection; secondary-site recovery capacity; platform restoration procedures; failover, validation and controlled failback.

BC keeps the business running. DR restores technology and service after disruption. Both are addressed throughout this Plan.

4. Recovery Principles

- Safety, personnel welfare, and containment take precedence over restoration speed.
- Information-security continuity is considered together with operational continuity.
- Recovery objectives are management targets validated by testing, not unconditional guarantees — achievement can depend on event scope, spare capacity, hardware lead times, colocation conditions, and third-party performance.
- A resource is not considered recovered merely because it powers on or becomes reachable; return to service requires defined technical validation (Section 10).
- Eliminate avoidable single points of failure across power, cooling, network, and control plane.

- Recovery priority is set by business impact and contractual SLA, and workloads are reallocated onto reserved or rebalanced GPU capacity rather than idle 100% duplication in every tier.

5. Risk Scenarios Covered

Category	Representative events
Utility / Power	Grid outage, UPS/PDU failure, generator or fuel disruption
Cooling / Liquid	Chiller, pump, CDU, or facility water-loop failure
Network	Carrier outage, fiber cut, routing or security-device failure
Compute / Storage	Server, GPU, storage, or control-plane failure
Cyber	Ransomware, credential compromise, destructive attack
Site / People	Fire, flood, natural hazard, access or staffing disruption

6. Infrastructure Resilience

Facility-level controls reduce the probability that a component failure becomes a service outage.

Power

- Redundant utility / distribution paths where facility design supports them
- UPS ride-through and generator backup, with fuel planning, inspection and load testing
- A/B power feeds for critical IT equipment

Cooling & liquid

- Redundant cooling components aligned to site design
- CDU / pump monitoring and alarm escalation, with leak detection and isolation procedures
- Thermal response and controlled workload reduction ahead of a hard fault

Network

- Redundant leaf/spine and management paths, with multiple upstream carriers where available
- Diverse routing / fiber strategy and HA firewall/router design
- Out-of-band management independent of the production network

7. Multi-Site Recovery Architecture

Recovery for site-level events depends on a second, geographically separated site holding replicated data and, for reserved workloads, standby capacity. Final implementation depends on site capacity commitments and customer SLA.

	Primary site	Recovery / capacity site
Holds	Production AI/GPU cluster, customer workloads, local high-speed storage, management and control plane.	Secondary capacity, replicated critical data, recovery orchestration, priority workload restart.

A global service layer (DNS/routing, identity, monitoring, customer access) and a protected data layer (backups, replication, snapshots, training checkpoints) span both sites.

Recovery sequence: site failure → isolate → protect data → allocate recovery capacity → restart/redirect → validate → fail back.

8. Recovery Objectives (RTO / RPO)

RTO (Recovery Time Objective) is the target time to restore a service after disruption. RPO (Recovery Point Objective) is the maximum tolerable period of recoverable data or configuration loss. Objectives are set per asset class — a single figure applied uniformly would be wrong in both directions, since GPU hardware replacement is constrained by a vendor cycle that no operational action can shorten.

Asset class	RTO	RPO	Delivering mechanism
AI infrastructure platform (compute, network, storage)	4 h	15 min	Fabric and facility redundancy for component failure; replication cadence for the recovery point.
Tier III+ facility (site, power, cooling)	4 h	15 min	Colocation provider redundancy and SLA; continuous thermal and power telemetry.
NVIDIA Blackwell GPU hardware	8 h	1 h	Vendor replacement SLA; predictive detection starts the replacement cycle before failure is total.
Portal metadata and provisioning state	4 h	1 h	Continuous automated database backup.
Identity and access	4 h	Inherited	Microsoft availability commitments.
Replication and recovery control (on the recovery path)	1 h	0	Redeploy from the packaged chart against the surviving cluster; replication records live in that cluster, not the site that was lost.
Failure prediction and quarantine	4 h	1 h	Its absence raises the rate of unplanned outages; it does not prevent a recovery already under way.
Evidence, posture and SLA reporting	24 h	4 h	Rebuild from version-controlled infrastructure code; its absence is a reporting gap, not an outage.

"Inherited" denotes an objective derived from a third-party agreement (e.g., Microsoft) rather than from infrastructure under CNEX control. The replication-control row reflects that this function now sits on the recovery path itself: the control that executes a recovery must return faster than the assets it recovers, so its objective is intentionally shorter than the platform's.

Illustrative pattern by service tier

For customers structuring their own service tiers, the following pattern illustrates how recovery approach and objective typically scale with criticality:

Service class	Typical pattern	Illustrative RTO	Illustrative RPO
Critical API / inference	Active capacity + automated redirect	Minutes	Near-zero to minutes
Platform / control plane	Redundant services + rapid restore	< 1–4 hours	Minutes
Training workloads	Restart from checkpoint	Hours	Checkpoint interval
Archive / non-critical	Backup restore	Hours–days	Hours–24h

Recovery targets should be contractually defined by service tier, not assumed uniformly across all workloads.

9. How the Recovery Point Is Determined

The recovery point is set by the link between sites, not by the replication software. Measured replication throughput was 0.68-1.09 GB/s (mean 0.98 GB/s), with snapshot and clone operations completing in approximately 20 and 50 milliseconds. Those two operations are constant-time, they do not grow with dataset size, so only the byte transfer itself scales with change volume.

Change per 15-min interval	Transfer time (measured mean)	Outcome
200 GB	~3.4 min	Within interval
600 GB	~10.2 min	Within interval
900 GB	~15.3 min	At the edge of the interval
1,200 GB	Exceeds 15 min	Achievable recovery point widens

Transfer times derived from measured throughput of 0.68-1.09 GB/s, real ZFS send/receive across 1, 4, 8 and 16 GiB datasets, integrity verified by checksum on 8 of 8 runs. Measured against storage middleware rather than the production appliance; to be re-measured on CNEX hardware. Change volumes shown are illustrative and workload-dependent.

The recovery point commitment for a given customer is set after measuring change rate on the actual workload and the actual inter-site link, not assumed in advance. The platform reports the interval it achieved on every replication, so drift from target is visible rather than assumed.

10. Standby Capacity and Recovery Time

Replication determines how much work is lost. It does not determine how quickly work resumes, that is set by whether hardware capable of running the workload is available at the recovery site at the moment it is needed. For general compute this distinction is minor, because capacity is fungible and can be rented within minutes. For GB300 NVL72 it is the dominant term: accelerated infrastructure of this class is procured on a lead time measured in weeks to months, so a recovery site with no reserved capacity has a recovery time equal to that lead time, regardless of how current the replicated data is.

Reserved capacity at recovery site	Description	Resulting RTO
Hot standby	100% (matched capacity, powered and current)	Minutes
Warm standby (tiered)	25–35% (covers workloads designated tier 1)	Hours
Pilot light	~10% (control plane and storage only)	Days
Procure on demand (current default posture)	Storage only; compute ordered after the event	Weeks

The replicated data is equally current in all four rows above, only the time to resume work differs, and it differs by three orders of magnitude. Which workloads warrant reserved capacity, and at what ratio, is a budget decision customers make with CNEX; it is not something any configuration can avoid.

Where the recovery clock is actually spent

Continuous replication removes data movement from the recovery clock. What remains is verification, promotion, and workload start, not restoring the working set.

Scenario	Recovery time	What the clock is spent on
Predicted GPU degradation	No outage	Planned drain and resume on healthy hardware.
Component or node failure	Minutes	Reschedule onto in-site spare capacity; data never moves.
Site loss, capacity reserved	Hours	Verification, promotion and workload start (not data movement, since data is already at the recovery site).
Site loss, no capacity reserved	Weeks	Data is already there; nothing can start until hardware exists to run it (procurement lead time).

11. AI / GPU Workload Recovery Approach

GPU disaster recovery balances recovery objectives against the high cost of maintaining idle accelerator capacity. CNEX follows a consistent six-step approach:

- 1. Protect: Replicate customer data, platform configuration and checkpoints.
- 2. Detect: Identify site/service impact through monitoring and incident command.
- 3. Prioritize: Classify workloads by SLA, customer criticality and recovery tier.
- 4. Reallocate: Reserve, rebalance or provision capacity at the recovery site.
- 5. Restart: Restart training from checkpoint; redirect inference/API traffic.
- 6. Validate: Confirm data integrity, performance, security and customer access.

Scenario reference

The table below summarizes how each recognized failure scenario is addressed and what governs its recovery time.

Scenario	How it is addressed	Time	Bounded by
Predicted GPU degradation	Device quarantined; workload drains on a planned schedule and resumes on healthy hardware.	No outage	Warning horizon (varies by failure mode)
Unpredicted GPU failure	Quarantine and reschedule onto in-site spare capacity.	Minutes	Spare capacity, then checkpoint interval
Node / component failure	Workload relocated in-site; hardware replaced under vendor SLA.	Hours	Vendor replacement SLA
Network uplink failure	BGP failover to secondary uplink; out-of-band path for manual failover.	Seconds–minutes	Facility and carrier
Control plane loss	Customer workloads continue on isolated networks; control redeployed from packaged chart.	< 1 hour	Chart redeploy (no customer workload waits on it)
Ransomware / data corruption	Recovery from a retained snapshot predating the event, not from the replica (see Section 12).	Hours–days	Snapshot retention depth

Scenario	How it is addressed	Time	Bounded by
Facility partial outage	Facility redundancy; continuous thermal telemetry.	Contained	Colocation provider SLA
Site loss, capacity reserved	Fence primary, promote recovery site, start workloads against data already present.	Hours	Verification and workload start
Site loss, no capacity reserved	Data is already present at the recovery site; hardware must still be procured.	Weeks	Procurement lead time

12. Cyber Incident and Ransomware Recovery

Replication copies corruption as faithfully as it copies good data, so a ransomware event reaches the recovery site too. Recovery in that case comes from a retained point-in-time snapshot predating the event (not from the live replica) which makes retention depth, not replication cadence, the control that matters. Retention depth is set against how long an intrusion might realistically go undetected, not against storage cost alone.

Phase	Controls
Protect	Role-based access and MFA; network segmentation and least privilege; backup encryption and retention policies; immutable/offline recovery copies where required.
Detect & contain	Security monitoring and alert escalation; isolate affected systems/credentials; preserve evidence; coordinate legal, customer, and vendor response.
Recover	Restore from known-good recovery points predating the event; credential and key recovery; validate integrity before reconnecting; post-incident remediation and lessons learned.

13. Incident Response, Activation, and Recovery Workflow

CNEX activates coordinated BC/DR response when a disruptive event occurs or is imminent and requires multi-function recovery beyond routine incident response. For example, an unresolved SEV1 incident, an event likely to cause SLA breach across more than one customer, near-complete loss of production infrastructure management capability, a colocation-facility emergency, or a confirmed destructive security event.

Workflow: Detect (monitoring/alert) → Assess (impact & severity) → Declare (BC/DR activation) → Contain (isolate failure) → Recover (failover/restore) → Validate (service & data) → Communicate (customer updates) → Fail back (controlled return).

Activation is led by a designated Incident Commander with defined succession, supported by function leads across facilities, network, platform, security, and customer communications. Key decision gates (declaring DR, selecting the recovery site, approving customer cutover, and validating failback) require explicit sign-off rather than proceeding automatically.

#	Action	Target
1	Declare activation; record trigger, time, and scope; open BC/DR event log.	Immediate
2	Notify the BC/DR team; confirm availability and assignments.	≤15 min
3	Post initial status update where customer-facing service is affected.	≤30 min
4	Send initial customer notification when impact is confirmed or likely within the hour.	≤1 hour

#	Action	Target
5	Initiate applicable recovery procedure(s).	Concurrent
6	Maintain timestamped actions, decisions, and communications for the event record.	Continuous

Return-to-service gate

Before restored infrastructure is returned to customer service, CNEX confirms: hardware health, power, thermal and firmware state are normal; expected GPU/Grace compute resources are present and diagnostics pass; NVLink/NVSwitch topology and link health are restored; network paths, routing and segmentation are normal; storage/NVMe integrity checks pass; customer network isolation is verified; monitoring, logging and security controls are operational; and no recurring hardware, link, or thermal errors are observed during validation.

14. Customer Communications During an Event

Customers receive proactive notification through two standard templates, issued within the timeframes in Section 13.

Service disruption notice

Subject: [URGENT] CambridgeNexus Service Disruption Notice on [Date]. States that CNEX is experiencing a service disruption affecting the customer's dedicated GPU infrastructure, with real-time updates at status.cambridgenexus.com and an urgent contact path via the customer's CNEX representative or support@cambridgenexus.com.

Security incident notice

Subject: [URGENT] CambridgeNexus Security Incident Notice on [Date]. Notifies the customer of a security incident that may affect their dedicated infrastructure environment, with the same escalation path.

15. Evidence, Testing, and Assurance

The failure continuity programs actually suffer is not a backup that fails, it is a backup that reports success and cannot be restored, or a replication that reports complete and moved nothing. Both look identical on a status page. CNEX addresses this in the data model rather than in a procedure: every replication record states whether real data moved, and each protected data class carries a grade that cannot exceed what has actually been demonstrated.

Grade	Meaning	How it is reached
Observed	Recovery has been demonstrated for this class.	Only by a restore or failover drill that passed, there is no other path.
Declared	Protection is configured but not demonstrated.	Configuration alone; stays here until a drill passes.
Unknown	Nothing has been established.	The default. A control that stops reporting degrades to Unknown rather than remaining green.

Aggregation is by weakest grade: a set containing one simulated transfer is reported as simulated, not averaged against the real transfers beside it.

What may be requested as evidence

- Per-class protection status with its grade, and for drilled classes, the measured recovery time
- The most recent restore and failover drill records, with dates and outcomes

- Replication records showing transfer mode and achieved recovery point
- Service level reporting against availability and incident response commitments
- Post-incident review from any prior plan activation

Testing cadence

Activity	Recommended cadence	Evidence produced
BC/DR tabletop exercise	At least annually; quarterly during initial production maturity	Scenario, participants, decisions, gaps, actions
Backup / restore validation	At least annually and after material backup changes	Restore result, measured RPO/RTO, exceptions
Break-glass credential test	At least annually	Access test, custody verification, remediation
Network failover exercise	At least annually where safe	Failover/failback results, traffic validation
Management-plane recovery exercise	At least annually	Rebuild/restore evidence, control validation
Technical recovery test	At least annually	Measured recovery time, diagnostics, acceptance gate

16. Control Mapping

Control state is declared at current operating status, so no row below requires discounting for a future state that hasn't been reached yet.

Control	Requirement	Supporting capability	State
ISO 27001 A.5.29	Security during disruption	Controls maintained through recovery; audit-record integrity verifiable after restore	Operating
ISO 27001 A.5.30	ICT readiness for continuity	Documented plan with defined objectives, named owners, mandatory activation criteria	Operating
ISO 27001 A.8.13	Information backup	Backup with verified restore; scheduling, retention and off-host copy	Phased
ISO 27001 A.8.14	Redundancy of processing	Facility, uplink and fabric redundancy; predictive workload relocation	Operating
ISO 27001 A.8.16	Monitoring activities	Continuous GPU, thermal and network telemetry with automated alerting	Operating
SOC 2 CC7.2	Anomaly detection	Nineteen-metric telemetry across eight GPU failure modes	Operating
SOC 2 CC7.4	Incident response	Activation criteria, named authority, notification timelines, event log	Operating
SOC 2 A1.2	Backup and recovery	Recovery procedures per scenario with defined objectives	Operating
SOC 2 A1.3	Recovery plan testing	Drills producing dated records with measured recovery times	Phased

"Phased" means the underlying control is performed today but its automation on a fixed cadence with machine-generated dated records is still being built out. A control performed and recorded manually is a real control with weaker evidence, and is marked accordingly rather than as fully automated.

17. Roles, Shared Responsibility, and Boundaries

Functional responsibility

Function	Responsibility
Data center facilities	Power, cooling, fire protection, physical access, facility incident response
Network operations	Carrier, routing, switching, firewall, out-of-band and connectivity recovery
Platform / cloud ops	Compute orchestration, scheduler, identity, monitoring and service restoration
Storage / data	Replication, backup, snapshots, restore validation and retention
Security	Cyber incident containment, forensics coordination and clean recovery
Customer / application	Application architecture, workload checkpoints, application-level HA and recovery requirements

Boundaries

Three constraints remain in place after everything above, and prospective customers should evaluate CNEX with these in view rather than assuming they are solved by the infrastructure layer.

Workload resilience remains with the customer. CNEX restores the infrastructure layer to a state where workloads can resume. Workload-level checkpointing and restart behavior within a dedicated environment remain the customer's responsibility. A four-hour infrastructure recovery time is time to usable infrastructure, time to productive output also depends on the customer's own checkpoint interval.

Failover between two sites is deliberately not automatic. With two sites and no independent third arbiter, no system can reliably distinguish a failed primary from an unreachable one. Automatic promotion during a network partition can produce two active primaries, and for cryptographically chained financial and audit records that yields two internally consistent histories that cannot afterward be reconciled. Promotion is therefore a deliberate decision with an explicit verification step, and that step fails closed: if the original primary cannot be proven inactive, promotion is refused. A refused promotion extends an outage; an unverified one can corrupt an audit record permanently, only one of those is recoverable, which is why CNEX accepts the former risk.

Inherited objectives are only as strong as their agreements. Identity, collaboration, and cloud connectivity objectives derive from third-party providers (e.g., Microsoft), and nothing in this Plan alters those providers' own commitments.

18. Maturity Roadmap

CNEX reports its BC/DR program honestly against a four-phase maturity model. Current phased items are tracked against Phase 4 completion.

Phase	Scope
Phase 1: Baseline	Business impact analysis; asset/service inventory; critical dependencies; initial RTO/RPO
Phase 2: Resilience	Facility & network HA; backup/replication; runbooks & escalation; recovery capacity plan

Phase	Scope
Phase 3: Multi-site	Cross-site data protection; workload recovery; traffic redirection; customer recovery tiers
Phase 4: Assurance	Scheduled exercises; measured RTO/RPO; audit evidence; continuous improvement

Goal: resilient infrastructure, protected data, repeatable recovery, and evidence that the plan works

19. Plan Governance

- Reviewed at least annually and after production go-live, material architecture change, major vendor change, significant incident, or exercise finding.
- Technical runbooks kept synchronized with deployed hardware, firmware, network topology, and vendor support procedures.
- Emergency contacts, escalation paths, and customer communication channels validated on the same cycle.
- Material changes require management approval; prior approved versions are retained per document-retention requirements.

Document Control

Field	Value
Policy owner	VP of Operations & Delivery
Effective date	May 1, 2026
This revision	11 August 2026
Sources compiled	CambridgeNexus BC/DR Plan (internal, Rev. 1.x); ProphetStor Federator.ai Cortex Customer Assurance Brief; CambridgeNexus DR/Business Continuity due-diligence deck

Detailed internal runbooks, named on-call assignments, and the dependency/escalation register remain in the internal-only BC/DR Plan and are not reproduced in this customer edition.