NexaCore Solutions: Cloud Resource Allocation Plan (FY2026)

Department: Operations (SRE/Infrastructure Team)
Plan Owner: [Head of Operations]
Date: 2025-09-15

1. Executive Summary

This plan outlines the resource allocation strategy for the upcoming fiscal year, focusing on controlling cloud expenditure while meeting the scaling demands of our target enterprise clients. The primary strategy is shifting from on-demand compute to a blended model utilizing **Reserved Instances (RIs)** and **Spot Instances** where appropriate, with a target **3-year cost saving of 18%**.

2. Allocation Principles

1. **Cost Centrality:** Every service must be tagged with a Cost Center (e.g., cc-RND, cc-Prod, cc-Ops).
2. **Rightsizing:** Quarterly review of all non-production resources to ensure they are appropriately sized for the workload (e.g., moving underutilized QA environments to smaller instance types).
3. **Automation:** Prioritize Infrastructure as Code (IaC) via Terraform for predictable, repeatable provisioning and de-provisioning, particularly for ephemeral testing environments.

3. Resource Budget Breakdown (FY2026 Estimate)

Cost Center Tag	Description	FY2025 Spend (Baseline)	FY2026 Budget (Target)	Allocation %
CC-PROD-CORE	Core Production Environment (Client-facing services)	$4,200,000	$4,800,000	60%
CC-RND-DEV	R&D/Development & Feature Testing Environments	$1,500,000	$1,300,000	16%

CC-DATA-LAKE	Centralized Logging, Analytics, and AI Training Data Store	$800,000	$1,000,000	12.5%
CC-OPS-TOOLS	Monitoring, CI/CD, and Internal IT Services	$500,000	$500,000	6.25%
CC-BACKUP-DR	Disaster Recovery Sites and Immutable Backups	$400,000	$400,000	5%
TOTAL		**$7,400,000**	**$8,000,000**	**100%**

Note: The increase in PROD-CORE and DATA-LAKE reflects expected client acquisition and increased AI model training requirements.

4. Scaling Strategy for Client Growth

To manage unpredictable enterprise client growth, we will utilize a two-tier scaling strategy:

Tier 1: Auto-Scaling & Vertical Scaling

- All customer-facing microservices within the K8s clusters will utilize horizontal Pod Autoscalers (HPA) based on CPU/Memory utilization, maintaining a target utilization of **65%**.
- Database clusters (e.g., PostgreSQL RDS) will be configured for automated vertical scaling (Storage) and read replicas (Horizontal scaling).

Tier 2: Pre-Commitment (Reserved Instances)

- **Target Coverage:** Aim to cover **75%** of the expected base load (calculated from FY2025 usage + forecasted small client base) with 1-year or 3-year Reserved Instances (RIs). This provides immediate guaranteed capacity at a discounted rate.
- **Justification:** Given the long-term contracts (3-5 years) with our enterprise clients, RIs offer stability and significant cost savings over on-demand pricing for predictable workloads.

5. Implementation Roadmap

Quarter	Focus Area	Key Action Items	Budget Impact
Q1	Resource Rightsizing & Tagging	Implement mandatory resource tagging script across all accounts. Rightsize 90% of Dev/QA environments.	Moderate savings (Immediate)
Q2	RI Commitment	Purchase 3-year RIs to cover 50% of forecastable base load for core services.	High up-front cost, immediate long-term savings
Q3	Autoscaling Optimization	Fine-tune HPA configurations to prevent over-provisioning during peak hours. Evaluate serverless options for batch processing.	Moderate savings (Ongoing)
Q4	FY2027 Forecasting	Begin capacity planning and budget forecasting for the next fiscal year based on Q1-Q3 usage patterns.	Planning