Devopstrio

# AI Infrastructure Support (GPU, Azure AI, AWS SageMaker, etc.)

Accelerate. Automate. Scale

# Introduction

Devopstrio

The primary objective of this initiative is to establish a unified and scalable AI Infrastructure Support Framework that enables seamless development, deployment, and management of AI/ML workloads across hybrid and multi-cloud environments. This framework focuses on optimizing GPU utilization, accelerating model training, ensuring cost efficiency, and maintaining enterprise-grade governance across AI platforms like Azure AI, AWS SageMaker, and on-prem GPU clusters.

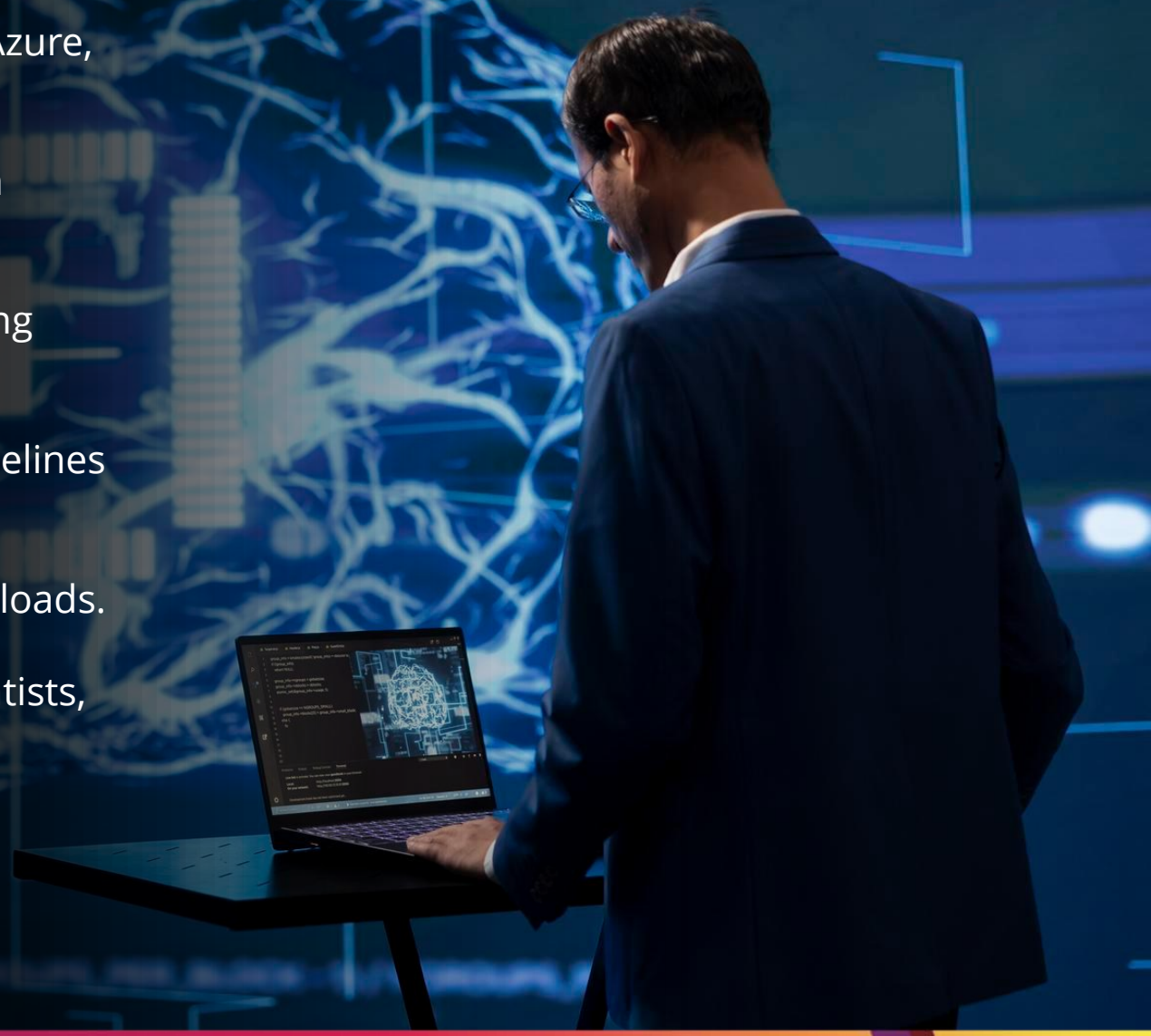
**Key Drivers:**

- Rapid enterprise adoption of AI and machine learning workloads requiring high-performance compute.
- Increasing demand for GPU-optimized infrastructure for model training and inferencing.
- Rising cloud costs due to unmonitored GPU consumption and idle instances.
- Need for standardized MLOps pipelines across multiple cloud environments.
- Strategic goal to build a resilient, scalable, and cost-efficient AI foundation.

# Why Establish AI Infrastructure Support?

Devopstrio

- ✓ Streamline management of AI/ML environments across Azure, AWS, and on-prem.
- ✓ Optimize GPU resource allocation and utilization through automation and scheduling.
- ✓ Ensure high availability, security, and scalability for training workloads.
- ✓ Reduce time-to-market by enabling pre-configured AI pipelines and environments.
- ✓ Integrate cost control and FinOps practices for GPU workloads.
- ✓ Enable cross-functional collaboration between data scientists, DevOps, and infrastructure teams.



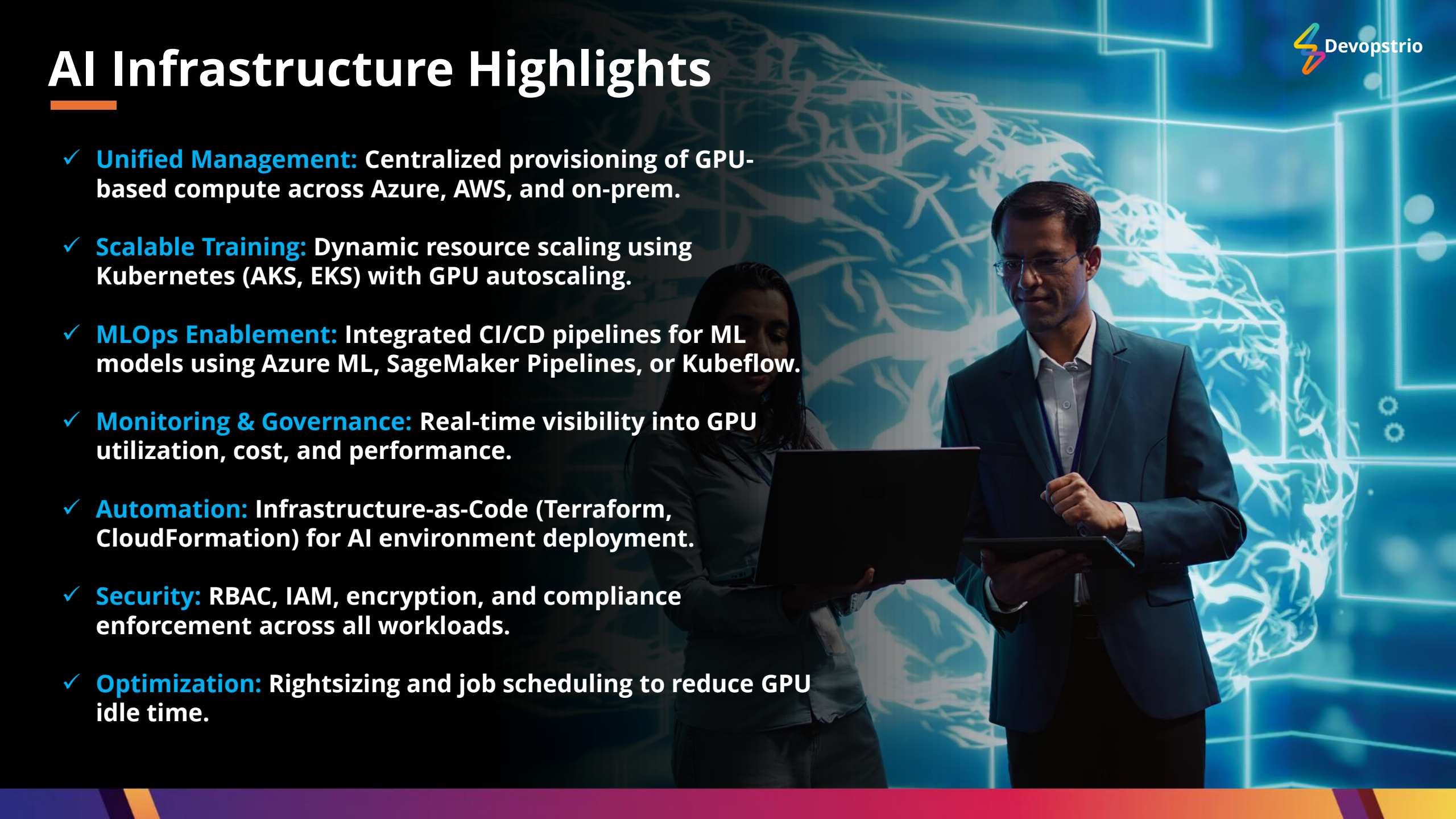
# Current Environment Overview

- ✓ AI workloads deployed inconsistently across multiple cloud accounts.
- ✓ Limited visibility into GPU usage, cost, and job performance.
- ✓ Manual setup of AI/ML environments with varying dependencies and tool versions.
- ✓ Lack of automated MLOps pipelines for model training and deployment.
- ✓ Data scientists using independent Jupyter or SageMaker notebooks without governance.
- ✓ No centralized monitoring or cost tracking for GPU-based workloads.

Devopstrio

# AI Infrastructure Highlights

- ✓ **Unified Management:** Centralized provisioning of GPU-based compute across Azure, AWS, and on-prem.
- ✓ **Scalable Training:** Dynamic resource scaling using Kubernetes (AKS, EKS) with GPU autoscaling.
- ✓ **MLOps Enablement:** Integrated CI/CD pipelines for ML models using Azure ML, SageMaker Pipelines, or Kubeflow.
- ✓ **Monitoring & Governance:** Real-time visibility into GPU utilization, cost, and performance.
- ✓ **Automation:** Infrastructure-as-Code (Terraform, CloudFormation) for AI environment deployment.
- ✓ **Security:** RBAC, IAM, encryption, and compliance enforcement across all workloads.
- ✓ **Optimization:** Rightsizing and job scheduling to reduce GPU idle time.



# Support Goals

- ✓ Achieve >90% GPU utilization efficiency across all environments.
- ✓ Reduce AI infrastructure cost by 25–30% through optimization and automation.
- ✓ Enable end-to-end visibility into model training, cost, and performance metrics.
- ✓ Establish MLOps best practices for repeatable and automated workflows.
- ✓ Ensure compliance with enterprise data security and AI governance standards.

# Tools & Techniques

Devopstrio

- ✓ **AI Platforms:** Azure Machine Learning, AWS SageMaker, GCP Vertex AI.
- ✓ **Compute & Orchestration:** AKS, EKS, Kubernetes, Docker, and GPU scheduling tools.
- ✓ **Infrastructure Automation:** Terraform, Bicep, CloudFormation.
- ✓ **Monitoring & Analytics**: Azure Monitor, CloudWatch, Grafana, Prometheus.
- ✓ **MLOps & Pipelines:** Kubeflow, MLflow, GitHub Actions, Azure DevOps.
- ✓ **Cost & Optimization Tools:** Azure Cost Management, AWS Budgets, CloudHealth.
- ✓ **Security & Compliance:** IAM, Azure Policy, KMS, and VPC-based isolation.

# Migration Process

Devopstrio

## 1.Assessment & Planning

- Evaluate current AI workloads, GPU needs, and cost trends.

## 2. Architecture & Design

- Define target GPU infrastructure and cloud integration model.

## 3. Proof of Concept (PoC)

- Test hybrid deployment and measure training performance improvements.

## 4.Automation & Deployment

- Implement IaC templates, scaling policies, and job scheduling automation.

## 5. Monitoring & Governance

- Enable dashboards for cost, usage, and GPU performance.

## 6. Continuous Optimization

- Review workloads, tune performance, and refine model deployment pipelines.

# Expected Outcomes

- ✓ Accelerated AI model development with scalable GPU provisioning.
- ✓ Reduced cloud spend through automated rightsizing and cost governance.
- ✓ Enhanced collaboration between data science and infrastructure teams.
- ✓ Improved visibility into workload efficiency and performance metrics.
- ✓ Unified AI operations framework across multi-cloud environments.
- ✓ Enterprise-grade governance ensuring compliance and cost accountability.


Devopstrio

# Challenges and How We Address Them

Devopstrio

**GPU Resource Bottlenecks**

- Implement auto-scaling clusters and GPU job scheduling.

**Cost Overruns on Cloud GPUs**

- Integrate FinOps monitoring and spot instance utilization.

**Fragmented AI Workflows**

- Standardize pipelines via MLOps frameworks.

**Security & Compliance Concerns**

- Apply IAM roles, encryption, and isolated VPCs.

**Model Deployment Complexity**

- Automate CI/CD workflows for model lifecycle.

**Multi-Cloud Management Overhead**

- Use unified dashboards and IaC for consistent provisioning.


# Deliverables

Devopstrio

- ✓ AI Infrastructure Assessment Report – Current workloads, GPU utilization, and optimization potential.
- ✓ Target AI Architecture Blueprint – Hybrid topology, orchestration layers, and data flow diagrams.
- ✓ MLOps Implementation Plan – Automated model lifecycle and CI/CD workflow design.
- ✓ Pilot Performance Report – GPU efficiency metrics and cost reduction outcomes.
- ✓ Automation Templates & Scripts – Terraform / CloudFormation for environment provisioning.
- ✓ Monitoring & Governance Dashboards – GPU utilization, spend tracking, and alerts.
- ✓ Operational Runbooks & SOPs – AI workload management, scaling, and troubleshooting procedures.
- ✓ Training & Knowledge Transfer Materials – For DevOps, Data Science, and Cloud teams.
- ✓ Final ROI & Optimization Report – Cost savings, performance benchmarks, and governance maturity.

# Trusted by


Microsoft


BT


GXO


airbnb


METRO
BANK


COSTCO
WHOLESALE


Virgin
media


servicenow


TOPLAND
GROUP


paymentology


bp


GoDaddy


# Global HQ

**United Kingdom,**
**128 City Road, London, EC1V 2NX**

**Global Presence & Delivery Centers**
- **HQ:** London, United Kingdom
- **Offices:** New York (USA), Chennai (India)
- **Delivery Model:** Hybrid (Onshore + Offshore)
- **Support Coverage:** 24x7 Global Operations
- **Compliance:** GDPR, HIPAA, ISO 27001

Company Registered in England & Wales Number 13024867

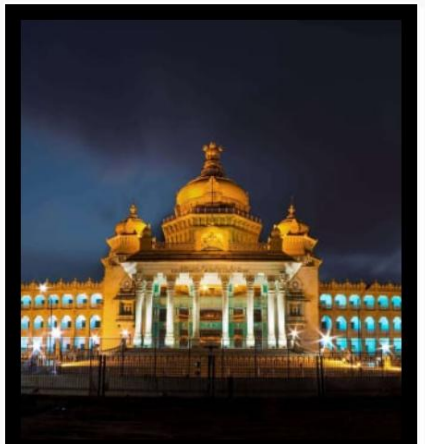
# Global Presence

LONDON
UNITED KINGDOM

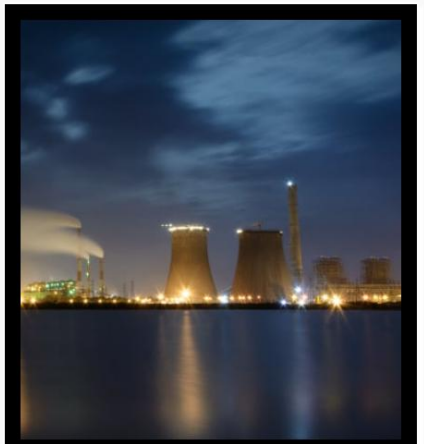
TENNESSEE
UNITED STATES

# India offices

CHENNAI
INDIA



BANGALORE
INDIA



PASUVANTHANAI
INDIA

# Thank You!

## Let’s Discuss

## How We Can Accelerate Your Transformation Journey

**Contact:**

info@devopstrio.co.uk

www.devopstrio.co.uk

LinkedIn