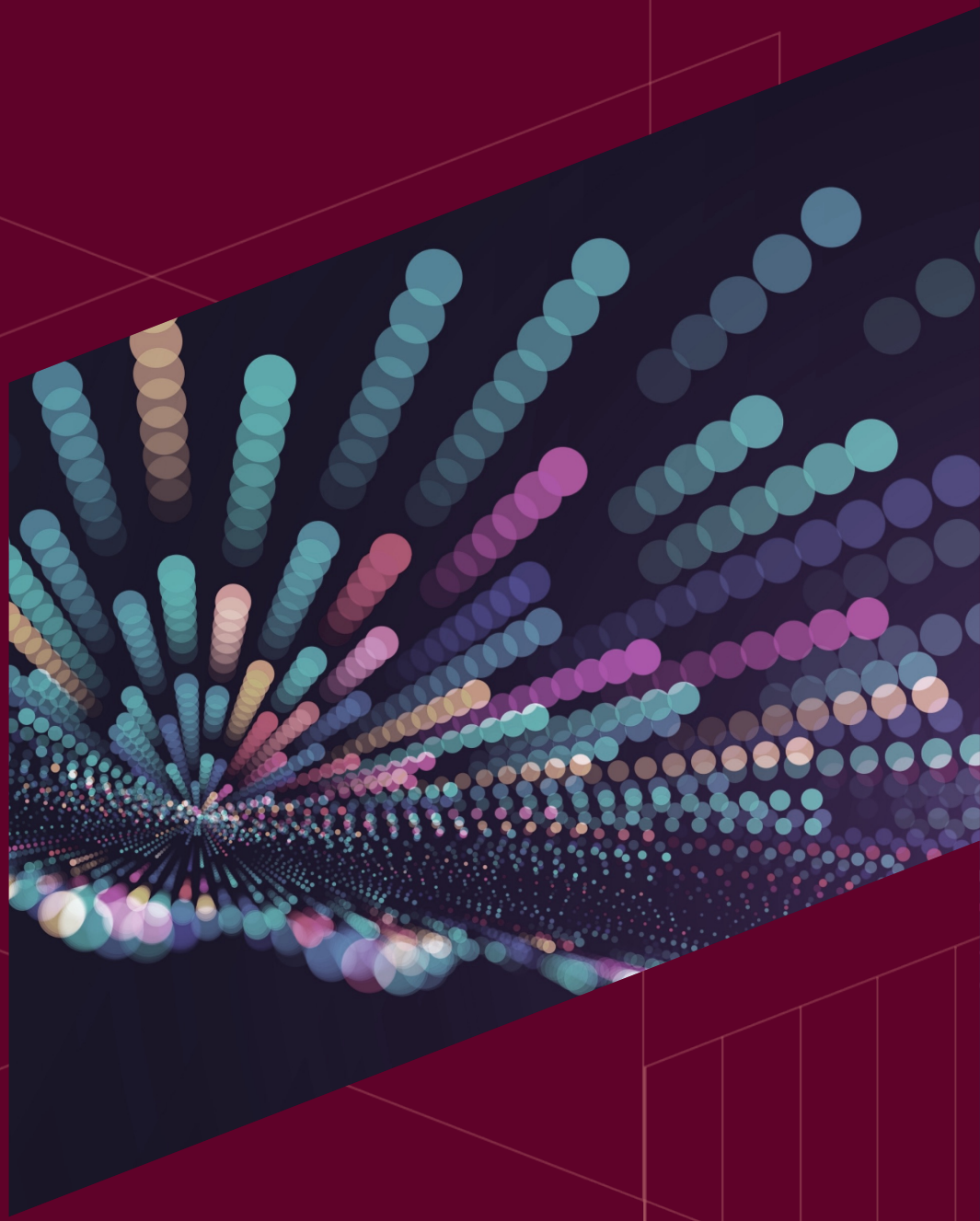
THIRD EYE ADVISORY

TECH mahindra

Whitepaper

# Cloud Services Cost Optimization - Beyond FinOps: Tech Mahindra's Perspective



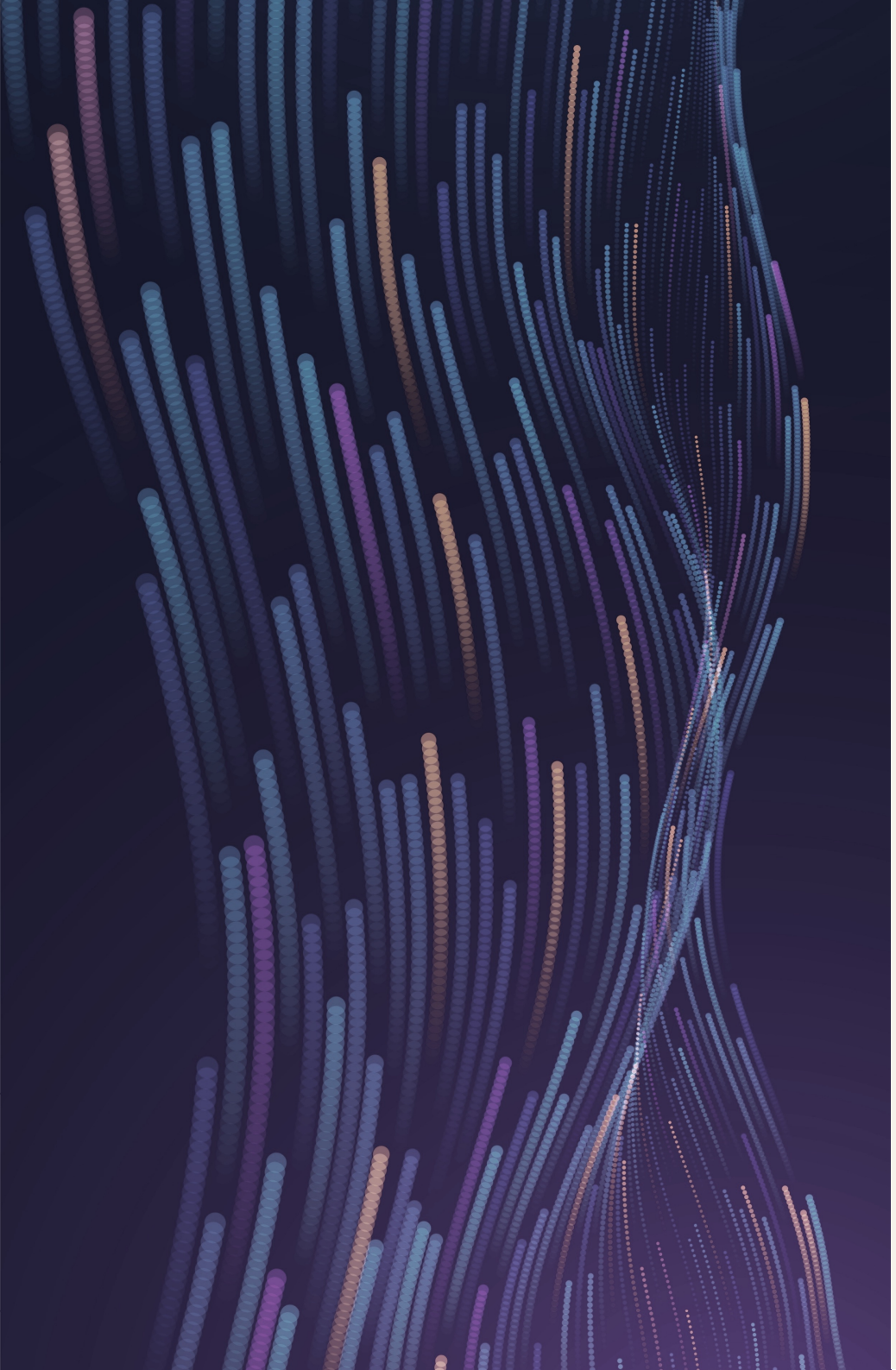

Authors

Dr. Sakthivel Ramaiah, Tech Mahindra

Amar Samarth, Tech Mahindra

PRABHAS HARLAPUR, Intel Corporation

MOUSUMI DASH, Intel Corporation



# Executive Summary

Organizations adopted public cloud services with intention to cut down on their on-premises spend and shift from a ‘CapEx’ to ‘OpEx’ model. However, post cloud adoption, they experienced cost overruns, leading them to explore options for optimization and reducing cloud spend. Typically, conversations around this revolved around a particular FinOps tool. However, from our perspective, cost optimization is not about a tool. It is a process and approach before and after implementation / moving to cloud.

Typically, any organization approach starts their cloud journey by adopting infrastructure as a service (IAAS) cloud services, which mostly involved a lift and shift sort of migration from on-premises to cloud. Later, app modernization phase involved adoption of Microservices and containerization of applications. Also, current trend of shifting from IAAS services to PAAS services is adding difficulty in standard cost management approaches.

Cost control starts with the architecture design level process within organization and adoption of the tools. The idea of this white paper is an effort to highlight the options and approach of cost reduction / optimization beyond FinOps tool.

# Table Of Contents

# Concerns of Cloud Customers

According to the latest Flexera 2025 State of the Cloud report, more than half of all data now resides in public clouds, and 84% of respondents cite cloud spend as their top challenge. With cloud spending expected to increase by 28% in the coming year, these findings indicate that organizations are rethinking cloud cost management strategies to manage increasing complexity while maintaining cost-effectiveness and financial control.

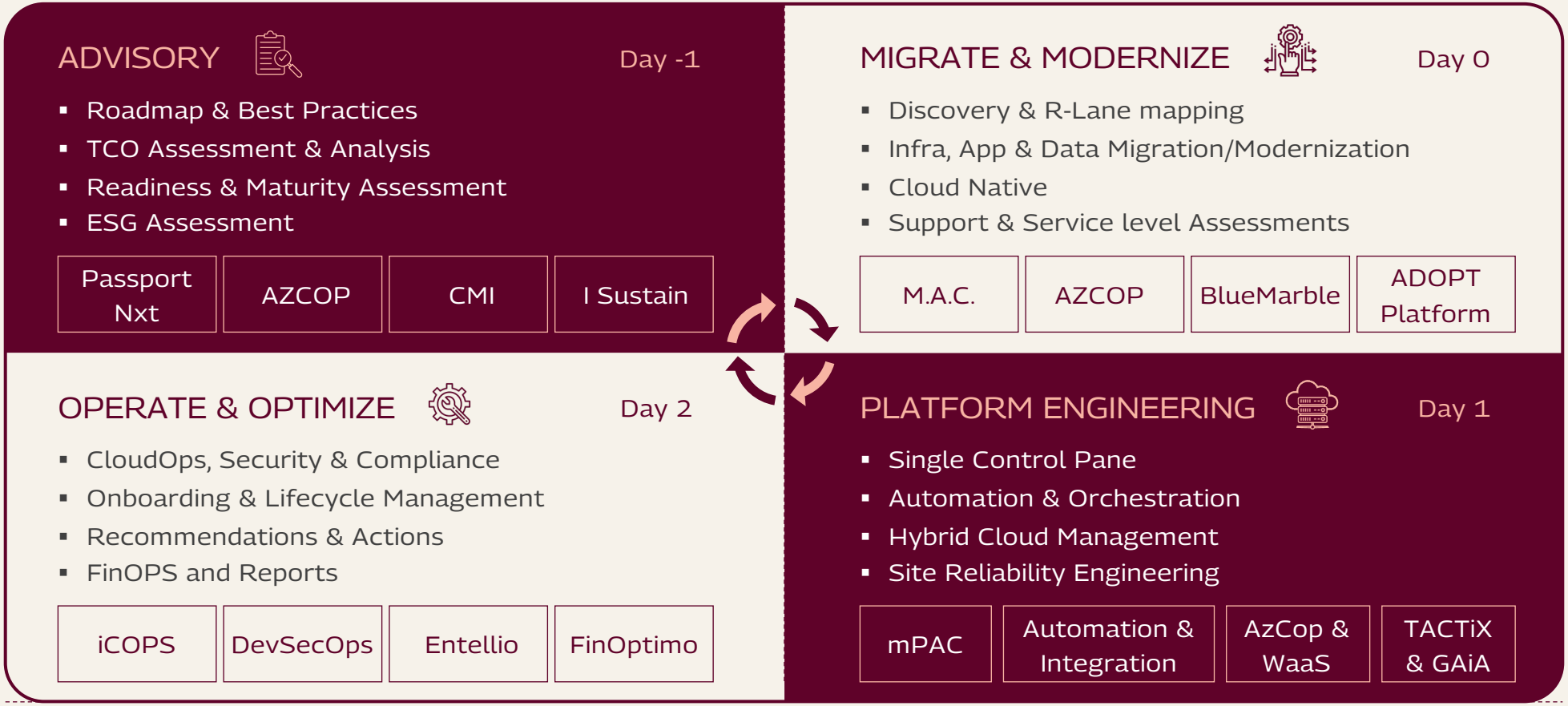
Gartner estimates worldwide public cloud end-user spending at $723 billion, which is up from $595.7 billion in 2024, highlighting the continued shift of enterprise IT investments to the cloud. By 2027, 90% of organizations are expected to adopt hybrid cloud models, with double-digit growth driven by AI and GenAI adoption, integrated cloud infrastructure and platform services, and distributed multi-cloud architectures.

As public cloud adoption accelerates and cloud spend remains a top priority, organizations are increasingly focused on improving visibility, governance, and financial accountability across their cloud environments. Tech Mahindra's Cloud BlazeTech (CBT) platform supports this need by enabling enterprises to analyze, manage, and optimize cloud environments, with FinOps serving as a core focus area for disciplined cloud cost governance.

# Tech Mahindra Cloud BlazeTech Platform

Tech Mahindra Cloud BlazeTech (CBT) is a platform that delivers an end-to-end cloud Solution to customers, right from consulting and defining the cloud roadmap, to setup, manage and automate their future hybrid integrated cloud environment.

This platform was built with a vision of delivering a Unified Cloud Platform, with right blend of frameworks and industry-quality standards that simplify an enterprise's path to orchestrating and achieving customers cloud journey.



In “Platform Engineering” phase Cloud BlazeTech uses Intel powered VMs to provisions optimized workloads and save on cost across hybrid cloud environments”.
In “Operate and Optimize” phase, cloud BlazeTech helps monitor and deliver continuous improvement in cloud cost spend and optimize utilization leveraging FinOps approach.

# FinOps Approach

FinOps is an operational framework and cultural practice which maximizes the business value of cloud and technology, enables timely data-driven decision making, and creates financial accountability through collaboration between engineering, finance, and business teams.

Business and technology strategy of an organization defines the scope of the FinOps, which gives context to the FinOps practitioners. It revolves around understanding the cloud usage and pattern, then quantifying it in terms of budgeting and forecasting and finally optimizing on the usage and Cost. During provisioning phase, leverage Intel-powered VM SKUs offers better performance-per-dollar ratios, especially for compute-intensive workloads

# Typical FinOps Tool

Focuses on the spend, monitors usage and remediation is in a reactive form, wherein actions are taken after the spend. However, it requires a pro-active approach like selecting right resources before provisioning a workload.

However, the result does not meet organizations expectations as they expect additional 10% - 15% reduction in cloud spend every year. This is forcing them to explore additional measures and methods to cut down on their cloud monthly spend.

Reasons for cost overrun could be due to their approach and not adhering to best practices for resource provisioning and governance. Organizations adopt cloud and rely on FinOps tool to control their cloud spend. They leverage cloud native / third party FinOps tool like Densify, FinOptimo, etc. and rely on this tool to help reduce cloud cost.

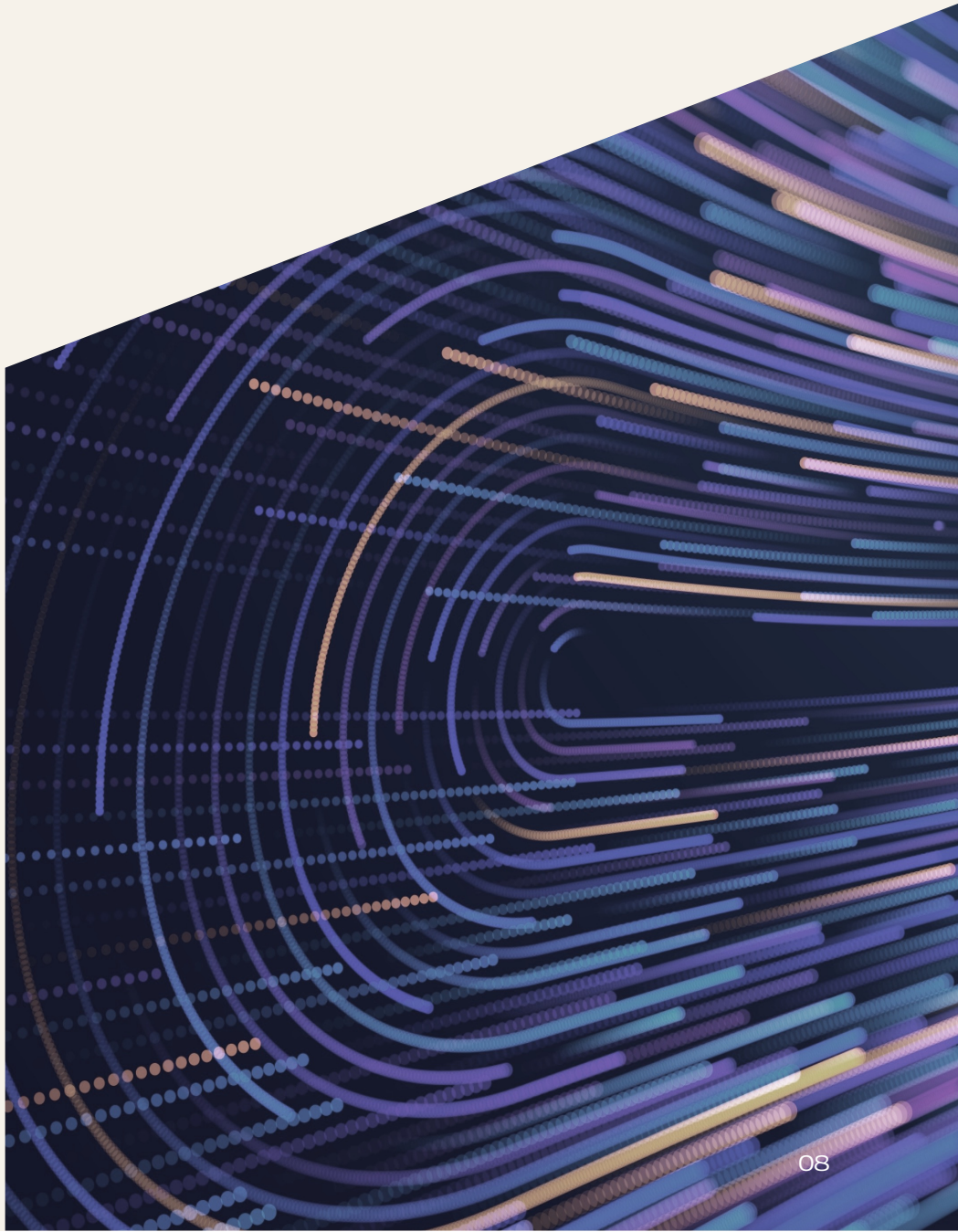
Most of the FinOps tool delivers below functionality and help manage cloud spend at resource (VM) Level.

- Cloud cost monitoring and dashboards
- Resource (CPU / RAM / Disk) utilization
- Recommendation and Remediation: Resizing, Convert to RI / Spot Instance, etc.
- Budget and quota management
- Scheduling customized cloud spend report
- Support Integration with ITSM, Billing, etc.

FinOps service provided by Service providers like Tech Mahindra, include implementation, Configuration and Management of FinOps tool. They bring-in their expertise in implementing, configuring (Power scheduling, Tagging Process, Responsibility Mapping against Budget, etc.) and managing the Tool.

# Cloud Subscription & O/S License

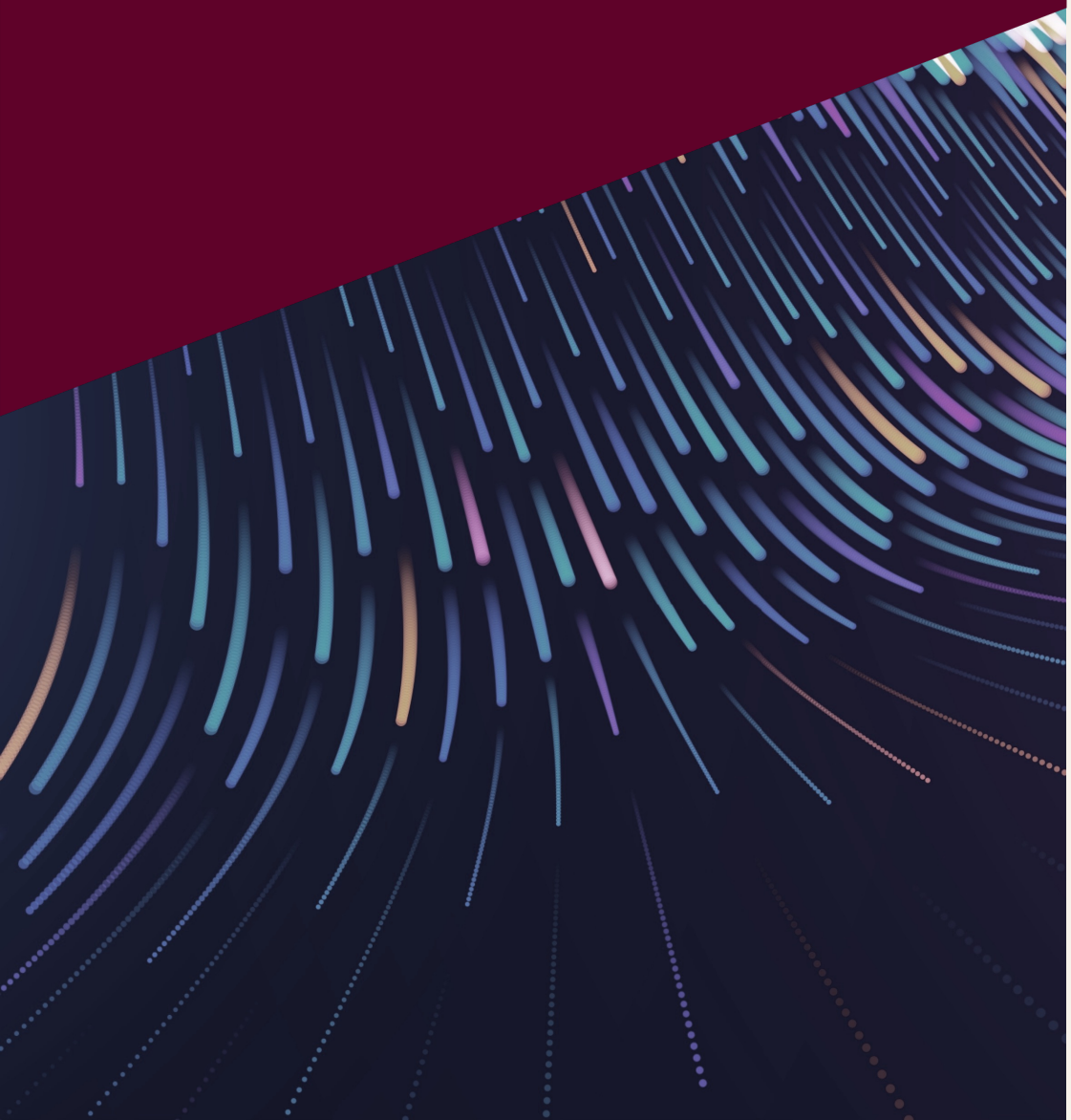
Every Organization should start their cloud cost management by analyzing their spend on cloud subscription and O/S licenses.  Organizations usually sign an (Enterprise Agreement) with hyper-scalers to get standard discounts on committed spend. However, if organization is not into an 'Enterprise Agreement' with their CSP vendor, then opting for license and support from an GSI Organization like Tech Mahindra, can be overall better pricing and many additional benefits which can further help reduce their monthly spend. Intel-based platforms offer compatibility and enhanced performance for BYOL strategies, especially for Windows Server and SQL Server workloads. Refer below comparison table to check on the benefits which can be leveraged by switching from an EA to CSP type Subscription & Support services and save up to 12% on their Monthly bill.



## Subscription from GSI (Tech Mahindra) over EA (Hyper Scalers)

| Scope / Features | GSI (Tech Mahindra) | EA (Enterprise Agreement) |
|---|---|---|
| Increase Workloads (True-Up) | Any Time | Any Time |
| Decrease Workloads (True-Down) | Any Time | Annually with 30 days advanced notice |
| Requirements to maintain minimal resource count (commitments) | Not Required | Required |
| Discounts | Various incentive programs available from Microsoft to discount Azure costs dependent on monthly consumptions. Provider can offer competitive pricing | Discretionary discount by Microsoft Field Seller (5-10% based on annual consumption) |
| Length of Contracts (Commitments) customer can cancel any time | Flexible Options - There is no commitments and customer can cancel anytime | 3 Years |
| Payments | Monthly or Annual Payments | Full (upfront) or Annual Payments |
| Support | Advanced support for Partners (ASfp) or Premier support Basic support and additional charges for Microsoft for Partners (PSfp) | Basic support and additional charges for Microsoft Premier Support |
| Cost | Calculated based on overall subscriptions and consumptions | Fixed cost per VM |
| Coverage | Most of the Microsoft Products covered (Azure, Microsoft 365, Dynamic 365) | All Microsoft Products covered |
| FinOps Support/Administration | CSP can leverage their support team knowledge on customer environment, Usage Summary, Customer's Strategy and recommend periodic optimization plans to reduce the consumption and operating cost | Dependent on Microsoft's recommendations and Customer SME's recommendations |

# Operating System License (BYOL & AHUB)



As a part of analysis, evaluation of the O/S Licenses (excluding open-source OS) used for on-premises resources which are to be migrated to public cloud can give drastic cost benefits by leveraging BYOL & AHUB (for Azure Cloud Only) features.

- Leveraging BYOL helps on saving License cost
- Organization can save up to 85% over standard pay-as-you-go rate leveraging Windows Server & SQL Server Licenses with Azure Hybrid benefit.
- Azure Hybrid benefit for Linux helps consolidate RedHat & SUSE O/S Subscription under one and leverage co-located technical support for all 3 O/S
- One on-premises license to be used for up to four cloud resource O/S

A-HUB benefit plan (Azure Hybrid Use Benefit): For each Windows Server dual processor license with Software Assurance, customers can run two virtual machines with up to 8 cores or a 16-core virtual machine, whichever is lower in price.

# Evaluate Existing Architecture

Review architecture from a cost perspective can help conclude on the areas where we can reduce cloud monthly spend. Some of the areas which can be reviewed are:

- Resource (VM) – Redundancy Required (LRS / GRS), Disk Type,
- Storage – Storage Tier, GRS to LRS,
- Backup – On Cloud, Leverage non-native storage solutions
- Ingress & Egress Cost Analysis
- Region Comparison – Resource Cost varies in different regions

Few configurations example is as mentioned below.

Example of enforcing Azure Policy and other methods for cost optimization:

- Restrict VM SKUs: Prevent deployment of expensive VM sizes. policy: "Allowed virtual machine SKUs"
- Enforce Auto-shutdown: Automatically shut down dev/test VMs during off-hours.
- Policy: "Configure auto-shutdown on virtual machines"
- Tagging for Cost Tracking: Policy: "Require a tag on resources" - Enforce tagging (e.g., Environment, Owner, Cost Center) to track and allocate costs.

## Specific Optimization Tactics by Resource Type

### Virtual Machines

- Use Reservations (RI) or Savings Plans
- Use Auto-scale and Auto-shutdown
- Migrate to B-series or Spot VMs for dev/test
- Intel Xeon Scalable processors provide built-in accelerators and workload-optimized SKUs to help achieve FinOps targets. With built-in telemetry and NUMA-aware architecture enable fine-grained workload tuning, unlocking up to 50% cost savings.

### Containers (AKS)

- Use Cluster Auto-scaler
- Use Virtual Nodes for burst workloads
- Optimize node pool sizes and pod density

### Storage

- Move infrequently accessed data to Cool or Archive tiers
- Enable lifecycle management policies
- Use Premium storage option only when needed

### Databases

- Use SQL Serverless for intermittent workloads
- Scale down or pause DBs when idle
- Use Elastic Pools for multiple small databases
- Intel-optimized database configurations can reduce query latency and improve throughput, especially when using Intel-validated tuning scripts.

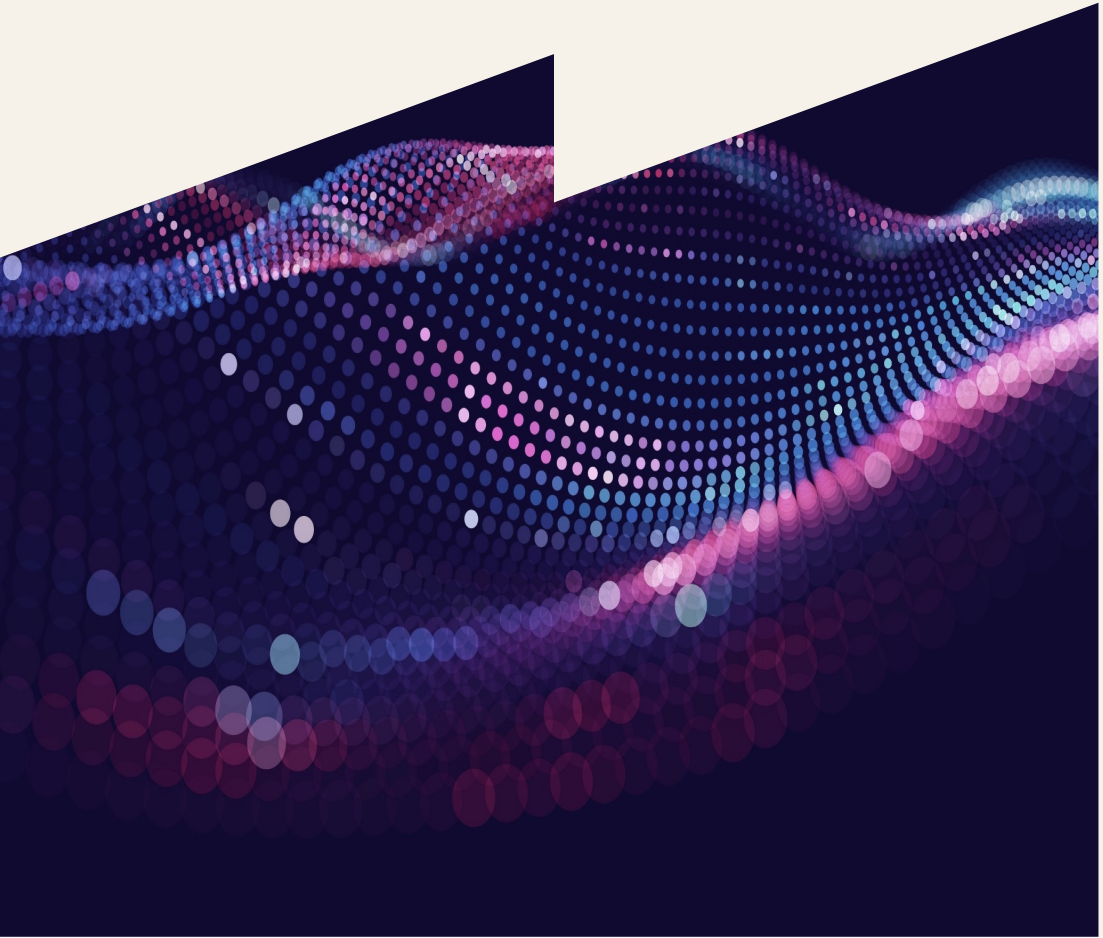
After reviewing and implementing cost optimization measures at license, process and architecture level, then organization can leverage FinOps tool to check cost optimization at utilization level.

Organization can select right FinOps tool based on their need for example:

- Cloud based or on-premise tool
- Supporting single or multi-cloud
- Time to implement and duration for which we need to use that tool
- Tool cost and its pricing model
- Agent or agentless
- Overall saving (including Tool cost)

# Review Cost Optimization at Kubernetes Nodes & PODs Level

Most of the FinOps tool monitor and manage utilization at virtual machine / compute node level. However, if we look at the real-time cost insights into a Kubernetes Cluster Nodes, we can see a scope in optimizing at Pod level, which are run in Compute Nodes of an K8s cluster.



Organization can leverage open-source tools like Kubex / Kubecost / Karpenter to optimize at Pod level and automation to further reduce on cloud costs.

Prime focus of this approach will be to

- Optimize and manage cost at POD level in an K8s environment
- Integrate with FinOps Tool to get single view on optimization at Node as well as POD level

Further Optimization at Pod level can be achieved by leveraging tools like Kubegreen. Kubegreen acts as an additional layer to complement Kubecost, by creating a 'Sleep Info' policy to automatically scale down workloads during off-hours to further reduce on cloud costs.

Processor-Level Optimization for Cloud Workloads: After optimizing at the FinOps levels, the next frontier is processor-level tuning, which can unlock 15% to 50% cost savings and significant performance gains when aligned with workload characteristics.

## Key Optimization Strategies

- Processor Vendor Selection: Evaluate performance-per-dollar using Built-in accelerators and better NUMA handling
- Generation Matching: Use the latest CPU generations to leverage architectural improvements
- SKU Optimization per cloud provider, select SKUs based on: Workload type, performance benchmarks
- Leverage benchmarking data: Use industry standard tools to simulate and validate workload behavior before deployment

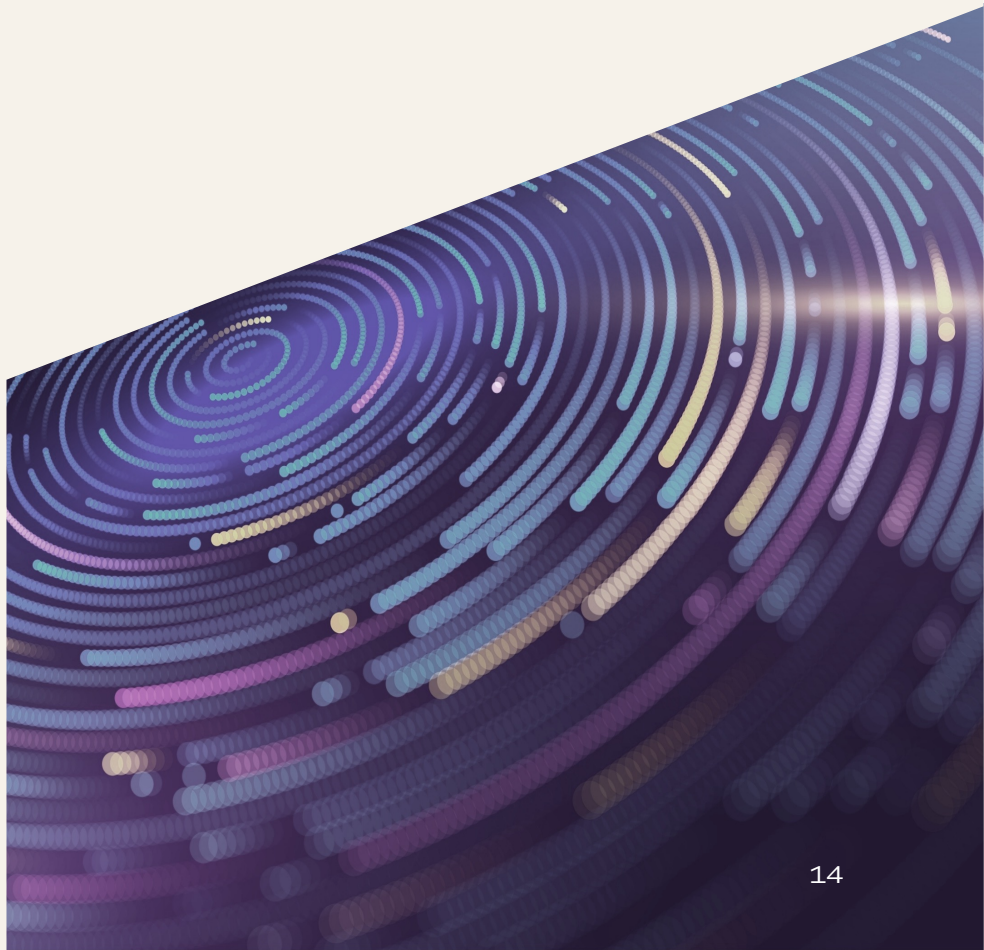
## Workload Optimization Using Automation Scripts Optimization Techniques

- Dynamic Tuning Scripts: Adjust kernel parameters, CPU affinity, and I/O schedulers based on workload profiles.
- NOPM/TPM Boosting: For transactional systems (e.g., databases), tuning can yield up to 40% performance gains.
- Down-tiering Strategy: If performance is not critical, run workloads on lower-tier SKUs with tuning to save 10-20% in costs.

## Tech Mahindra + Intel: OCR/OCM Framework

Optimize Cloud Recipes (OCR) and Optimized Cloud modules (OCM) are pre-tested scripts for:

- Databases (PostgreSQL, MySQL, Oracle)
- MS-SQL , MariaDB
- Web and Application Servers



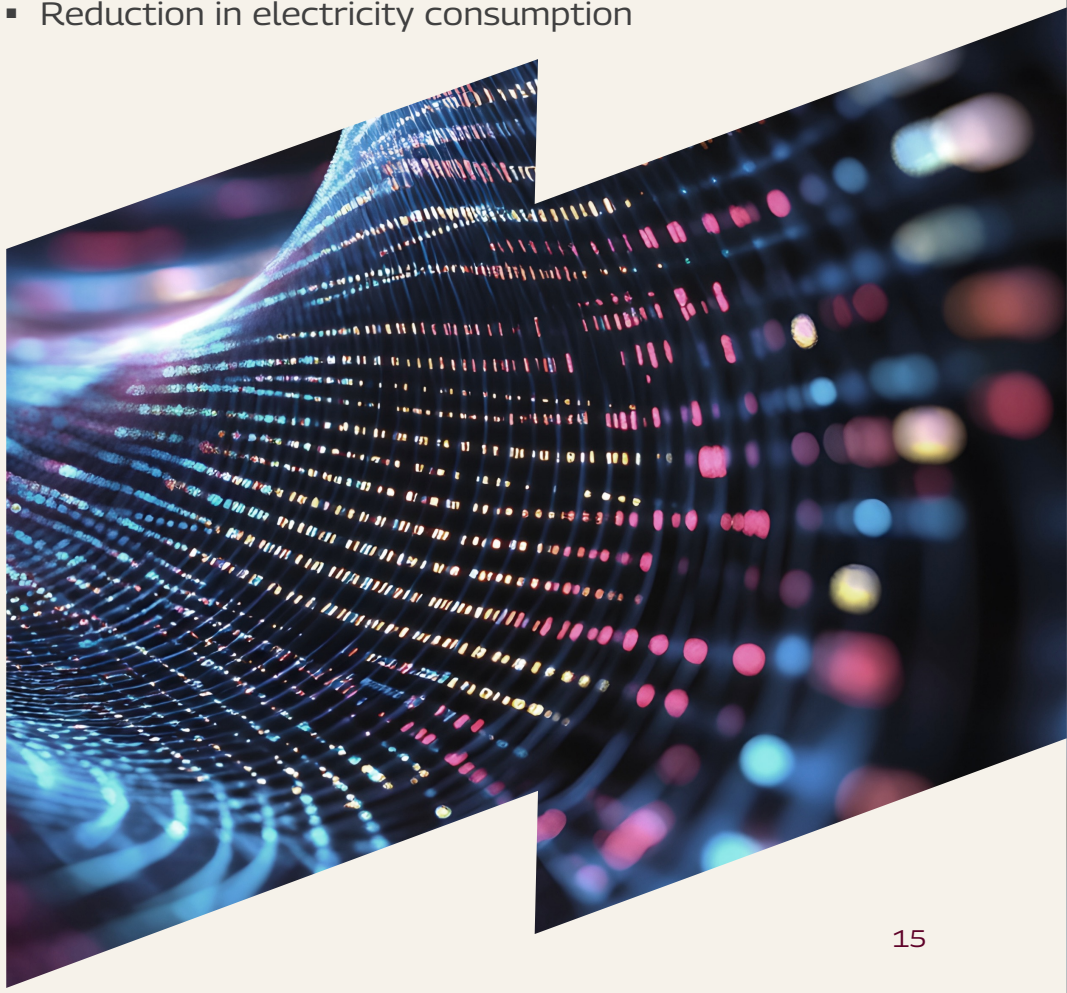
These modules are validated on Intel platforms and benchmarked using tools like YCSB, & HammerDB.

Control Power consumption at Processor Level: After we complete optimization at VM, POD, Workload and Processor level, we can still explore further reduction in cost, by means of controlling the power consumption of resources, primarily for on-premises DC / private Cloud.

- Reduces energy consumption in data centers by Intel's advanced power management features, including P-states and C-states, contribute significantly to energy savings in private cloud and on-prem environments. intelligently managing CPU power states (P-states) and idle states (C-states). It continuously adjusts these states based on real-time server utilization, offering substantial energy and cost savings without affecting performance.
- This helps lower organizations operational costs, reduce energy usage, and contribute to sustainability goals.
- This in turn reduces cooling requirement which further saves on energy cost.
- Up-to 30% saving in energy consumption (electricity).

Tech Mahindra as an organization has expertise and strong capability across all the above-mentioned areas of cost optimization, which can collectively assure customers of a minimum 20% savings along with the benefits like:

- Increase in performance (network, processor)
- Reduction in Cloud spend
- Reduction in electricity consumption



# About the Author

With over 25 years of experience across industries, specializes in IT infrastructure, applications, and cloud transformations. As VP/Global Head of Cloud Platforms, he drives competitive strategies for hybrid multi-cloud practice covering strategy, assessment, migration, Cloud Ops, FinOps, DevOps, and application modernization.

## Dr. Sakthivel Ramaiah

Group Practice Head,
Platforms and Solutions,
Tech Mahindra

With over 20 years of experience in building capabilities across various technologies in on-premises, private and public cloud domains. Currently responsible for partner management, automation, integration, for hybrid multi cloud environments while developing new value-added services using emerging technologies

## Amar Samarth

Principal Solution Architect,
Tech Mahindra

# About the Author

A Seasoned Solution Architect with deep expertise in Datacenter, Cloud, and AI technologies. He specializes in next-generation GenAI solutions using Intel Xeon and Gaudi platforms, covering the full lifecycle from inference services to fine-tuning and training. He is also focused on cloud resource optimization across that reduce total cost of ownership (TCO) and improve workload efficiency.


## Prabhas Harlapur

Lead Solution Architect and Industrial Technical Specialist, Intel Corporation


Experienced Senior Solution Architect with expertise in Cloud Computing and Data Center AI solutions. Strong background in designing and architecting scalable, secure, and performance-optimized cloud , data center infrastructures and AI workloads across diverse industries. Expertise in driving technical sales, building relationships with key stakeholders, and collaborating with cross-functional teams to deliver tailored solutions.


## Moushumi Dash

Senior Solution Architect, Intel Corporation

## About Intel

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

## About Tech Mahindra

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 149,000+ professionals across 90+ countries helping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies. For more information on how TechM can partner with you to meet your Scale at Speed™ imperatives, please visit https://www.techmahindra.com/.

*Figures as per Q3, FY 26.