Universal Library of Engineering Technology Research Article

ISSN: 3064-996X | Volume 2, Issue 3

Open Access | PP: 44-49

DOI: https://doi.org/10.70315/uloap.ulete.2025.0203009

Strategies for Building Distributed IaaS Infrastructures for Medium-Sized Enterprises

Alexander Andreyev

Solution Engineer Independent IT Contractor, Seattle, USA.

Abstract

This article examines a set of strategies for constructing distributed IaaS infrastructures—encompassing multi-cloud, hybrid, and edge-cloud models—for medium-sized enterprises. The purpose of the study is to analyze and comparatively evaluate key architectural approaches to deploying a distributed cloud platform, taking into account requirements for latency, cost, regulatory constraints, and the maturity of DevOps processes. Infrastructure spending by mid-sized enterprises on cloud services has ramped up rapidly. This is in response to the new workloads that require rapid scaling with tight RTT requirements of 30–40 ms for business-critical applications. Analytical reports by Gartner, IDC, Flexera, Datadog, and the FinOps Foundation have helped this paper share practical recommendations for striking a balance between CAPEX and OPEX, with a focus on fault tolerance and security at the core. An aspect from which novelty emanates in this work is its integrated approach, encompassing Multi-Cloud, Hybrid, and Edge-Cloud models. It has been uniquely applied toward unified engineering principles: shared-nothing architectures, automation via IaC (Terraform, Pulumi, CDK), a unified SD-WAN transport plane, a multi-tiered ring fault-tolerance model, Zero Trust security, and FinOps cost management. The hybrid container and serverless platform has also been validated in the article regarding its efficacy as an economically predictable and scalable solution while working alongside peak loads. Multi-cloud increases flexibility and ensures maximum high availability via active-active implementations among various providers. Hybrid proves best whenever stringent data protection and local storage mandates are in force, while the edge-cloud decreases latency by bringing computing resources much closer to the user. This article will target readership among IT directors, cloud architects, and project managers driving digital transformation initiatives and distributed cloud platform projects within mid-sized enterprises.

Keywords: *IaaS, Distributed Infrastructure, Medium-Sized Enterprises, Multi-Cloud, Hybrid Model, Edge-Cloud.*

INTRODUCTION

Medium-sized enterprises now regard the cloud not as an experiment, but as an indispensable element of strategic advantage. According to Gartner, global IT spending will reach USD 5.43 trillion in 2025, with infrastructure services driving the highest growth rate (Hale, 2025). Simultaneously, 76% of small and medium-sized enterprises plan to increase their IT budgets in the coming year, demonstrating their readiness to invest in digital platforms to compete with larger players (Blackwell et al., 2024).

The primary growth drivers are new workload types. The demand for generative AI, streaming analytics, and digital products requires rapidly scalable compute and access to specialized GPU nodes; as a result, spending on data centers—a category that includes IaaS services—is growing at 42.4% year-over-year, outpacing all other segments (Hale, 2025). IDC notes that 35% of medium-sized enterprises have already included AI in their list of priority investments. By 2027, half will adjust their budgets to allocate dedicated line items for AI services that are unattainable without cloud infrastructure (Blackwell et al., 2024).

However, compute capacity alone is insufficient: modern client scenarios impose stringent latency requirements. Network operator practice indicates that any RTT above 100 ms is perceptible to the user, while the optimal range lies between 30 and 40 ms (IR Media, 2023). Achieving such performance without geodistributed nodes and proximity to end-users is virtually impossible, making a distributed IaaS architecture a technical necessity rather than an excess luxury.

Finally, the distributed model helps balance CAPEX and OPEX; however, success depends on the maturity of cost-control processes. Analysis by the FinOps Foundation shows that the median Effective Savings Rate on AWS remains at zero, and even the 75th percentile achieves only 23% savings compared to on-demand rates, indicating a

Citation: Alexander Andreyev, "Strategies for Building Distributed IaaS Infrastructures for Medium-Sized Enterprises", Universal Library of Engineering Technology, 2025; 2(3): 44-49. DOI: https://doi.org/10.70315/uloap.ulete.2025.0203009.

significant efficiency reserve (FinOps Foundation, 2024). For medium-sized enterprises, this argues in favor of adopting a FinOps approach: systematic management of reservations, spot policies, and inter-cloud traffic turns distributed IaaS from a budgetary risk into a tool for sustainable growth.

MATERIALS AND METHODOLOGY

The study of strategies for building distributed IaaS infrastructures for medium-sized enterprises is based on the analysis of 15 sources, including analytical agency reports, industry research, user surveys, and provider documentation. Initial data comprised estimates of total IT spending and infrastructure services growth rates (Hale, 2025), forecasts of global spending on edge solutions (IDC Research, 2025), and market development for Infrastructure-as-Code (Globe Newswire, 2024). Surveys by Flexera (Flexera, 2024), Datadog (Datadog, 2025), and CNCF (Hendrick, 2025) provided quantitative and qualitative insights into multi-cloud and serverless practices, while the FinOps Foundation report (FinOps Foundation, 2024) supplied data on cost-savings metrics in cloud platform usage.

The theoretical framework comprised works devoted to distributed infrastructure models, including multi-cloud, hybrid architectures, and edge-cloud solutions. The study by Hale (2025) drew on Gartner data regarding overall market size, and Blackwell et al. (2024) demonstrated the growth of AI workloads in medium-sized enterprises. Special attention was paid to user latency requirements, as IR Media (2023) demonstrated that a 30–40 ms RTT threshold is critical for user experience.

The methodology embraced several steps. First, it compared the three architectural models based on major metrics—RTT, CAPEX/OPEX ratio, FinOps efficiency, and DevOps process maturity—primarily drawing on IR Media (2023) and the FinOps Foundation (2024). Second, market forecasts for IaC and edge technologies from Globe Newswire (2024) and IDC Research (2025) helped define investment attractiveness and growth rates in quantitative terms. Third, Nutanix surveys and case studies from Flexera (2024), Datadog (2025), and HashiCorp (2024) content analysis revealed actual workload dispositions, along with Kubernetes cluster management practices.

Ultimately, a thorough examination of provider documentation (AWS, 2024; Susnjara, 2025; HashiCorp, 2024) enabled the compilation of a list of engineering steps. These include setting up an SD-WAN transport plane, utilizing Terraform/Pulumi/CDK for automation, and implementing a multi-tiered ring shaft fault-tolerance model. Security aspects were studied through the implementation of Zero Trust (DORA, 2025) and centralized IAM, completing the methodological foundation of research.

RESULTS AND DISCUSSION

The choice among multi-cloud, hybrid, and edge-cloud defines the architecture of a distributed IaaS and directly impacts cost, latency metrics, and service resilience. All three models employ the same basic building blocks—virtualized compute, software-defined networking, and an IaC approach—but balance control over infrastructure and access to the global provider ecosystem differently.

The multi-cloud strategy, in which workloads are concurrently deployed across AWS, Azure, and GCP, has become the de facto standard. According to Flexera, 89% of surveyed companies adopt it, and half employ active-active distribution for fault-tolerance purposes (Flexera, 2024). While data warehouses, containers, and serverless functions each jumped by nine percentage points in usage year-over-year, ML/AI—up five points—stands out for having the highest experimentation (32%) and planned adoption (17%), highlighting its role as the most strategically prioritized PaaS offering, as shown in Figure 1.

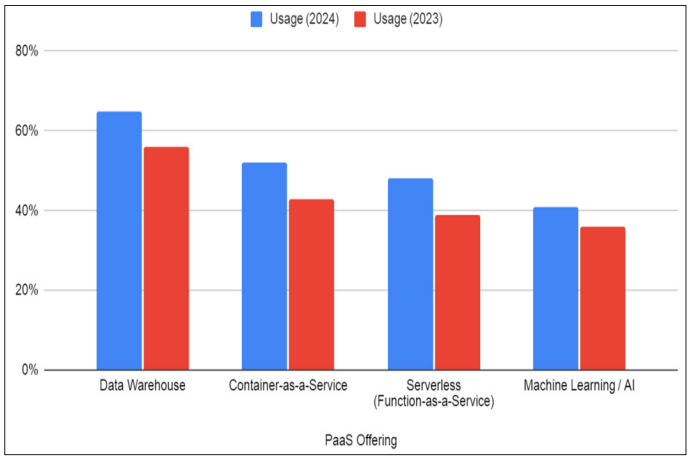

Fig. 1. YoY Usage Growth and Adoption Plans for PaaS Offerings (Flexera, 2024)

This model minimizes vendor lock-in, enables the selection of optimal services from each platform (GPU instances in AWS, analytics services in GCP, and integration with Microsoft 365 in Azure), and sets the upper SLA threshold—when networks are properly organized, migration between regions and providers remains transparent to end-users.

The hybrid model combines public cloud and on-premises sites, retaining control over sensitive data and specialized hardware. An IBM Transformation Index study finds that over 77% of companies already operate in a hybrid configuration, viewing it as the optimal compromise between flexibility and regulatory constraints (Susnjara, 2025). Its key advantage is maintaining critical databases alongside legacy systems, while offloading peak or AI workloads to the cloud as needed, all using a unified containerization stack and automated CI/CD pipelines.

Edge-cloud extends the hybrid model to branches and IoT devices by relocating compute to nodes located tens of kilometers from end-users. IDC projects that global spending on edge solutions will reach USD 261 billion by 2025, underscoring their rapid commercialization, as shown in Figure 2 (IDC Research, 2025).

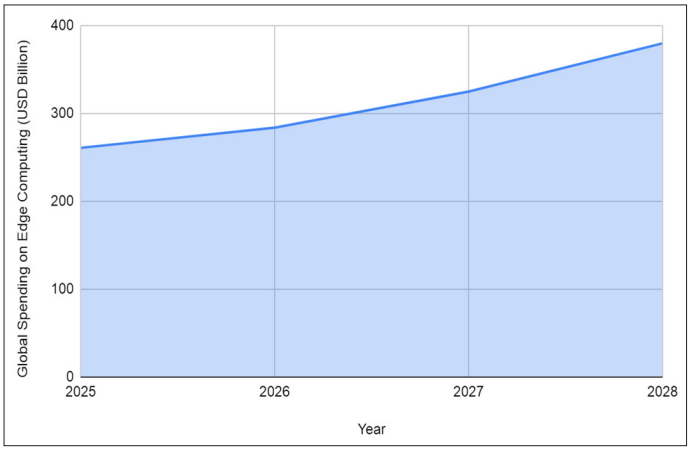

Fig. 2. Projected Global Edge Computing Spending (IDC Research, 2025)

For medium-sized enterprises, this approach reduces latency for point-of-sale terminals or computer vision systems and simultaneously offloads interregional links: data are aggregated locally and sent to centralized storage only after preprocessing.

Five factors drive model choice: required transaction latency, data sensitivity and regulatory frameworks (such as GDPR, PCI DSS, and HIPAA), CAPEX/OPEX ratio and FinOps readiness, DevOps process maturity and team expertise, and dependence on partner ecosystems, including channel integrations and software licensing. Consequently, multi-cloud suits rapid international scaling, hybrid fits industries with strict compliance demands, and edge-cloud is appropriate where business logic is sensitive to microsecond-level latency and unstable external connections.

Despite differences among multi-cloud, hybrid, and edge scenarios, their successful implementation rests on a unified set of engineering principles. These dictate service interactions, infrastructure encoding, end-to-end network construction, and the level at which fault tolerance is embedded. Without these fundamentals, even the most elegant distribution model quickly degrades into isolated islands, resulting in increased latency, operational risks, and costs.

Experience shows that the decisive factor for flexibility is the loose coupling of components and a shared-nothing design. DORA highlights this architecture as a key capability: isolated services can be modified and deployed independently, and the failure of one module does not cascade into neighboring services (DORA, 2025). That is why, in the 2024 benchmark, elite teams restore production in under an hour, whereas low-performing teams endure week-long outages—the variance in MTTR largely reflects the level of service coupling (Kosta Mitrofanskiy, 2024).

The next layer is infrastructure automation. Infrastructure-as-Code, implemented via Terraform, Pulumi, or CDK, transforms topology from a manual artifact into a verifiable, reproducible source of truth. Market dynamics underscore the maturity of this approach: the IaC segment grew from USD 917 million in 2023 to a projected USD 5.87 billion by 2032, with a compound annual growth rate of approximately 23% (Globe Newswire, 2024), as shown in Figure 3.

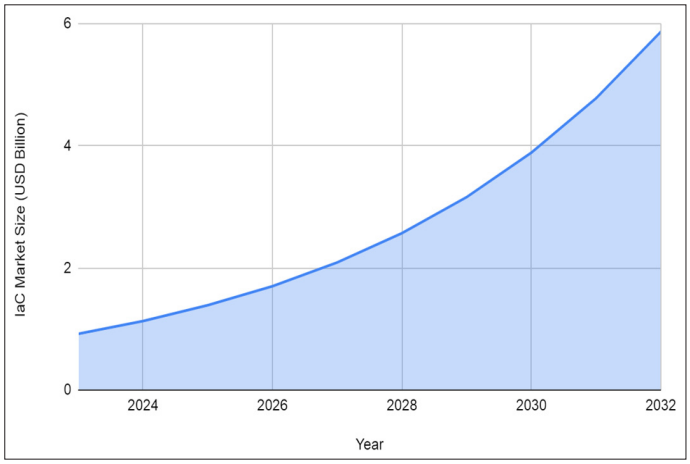

Fig. 3. Projected Infrastructure-as-Code Market Size (Globe Newswire, 2024)

According to the HashiCorp survey, three-quarters of companies already consider automation critically important for realizing their cloud strategy, directly linking it to operational efficiency and security (HashiCorp, 2024).

Any distributed topology requires a unified transport plane. SD-WAN over private links and direct provider connections establishes a predictable data path and reduces the cost of inter-cloud traffic. For mid-market businesses, this means the ability to secure backbone-grade SLAs without the capital expenditure associated with proprietary channels.

Finally, the resilience of distributed IaaS is constructed according to the principle of ring ramparts: initially via independent Availability Zones, then through geographically dispersed regions, and, where necessary, across multiple clouds. AWS documentation emphasizes that the logical and physical isolation of regions prevents correlated failures, and that a multilayered design allows organizations to balance cost against target RTO/RPO (AWS, 2024). Practically speaking, this entails active-active configurations for transactional services and active-passive setups where the business can tolerate minute-long outages. As requirements grow, a multi-cloud layer is added to ensure provider independence and compliance with regulations. Such a hierarchy transforms distributed infrastructure from a mere collection of nodes into an integrated, predictable platform suitable for rapid releases and stringent financial metrics.

Within a distributed IaaS architecture, the platform layer dictates how easily the engineering team can migrate workloads among regions, providers, and edge nodes. This layer must remain uniformly accessible across multi-cloud, hybrid, and edge environments; otherwise, each tier would necessitate its orchestration and monitoring stack, sharply increasing operational overhead.

Container orchestration via Kubernetes has become the de facto standard: in the latest annual CNCF survey, 93% of organizations reported that they are using Kubernetes in production, piloting it, or actively evaluating it, and this uptake is virtually independent of company size—equally prevalent among startups and large enterprises (Hendrick, 2025). From widespread adoption, the logical next step is operating multiple clusters across different clouds: the Nutanix ECI-2025 analysis found that 98% of respondents maintain at least one Kubernetes environment, and nearly 80% manage two or more, with the most common scenario being two to three clusters distributed by provider or site (Nutanix, 2025). For mid-market firms, this enables the centralization of CI/CD, observability, and security policies regardless of the physical deployment location of nodes.

However, not all workloads are economically justified in long-lived containers. Event-driven and unpredictable tasks—such as one-off data transformations, HTTP webhooks, or report generation—are both cheaper and faster to execute in a serverless model. According to Datadog, over 70% of AWS customer organizations, 60% in Google Cloud, and 49% in Azure now employ at least one serverless technology; moreover, over the past year, growth in Azure and GCP exceeded 6–7 percentage points, and in AWS it rose by 3 points, confirming a sustained shift toward events as the primary compute trigger (Datadog, 2025). For mid-market companies, this economic breakthrough is critical: they pay only for the actual execution time of functions, rather than idle virtual machines, and scaling occurs instantly and automatically.

The combination of Kubernetes and serverless yields a hybrid on-demand capacity model. Long-lived services and databases are maintained in containers to preserve predictability and fine-tuning capabilities. At the same time, peak invocations are routed to functions hosted in the region nearest to the user. A unified access-control system, end-to-end request tracing, and the aforementioned network layer render the transition between paradigms seamless for both developers and API consumers. This clear delineation of platform responsibilities enables mid-market firms to control costs, meet latency requirements, and, when necessary, scale any application component within minutes without altering its logic.

Consequently, the roadmap to distributed IaaS almost invariably begins with deploying a Kubernetes cluster as the foundation for state management, followed by the integration of serverless providers for event-based metered compute. Together, these technologies establish an elastic, provider-neutral, and economically predictable platform that underpins further advances in automation, observability, and FinOps governance.

The broader the distribution of infrastructure, the more imperative it becomes to enforce security according to a zero trust paradigm. Zero Trust rejects any distinction between internal and external perimeters; access is granted only upon successful verification of the subject's identity, device integrity, and the context of the request. This segmentation is particularly critical in multi-cloud and edge environments, where traffic continuously traverses provider and regional boundaries. Under Zero Trust, application and platform services expose only the minimal necessary port set, and participant authenticity is confirmed via mutual certificate-based authentication or hardware roots of trust.

A centralized access-management mechanism is anchored in a unified identity system. Instead of long-lived keys, short-lived tokens issued by a federated provider—combining multi-factor authentication with automatic revocation at the slightest indication of compromise—are employed. In Kubernetes clusters migrated across clouds, such IAM obviates the need for manual synchronization of roles and policies, enabling developers to operate with high-level abstractions—service and group names—rather than provider-specific credentials.

The threat of distributed denial-of-service attacks remains pertinent regardless of where computing nodes are hosted. Infrastructure providers supply multilayered firewalls, but an effective strategy requires integrating these with in-house filtering and caching mechanisms. The ingress point to the public network is terminated by a managed load balancer that permits only traffic that has passed reputation and behavioral analysis checks. Within the private context, SD-WAN channels isolate workloads from the public Internet, and at the edge, edge nodes accept only pre-authorized connections, thereby minimizing the attack surface.

Robust data encryption completes the security model. Storage systems, message queues, and replication channels enable encryption by default, and the transport layer is applied to all requests, including inter-cluster system traffic. To comply with European and international regulations, an immutable event log is implemented, recording key management operations, access to confidential objects, and privilege escalation attempts. Collectively, these measures transform a geographically distributed IaaS platform into a predictable and auditable environment that can be scaled and optimized without fear that new sites or providers will increase business risk.

Thus, selecting the optimal distributed IaaS infrastructure model for mid-market companies reduces to aligning latency requirements, regulatory mandates, budget constraints, and the maturity of DevOps processes. Multi-cloud offers flexibility and resilience, hybrid preserves control over critical data, and edge-cloud minimizes peripheral latency. Regardless of the scenario, the foundational principles remain a shared-nothing architecture, automation via Infrastructure as Code, a unified SD-WAN transport plane, multilayered resilience, and stringent Zero Trust security with centralized IAM and ubiquitous encryption. Integrating Kubernetes and serverless paradigms enables combining

the predictability of long-lived services with cost-effective compute-as-you-go for bursty workloads. At the same time, a FinOps approach ensures transparency and optimization of spending. By applying these practices in concert, mid-market businesses gain not only a distributed infrastructure but also a scalable, reliable, and economically efficient platform for sustainable growth.

CONCLUSION

The study has demonstrated that a distributed IaaS infrastructure constitutes a multilayered, modular platform that unites virtualized compute, software-defined networking, and Infrastructure-as-Code principles. The choice among multi-cloud, hybrid, and edge-cloud models is primarily determined by latency requirements, regulatory constraints, economic considerations, and the maturity of DevOps processes. A multi-cloud strategy delivers maximum flexibility and fault tolerance through active-active deployments across different providers. A hybrid combination of public and on-premises resources offers an optimal compromise for industries with stringent security and local data requirements. The edge-cloud can reduce RTT to target levels of 30–40 ms for latency-sensitive user scenarios.

Central to the successful implementation of any model is adherence to the engineering tenets of a shared-nothing architecture and loose coupling of components: service isolation by DORA principles minimizes catastrophic failures and sustains a low MTTR. Automation via Terraform, Pulumi, or CDK renders the infrastructure a reproducible and verifiable environment—a single source of truth for teams—which is corroborated by the rapid expansion of the IaC market. A unified SD-WAN transport plane ensures predictability and cost savings for inter-cloud traffic. At the same time, a multilayered ring-rampart resilience model—from AZs to multi-cloud—provides a robust foundation for meeting target RTO/RPO.

At the architectural core of the distributed platform lies Kubernetes container orchestration and the serverless paradigm: combining long-lived clusters with instantly scalable functions enables mid-market businesses to control costs, meet latency requirements, and respond swiftly to load spikes. Meanwhile, a FinOps approach—encompassing reservation management, spot instances, and inter-cloud traffic optimization—transforms potential budgetary risks into factors for sustainable growth, unlocking up to 23% savings compared to on-demand tariffs.

Finally, security is addressed through the adoption of a Zero Trust model with centralized IAM and end-to-end data encryption. This approach neutralizes the risks of distributed DoS attacks and simplifies access management in heterogeneous environments where services continuously migrate among clouds, the edge, and on-premises sites.

REFERENCES

1. AWS. (2024). *AWS Prescriptive Guidance*. AWS. https://docs.aws.amazon.com/prescriptive-guidance/latest/aws-multi-region-fundamentals/introduction.html
2. Blackwell, J., Clemmons, E., Deka, S., Evans, K., Longo, M., & Wilson, S. (2024). *IDC FutureScape IDC FutureScape: Worldwide Small and Medium-Sized Business*. Bitpipe. https://media.bitpipe.com/io_32x/io_326214/item_2879468/US52638024.pdf
3. Datadog. (2025). *The State of Serverless*. Datadog. https://www.datadoghq.com/state-of-serverless/
4. DORA. (2025). *Capabilities: Loosely Coupled Teams*. DORA. https://dora.dev/capabilities/loosely-coupled-teams/
5. Finops Foundation. (2024). *How to Calculate Effective Savings Rate*. Finops Foundation. https://www.finops.org/wg/how-to-calculate-effective-savings-rate-esr/
6. Flexera. (2024). *State of the Cloud Press Release*. Flexera. https://www.flexera.com/about-us/press-center/flexera-2024-state-of-the-cloud-managing-spending-top-challenge
7. Globe Newswire. (2024, December 16). *Infrastructure as Code Market to Reach USD 5869.3 Million by 2032*. Globe Newswire. https://www.globenewswire.com/news-release/2024/12/16/2997578/0/en/Infrastructure-as-Code-Market-to-Reach-USD-5869-3-Million-by-2032-Driven-by-Increased-Adoption-of-Automation-and-Cloud-Computing-Research-by-SNS-Insider.html
8. Hale, C. (2025, July 16). *Global AI adoption is expected to push IT spending beyond $5.4 trillion in 2025*. TechRadar. https://www.techradar.com/pro/global-ai-adoption-to-push-it-spending-beyond-usd5-4-trillion-in-2025
9. HashiCorp. (2024). *HashiCorp State of Cloud Strategy Survey*. HashiCorp. https://www.hashicorp.com/en/state-of-the-cloud
10. Hendrick, S. (2025). *Valerie Silverthorne, Cloud Native Computing Foundation, Cloud Native 2024 Approaching a Decade of Code, Cloud, and Change*. CNCF. https://www.cncf.io/wp-content/uploads/2025/04/cncf_annual_survey24_031225a.pdf
11. IDC Research. (2025). *IDC Estimates Global Spending on Edge Computing to Grow*. IDC Research. https://my.idc.com/getdoc.jsp?containerId=prUS53261225
12. IR Media. (2023). *Network Latency - Common Causes and Best Solutions*. IR Media. https://www.ir.com/guides/what-is-network-latency
13. Kosta Mitrofanskiy. (2024, September 5). *DORA Metrics: How You Can Measure DevOps Success*. Intellisoft. https://

intellisoft.io/dora-metrics-how-you-can-measure-devops-success/

14. Nutanix. (2025). *Enterprise Cloud Index*. Nutanix. https://www.nutanix.com/enterprise-cloud-index
15. Susnjara, S. (2025, February 10). *What is Cloud Computing?* IBM. https://www.ibm.com/think/topics/cloud-computing