

DOI:10.29013/EJTNS-26-5-23-29



ECONOMIC EFFICIENCY OF IT INFRASTRUCTURE ENGINEERING OPTIMIZATION IN COMMERCIAL COMPANIES

*Ivan Otkidach*¹

¹ DevOps Engineer, WB Tech LLC, Koledino, Russia

Cite: Otkidach I. (2026). *Economic efficiency of IT infrastructure engineering optimization in commercial companies. European Journal of Technical and Natural Sciences 2026, No 5.* <https://doi.org/10.29013/EJTNS-26-5-23-29>

Abstract

This article examines the economic efficiency of engineering optimization of IT infrastructure in commercial companies. It analyzes the impact of technical solutions on operating costs, service availability, computing resource utilization, and losses associated with infrastructure failures and downtime. Modern optimization methods are considered, including virtualization, rightsizing, autoscaling, Infrastructure as Code, observability, and Site Reliability Engineering. Particular attention is given to current approaches such as FinOps, AIOps, platform engineering, and intelligent energy management. The role of engineering solutions in improving business process resilience and achieving an economically justified balance between infrastructure costs, performance, and service reliability is also examined.

Keywords: *IT infrastructure, economic efficiency, engineering optimization, FinOps, AIOps, observability, autoscaling*

Introduction

Digital infrastructure has become an integral part of commercial organizations, since the operation of a significant share of internal and customer-facing processes directly depends on its stability and scalability. At the same time, more complex architectures, growing volumes of processed data, and the spread of cloud services increase the requirements for the quality of IT system operation. Under these conditions, infrastructure is no longer considered only as a technical environment and is increasingly becoming an object of management decisions related to cost control, risk management, and service quality.

The transition from mainly static infrastructures to dynamic environments, where resource usage, service configurations, and workload patterns can change continuously, increases the relevance of engineering optimization. This requires closer integration of technical decisions with economic criteria and supports approaches based on automation, forecasting, and better alignment of infrastructure resources with business needs. The goal of this research is to identify the main areas of engineering optimization of IT infrastructure in commercial companies and assess their impact on the economic indicators of an organization, taking into account modern approaches to resource management and operating costs.

Main part. Economic aspects of IT infrastructure operation in commercial companies

Generally, IT infrastructure in modern commercial companies is not only the technical basis of information systems but also a factor that directly affects operating costs, business process continuity, and financial performance. As digitalization increases dependence on computing resources, failures of servers, network infrastructure, or corporate applications can interrupt sales and customer service, reduce employee productivity, and create additional costs. Therefore, IT infrastructure efficiency should be assessed in terms of operating costs, resource utilization, and the required level of service availability.

The economic assessment of infrastructure traditionally includes capital and operating expenses. **Capital expenditures (CAPEX)** include the purchase of server and network equipment, data storage systems, construction or modernization of facilities, and other long-term investments. **Operating expenses (OPEX)** include electricity costs, maintenance, software licensing, cloud service payments, IT staff costs, and other ongoing expenses. However, the transition from CAPEX to OPEX does not itself mean higher economic efficiency. One of the empirical studies showed a positive relationship between cloud computing adoption and company profitability and market value, although the nature and size of the effect differed in the short and long term (Chen X. et al., 2022). At the same time, the results of a banking sector study show a more complex picture: the use of cloud technologies was associated with higher profit efficiency, but also with lower cost efficiency and higher operational risk (Cheng M. et al., 2022).

A more complete assessment of costs can be provided by **the Total Cost of Owner-**

ship (TCO) indicator. Unlike a comparison of equipment purchase costs or a cloud provider's monthly bill, TCO covers costs throughout the considered infrastructure lifecycle.

Actual **resource utilization** is an important factor in TCO. Overprovisioning helps systems handle peak loads but leads to payments for idle CPU, memory, and storage resources, while underprovisioning can reduce direct costs at the expense of performance and service quality. One study based on large-scale cluster data showed that when different types of workloads were executed separately, the average CPU utilization of the studied servers was only 29.29% for batch jobs and 7.4% for service jobs. Co-locating the workloads increased average utilization to 39.26%. The estimated TCO savings for the studied infrastructure segment, which included more than 4,000 machines, were about \$600,000, while scaling the same approach to a much larger server infrastructure could result in savings of tens of millions of dollars (Everman B. et al., 2021).

One of the important components of economic efficiency is **the availability of IT services.** For digital business processes, an infrastructure failure creates costs that may significantly exceed the cost of the failed equipment itself. Losses arise from lost revenue, employee downtime, recovery of data and systems, involvement of technical specialists, penalties for violations of service level agreements, as well as reputational consequences (Perelekhov I., 2026). The reliability of IT infrastructure can be assessed using several operational parameters that also have direct economic significance. The main indicators usually include Mean Time Between Failures (MTBF), Mean Time to Recovery (MTTR), Service Level Agreement (SLA) and Service Level Objective (SLO) (table 1).

Table 1. Reliability parameters of IT infrastructure and their economic significance

Parameter	Meaning	Economic significance
Availability	Share of time when the service is available.	Directly affects the probability and duration of business downtime.
MTBF	Mean time between failures.	Characterizes the frequency of infrastructure failures.

Parameter	Meaning	Economic significance
MTTR	Mean time to recovery.	Longer recovery increases losses caused by downtime.
SLA	Agreed service level.	Its violation may result in financial penalties and customer losses.
SLO	Internal target for service quality.	Helps define the required level of reliability.
Redundancy	Use of additional infrastructure resources for fault tolerance.	Reduces expected outage losses but increases infrastructure costs.

The economic importance of downtime is confirmed by industry statistics. According to 2024 data (Uptime Institute, 2024), 54% of respondents who had experienced a significant, serious, or severe outage during the previous three years estimated the cost of their most recent incident at more than \$100,000, while 20% estimated it at more than \$1 million. The assessment included direct costs, lost opportunities, and reputational losses. At the same time, the growing dependence of commercial processes on digital services increases the economic cost of even relatively short periods of unavailability.

From an economic point of view, not only the availability percentage itself is important, but also the cost of increasing it. For instance, moving from 99.9% to 99.99% availability requires additional redundancy of equipment, communication channels, computing resources, and organizational recovery mechanisms. Therefore, maximizing availability at any cost is not a rational goal. The engineering task is to achieve a level of reliability at which additional fault-tolerance costs are balanced against the expected reduction in losses from incidents.

Reducing infrastructure costs without considering performance and reliability can worsen overall business results, while excessive redundancy creates unjustified fixed costs. Engineering optimization should therefore focus on an economically justified balance between infrastructure cost, performance, and resilience.

Engineering methods for IT infrastructure optimization and their impact on economic indicators

Engineering optimization of IT infrastructure is aimed at providing the required performance and service availability while using computing and financial resources efficiently. Basic methods include server consolidation through virtualization, which makes it possible to increase equipment utilization and reduce costs for its purchase, power supply, cooling, and maintenance (fig. 1).

However, excessive consolidation can reduce performance because virtual machines compete for shared resources. In cloud and virtualized environments, rightsizing helps align allocated CPU, RAM, and other resources with actual demand while preserving the required performance margin.

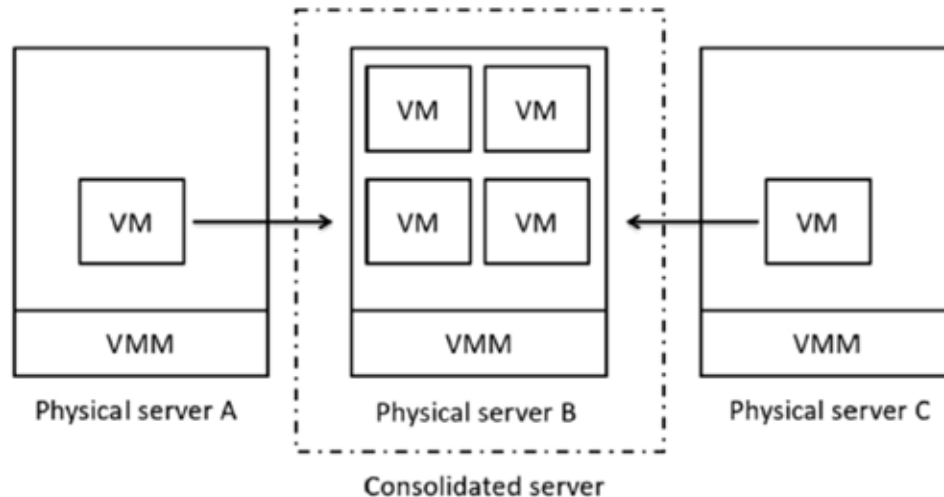
Containerization and orchestration systems provide more flexible resource management and allow individual application components to scale independently. The greatest economic effect is achieved when containerization is combined with autoscaling. One study proposed joint horizontal and vertical scaling of containers and virtual machines while considering placement costs (Entrialgo J. et al., 2024). In experiments on a real application, the resulting configurations were approximately 20% cheaper than those produced by the studied heuristic resource allocation algorithms.

Another study showed that the use of ML for cloud resource management reduced the number of QoS requirement violations by up to 80% compared with previously used algorithms, while virtual machine usage decreased by 3–50% depending on the experimental scenario (Pintye I. et al., 2024).

Thus, workload forecasting makes it possible to reduce both the amount of excessively reserved capacity and the probability of service degradation. Component redundancy, load

distribution, data replication, and automatic failover reduce expected losses from failures, although they require additional infrastructure resources.

Figure 1. *Traditional server consolidation (Bermejo B. & Juiz C., 2022)*



Infrastructure management automation also has a significant impact on operating costs. The **Infrastructure as Code (IaC)** concept involves describing infrastructure through machine-readable configurations, which makes its creation and modification automated and reproducible, while version control and testing can be applied to the configurations. The economic component is gradually being integrated directly into such processes. In a 2024 study, authors analyzed more than 152,000 repositories and then examined cost-related changes in projects using Terraform. The results showed that developers take the cost of cloud resources into account both before and after deployment and optimize it through service selection, resource allocation, and configuration changes (Feitosa D. et al., 2024).

For distributed and microservice systems, **observability** is especially important, combining the analysis of metrics, event logs, and distributed traces. Its economic value is mainly related to reducing the time required to detect and resolve failures and, therefore, reducing the duration of the resulting downtime.

Observability is closely related to the **Site Reliability Engineering (SRE)** approach, which involves reliability management based on measurable service quality objectives. Its economic meaning is to pro-

vide the required rather than the maximum possible level of availability, since additional reliability also requires additional infrastructure costs.

Thus, modern engineering optimization methods affect economic indicators in different ways. Consolidation and rightsizing reduce excess capacity, while containerization, autoscaling, and predictive algorithms help adapt resources to workload changes. Redundancy reduces potential losses from failures, while Infrastructure as Code, observability, and SRE lower manual effort, improve reproducibility, and support faster and more economically justified reliability management.

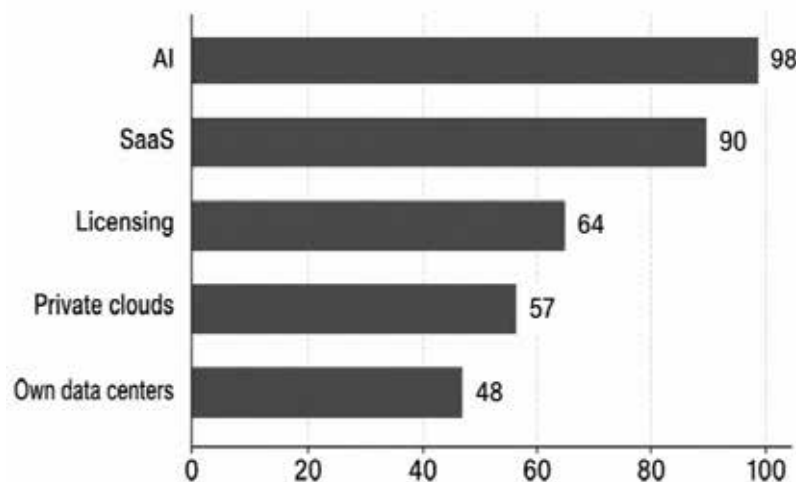
Modern approaches to intelligent and economic optimization of IT infrastructure

The development of cloud technologies, containerization, and distributed architectures has changed the approach to IT infrastructure optimization. Along with conventional methods of reducing operating costs, continuous infrastructure management based on data about its cost and technical condition is becoming increasingly relevant. Important areas include FinOps, the use of artificial intelligence (AI) in IT operations, the development of internal engineering platforms, and intelligent energy management.

One of the key modern approaches is **FinOps**, which involves joint management of technology costs by engineering and business departments. Its task is to reduce costs and relate infrastructure costs to the business val-

ue created. According to the State of FinOps 2026, its practice already goes beyond public clouds and covers cost management for AI, Software as a Service (SaaS), licensing, private clouds, and own data centers (fig. 2).

Figure 2. Areas of cost management in FinOps practice (FinOps, 2025)



An important development of FinOps is **the use of unit economics**, where the cost of not the entire infrastructure, but a separate unit of created result is evaluated. This approach becomes notably relevant for AI infrastructure, since its cost is determined by GPU usage, training and inference requests, the volume of processed data, and other parameters.

Another rapidly developing area is **AI for IT Operations (AIOps)**, which means the application of machine learning and other AI methods to IT infrastructure operations. The AIOps practice is used for anomaly detection, failure prediction, root cause identification, and support of recovery actions (Díaz-de-Arcaya J., 2024). Its economic effect is mainly associated with reducing specialist labor costs and incident duration. Automated analysis of metrics, logs, and traces makes it possible to detect deviations faster and move from responding to an already occurred failure to predicting potential problems.

The current stage of AIOps development is also associated with **the use of large language models and autonomous AI agents**. Intelligent agents can interact with cloud infrastructure, localize intentionally created failures, and take actions to eliminate them (Khlystun M., 2026). Fully autonomous management is still associated with

significant risks, so for critical infrastructure a more realistic model is one in which AI performs analysis and standard verifiable operations, at the same time potentially dangerous changes remain under the control of a specialist.

The reduction of operating labor costs is also connected with the development of **platform engineering**. This approach involves creating an internal platform that provides product teams with standardized access to computing resources, storage, databases, CI/CD, observability, and other infrastructure services.

Another direction is the combination of economic and energy optimization. Within **Green Ops**, reducing the number of idle resources, consolidating computing workloads or dynamically managing capacity make it possible to reduce both energy consumption and operating costs. Workload forecasting, dynamic consolidation and migration of virtual machines, as well as machine learning methods can be used to manage the energy consumption of cloud data centers.

Modern IT infrastructure is increasingly optimized continuously based on technical and economic data. Observability, AIOps, automated platforms, and FinOps make it possible to connect the state of the system, engineering actions, and their financial effect,

while their use is justified only when the final economic result is positive.

Conclusion

The economic efficiency of IT infrastructure is determined not only by the amount of direct costs for equipment, cloud resources, and maintenance, but also by how efficiently computing capacity is used, how quickly the system adapts to workload changes, and how stable critical services are. Engineering optimization makes it possible to influence these parameters through consolidation and rightsizing of resources, autoscaling, fault-tolerant architectures, IaC, observability, and SRE. Excessive reduction of resources can increase downtime risks, while excessive

redundancy leads to unjustified growth of operating costs.

Modern approaches develop this logic through a closer connection between technical and economic data. FinOps increases the transparency of technology costs and makes it possible to relate them to the business value created, AIOps expands the possibilities for workload forecasting and incident automation, while platform engineering helps standardize and reduce repetitive engineering operations. As a result, IT infrastructure optimization is gradually becoming a continuous process in which decisions are made taking into account cost, utilization, reliability, and business process requirements.

References

- Bermejo, B., & Juiz, C. (2022). A general method for evaluating the overhead when consolidating servers: Performance degradation in virtual machines and containers. *Journal of Supercomputing*, – 78. – P. 11345–11372. URL: <https://doi.org/10.1007/s11227-022-04318-5>
- Chen, X., Guo, M., & Shangguan, W. (2022). Estimating the impact of cloud computing on firm performance: An empirical investigation of listed firms. *Information & Management*, – 59 (3). – 103603 p. URL: <https://doi.org/10.1016/j.im.2022.103603>
- Cheng, M., Qu, Y., Jiang, C., & Zhao, C. (2022) Is cloud computing the digital solution to the future of banking? *Journal of Financial Stability*, – 63. – 101073 p. URL: <https://doi.org/10.1016/j.jfs.2022.101073>
- Díaz-de-Arcaya, J., Torre-Bastida, A. I., Zárate, G., Miñón, R., & Almeida, A. (2024). A joint study of the challenges, opportunities, and roadmap of MLOps and AIOps: A systematic survey. *ACM Computing Surveys*, – 56 (4). – 84. – P. 1–30. URL: <https://doi.org/10.1145/3625289>
- Entrialgo, J., García, M., García, J., López, J. M., & Díaz, J. L. (2024). Joint autoscaling of containers and virtual machines for cost optimization in container clusters. *Journal of Grid Computing*, – 22 (1). – 17 p. URL: <https://doi.org/10.1007/s10723-023-09732-4>
- Everman, B., Rajendran, N., Li, X., & Zong, Z. (2021). Improving the cost efficiency of large-scale cloud systems running hybrid workloads: A case study of Alibaba cluster traces. *Sustainable Computing: Informatics and Systems*, – 30. – 100528 p. URL: <https://doi.org/10.1016/j.suscom.2021.100528>
- Feitosa, D., Penca, M.-T., Berardi, M., Boza, R.-D., & Andrikopoulos, V. (2024). Mining for cost awareness in the infrastructure as code artifacts of cloud-based applications: An exploratory study. *Journal of Systems and Software*, – 215. – 112112 p. URL: <https://doi.org/10.1016/j.jss.2024.112112>
- Khlystun, M. (2026). LLM inference in high-load infrastructure: Architectural approaches and practical constraints. *Cold Science*, – 29. – P. 57–68.
- Perelekhov, I. (2026). Designing fault-tolerant e-commerce systems: Architectural patterns and practical solutions. *The Scientific Heritage*, – 189. – P. 58–63. URL: <https://doi.org/10.5281/zenodo.21617973>
- Pintye, I., Kovács, J., & Lovas, R. (2024). Enhancing machine learning-based autoscaling for cloud resource orchestration. *Journal of Grid Computing*, – 22 (4). – 68 p. URL: <https://doi.org/10.1007/s10723-024-09783-1>

State of FinOps 2025 Report. (2025). *Fin Ops*. [cited: 10.08.2026]. Available from URL: <https://data.finops.org/2025-report>
Uptime Institute. (2024). Uptime Institute Global Data Center Survey 2024. *Uptime Institute Intelligence*. – 31 p.

submitted 15.09.2026;
accepted for publication 29.09.2026;
published 30.09.2026
© Otkidach I.
Contact: ivan.research.mail@gmail.com