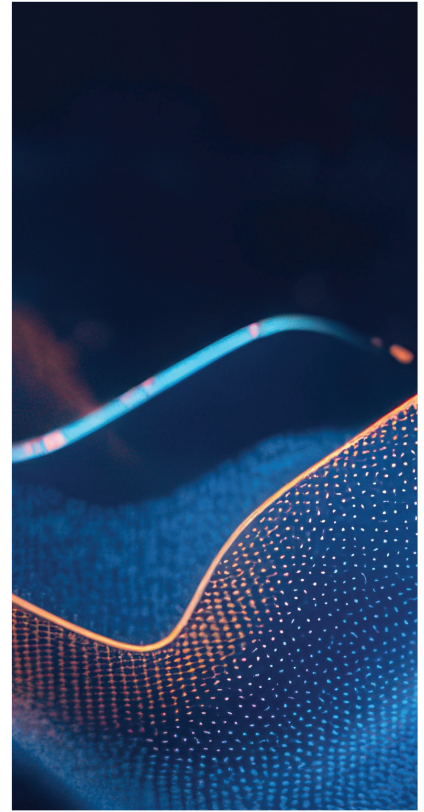
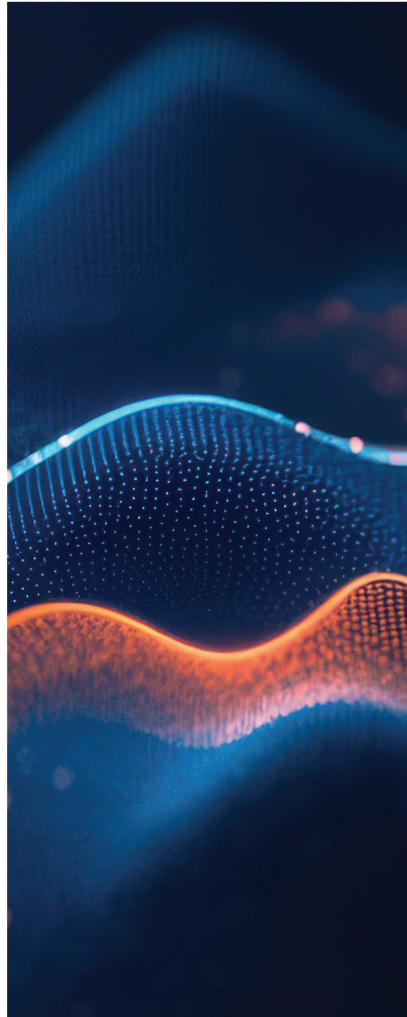


FROST & SULLIVAN
BEST PRACTICES



2026

NORTH AMERICAN CLOUD
COST OPTIMIZATION
PLATFORM

**TECHNOLOGY INNOVATION
LEADERSHIP**



Table of Contents

<i>Best Practices Criteria for World-Class Performance</i>	<i>3</i>
The Transformation of the Cloud Cost Optimization Platform Industry.....	3
Expanding Innovation Across Enterprise Workloads	5
Scalable Customer Adoption through Operational Excellence	7
Sustained Growth Driven by Intelligent Economics.....	8
<i>Conclusion</i>	<i>9</i>
<i>What You Need to Know about the Technology Innovation Leadership Recognition</i>	<i>10</i>
Best Practices Recognition Analysis.....	10
Technology Leverage	10
Business Impact	10
<i>Best Practices Recognition Analytics Methodology.....</i>	<i>11</i>
Inspire the World to Support True Leaders	11
<i>About Frost & Sullivan</i>	<i>12</i>
The Growth Pipeline Generator™	12
The Innovation Generator™	12

Best Practices Criteria for World-Class Performance

Frost & Sullivan applies a rigorous analytical process to evaluate multiple nominees for each recognition category before determining the final recognition recipient. The process involves a detailed evaluation of best practices criteria across two dimensions for each nominated company. CAST AI Group Inc. excels in many of the criteria in the cloud cost optimization platform space.

RECOGNITION CRITERIA	
<i>Business Impact</i>	<i>Technology Leverage</i>
Financial Performance	Commitment to Innovation
Customer Acquisition	Commitment to Creativity
Operational Efficiency	Stage Gate Efficiency
Growth Potential	Commercialization
Human Capital	Application Diversity

The Transformation of the Cloud Cost Optimization Platform Industry

The North American cloud cost optimization market is undergoing a period of rapid transformation, driven by the convergence of cloud-native adoption and the exponential rise of artificial intelligence (AI) workloads. Enterprises significantly increased their reliance on Kubernetes-based, containerized environments to support modern applications, yet this shift introduces new layers of operational complexity. Traditional approaches to cloud management often rely on manual provisioning, periodic optimization, and reactive monitoring, and thereby struggle to keep pace with dynamic and unpredictable workload behavior. As a result, organizations frequently over-provision resources to maintain performance and reliability, leading to persistent inefficiencies and substantial cost leakage across compute and storage environments. Frost & Sullivan points out that the limitations of human-led operations further compound these challenges, where teams lack the ability to continuously monitor and optimize infrastructure in real time, creating gaps between actual demand and allocated resources.

Simultaneously, the rapid expansion of AI and graphic processing unit (GPU)-intensive workloads introduce a new dimension of cost and efficiency challenges. Enterprises invest heavily in GPUs and large-scale AI infrastructure - yet utilization rates often remain extremely low due to fragmented workloads and

“Through its continuous innovation pipeline, which includes new agent-based solutions and expanded capabilities for AI-driven environments, CAST AI improves cloud efficiency while shaping the future of infrastructure management. By combining deep technical innovation with measurable operational impact, the company leads the transformation of cloud operations into an autonomous, intelligent, and self-optimizing ecosystem.”

**- Soumyadeep Roy Chowdhury,
Senior Industry Analyst, GOA**

lack of intelligent orchestration. This inefficiency is particularly critical as organizations shift toward large language models (LLM) and token-based consumption models, where infrastructure costs directly influence compute usage and throughput. Additionally, many cloud cost management tools remain focused on static optimization or visibility rather than active, real-time decision-making, preventing organizations from scaling resources to meet demand. As a result, the industry recognizes the need for autonomous, intelligence-driven optimization platforms that continuously manage infrastructure, eliminate waste,

and ensure performance stability in an environment defined by scale, complexity, and accelerating cost pressures.

Founded in 2019 and headquartered in Miami, Florida, CAST AI Group Inc. (CAST AI) is a leading innovator in autonomous cloud and AI infrastructure management. The company redefines cloud operations through its agent-based platform that delivers real-time optimization across Kubernetes environments, enabling enterprises to improve performance, eliminate waste, and ensure infrastructure stability at scale. By extending its capabilities into GPU and token optimization, Frost & Sullivan’s own research concludes that CAST AI properly addresses the most critical cost and efficiency challenges in modern AI-driven environments.

Autonomous AI Infrastructure Leadership

CAST AI evolved from its origins as a Kubernetes cost optimization platform into an application performance automation platform, redefining how enterprises manage cloud-native and AI-driven infrastructure. Rather than relying on dashboards and alerts that require human intervention, the platform proactively identifies and resolves inefficiencies and risks before they escalate into operational issues. This autonomous decision-making model enables enterprises to achieve greater stability and reliability, while eliminating inefficiencies such as over-provisioning and resource waste. By embedding intelligence directly into the infrastructure layer, CAST AI effectively transforms cloud operations from a human-led, periodic process into a continuous, machine-driven system capable of making decisions at a scale and speed unattainable by traditional developer operations (DevOps) or site reliability engineering teams.

This orchestration capability extends across the full compute stack, encompassing central processing units (CPU), GPU, memory, storage, and network resources. The company purpose-built the platform for Kubernetes and cloud-native environments, supporting major hyperscalers, as well as hybrid and multi-cloud deployments. Its seamless operation across these environments ensures enterprises can apply consistent optimization strategies regardless of infrastructure complexity, significantly enhancing scalability and operational consistency.

A defining aspect of CAST AI’s innovation is its ability to manage diverse workload types autonomously, including containerized applications, databases, and AI and machine learning (ML) workloads. The

platform's advanced automation engine incorporates capabilities such as real-time autoscaling, resource bin-packing, Spot Instance orchestration, and drift remediation. These capabilities are further reinforced by a proactive intelligence layer that anticipates and prevents system failures through functionalities such as out-of-memory prevention, policy-based enforcement, security patching, and predictive infrastructure adjustments. This shift from reactive monitoring to predictive, self-healing infrastructure marks a fundamental advancement in cloud operations.

CAST AI's leadership in GPU and AI workload optimization reflects a robust commitment to innovation, a critical area in the current computing landscape. Despite being one of the most expensive aspects of computing, enterprise GPU environments often operate at just 5% of their GPU capacity before optimization.¹ The company addresses this inefficiency by continuously aligning resource allocation with real-time workload demand, increasing utilization and improving overall infrastructure performance.

Building on this foundation, the company is advancing into token optimization and AI economics, extending its optimization model beyond infrastructure into the operational cost structure of LLMs. Through innovations such as its coding agent solution, CAST AI significantly reduces token-related costs while increasing throughput. These advancements position the company at the forefront of a new phase of infrastructure innovation, where managing the efficiency of AI consumption is as critical as managing compute resources.

Through its continuous innovation pipeline, which includes new agent-based solutions and expanded capabilities for AI-driven environments, Frost & Sullivan recognizes how CAST AI improves cloud efficiency while shaping the future of infrastructure management. By combining deep technical innovation with measurable operational impact, the company leads the transformation of cloud operations into an autonomous, intelligent, and self-optimizing ecosystem.

Expanding Innovation Across Enterprise Workloads

CAST AI operates seamlessly across the full spectrum of modern enterprise workloads, spanning containerized applications and databases to increasingly complex AI and ML environments. The company built its platform architecture for Kubernetes and cloud-native environments to deliver consistent optimization outcomes across workloads and infrastructure. The platform supports deployments across Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and hybrid or on-premises environments, enabling enterprises to standardize performance optimization across diverse operating landscapes without compromise.

The platform manages mission-critical applications across industries, including telecommunications, financial services, automotive, and digital platforms. CAST AI's agents operate in environments that demand continuous uptime and high reliability like large-scale enterprise systems and user-facing applications supporting millions of users. This capability demonstrates technical robustness and adaptability to the unique operational requirements of each industry, reinforcing the platform's broad applicability and enterprise readiness.

¹ Provided by CAST AI.

“CAST AI’s rapid time to value realization, scalable expansion strategy, and deeply embedded operational efficiency creates a powerful customer acquisition engine. The platform’s ability to deliver immediate and sustained impact accelerates adoption and drives long-term customer growth, positioning the company as a trusted and integral partner in optimizing modern cloud and AI infrastructure.”

**- Soumyadeep Roy Chowdhury,
Senior Industry Analyst, GOA**

CAST AI’s growing specialization in AI and GPU-intensive workloads highlights its ability to address one of the most demanding and rapidly evolving application categories. The platform optimizes GPU utilization and dynamically allocates resources based on real-time demand, supporting workloads such as LLM inference, training environments, and AI-driven applications. Given the high cost and scarcity of GPU resources, this capability enables enterprises to maximize performance while minimizing unnecessary expenditure.

Beyond infrastructure optimization, CAST AI expands its application footprint into emerging domains through continuous innovation. The new agent-based solutions, including the coding-focused agent, reflect a deliberate expansion of its core architecture into adjacent use cases beyond traditional infrastructure management. By applying the same autonomous, policy-driven intelligence to developer workflows, the platform creates new pathways for efficiency and cost optimization at the application layer. The company’s approach redefines the role of agents beyond operations, extending their impact across broader enterprise processes.

Early leadership in token optimization further demonstrates its innovative approach at the intersection of infrastructure and AI economics. As enterprises scale AI adoption, token usage costs become a critical factor in scalability and feasibility. CAST AI addresses this challenge by optimizing token processing and routing, enabling cost reduction and throughput improvement. By extending its capabilities into this emerging layer, the company shapes industry direction and positions itself at the forefront of next-generation AI infrastructure management.

CAST AI further demonstrates its creativity by orchestrating across multiple operational domains, including DevOps, security operations, AIOps, and MLOps. The platform’s agent-based model bridges traditionally siloed functions, enabling coordinated decision-making across the full application lifecycle. This integrated approach improves efficiency and reduces operational complexity, allowing enterprises to manage diverse applications through a unified framework.

Through direct engagement with enterprise users and developers, CAST AI continuously refines its platform to address real-world challenges and emerging use cases. The company’s feedback-driven model ensures that new capabilities are technically advanced and practically relevant, accelerating adoption across diverse environments.

Together, CAST AI’s broad application coverage and forward-looking innovation strategy positions the platform as a flexible, intelligent layer that adapts to the evolving enterprise demands. By pairing broad applicability with creative expansion into new domains such as AI economics and developer workflows, the company differentiates itself as a dynamic and future-ready innovation leader in cloud and AI infrastructure.

Scalable Customer Adoption through Operational Excellence

CAST AI's customer acquisition is intrinsically linked to the immediate and measurable value delivered by its platform, supported by a highly efficient, product-led growth strategy. Rather than relying on extended sales cycles or heavy upfront commitments, the company enables enterprises to rapidly deploy its autonomous agents directly into their cloud environments and realize optimization outcomes within minutes. This rapid time to value lowers barriers to entry, accelerates proof-of-concept timelines, and drives faster customer conversion, particularly among large enterprises managing complex cloud-native infrastructures.

The company's land-and-expand model further strengthens its acquisition strategy by converting initial engagements into long-term, high-growth relationships. Enterprises typically begin with a limited deployment or targeted use case, after which the platform's demonstrated effectiveness drives broader adoption across additional workloads, teams, and business units. Consistent efficiency and stability improvements at scale drive this expansion. As a result, CAST AI achieves robust growth within existing accounts, reflecting clear product-market fit and a deep integration into enterprise operations.

CAST AI's operational model is a critical driver of its acquisition success, aligning closely with customer outcomes. The platform's autonomous agents eliminate the need for continuous human oversight, replacing manual, periodic optimization processes with real-time, machine-driven decision-making. Shifting to an autonomous, continuous optimization model dramatically reduces operational complexity while enabling performance at a scale beyond traditional DevOps or site reliability engineering approaches. The platform optimizes infrastructure performance and allows engineering teams to focus on higher-value strategic initiatives.

CAST AI's operational efficiency extends beyond the product into its customer engagement model. The company adopts a highly integrated approach, embedding developers and technical teams directly into customer interactions, including onboarding and optimization processes. As a result, it minimizes gaps between product capabilities and customer requirements, accelerates issue resolution, and ensures rapid integration of feedback into ongoing platform development.

The platform's ability to autonomously manage infrastructure at scale drives significant efficiency gains at the system level. By continuously analyzing workload behavior and dynamically adjusting resource allocation, CAST AI eliminates over-provisioning while maintaining performance stability. The solution optimizes resources across CPU, memory, and increasing GPU environments, which are critical for AI workloads. These improvements run continuously in real time, allowing organizations to maintain optimal performance as application demands evolve.

In parallel, the platform's proactive intelligence prevents operational issues, such as memory constraints or performance bottlenecks, by addressing them before they impact application performance. The company's preventive approach improves system reliability and reduces the need for reactive troubleshooting, further contributing to overall operational efficiency. By shifting enterprises from reactive management models toward continuous, autonomous optimization, CAST AI enables a more resilient and efficient infrastructure environment.

CAST AI's rapid time to value realization, scalable expansion strategy, and deeply embedded operational efficiency creates a powerful customer acquisition engine. The platform's ability to deliver immediate and sustained impact accelerates adoption and drives long-term customer growth, positioning the company as a trusted and integral partner in optimizing modern cloud and AI infrastructure.

Sustained Growth Driven by Intelligent Economics

CAST AI's growth trajectory reflects strong market alignment, scalable business economics, and an effective commercialization model, positioning it for sustained expansion within the evolving cloud and AI infrastructure landscape. The company operates at the intersection of two major technology shifts (i.e., cloud-native computing and AI-driven workloads) both experiencing rapid enterprise adoption. By addressing the growing complexity and cost of these environments, CAST AI delivers a platform that grows more essential as organizations scale their infrastructure and AI initiatives.

Strong enterprise traction highlights this growth potential, with large organizations generating the majority of revenue. CAST AI secures and expands these relationships by delivering measurable value in mission-critical environments. Enterprises typically adopt the platform to address immediate inefficiencies, then expand usages across workloads teams, and business units as results compound. This pattern demonstrates platform scalability and alignment with core operational priorities.

CAST AI pairs rapid growth with strong underlying economics. The company continues to expand annual recurring revenue while maintaining high gross margins comparable to software-as-a-service platforms. Many AI-driven infrastructure competitors struggle with margin constraints due to high compute costs. Conversely, CAST AI preserves capital efficiency by deploying lightweight agents within customer environments rather than owning or reselling infrastructure. This approach supports a structurally advantaged economic model while delivering high-value outcomes to customers. Strong net revenue retention further reinforces the strength of the company's financial performance, as existing customers consistently increasing spend over time. Deep integration into customer operations. Once deployed, the platform becomes an integral part of how infrastructure drives this expansion, making the platform highly durable within enterprise environments.

The company's commercialization approach reinforces this foundation by aligning product value with clear, measurable financial outcomes. CAST AI's platform directly reduces infrastructure costs, improves resource utilization, and increase operational stability, benefits that translate into immediate and quantifiable return on investment. This value-centric positioning simplifies enterprise purchasing decisions for enterprises, shortens sales cycles and accelerates adoption without heavy reliance on traditional sales processes. Continued innovation in areas such as GPU and token optimization further strengthen its growth outlook by capturing demand from expanding AI workflows. Additionally, with strong backing from leading investors and its unicorn valuation, CAST AI combines financial discipline, scalable growth, and ongoing innovation. By aligning its commercialization strategy with measurable customer outcomes and maintaining disciplined economic fundamentals, the company demonstrates a clear path toward sustained growth and long-term market leadership.

Conclusion

CAST AI establishes itself as a transformational leader by redefining cloud and artificial intelligence (AI) infrastructure through a truly autonomous, agent-driven platform that unifies performance, efficiency, and scalability across the full compute stack. Its ability to extend this core innovation across diverse enterprise applications, while pioneering advancements in graphic processing unit and token optimization, demonstrates technical depth and forward-looking vision at the frontier of AI economics. Coupled with a powerful land-and-expand model and deeply embedded operational efficiency, the company consistently translates innovation into measurable customer value and accelerating enterprise adoption. Together, these capabilities position CAST AI as a category innovator and a defining force shaping the future of intelligent, self-optimizing infrastructure on a global scale.

With its strong overall performance, CAST AI Group Inc. earns the 2026 Frost & Sullivan North American Technology Innovation Leadership Recognition in the cloud cost optimization platform industry.

What You Need to Know about the Technology Innovation Leadership Recognition

Frost & Sullivan's Technology Innovation Recognition identifies the company that has introduced the best underlying technology for achieving remarkable product and customer success while driving future business value.

Best Practices Recognition Analysis

For the Technology Innovation Leadership Recognition, Frost & Sullivan analysts independently evaluated the criteria listed below.

Technology Leverage

Commitment to Innovation: Continuous emerging technology adoption and creation enables new product development and enhances product performance

Commitment to Creativity: Company leverages technology advancements to push the limits of form and function in the pursuit of white space innovation

Stage Gate Efficiency: Technology adoption enhances the stage gate process for launching new products and solutions

Commercialization: Company displays a proven track record of taking new technologies to market with a high success rate

Application Diversity: Company develops and/or integrates technology that serves multiple applications and multiple environments

Business Impact

Financial Performance: Strong overall business performance is achieved in terms of revenue, revenue growth, operating margin, and other key financial metrics

Customer Acquisition: Customer-facing processes support efficient and consistent new customer acquisition while enhancing customer retention

Operational Efficiency: Company staff performs assigned tasks productively, quickly, and to a high-quality standard

Growth Potential: Growth is fostered by a strong customer focus that strengthens the brand and reinforces customer loyalty

Human Capital: Leveraging innovative technology characterizes the company culture, which enhances employee morale and retention

Best Practices Recognition Analytics Methodology

Inspire the World to Support True Leaders

This long-term process spans 12 months, beginning with the prioritization of the sector. It involves a rigorous approach that includes comprehensive scanning and analytics to identify key best practice trends. A dedicated team of analysts, advisors, coaches, and experts collaborates closely, ensuring thorough review and input. The goal is to maximize the company's long-term value by leveraging unique perspectives to support each Best Practice Recognition and identify meaningful transformation and impact.

VALUE IMPACT			
STEP		WHAT	WHY
1	Opportunity Universe	Identify Sectors with the Greatest Impact on the Global Economy	Value to Economic Development
2	Transformational Model	Analyze Strategic Imperatives That Drive Transformation	Understand and Create a Winning Strategy
3	Ecosystem	Map Critical Value Chains	Comprehensive Community that Shapes the Sector
4	Growth Generator	Data Foundation That Provides Decision Support System	Spark Opportunities and Accelerate Decision-making
5	Growth Opportunities	Identify Opportunities Generated by Companies	Drive the Transformation of the Industry
6	Frost Radar	Benchmark Companies on Future Growth Potential	Identify Most Powerful Companies to Action
7	Best Practices	Identify Companies Achieving Best Practices in All Critical Perspectives	Inspire the World
8	Companies to Action	Tell Your Story to the World (BICEP*)	Ecosystem Community Supporting Future Success

*Board of Directors, Investors, Customers, Employees, Partners

About Frost & Sullivan

Frost & Sullivan is the Growth Pipeline Company™. We power our clients to a future shaped by growth. Our Growth Pipeline as a Service™ provides the CEO and the CEO's growth team with a continuous and rigorous platform of growth opportunities, ensuring long-term success. To achieve positive outcomes, our team leverages over 60 years of experience, coaching organizations of all types and sizes across 6 continents with our proven best practices. To power your Growth Pipeline future, visit Frost & Sullivan at <http://www.frost.com>.

The Growth Pipeline Generator™

Frost & Sullivan's proprietary model to systematically create ongoing growth opportunities and strategies for our clients is fueled by the Innovation Generator™.

[Learn more.](#)

Key Impacts:

- **Growth Pipeline:** Continuous Flow of Growth Opportunities
- **Growth Strategies:** Proven Best Practices
- **Innovation Culture:** Optimized Customer Experience
- **ROI & Margin:** Implementation Excellence
- **Transformational Growth:** Industry Leadership



The Innovation Generator™

Our 6 analytical perspectives are crucial in capturing the broadest range of innovative growth opportunities, most of which occur at the points of these perspectives.

Analytical Perspectives:

- Megatrend (MT)
- Business Model (BM)
- Technology (TE)
- Industries (IN)
- Customer (CU)
- Geographies (GE)

