Redefining Value: Pricing in the Age of AI

The explosion of AI is redefining software economics. As enterprise application providers are integrating increasingly complex AI-based capabilities into their platforms, traditional seat-based licensed pricing models fall short because of the variable, dynamic and consumption-intensive nature of AI cost structures. This paper explores why "as-a-service" pricing must evolve to incorporate usage-based, outcome-based and hybrid models that better align with AI costs and value delivered.

We outline why "as-a-service" platform vendors will benefit from adopting new billing platforms that deliver real-time charging and dynamic pricing at massive scale.

Drawing from real-world examples, we provide strategic guidance for designing transparent, flexible pricing frameworks that reflect the operational realities and commercial opportunities of the AI era.

The AI Disruption to Enterprise Pricing

Enterprise software is undergoing a seismic shift. Applications are no longer static repositories of business logic but increasingly are being infused with AI to deliver smarter insights, automate workflows and even act autonomously through agentic capabilities. Leading vendors like Salesforce, ServiceNow, Informatica and others are rapidly integrating generative and agent-based AI across their platforms.

Yet while intelligent functionality has advanced, pricing models have not necessarily kept pace. Per-seat or per-user license prices are still prevalent, and this model is ill-suited to the elastic and compute-intensive nature of AI workloads. As enterprise customers lean more heavily on AI, the mismatch between resource utilization and cost undermines business alignment and profitability.

The Shifting Economics of AI in Enterprise Platforms

AI workloads are not just computationally expensive — they are unpredictable. An AI-enabled query could generate a one-line summary of a knowledge base article, the full transcript of a customer service call or a detailed multi-paragraph report of a customer's behavior and activity across multiple channels. These different activities consume vastly different resources. Enterprise software providers should look to the evolution of the telecommunications industry, where the real-time dynamic billing models that evolved over decades are well-suited for the emerging AI market. As modern telco networks do, AI resources will be flexibly deployed, must scale on-demand and should adapt in response to real-time cost insights and value delivered.

Enterprise "as-a-service" providers must now factor in:

- **High variability of compute:** GPU utilization is critical and varies widely across the "AI life cycle" from training to inference to massive parallel processing of data and tasks.
- **Third-party infrastructure costs:** Applications providers might incur real-time costs from AI providers, large language models (LLMs) and clouds like OpenAI, AWS, Azure or Google Cloud Platform (GCP).
- **Spiky, activity-based usage patterns:** AI services aren't used evenly; they spike based on manual user or batch activity, time of day, day of week or seasonal demand.

Some current AI pricing examples:

- **Salesforce:** Introduced Agentforce Flex Credits to price AI actions individually. Customers pay for the specific AI-powered actions they consume, such as recommendations or workflow automation, creating a direct tie between usage and cost.
- **Informatica:** Uses Informatica Processing Units (IPUs) to count and charge for AI services. This model gives customers a pre-purchased block of credits to spend flexibly across services, aligning cost with AI consumption.
- **ServiceNow:** Adds usage-based pricing on top of seat licenses to manage high-token Now Assist LLM operations. This allows users to invest in AI features without overcommitting, especially for scenarios involving generative responses.
- **Microsoft:** Copilot Chat in Microsoft 365 is available for businesses on a consumption-based model. Charges are determined according to the usage of generative AI features, helping customers manage spend by controlling access and activity.
- **AWS Bedrock:** Token charges based upon inference inputs and outputs, with costs dependent on the AI model invoked, region and optional charges for data transfer and storage.

- **Globant:** Offers AI Pods with token-based subscription tiers. Businesses can choose a tier based on expected AI use, providing both flexibility and predictability in pricing.
- **Coupa:** Coupa has launched consumption-linked AI models that scale with business value. Pricing is aligned to usage of AI-driven recommendations and insights, ensuring customers pay in proportion to the outcomes generated and don't have to pay for "bad recommendations."

In summary, usage-based AI pricing aligns customer cost with resource consumption, allowing them to pay based on tokens processed, AI queries made and their complexity, jobs triggered, actions executed, data processed or time consumed. These pricing models support business experimentation and will facilitate adoption. In addition, outcome-based pricing ties cost to business value or guaranteed service levels, such as charging for successful call outcomes or threat detection within a specified timeframe. Finally, hybrid models deliver both predictability of spend for the base price and the elasticity to accommodate spikes in usage over the base threshold.

Charging and billing for AI-embedded workflows will demand the scalability and reliability of billing systems that have been tried and tested on the world's largest telecommunications networks.

The Coming Wave of AI Bill Shock

While token-based AI pricing will help accelerate adoption by making it easy for customers to experiment, it also introduces new risks, foremost among them, AI bill shock. As close observers of the telco industry can relate from experience, early adopter enterprises may find that these initial token-based offers obscure rather than clarify cost. Customers with pre-purchased bulk credits are not likely to understand how quickly they are consumed by complex queries, data processing or autonomous agent actions.

Anecdotal evidence from early deployments suggests a common theme: usage explodes, bills follow and customers are caught off guard. The lack of transparency into the actual cost per AI action, especially in agent-based models, makes it difficult to forecast spend or assess ROI. Worse, many enterprises may discover that what appeared to be an efficiency play (e.g., replacing call center agents with AI) may not yet offer a cost advantage, particularly in regions with lower labor costs.

This pricing opacity will likely affect the software providers themselves, who will have to account for real-time usage fees from OpenAI, AWS, GCP or other infrastructure and LLM partners. In this context, real-time monetization systems aren't just about billing; they're a strategic necessity. The ability to meter usage, calculate cost and convey pricing in real-time helps prevent customer backlash and ensures that providers can maintain margins while scaling AI services.

Trust and Transparency: The Importance of Real-Time Observability

Critical requirements to manage and avoid AI customer bill shock include:

Provision of real-time usage dashboards and cost trackers	Notifications when thresholds are reached	Usage and billing projections based upon historical data and consumption in the current period	Dynamic recommendations for the best plans

All of these capabilities rely on real-time usage charging and dynamic monetization systems.

AI Pricing Case Study

This case study illustrates the example of a call center vendor that has historically provided businesses with outsourced human agents, connectivity, process automation and workflow solutions to handle customer interactions over the phone. These interactions include inquiries, sales, payments, issues and returns. Traditional outsourced call center staff and applications have been billed on a per-seat basis, wherein businesses choose the number of agents designated to handle phone calls during specific periods of the day.

An initial wave of call center AI tools has been launched that provide complementary services for those human agents, supporting them by answering basic questions (store hours, directions, parking availability), note-taking, generating and sending responses via email or text message. As these services are being introduced, we have observed that initial price plans for the AI add-on features mirror the human agent seat-based price plans:

BRONZE AI PACKAGE

$15

AGENT SEAT/MONTH

- ✓ Automatic call transcripts
- ✓ Automatic call/meeting summaries

SILVER AI PACKAGE

$30

AGENT SEAT/MONTH

- ✓ Automatic call transcripts
- ✓ Automatic call/meeting summaries
- ✓ Automated call routing to best human agent
- ✓ Real-time note-taking
- ✓ Generates action items

GOLD AI PACKAGE

$45

AGENT SEAT/MONTH

- ✓ Automatic call transcripts
- ✓ Automatic call/meeting summaries
- ✓ Automated call routing to best human agent
- ✓ Real-time note-taking
- ✓ Generates action items
- ✓ Automatic handling of basic FAQ from a knowledge base
- ✓ Generates follow-up emails and SMS messages
- ✓ Detailed call analytics and reporting

Pricing AI agents based on human seats neither aligns with the resources required for the AI agents to perform their functions nor with the incremental value that the business receives. As AI agents become more capable and assist with more tasks, a better pricing method assigns a value to each task performed. By bundling tasks to match varying call center needs, charging in line with AI-assisted task value can improve customer experience, build trust and align provider revenue with resources used and value delivered.

Some examples of value-based metrics for AI tasks include:

- Automatic routing of an incoming call based on agent skills and availability — priced per call
- Generating call transcripts — priced per number of words, e.g, a charge for each 100 words transcribed
- Gathering, summarizing and presenting customer history to a human agent when handling an inbound call — charged per volume of historical data processed

The following offers illustrate how value-centric pricing can be incorporated with various metrics and volume tiers:

BRONZE CALL CENTER BUNDLE		
Automatic call routing	1-200/day	$ 0.50
	200+/day	$ 0.40
FAQ per response		$ 0.15
Call transcripts per 100 words	Up to 50k words/day	$ 0.70
	50k+ words/day	$ 0.60
Automatic appointments with 2FA	Not included	
Appointment reminders	1–20/day	$ 0.50
	21+/day	$ 0.30
Customer summaries	Not included	
Localized greetings	Not included	
Simultaneous call limit		20
100% call responses guarantee	Not included	

SILVER CALL CENTER BUNDLE		
Automatic call routing	1-200/day	$ 0.50
	200+/day	$ 0.40
FAQ per response		$ 0.15
Call transcripts per 100 words	Up to 50k words/day	$ 0.70
	50k+ words/day	$ 0.60
Automatic appointments with 2FA	1–20/day	$ 1.00
	21+/day	$ 0.90
Appointment reminders	1–20/day	$ 0.50
	21+/day	$ 0.30
Customer summaries	1-200/day	$ 2.50
	200+/day	$ 2.10
Localized greetings	Monthly fee per region	$ 100.00
	Per call, 1-200/day	$ 0.50
	Per call, 201+/day	$ 0.40
Simultaneous call limit		50
100% call responses guarantee	8am–5pm	$ 1000.00
	24 hours	$ 1500.00

GOLD CALL CENTER BUNDLE		
Automatic call routing	1-200/day	$ 0.50
	200+/day	$ 0.40
FAQ per response		$ 0.15
Call transcripts per 100 words	Up to 50k words/day	$ 0.70
	50k+ words/day	$ 0.60
Automatic appointments with 2FA	1–20/day	$ 0.95
	21+/day	$ 0.87
Appointment reminders	1–20/day	$ 0.50
	21+/day	$ 0.30
Customer summaries	1-200/day	$ 2.50
	200+/day	$ 2.10
Localized greetings	Monthly fee per region	$ 90.00
	Per call, 1-200/day	$ 0.50
	Per call, 201+/day	$ 0.40
Simultaneous call limit		100
100% call responses guarantee	8am–5pm	$ 2000.00
	24 hours	$ 4500.00

Strategic Considerations for Enterprise Software Providers

To compete and scale profitably, enterprise software providers must:

- Rethink pricing to reflect AI-specific cost/value dynamics
- Evaluate billing systems for their ability to support usage processing at scale, their architecture to flexibly fit into your ecosystem, and their support of usage and pricing observability at scale
- Package AI features transparently to foster customer trust
- Automate real-time monetization to integrate the AI life cycle from utilization to billing

Designing for the AI Future

AI will demand new contracts between application providers and their enterprise buyers. As usage explodes and the resources required to meet the demand skyrocket, flat-rate licensing is increasingly unfit for purpose. The future of AI-embedded application logic lies in:

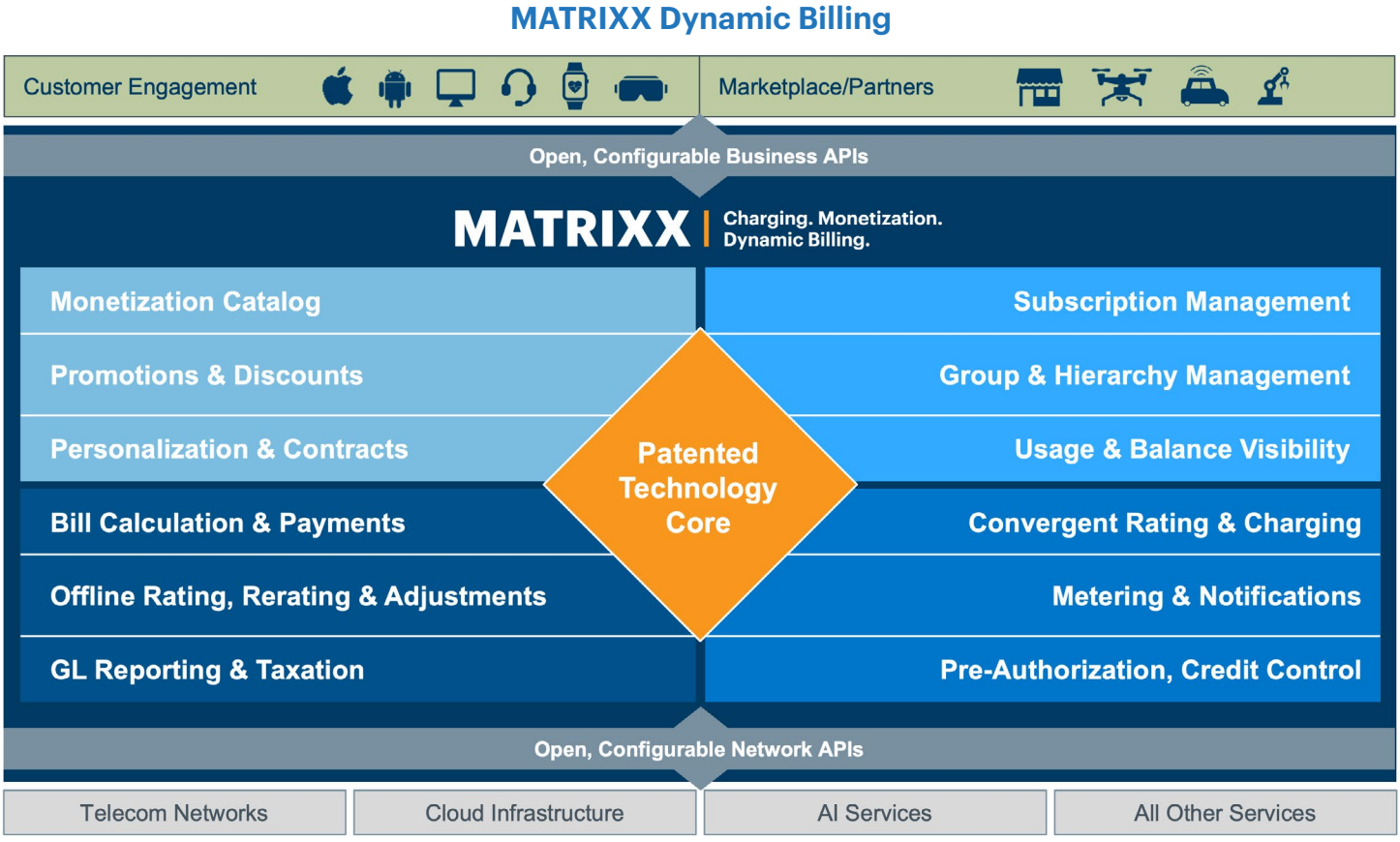
- Usage-aware pricing that scales with demand
- Outcome-based models aligned to business value
- Real-time, transparent observability and billing

Application providers must adapt quickly to new pricing and billing models based on value, and to do so, they need dynamic, real-time and highly performant monetization solutions. MATRIXX Dynamic Billing is a patented real-time rating, charging and monetization engine that has supported flexible value chains including B2C, B2B and wholesale, delivering transformed operations and enhanced customer experiences to providers around the world. By adapting dynamic, value-based pricing models, AI-enabled software vendors will encourage adoption, unlock new revenue streams and drive higher margins that will ensure they recover infrastructure costs.

MATRIXX Dynamic Billing for AI Agents

MATRIXX Dynamic Billing streamlines and unifies once disparate processes for charging and billing for any activity — a session on a telco network, a reading from an IoT sensor or the invocation of an AI agent. By consolidating and simplifying these processes, all charges are generated, managed and available to users in real-time. Most critically, MATRIXX Dynamic Billing assigns a value to an activity based on any single or combination of parameters attributed to the event, including context and the value of its outcomes. The result is complete transparency of all incurred costs for the user and full visibility of all revenue for the enterprise software provider at every point in the AI life cycle.

MATRIXX Dynamic Billing handles the biggest challenges in pricing complexity and scales to support millions of users and billions of actions in real-time. The unique, flexible and high-performance MATRIXX architecture benefits any provider pursuing new and evolving AI-centric business models in a way that encourages adoption, demonstrates value and keeps pace with the rapid rate of innovation.



About MATRIXX Software

MATRIXX Software delivers a dynamic billing, monetization and charging solution proven at scale. Global service providers like Telefónica, IoT providers like Tata Communications and network-as-a-service providers like DISH rely on MATRIXX to overcome the limitations of existing billing applications. MATRIXX provides a unified platform that transforms and simplifies billing operations across consumer, enterprise and wholesale businesses. With MATRIXX, operators can rapidly configure, deploy and monetize personalized offerings, enabling commercial innovation and real-time customer experiences that drive revenue and growth.

matrixx.com