

Three Eras of Enterprise Technology Pricing: On-Premises Infrastructure, Cloud Computing, and the AI Factory

Rajan Seth

Independent Researcher, USA

Received: 05.02.2026

Accepted: 08.02.2026

Abstract

Enterprise technology pricing has evolved through distinct phases. Each phase reflects fundamental shifts in how value gets perceived across technology industries, and the pre-cloud period established ownershipcentric frameworks. Physical hardware attributes governed pricing structures through per-socket and perdevice metrics during earlier decades. Capital expenditure inefficiencies characterised computing environments of the ownership era. Organisations faced chronic underutilisation stemming from peak capacity provisioning requirements. Shelfware accumulation represented the software manifestation of resource inefficiency. Cloud computing introduced consumption-based economics as a transformative alternative. Capital outlays are converted into operational expenses through elastic provisioning mechanisms. Hyperscale providers abstracted infrastructure ownership entirely from enterprise customers. Per-vCPU-hour and per-gigabyte-month metrics aligned costs with actual utilisation patterns. Software-as-a-service platforms popularised subscription-based per-seat models across enterprise markets. Value gaps persisted between power users and casual consumers paying equivalent subscription rates. The emerging AI factory era represents a radical transformation in pricing paradigms. Previous commoditisation trends have inverted completely due to accelerated demand. Token-based economics establishes novel pricing foundations for generative artificial intelligence services. Outcome-based models emerge as definitive solutions for modern enterprises. Payments now align with business results rather than input consumption metrics. Customer support operations are beginning to transition toward per-resolution charges. Financial services adopt AI agent platforms with task-based pricing structures. Automatic model tuning services demonstrate platform-level value capture through managed complexity reduction.

Keywords: Enterprise Pricing Evolution, Cloud Computing Economics, Outcome-Based Pricing, Token Economics, Tensor Core Architecture, Consumption Models

I. Introduction

The pricing structures employed within enterprise technology markets represent philosophical statements regarding value location rather than mere numerical calculations. Across three decades of technological advancement, fundamental reconceptualisations have occurred regarding how organisations capture and pay for technology value. The trajectory has progressed from purchasing physical infrastructure through renting computational time toward the emerging paradigm of paying for achieved results. Strategic pricing research demonstrates that value-based approaches systematically outperform cost-plus methodologies in revenue generation and market positioning, with organisations employing customer value-based pricing achieving superior financial performance across multiple industry sectors [1]. Customer willingness to pay

10.48047/jocaaa.2026.35.02.11

correlates directly with perceived value rather than production costs, yet many organisations continue anchoring prices to internal cost structures, creating fundamental tensions when pricing metrics measure technical specifications rather than business outcomes [1].

The pre-cloud era established pricing mechanisms anchored to physical hardware characteristics. Organisations acquired, housed, cooled, and secured computational infrastructure, with pricing units denominated in servers, racks, and processor sockets. Infrastructure software licensing attached to underlying hardware through per-socket metrics that doubled costs when organisations deployed more powerful multi-socket servers, regardless of software efficiency improvements [2]. Application software distributed via physical media commanded per-desktop or per-license-key pricing, with traditional software pricing relying heavily on one-time license fees coupled with annual maintenance charges, creating upfront revenue concentration but limited ongoing customer engagement [2]. The fundamental inefficiency of this era manifested through capital expenditure patterns where organisations provisioned for peak capacity requirements, resulting in substantial shelfware accumulations and chronically underutilised silicon investments.

Cloud computing fundamentally shattered ownership models by introducing consumption-based economics where hyperscale providers owned hardware while enterprises rented capabilities. Cloud computing emerged as a transformative paradigm by offering elasticity, allowing computational resources to scale dynamically based on demand patterns rather than fixed capacity provisioning [3]. The utility pricing model inherent in cloud services fundamentally differs from traditional software licensing by charging customers based on actual consumption rather than potential usage rights [3]. This transformation converted capital expenditures into operational expenses, fundamentally altering technology acquisition economics and financial planning processes. Simultaneously, software-as-a-service platforms popularised per-user-per-month subscription models, though value gaps persisted where power users consuming substantial resources paid identical rates to casual users accessing platforms infrequently.

The AI factory era emerging from 2024 represents the most radical pricing transformation. Graphics processing unit architectures incorporating tensor cores have achieved substantial performance improvements for matrix operations fundamental to deep learning workloads, with modern accelerators demonstrating order-of-magnitude throughput increases for artificial intelligence computations [4]. Token-based economics have emerged as the fundamental pricing mechanism for generative artificial intelligence services, where optimal pricing strategies must account for the unique characteristics of language model consumption patterns [11]. The definitive solution emerging involves outcome-based pricing that aligns payments with business results, with AI agent platforms demonstrating autonomous task completion capabilities that justify pricing based on achieved outcomes rather than computational resource consumption [12]. This evolution removes friction between price and value by transitioning from asking how many boxes organisations want to own, through how much time organisations want to rent, toward what results organisations want to achieve.

II. Related Work

The intersection of pricing strategy and technological evolution has received sustained scholarly attention, though comprehensive frameworks addressing metric adaptation across major paradigm transitions remain underdeveloped. Value-based pricing literature establishes fundamental principles for aligning prices with customer willingness to pay rather than internal cost structures [1]. Research demonstrates that organisations implementing value-oriented approaches achieve superior financial performance compared to cost-plus methodologies, with evidence indicating that customer value-based pricing correlates positively with firm profitability across diverse market conditions [1]. The gap between theoretical frameworks and

10.48047/jocaaa.2026.35.02.11

operational execution becomes especially pronounced during technological disruptions that fundamentally alter relationships between measurable characteristics and delivered benefits. Value quantification emerges as essential for establishing pricing metrics that capture genuine customer benefits rather than technical implementation details [1].

Cloud computing economics research examines utility pricing models enabling conversion of capital expenditures into operational expenses through pay-per-use mechanisms [3]. The elasticity inherent in cloud platforms enables dynamic resource scaling without upfront infrastructure investments, creating cost flexibility aligned with actual demand patterns rather than peak capacity planning [3]. Infrastructure-as-a-service providers introduced resource pooling mechanisms where multiple customers share underlying hardware through virtualisation, necessitating pricing approaches that account for resource allocation efficiency and multi-tenant architectures [3]. Platform and software service layers build upon these infrastructure foundations, inheriting many pricing characteristics while introducing additional considerations around application-specific value metrics and user engagement patterns [3]. The Berkeley perspective on cloud computing identified pay-as-you-go economics as fundamental to utility computing adoption, enabling organisations to convert fixed infrastructure investments into variable operational costs scaling with business activity [9].

Virtualisation economics literature examines cost reduction benefits from server consolidation and improved hardware utilisation rates [6]. The hypervisor design enabling multiple virtual machines on shared physical resources raises performance isolation concerns requiring sophisticated scheduling mechanisms [6]. Paravirtualisation techniques reduce overhead by facilitating communication between guest operating systems and virtualisation layers, with memory management presenting particular challenges as hypervisors must maintain isolation while minimising address translation overhead [6]. Virtualisation performance depends critically on efficient processor scheduling across multiple virtual machines competing for execution time, with device input/output operations creating additional complexity through abstraction layers that traditionally imposed substantial performance penalties [6]. Existing literature inadequately addresses how licensing models responded to multi-core processor adoption and the revenue implications of delayed pricing restructuring.

Multi-core architectures became the dominant response to power-limited scaling, distributing computational work across multiple lower-frequency cores rather than pursuing single-threaded performance through clock speed increases [7]. Heterogeneous computing architectures incorporating specialised accelerators alongside general-purpose cores represent further evolution, optimising energy efficiency by matching workload characteristics to appropriate execution units [7].

Software licensing research segments traditional models, including perpetual licenses, subscription services, and usage-based pricing across dimensions like named users, concurrent sessions, and device installations [2]. The shift toward software-as-a-service platforms converted one-time purchases into recurring revenue models with significant implications for vendor and customer financial forecasting [2]. Per-user pricing models dominate enterprise software markets due to intuitive correlation between user count and organisational value extraction, though this metric faces challenges in scenarios where small user populations generate disproportionate business impact [2]. Feature-based pricing strategies segment offerings into tiered packages, enabling vendors to capture value differentiation across customer segments with varying functional requirements and willingness to pay [2]. Open innovation frameworks recognise that valuable knowledge distributes broadly across ecosystems rather than concentrating within organisational boundaries, influencing platform business models facilitating third-party integrations and complementary service development [8].

III. Era One: The Pre-Cloud Period (1990s–2010)

A. The Ownership Paradigm

The pre-cloud era established value as fundamentally physical. Computational power required acquisition, housing, cooling, and security rather than rental or subscription. Hardware dominated pricing metrics, with raw CPU servers, spinning hard disks, and specialised storage arrays commanding prices denominated in boxes and racks [9]. The pricing unit reflected physical infrastructure ownership where organisations maintained complete control over deployed assets while bearing full responsibility for capacity planning, maintenance, and eventual replacement cycles. The elimination of upfront costs for computer hardware represented one of the primary economic advantages that would later drive cloud adoption, highlighting the substantial capital requirements characterising pre-cloud infrastructure acquisition [9].

Infrastructure software is attached to underlying hardware through licensing mechanisms that measure physical characteristics rather than computational output. The per-socket metric dominated enterprise software pricing, where database licenses, virtualisation platforms, and middleware products were charged based on processor socket counts within licensed servers [6]. This metric aligned reasonably with value delivery when each socket contained single processing cores with relatively fixed computational capacity, enabling straightforward cost prediction through physical socket enumeration [6]. However, this measurement approach contained fundamental vulnerabilities that subsequent multi-core processor evolution would expose dramatically.

Application software distribution relied upon physical media, including CD-ROMs and limited download mechanisms that constrained deployment flexibility. Products, including productivity suites and creative software packages sold through per-desktop or per-license-key models, establishing perpetual ownership rights coupled with optional maintenance agreements [2]. Software vendors faced fundamental decisions regarding pricing model selection, with choices ranging from perpetual licensing to emerging subscription-based approaches, each carrying distinct implications for revenue predictability and customer relationships [2]. The distribution bottleneck inherent in physical media created natural enforcement mechanisms while simultaneously limiting vendor visibility into actual software utilisation patterns across customer deployments.

B. Capital Expenditure Inefficiencies

The fundamental flaw characterising pre-cloud pricing manifested through capital expenditure patterns requiring organisations to provision for peak capacity requirements. Infrastructure investments sized for maximum anticipated demand resulted in chronic underutilisation during normal operational periods [9]. The Berkeley analysis of cloud computing economics specifically identified the risk of under-purchasing and under-utilising resources as primary motivations for transitioning away from ownership models, where organisations either purchased insufficient capacity for peak demands or excessive capacity that remained idle during typical operations [9]. Organisations effectively purchased high-performance capabilities to address occasional peak requirements while experiencing minimal utilisation during typical workloads.

Shelfware accumulations represented the software manifestation of this inefficiency, where perpetual licenses purchased exceeded actual deployment requirements [2]. Volume licensing agreements encouraged bulk acquisitions offering per-unit discounts, yet organisations frequently overestimated deployment needs or experienced changing requirements that left substantial license inventories unused. The complexity inherent in multi-dimensional pricing created potential for optimisation but risked customer confusion when pricing structures became overly intricate [2]. The combination of hardware overprovisioning and software overacquisition created systematic value leakage where organisations paid for capabilities never utilised, while vendors captured revenue disconnected from actual value delivery. Technology refresh cycles compounded these inefficiencies through depreciation schedules misaligned with capability evolution.

10.48047/jocaaa.2026.35.02.11

Organisations committed to multi-year hardware deployments faced rapid obsolescence as processor performance, storage density, and networking capabilities advanced faster than replacement cycles permitted [7]. The capital investment model locked organisations into aging infrastructure while competitors deploying newer systems gained performance advantages, creating pressure for accelerated replacement schedules that further increased total cost of ownership without proportional value improvements.

Pricing Dimension	Measurement Basis	Value Alignment Characteristics	Fundamental Limitations
Hardware Infrastructure	Per-server, perrack, physical socket count	Direct correlation with acquired physical capacity	Peak provisioning requirements created chronic underutilisation
Database Software	Per-processorsocket licensing	Reasonable alignment when sockets contained single cores	Multi-core evolution severed the connection between sockets and capacity
Application Software	Per-desktop, perlicense-key ownership	Clear deployment point enumeration	Shelfware accumulation from volume licensing overacquisition
Storage Systems	Per-terabyte capacity acquisition	Transparent capacity-based pricing	Growth requirements demanded repeated capital investments
Maintenance Agreements	Percentage of license value annually	Predictable ongoing cost structure	Continued payments regardless of actual support utilisation

Table I: Pre-Cloud Era Pricing Mechanisms and Limitations [2, 6, 9].

IV. Era Two: The Cloud Computing Period (2010–2023)

A. Infrastructure Transformation

The emergence of hyperscale cloud providers fundamentally shattered ownership models by abstracting infrastructure into service offerings. Amazon Web Services, Google Cloud Platform, and Microsoft Azure purchased hardware at unprecedented scale while enterprises rented capabilities without physical asset ownership [3]. Cloud computing introduced transformative economic models by converting capital expenditures into operational expenses, fundamentally altering technology acquisition patterns for enterprises seeking computational capabilities [9]. The infrastructure shift replaced boxes with instances, enabling granular resource provisioning aligned with actual requirements rather than peak capacity estimates.

Compute pricing transitioned from physical server acquisition to virtual machine rental with hypergranular measurement capabilities. The utility computing paradigm enables organisations to provision resources dynamically without upfront infrastructure investments, creating cost elasticity that scales with actual demand rather than peak capacity planning [9]. Pay-as-you-go pricing represents the purest expression of consumption-based economics, charging customers only for resources actually utilised during specific time periods [3]. This elasticity represented the defining characteristic of cloud computing, allowing customers

10.48047/jocaaa.2026.35.02.11

to scale resources dynamically without hardware procurement delays while converting capital expenditures into operational expenses aligned with business activity levels. The illusion of infinite capacity eliminated procurement bottlenecks while introducing billing variability that challenged traditional budgeting processes accustomed to fixed infrastructure costs [9].

Storage services eliminated hardware acquisition through consumption-based pricing denominated in gigabytes per month [5]. Organisations increasingly shifted away from purchasing hard drives and storage arrays, instead provisioning capacity dynamically as requirements evolved. Storage system performance exhibits considerable variation across providers in both throughput and latency characteristics, with substantial differences in input/output operations for comparable pricing tiers creating complexity in vendor selection [5]. Tiered storage offerings enabled cost optimisation through data lifecycle management, with frequently accessed data commanding premium pricing while archival storage offered substantially reduced rates for infrequently retrieved information.

B. The Private Cloud Conflict

Legacy virtualisation vendors faced substantial challenges adapting per-socket licensing to cloud-native environments characterised by dense multi-core processors [6]. The socket-based metric that functioned adequately when processors contained single cores became severely misaligned as semiconductor manufacturing advanced to integrate multiple cores within individual sockets [7]. Each processor generation delivered exponentially greater computational capacity while socket counts—and therefore licensing costs—remained constant, creating systematic value capture failures for software vendors. The transition from single-core to multi-core architectures created massive value-pricing disconnects. The breakdown of Dennard scaling principles forced processor architects toward parallel execution strategies, with multi-core integration becoming the primary method for continued performance improvements [7]. Dark silicon phenomena emerged where thermal constraints prevented simultaneous activation of all available transistors, further accelerating the transition toward distributed processing across multiple cores [7]. Customers deploying newer hardware received dramatically greater computational resources without proportional pricing increases, essentially receiving free capacity improvements through hardware upgrades while vendors lost revenue opportunities scaling with expanding capability gaps.

Recognition of this misalignment eventually forced transitions toward core-based licensing models. By shifting pricing units from physical sockets to individual processing cores, vendors re-established a correlation between pricing and computational capacity [6]. However, processor architecture diversity complicates core-based licensing implementation, as different microarchitectures deliver varying performance levels per core, creating potential fairness issues in pricing structures [7]. Delayed implementation resulted in prolonged revenue losses during extended transition periods, while migration complexities, including customer resistance to perceived price increases and technical challenges in accurate core measurement across diverse configurations, extended correction timelines substantially.

C. The Software-as-a-Service Revolution

Simultaneously with infrastructure transformation, application software migrated from desktop installations to browser-based delivery through software-as-a-service platforms. The per-user-per-month subscription became the dominant pricing metric for enterprise applications spanning customer relationship management, collaboration tools, and productivity platforms [2]. Software pricing strategy encompasses multiple dimensions, including licensing scope, user permissions, functional capabilities, and deployment scale, with vendors determining whether pricing aligns with installation instances, concurrent users, named users, or enterprise-wide agreements [2]. This seat-based economy converted one-time purchases into recurring revenue streams, fundamentally altering customer relationships and vendor economics while enabling continuous feature delivery without version upgrade friction.

10.48047/jocaaa.2026.35.02.11

The subscription model represented a substantial improvement over per-key licensing through ongoing revenue predictability and reduced adoption barriers, eliminating large upfront payments [2]. Freemium strategies further lowered initial friction by providing basic functionality without charge while monetising premium features and expanded capabilities. Platform ecosystems emerged where third-party developers created complementary applications and integrations, with open innovation principles recognising that valuable knowledge distributes broadly across ecosystems rather than concentrating within organisational boundaries [8]. The open model adoption acknowledges that critical information exists beyond individual organisations, distributed across developer communities, open-source contributions, and platform participants [8].

However, per-seat pricing retained value gaps despite improvements over ownership models. Power users consuming platform capabilities for extended daily periods paid identical rates to casual users accessing systems infrequently [2]. Per-user pricing models dominate enterprise software markets due to intuitive correlation between user count and organisational value extraction, though this metric faces challenges in scenarios where small user populations generate disproportionate business impact [2]. This pricing equivalence failed to capture value differences across usage intensity segments, where organisations paying for numerous seats might experience minimal utilisation from substantial user populations, while heavy users extracted disproportionate value relative to subscription costs.

Pricing Category	Era One Metric	Era Two Metric	Transformation Characteristics
Compute Resources	Per-server acquisition	Per-vCPU-hour consumption	Physical ownership replaced by elastic rental with granular billing
Storage Capacity	Per-terabyte hardware purchase	Per-GB-month consumption	Capacity provisioning decoupled from physical asset acquisition
Infrastructure Software	Per-socket perpetual license	Per-core subscription or consumption	Socket-based metrics adapted to multicore density realities
Application Software	Per-desktop perpetual ownership	Per-seat monthly subscription	One-time purchases converted to recurring revenue streams
Financial Model	Capital expenditure dominance	Operational expenditure alignment	Technology costs aligned with business activity patterns
Capacity Planning	Peak provisioning requirements	Elastic scaling capabilities	Overprovisioning eliminated through dynamic resource adjustment

Table II: Pricing Metric Evolution from Pre-Cloud to Cloud Eras [2, 3, 5, 6, 9].

V. Era Three: The AI Factory Period (2023–Present)

A. Inverted Hardware Economics

The emergence of artificial intelligence workloads has fundamentally altered hardware economics, inverting the commoditisation trends that characterised cloud computing evolution. Graphics processing unit architectures incorporating tensor cores have achieved substantial performance improvements for matrix operations fundamental to deep learning workloads [4]. Accurate performance modelling of tensor core operations reveals that modern accelerators deliver throughput improvements of two to three orders of

10.48047/jocaaa.2026.35.02.11

magnitude compared to general-purpose CPUs for artificial intelligence computations [4]. Unlike the cloud era, where hardware abstraction enabled competitive pricing pressure driving costs downward, AI-optimised accelerators face supply constraints that maintain elevated price points despite increasing production volumes.

GPU performance trajectories demonstrate improvement rates substantially exceeding traditional processor advancement patterns, with tensor core architectures specifically optimised for the matrix-multiply-accumulate operations fundamental to neural network training and inference [4]. This architectural specialisation creates pricing complexity where hardware generations differ dramatically in performance-per-dollar metrics, requiring pricing frameworks capable of accommodating rapid capability evolution while maintaining customer cost predictability.

The scalable gradient-free optimisation approaches employed by such platforms reduce the expertise required for effective model development, justifying pricing premiums beyond underlying compute instance costs [10]. Ensemble methods combining multiple optimisation strategies further enhance tuning efficiency, with platforms capturing value through managed complexity reduction rather than raw computational resources [10].

B. Token Economics and Margin Pressures

The atomic unit of the AI era has emerged as the token—the fundamental measurement of generative artificial intelligence consumption and production. Optimal pricing mechanisms for large language model services must account for the unique characteristics of token-based consumption, where traditional per-unit pricing frameworks require adaptation to accommodate linguistic rather than computational units [11]. Token pricing theory establishes that efficient mechanisms must balance provider cost recovery with consumer value extraction, creating novel economic structures distinct from traditional software or infrastructure pricing [11]. Billions of tokens flow through enterprise AI implementations daily, creating consumption volumes requiring sophisticated measurement and billing infrastructure exceeding traditional software metering complexity.

Enterprises integrating generative AI capabilities face margin compression pressures from token cost structures [11]. Unlike traditional software, where marginal usage costs approach zero after development investment, each AI interaction consumes computational resources with measurable costs requiring recovery through pricing mechanisms. The analysis of token-based pricing demonstrates that optimal mechanisms differ substantially from traditional per-seat or per-instance approaches, requiring consideration of generation costs, prompt lengths, and output complexity [11]. This variable cost structure conflicts with enterprise preferences for predictable expenditure patterns, enabling annual budgeting and cost allocation across business units.

The fixed-price trap emerges when software-as-a-service providers charge flat subscription rates while embedding AI capabilities with variable consumption costs [11]. Heavy users generating substantial token volumes can consume costs exceeding subscription revenue, destroying margins for providers absorbing consumption variability. Conversely, the metered trap arises when providers pass token costs directly to customers, generating resistance from enterprise buyers unable to budget for unpredictable consumption patterns. Neither extreme provides sustainable economics, requiring hybrid approaches that balance cost recovery with customer experience expectations. Token pricing optimisation must therefore consider both provider sustainability and customer adoption incentives to achieve market equilibrium [11].

C. Outcome-Based Pricing Emergence

The definitive solution emerging from AI era pricing challenges involves outcome-based models aligning payments with business results rather than input consumption. This evolution represents the logical culmination of value-based pricing principles, where customers pay for achieved objectives rather than

10.48047/jocaaa.2026.35.02.11

resources consumed in pursuit of those objectives [1]. Value-based pricing requires a systematic understanding of how customers perceive and measure value from offerings, with outcome orientation representing the purest expression of this principle [1]. The fundamental insight recognises that users care about completed work rather than token counts, instance hours, or seat licenses—metrics measuring inputs rather than outputs.

AI agent platforms demonstrate the viability of outcome-oriented approaches through autonomous task execution capabilities. The FinRobot platform illustrates how AI agents can perform complex financial analysis tasks, including market prediction, document analysis, and trading strategy development with minimal human intervention [12]. Such platforms integrate multiple specialised AI models coordinated through unified interfaces, enabling end-to-end task completion that justifies pricing based on delivered outcomes rather than computational consumption [12]. The multi-agent architecture employed allows different components to handle specialised subtasks while presenting unified results to users [12]. Financial services applications particularly demonstrate outcome-based pricing potential through quantifiable result measurement. AI agents performing market analysis, risk assessment, and trading execution produce measurable outcomes, including prediction accuracy, portfolio returns, and risk-adjusted performance metrics [12]. The open-source architecture of platforms like FinRobot enables customisation for specific institutional requirements while maintaining core capabilities for autonomous task completion [12]. This task-oriented approach naturally aligns with pricing models where payment correlates with successful outcome delivery rather than resource consumption during processing. Customer support operations demonstrate parallel evolution through per-resolution pricing mechanisms. Traditional models charged per-support-agent-seat, measuring headcount rather than issue resolution effectiveness. AI-enabled support systems enable per-resolution charges where payment occurs only upon successful ticket resolution without human intervention. This metric alignment ensures vendors capture value proportional to delivered outcomes while customers pay for results rather than capability access. The transition from input to outcome measurement requires robust verification mechanisms but ultimately creates sustainable economics for both providers and consumers.

Domain	Traditional Metric	AI Era Metric	Value Alignment Improvement
Customer Support	Per-supportagent seat	Per-resolution achieved	Payment occurs only upon successful issue resolution
Financial Analysis	Per-analyst seat	Per-insight or per-prediction delivered	Pricing reflects actionable findings and accuracy
Model Development	Per-datascientist seat	Per-model-deployed or pertuning-cycle	Automated tuning platforms capture complexity value
Content Generation	Per-creator seat	Per-content-piece produced	Payment scales with actual production volume
Trading	Per-trader seat	Per-trade-executed or	Costs align with transaction
Operations		performance-based	outcomes

10.48047/jocaaa.2026.35.02.11

Document Processing	Per-processor seat	Per-document-analysed	Value measured by extraction accuracy and throughput
---------------------	--------------------	-----------------------	--

Table III: Transition from Input-Based to Outcome-Based Pricing Metrics [10, 11, 12].

VI. Comparative Framework Analysis

A. Philosophical Evolution

The three-era evolution represents progressive refinement in aligning pricing with value location. Era one asked how many boxes organisations wanted to own, measuring value through physical asset accumulation [9]. Era two asked how much time organisations wanted to rent, measuring value through resource consumption duration [3]. Era three asks what results organisations want to achieve, measuring value through business outcomes delivered [1]. This philosophical progression removes successive layers of friction between payment and value, though each era introduced new alignment challenges requiring subsequent innovation.

The ownership paradigm located value in physical control, where possessing infrastructure conferred capability regardless of utilisation efficiency [9]. Organisations paid for potential rather than realised value, accepting capital expenditure inefficiencies as unavoidable costs of maintaining computational capabilities. The consumption paradigm relocated value to utilisation, where elastic resource access enabled payment proportional to actual usage patterns [3]. This improvement eliminated overprovisioning waste while introducing consumption variability that complicated financial planning. The outcome paradigm locates value in results, where demonstrated business impact determines payment rather than either ownership or consumption measurements [1].

B. Metric Inheritance Patterns

Each era inherited characteristics from predecessors while introducing transformative elements. Cloud infrastructure pricing inherited hardware-based metrics reflecting direct relationships between service costs and underlying physical specifications, with virtual machine pricing incorporating processor, memory, storage, and network dimensions mirroring hardware component structures [5]. Platform services are built upon infrastructure foundations while adding application-layer consumption dimensions, including API calls, transaction volumes, and managed service tiers. Automatic model tuning platforms exemplify this layering, capturing value through optimisation complexity management built upon underlying compute infrastructure [10].

The AI factory era inherits cloud consumption measurement capabilities while introducing outcomebased dimensions previously absent from technology pricing. Token economics build upon pertransaction measurement infrastructure developed for API billing, though token semantics differ substantially from traditional API calls in representing linguistic or content units rather than functional requests [11]. Outcome metrics require measurement infrastructure capturing business results rather than technical consumption, demanding integration with customer business systems to verify resolution completion, model deployment success, or financial prediction accuracy [12].

C. The Per-Seat Collapse

The per-seat model dominating era two faces fundamental challenges in the AI factory era that threaten its continued viability. This metric breaks down when AI agents perform work previously requiring multiple human users, creating economic misalignment where vendors lose seat revenue while delivering increased value through automation [12]. AI agent platforms capable of autonomous task execution across financial analysis, document processing, and decision support functions demonstrate capabilities previously requiring substantial human analyst teams [12]. If organisations reduce analyst headcount because AI

10.48047/jocaaa.2026.35.02.11

systems handle equivalent workloads, software vendors charging per-analyst-seat lose revenue despite providing greater value through enhanced capabilities.

This economic misalignment proves unsustainable for vendors and customers alike. Vendors face revenue erosion as customers reduce seat counts while extracting increased value from AI-enhanced capabilities. Customers face pressure to maintain unnecessary seat counts to preserve vendor relationships and support investment. The resolution requires transitioning to outcome-based metrics where vendor revenue scales with delivered value regardless of human headcount involved in achieving results [1]. Customer valuebased pricing demonstrates superior financial performance when properly implemented, though transitioning existing customer relationships requires careful change management [1]. Vendors succeeding in this transition will align revenue with customer success, creating sustainable economics supporting continued capability investment.

The following comparison synthesises the evolutionary trajectory of enterprise technology pricing across three distinct paradigms. Each era reflects fundamental shifts in value perception, measurement mechanisms, and economic models governing technology acquisition decisions. The transition from ownership through consumption toward outcome orientation demonstrates progressive alignment between payment structures and actual business value delivery.

Feature	Era 1: Pre-Cloud	Era 2: Cloud Era	Era 3: AI Factory
Timeline	1990s – 2010	2010 – 2023	2024 – Future
Core Philosophy	Ownership (Buying the Box)	Consumption (Renting the Time)	Outcome (Buying the Result)
Hardware Metric	Per-Socket / Per-Server (Physical Capacity)	Per-vCPU / Per-GB / Hr-Month (Elastic Usage)	GPU Premium (Performance Scarcity)
Software Metric	Per-Desktop / Per-Key (One-time License)	Per-Seat / Per-User (Subscription)	Per-Outcome / Work (Resolution, Lead, Fix)
Financial Model	CapEx (Heavy Upfront Investment)	OpEx (Pay-as-you-go)	Value-Based (Pay-for-Performance)
Primary Friction	Shelfware (Paying for unused capacity)	Unused Seats (Paying for inactive users)	Token Costs (High input costs vs. value)
Unit of Value	The Device	The Login	The Job Completed

Table 4. Comprehensive Comparison of Enterprise Technology Pricing Across Three Evolutionary Eras [1, 2, 3, 9, 11, 12].

Conclusion

The trajectory from ownership through consumption toward outcome orientation demonstrates progressive refinement in value alignment mechanisms. Socket-based licensing failures during multi-core processor adoption exposed critical vulnerabilities. Measuring physical characteristics rather than computational capacity created systematic revenue leakage. Delayed metric corrections resulted in prolonged erosion of vendor revenues across virtualisation markets. Per-seat subscription models improved substantially upon perpetual licensing structures. Recurring revenue streams replaced concentrated upfront payments. Customer engagement increased through continuous feature delivery without version upgrade friction. The

10.48047/jocaaa.2026.35.02.11

seat-based model now faces existential challenges from AI automation capabilities. Intelligent agents performing tasks previously requiring multiple human workers create fundamental economic misalignment. Vendors lose seat revenue while delivering enhanced value through automated task completion. Token economics introduce margin pressures unfamiliar to traditional software businesses. Variable consumption costs conflict sharply with enterprise budgeting preferences for predictable expenditures. Fixed-price subscriptions risk margin destruction when heavy users consume excessive tokens. Metered pricing approaches generate customer resistance due to cost unpredictability concerns. Hybrid mechanisms balancing cost recovery with adoption incentives represent emerging solutions for vendors. Outcome-based pricing aligns vendor success directly with customer achievement of business objectives. Per-resolution support charges and cost-per-software-unit development metrics exemplify value-aligned pricing structures. AI agent platforms in financial services demonstrate autonomous task completion, justifying outcome-oriented pricing. Successful vendors will implement systematic value measurement capabilities, enabling proactive restructuring. Adaptation speed emerges as the critical competitive differentiator. Organisations clinging to obsolete metrics will experience systematic value leakage. Competitors embracing outcome-aligned pricing innovation will capture emerging market opportunities across the AI factory era.

References

- [1] Andreas Hinterhuber, "What every manager should know about pricing," JOURNAL OF BUSINESS STRATEGY, 2025. [Online]. Available: <https://www.emerald.com/jbs/articlepdf/45/4/253/9501918/jbs-11-2022-0192.pdf>
- [2] Sonja Lehmann et al., "Pricing Strategies of Software Vendors," Business & Information Systems Engineering, 2009. [Online]. Available: <https://link.springer.com/content/pdf/10.1007/s12599-0090075-y.pdf>
- [3] Michael Armbrust et al., "A view of cloud computing," Communications of the ACM, 2010. [Online]. Available: <https://dl.acm.org/doi/pdf/10.1145/1721654.1721672>
- [4] Faizan A. Khattak and Mantas Mikaitis, "Accurate Models of NVIDIA Tensor Cores," arXiv, 2025. [Online]. Available: <https://arxiv.org/pdf/2512.07004>
- [5] Ang Li et al., "CloudCmp: Comparing Public Cloud Providers," ACM, 2010. [Online]. Available: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/cloudcmp-imec2010.pdf>
- [6] Paul Barham et al., "Xen and the Art of Virtualization," ACM, 2003. [Online]. Available: <https://cse.buffalo.edu/~stevko/courses/cse704/fall10/papers/2003-xensosp.pdf>
- [7] Shekhar Borkar and Andrew A. Chien, "The future of microprocessors," Communications of the ACM, 2011. [Online]. Available: <https://dl.acm.org/doi/pdf/10.1145/1941487.1941507>
- [8] Björn Remneland Wikhamn and Wajda Wikhamn, "Structuring of the Open Innovation Field," Journal of Technology Management and Innovation, 2013. [Online]. Available: <https://www.scielo.cl/pdf/jotmi/v8n3/art16.pdf>
- [9] Armando Fox et al., "Above the Clouds: A Berkeley View of Cloud Computing," ResearchGate, 2009. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/43018702/0c960525ae2cc30314000000libre.pdf?1456339668=&response-contentdisposition=inline%3B+filename%3DAbove_the_Clouds_A_Berkeley_View_of_Clou.pdf&Expires=1764236006&Signature=RVRmlfM~Qb9hr9TweIVRA1~LsCUDIsXW-

10.48047/jocaaa.2026.35.02.11

[JZl3NU9bsFupd4lNENlQ0TX4up6XkmXNLia-nMu5Kmc34e8tD9v4hIxm0oHRLXNBuuBv8lnixf~2p23IyKBGNlib8e8fsdDxjC0os2bK7PdheUhdjq9ZVQ4pHIQUv6TiDBRsB7OL5zLJBG5t2UYhC4vqw2iG4mNONtxZ4mvF~P8ncgSOgeopn3pETbZiszuEULTwB-Kz1EJ88zKGO8QyuN-7DSDFnmgf47XLAN06nFT8TNH8CdO6c2MttujM0BKlWMA saz6Kb1QSZtLmq~VvyJv987uDnHbVghIDYz7ukSEhjby68zQ &Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA](https://doi.org/10.48047/jocaaa.2026.35.02.11)

- [10] Valerio Perrone et al., "Amazon SageMaker Automatic Model Tuning: Scalable Gradient-Free Optimization," ACM, 2021. [Online]. Available: <https://dl.acm.org/doi/pdf/10.1145/3447548.3467098> [11] Weijie Zhong, "Token Is All You Price," arXiv, 2025. [Online]. Available: <https://arxiv.org/pdf/2510.09859>
- [12] Hongyang (Bruce) Yang et al., "FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models," arXiv, 2024. [Online]. Available: <https://arxiv.org/pdf/2405.14767?>