Received: 22 June 2025 | Revised: 17 October 2025 | Accepted: 14 November 2025 | Published online: 1 December 2025

RESEARCH ARTICLE

FinTech and Sustainable Innovation
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewFSI52026541

BON VIEW PUBLISHING

Similarities Between Cloud Computing and Insurance: Feasibility Study of Secondary Cloud Infrastructure

Xiaobo Cai[1,*] **and Haifei Qin**[1]

[1] *School of Mathematics and Computer Science, Chuxiong Normal University, China*

Abstract: This study reveals the fundamental isomorphism between cloud computing resource allocation and insurance risk dispersion mechanisms through mathematical modeling. Under the independent and identically distributed (i.i.d.) assumption (demand/loss variables $X_i \sim Poissson(\lambda_i)$), both systems share the core operational logic: $VaR(p) < nVaR_i(p)$, demonstrating that centralized management of resource/risk pools significantly enhances efficiency. Key findings include 1. Shared pricing strategy: The maximum price that users are willing to accept, calculated as $PM_i = \lambda_i(e^k - 1)/k$, is always greater than the provider's minimum price λ_i. This implies that, under the assumptions of this paper, cloud computing transactions are always achievable. 2. Secondary Cloud Infrastructure: Inspired by excess-of-loss reinsurance, we design a cross-cloud resource protocol where optimal threshold Ym_i minimizes costs $C_i = Ym_i + F_i$. This research provides a theoretical foundation for applying actuarial models to cloud computing, particularly for elastic resource pricing and disaster recovery.

Keywords: cloud-insurance isomorphism, risk aggregation var, cloud resource pricing, secondary cloud infrastructure

1. Introduction

Cloud computing and insurance are two concepts in the fields of information technology and finance, respectively.

1.1. Cloud computing

Cloud computing is a distributed computing paradigm based on the Internet plus. It abstracts physical resources (servers, storage, and network devices) into logical resource pools through virtualization technology and relies on wide area networks to achieve on-demand allocation and elastic supply of resources (as defined by the NIST standard). Its core breakthrough lies in the paradigm reconstruction of "computing power as a service"—users can break free from local hardware constraints and subscribe to standardized IT resources (such as computing power, storage, and software) in a self-service format, significantly lowering the threshold for informatization. From the perspective of technological lineage, cloud computing inherits and deepens three major historical paradigms:

Time sharing [1]: IBM CP-40 was the first to achieve multiterminal sharing of resources on a single host, laying the foundation for the concept of "resource reuse."

Utility computing [2]: Proposed a pay-per-use model, analogous to the water and electricity supply model.

Grid computing [3]: Initial practice of distributed scheduling in cross-institutional resource collaboration architecture (such as the SETI@home project) [4].

The maturity of virtualization technology (VMware ESX, 2001) marked a key turning point. It achieved multi-tenant resource isolation through a hardware abstraction layer, addressing the security bottleneck of large-scale resource sharing [5]. In 2006, Amazon launched its first commercial public cloud service (EC2/S3) [6], marking the transition of cloud computing from theoretical concept to industrialization. Over the subsequent decade, open source technologies (OpenStack, Kubernetes) propelled the popularization of hybrid cloud architectures, while containerization (Docker, 2013) and serverless computing (Serverless, AWS Lambda) further refined the granularity of resource scheduling, supporting the rise of the cloud-native application ecosystem [7, 8].

1.2. Insurance

Within the scope of this article, the term "insurance" specifically denotes nonlife insurance. Nonlife insurance encompasses categories designed to cover risks unrelated to human life, such as property damage, liability compensation, and consequential financial losses. Its historical evolution is deeply rooted in humanity's enduring need for mechanisms providing economic compensation against the adverse impacts of natural disasters, accidents, and unforeseen events. The core developmental trajectory can be delineated across several distinct epochs [9, 10]:

1) Early prototype (from pre-commercial civilization to the 17th century)

[*] **Corresponding author:** Xiaobo Cai, School of Mathematics and Computer Science, Chuxiong Normal University, China. Email: qhfei@cxtc.edu.cn

© The Author(s) 2025. Published by BON VIEW PUBLISHING. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

Foundations of Marine Insurance: The earliest documented precursors to formal insurance contracts can be traced to ancient Babylonian practices. The Code of Hammurabi (circa 2000 BC) explicitly recorded ship mortgages (Bottomry), widely regarded as the earliest forms of contractual risk transfer. The genesis of the modern nonlife insurance system, however, is typically associated with the first known ship voyage insurance policy, formalized by merchants in Genoa, Italy, in 1347. This pioneering contract explicitly covered perils inherent to maritime trade, including storms, piracy, and shipwreck.

Inception of Fire Insurance: The catastrophic Great Fire of London in 1666, which devastated approximately 13,200 structures, served as a pivotal catalyst. This disaster spurred Nicholas Barbon to establish the Fire Office in 1680, recognized as the world's first dedicated fire insurance company. The Fire Office pioneered the fundamental modern business model centered on the principles of "risk sharing and loss allocation," pooling premiums from many policyholders to indemnify the few who suffered losses.

2) Industrialization drives systematization (18th–19th centuries)

Specialization of insurance types: The transformative forces of the Industrial Revolution created novel risks demanding specialized coverage. This period witnessed the emergence of dedicated insurance products such as boiler explosion insurance (1854) and engineering insurance (1870). A significant regulatory milestone occurred in 1880 with the UK's Employers' Liability Act, which mandated employers to carry liability insurance. This legislative intervention effectively addressed the critical societal need for the socialization of work-related injury compensation, transferring financial risk from individuals to the insurance mechanism.

Introduction of actuarial science: A cornerstone development in the professionalization of insurance occurred in 1756 when James Dodson formulated a nonlife insurance actuarial model. Dodson's work systematically applied the Law of Large Numbers to premium pricing, utilizing empirical data like fire occurrence statistics. This scientific approach represented a profound breakthrough, superseding the limitations and inherent unpredictability of earlier experience-based pricing methods and establishing insurance underwriting on a sound mathematical footing.

3) Globalization and the complexity of risks (20th century) [11].

Catastrophe insurance mechanism: The devastating 1906 San Francisco earthquake, resulting in staggering losses estimated at $400 million (equivalent to billions today), exposed the limitations of individual insurers in covering massive, concentrated losses. This event directly motivated the establishment and maturation of sophisticated reinsurance systems, exemplified by firms like Munich Re. Reinsurance achieved cross-regional and cross-company diversification of catastrophic risks, enhancing the overall financial stability of the insurance sector.

Popularization of automobile insurance: Henry Ford's revolutionary implementation of assembly-line mass production (1913) dramatically increased automobile accessibility. This surge in vehicle ownership directly propelled motor vehicle insurance into becoming the single largest nonlife insurance category globally, consistently accounting for over 35% of total premiums, reflecting the profound societal impact of this technology.

The rise of policy insurance: Recognizing instances of market failure in covering certain pervasive risks, governments intervened to create public–private solutions. A landmark example is the United States Federal Flood Insurance Act of 1938, which established a government-led catastrophe insurance program. This initiative aimed to provide essential coverage for flood risks, which private insurers were often reluctant or unable to underwrite affordably due to the high correlation and potential magnitude of losses.

1.3. Current status

Recent years have witnessed a burgeoning interest in the intersectional research between cloud computing and insurance. The predominant focus within this domain, however, has largely positioned cloud computing as a novel technological tool deployed to optimize the core business processes and service delivery systems inherent to the insurance industry. The primary objective has been leveraging cloud capabilities to enhance existing insurance functions.

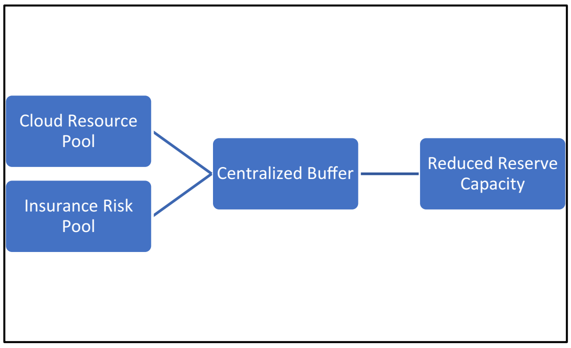
Empirical research demonstrates that real-time actuarial models harnessing the computational power and scalability of cloud platforms have successfully surpassed traditional pricing paradigms in terms of speed and complexity handling. For instance, Zhao et al. [12] developed an innovative cloud-edge collaborative architecture. This architecture effectively integrated dynamic, real-time vehicle telematics data directly into the motor insurance pricing process, yielding significant improvements in actuarial efficiency by enabling continuous risk assessment based on actual driving behavior. Similarly, the cloud-native catastrophe simulation platform (CATNet, 2020), co-constructed by Swiss Re and Amazon Web Services (AWS), unlocked unprecedented computational capabilities [13]. CATNet made the parallel computation of tens of thousands of distinct disaster scenarios feasible, a task computationally intractable with traditional on-premises infrastructure. This advancement led to a remarkable 40-fold increase in the efficiency of calculating claims reserves, which are critical for ensuring insurers' solvency after major disasters.

The domain of service innovation has also seen extensive applications of cloud computing, fostering entirely new insurance products and risk management approaches. Initiatives like the industrial metaverse insurance program launched by AIG in 2024, underpinned by Microsoft Azure's Digital Twins platform, illustrate the creation of novel insurance paradigms. This program pioneered coverage for virtual assets within industrial metaverses against potential damage or loss, addressing risks emerging from the digital transformation of industry [14].

Cloud technologies also achieved breakthroughs in the critical areas of risk control and regulatory compliance. A notable example is the medical insurance anti-fraud system implemented in 2021, built upon the high-throughput streaming data processing capabilities of the Alibaba Cloud Flink engine. This system achieved the real-time cross-verification of claim documents at the scale of hundreds of millions of transactions, elevating fraud detection accuracy to 92.3%. Simultaneously, advancements in secure cloud architectures are addressing stringent regulatory requirements. The deployment of an IBM Cloud Zero Trust Architecture by Generali in 2023 marked a milestone by successfully establishing a secure, closed-loop system for managing sensitive cross-border underwriting data flows. This system operated effectively under the rigorous data protection and privacy constraints mandated by the European Union's General Data Protection Regulation.

Notably, the ongoing integration of cloud platforms with other cutting-edge technologies is actively giving rise to new research focal points within the insurance sector. Federated learning techniques, which enable collaborative model training without sharing

Figure 1
Schematic diagram of dual-domain isomorphism



raw data, are facilitating the development of privacy-preserving data collaboration frameworks. Applications are emerging in areas like health insurance pricing models, as evidenced by pilot projects conducted by Ping An Insurance in 2023 [15]. Concurrently, the strategic (deployment/diffusion) of edge cloud infrastructure, bringing computational resources closer to data sources, is drastically reducing latency. This reduction is particularly impactful for usage-based insurance (UBI) models in motor insurance, exemplified by collaborations such as that between Huawei Cloud and PICC Property & Casualty in 2024, which achieved UBI data processing delays in the millisecond range, enabling truly real-time premium adjustments.

It is not difficult to find that previous cross-disciplinary research on cloud computing and insurance has mainly focused on optimizing insurance business and actuarial aspects using information tools such as cloud computing. The comparative study of the similarities between the two mentioned in this article has been initiated by using insurance models to solve problems in cloud computing. It is shown in Figure 1. This represents a distinct and complementary research vector: rather than applying cloud to insurance, this work leverages the mature mathematical frameworks and risk management strategies of insurance to innovate within cloud computing infrastructure and operations.

2. Mathematical Models of Structural Isomorphism

From the introduction in the previous section, cloud computing and insurance might initially appear to have no substantive connection. However, in this section, we will rigorously demonstrate that their fundamental mathematical models governing resource allocation and risk pooling are almost completely identical. This profound isomorphism stems from the core logic of aggregating individual, independent demands or losses into a centralized pool, exploiting statistical principles to achieve greater efficiency.

2.1. Model of cloud computing

This article assumes that the resources required by users, such as servers, storage, and network devices, can be arbitrarily divided. This divisibility assumption simplifies modeling and aligns well with the virtualized nature of cloud resources, where CPU cycles, memory, and storage can be allocated in fine-grained units. The critical modeling choice is to represent user demand as a stochastic process.

Let X_i be a random variable representing the resources required by the user i within a certain time frame. We assume that these resource demand variables across different users are independently and identically distributed (i.i.d.). This i.i.d. assumption implies that the resource consumption patterns of one user do not influence those of another, and that all users exhibit demand drawn from the same underlying statistical distribution, albeit potentially with different parameters (λ_i) reflecting their individual average demand levels. The choice of a discrete distribution is apt for modeling countable resource units (e.g., CPU hours, GB of storage consumed).

Assume $0 \leq p \leq 1$, then the p-quantile of X_i can be represented as

$$VaR_i(p) = Q_i(p) = \inf\{x \in Supp(X_i) | P(X_i < x) \geq p\} \tag{1}$$

This value-at-risk (VaR) metric quantifies the minimum resource capacity required to satisfy user i's demand with a probability of at least p. This means that if there is no cloud service, user i needs to purchase dedicated infrastructure capable of providing at least $VaR_i(p)$ units of service to satisfy their resource needs within a probability of p [16].

One extreme case is that if user i wants to fully satisfy their resource needs, they need to purchase infrastructure that can provide exactly $M_i = \max(Supp(X_i))$ units of service. This represents the maximum possible demand the user might ever generate. However, provisioning for this maximum is highly inefficient statistically.

The expectation (mean) of user i's resource demand is $E(X_i)$. The difference $VaR_i(p) - E(X_i)$ represents the "buffer" or "safety stock" of resources that, on average, do not contribute directly to utility for user i under normal demand conditions but are held idle to cover peak demand or variability. The magnitude of this buffer is directly related to the volatility of and the desired service level p. Higher volatility or a higher p necessitates a larger, more costly buffer.

If the aforementioned resources are provided by a cloud computing provider, the dynamics change fundamentally due to aggregation. The cloud provider serves a pool of users, enabling risk pooling.

Assume that there are n users with resource requirements. For user i, the demand during the certain time frame is a random variable X_i. Then, the total resources needed to be provided by the cloud computing provider during the same time period are a random variable $Y = \sum_{i=1}^{n} X_i$.

Assume $0 \leq p \leq 1$, then the p-quantile of Y can be represented as

$$VaR_Y(p) = Q_Y(p) = \inf\{y \in Supp(Y) | P(Y \leq y) \geq p\} \tag{2}$$

This defines the total capacity $VaR_Y(p)$ the cloud provider must provision to ensure that the aggregate demand of all n users is met with probability p.

The critical mathematical insight, underpinning the economic advantage of cloud computing, arises from the i.i.d. assumption. For the same probability level p, it consistently holds that:

$$VaR_Y(p) < \sum_{i=1}^{n} VaR_i(p) \tag{3}$$

This inequality $VaR_Y(p) < nVaR_i(p)$ (where $VaR_i(p)$ is the VaR for a representative user with average demand characteristics) demonstrates a fundamental efficiency gain [17]. The total buffer capacity required by a single unified cloud provider to serve all users at service level p is strictly less than the sum of the buffer capacities each user would need individually to achieve the same service level p on their own dedicated infrastructure. This efficiency gain implies that centralized management of resources through a cloud provider is significantly more efficient than when individual users providing resources independently to meet the same probability of demand fulfillment. The efficiency stems from the statistical phenomenon that the relative variability (e.g., coefficient of variation) of the aggregate demand Y decreases as N increases (Law of Large Numbers, Central Limit Theorem). Individual peaks and troughs in demand across users partially cancel each other out in the pool. While one user might experience high demand, another might be experiencing low demand. This diversification effect reduces the aggregate buffer required per unit of expected demand.

2.2. Model of insurance risk pooling

The mathematical structure governing insurance risk pools is remarkably similar, demonstrating the core isomorphism. Here, the "resource" being managed is capital reserved to cover losses.

Let X_i be a random variable representing the loss of user i within a certain time frame. These loss variables are also assumed to be independently and identically distributed (i.i.d.) across the pool of insureds. This assumes losses are independent events (e.g., one policyholder's house fire doesn't cause another's) and that the risk profiles are homogeneous or grouped into homogeneous risk classes.

Assume $0 \le p \le 1$, then the p-quantile of X_i can be represented as

$$VaR_i(p) = Q_i(p) = \inf\{x \in Supp(X_i) \,|\, P(X < x) \ge p\} \quad (4)$$

This means that without an insurance company, user i has a probability p of suffering a loss not exceeding $VaR_i(p)$. In other words, user i needs to personally set aside capital reserves of at least $VaR_i(p)$ to handle potential losses at probability p. This reserved capital acts as their personal safety net [18].

This means that without an insurance company, user i has a probability p of suffering a loss not exceeding $VaR_i(p)$. In other words, user i needs to prepare $VaR_i(p)$ to ensure that they can handle potential losses with probability p.

In an extreme case, if user i wants to be fully prepared for the worst-case scenario (i.e., cover any conceivable loss with $p = 1$), they need to prepare capital amounting to $M_i = \max(Supp(X_i))$. This represents the maximum possible loss they could incur.

The expected loss for user i is $E(X_i)$. The difference $VaR_i(p) - E(X_i)$ signifies the "risk premium" or "safety loading" in capital terms that, on average, remains unused by user i but is essential for them to withstand unexpected large losses. Holding this capital individually is costly, as it is tied up and cannot be used for other productive investments.

If an insurance company exists to pool these risks, the capital reserving requirement changes due to aggregation, analogous to the cloud model. The insurer manages a portfolio of risks.

Assume that there are n users at risk of loss. For user i, the loss during the certain time frame is a random variable X_i. Then, the risk of loss that the insurance company faces within a specified time period is a random variable $Y = \sum_{i=1}^{n} X_i$.

Assume $0 \le p \le 1$, then the p-quantile of Y can be represented as

$$VaR_Y(p) = Q_Y(p) = \inf\{y \in Supp(Y) \,|\, P(Y \le y) \ge p\} \quad (5)$$

This defines the total capital $VaR_Y(p)$ (or reserves) the insurance company needs to hold to ensure it can cover the aggregate los ses from all n policies with probability p, maintaining its solvency.

Under the i.i.d. assumption for the X_i, the same fundamental inequality holds for the same p:

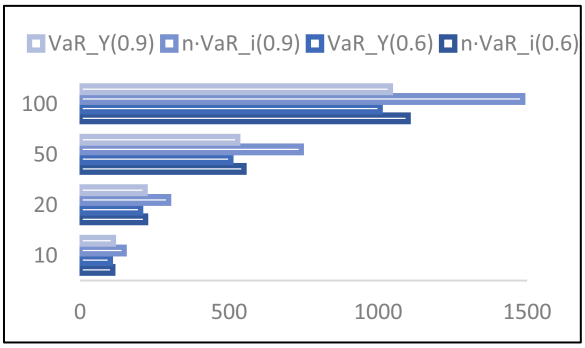
$$VaR_Y(p) < \sum_{i=1}^{n} VaR_i(p) \quad (6)$$

This demonstrates that an insurance company undertaking risk management collectively requires significantly less total capital reserve to achieve the same probability p of covering all claims than the sum of the reserves each individual user would need to cover their own potential losses at the same confidence level. This inequality $VaR_Y(p) < nVaR_i(p)$ highlights the superior efficiency of centralized insurance risk pools [19]. The collective buffer $(VaR_Y(p) - E(Y))$ is smaller per unit of expected loss than the sum of the individual buffers $(\sum VaR_i(p) - E(X_i))$ would be. This efficiency gain arises from the diversification of independent risks across the pool. Large losses are rare events. The probability that many insureds experience their maximum loss simultaneously is extremely low. The insurer only needs to hold enough capital to cover the aggregate loss corresponding to its desired solvency standard (p), which is significantly less than the sum of all potential maximum individual losses. This capital efficiency is a core economic justification for the existence of insurance.

Synthesis of the isomorphism based on the above analysis, it is evident that while cloud computing (managing computing resource demand) and insurance (managing financial loss risk) may seem unrelated at first glance, their underlying mathematical models governing the efficiency gains from centralized pooling are fundamentally aligned. Both systems leverage the statistical principle that aggregating independent, identically distributed random variables (X_i) into a central pool ($Y = \sum_{i=1}^{n} X_i$) reduces the relative variability of the aggregate. This reduction in relative variability translates directly into a reduced need for reserve capacity per unit of expected demand/loss when aiming for a specific service/solvency level p, quantified by the strict inequality $VaR_Y(p) < nVaR_i(p)$. The core operational logic of "risk aggregation reduces buffer requirements" is shared.

When every X_i follows a Poisson distribution with a parameter of 10, Figure 2 shows the difference between $nVaR_i(p)$ and $VaR_Y(p)$ under different user sizes and threshold p values. It can be seen that $VaR_Y(p)$ is significantly lower than $nVaR_i(p)$, and this advantage becomes more pronounced as the user size and threshold p increase.

Figure 2
Comparison chart of VaR buffer efficiency



Since the insurance industry is relatively more mature, having developed sophisticated models, pricing strategies, and mechanisms for managing pooled risks (especially via reinsurance), these established strategies can potentially be adapted and applied to enhance the design and operation of cloud computing systems. This paper primarily focuses on exploring the transferability of two key insurance concepts: pricing strategies based on risk and risk aversion, and the reinsurance mechanism, proposing its counterpart in cloud infrastructure.

3. Pricing Strategies: A Shared Economic Framework

The isomorphism extends beyond risk pooling into the economic principles governing transactions in both domains. Insurance pricing, particularly in nonlife insurance, primarily depends on the balance between the highest price the insured is willing to accept (reservation price) and the lowest price the insurance company is willing to offer (cost-based minimum price). This fundamental economic tension between demand (willingness to pay) and supply (cost of provision) is equally relevant to cloud computing service pricing. The mathematical model reveals a shared structure.

The highest price the insured is willing to accept (their reservation price) is mainly determined by the insured's expected risk and, crucially, their level of risk aversion. A risk-averse individual assigns a higher value to the certainty provided by insurance or cloud resources, increasing their willingness to pay. The lowest price the insurance company (or cloud provider) is willing to offer is primarily determined by their assessment of the expected risk (expected loss or expected resource demand) and their risk neutrality, focusing on covering expected costs.

This approach is also applicable to the pricing strategy for cloud computing services.

To formalize this shared pricing model, we make specific distributional assumptions suitable for countable events (resource requests, loss claims) occurring randomly over time. Assume X_i, representing the resource demand (cloud) or loss amount (insurance) for user i, follows a Poisson distribution with parameter λ_i. The Poisson distribution is characterized by $E(\lambda_i) = Var(\lambda_i) = \lambda_i$. This choice implies that demand/loss events are discrete, relatively rare (compared to the timeframe), and occur independently.

Central to modeling the demand side (user's willingness to pay) is the concept of utility, representing user preferences under uncertainty. Assume the utility function of all users exhibits constant absolute risk aversion (CARA), a common assumption for mathematical tractability and insights into risk behavior:

$$U(x) = 1 - e^{--kx} \tag{7}$$

Here $k > 0$ denotes the absolute risk aversion coefficient and reflects the user's risk aversion level. A larger k indicates a higher degree of risk aversion – the user is more sensitive to fluctuations in wealth or resource availability and derives greater value from certainty.

Considering the demand for resources (or the potential for loss) within the specified time range, the expected utility for user i *without* purchasing the service (insurance or cloud resources) is:

$$E\left(U_i^{no\ service}\right) = E(U(W_i - X_i)) \tag{8}$$

where W_i represents the total assets (or baseline utility level) of user i. Substituting the CARA utility function and using the properties of the Poisson distribution and the exponential function, the expected utility simplifies to:

$$E_i = E\left(U_i^{no\ service}\right) = 1 - \exp\left(\lambda_i e^k - kW_i - \lambda_i\right) \tag{9}$$

This expression quantifies the user's average welfare level facing the uncertain demand/loss λ_i without risk mitigation.

If user i is willing to pay the maximum amount PM_i (the Premium Maximum in insurance, or Price Maximum in cloud) to purchase the service (eliminating their direct exposure to X_i), then their utility becomes certain:

$$U(W_i - PM_i) = 1 - e^{-k(W_i - PM_i)} \tag{10}$$

The user's reservation price PM_i is determined by the point where they are indifferent between bearing the risk themselves and paying PM_i to transfer it. This is found by equating the certain utility of paying PM_i to the expected utility of bearing the risk:

$$U(W_i - PM_i) = E(U(W_i - X_i)) \tag{11}$$

Substituting the expressions derived above:

$$1 - e^{-k(W_i - PM_i)} = 1 - \exp\left(\lambda_i e^k - kW_i - \lambda_i\right) \tag{12}$$

$$e^{-k(W_i - PM_i)} = \exp\left(\lambda_i e^k - kW_i - \lambda_i\right) \tag{13}$$

$$-k(W_i - PM_i) = \lambda_i e^k - kW_i - \lambda_i \tag{14}$$

Solving for PM_i:

$$kPM_i = \lambda_i(e^k - 1) \tag{15}$$

$$PM_i = \frac{\lambda_i(e^k - 1)}{k} \tag{16}$$

Thus, we derive the key result for the user's maximum willingness to pay: $PM_i = \lambda_i(e^k - 1)/k$ [20].

On the supply side, cloud service providers or insurance companies can be reasonably modeled as risk-neutral entities. This assumption implies they focus solely on expected costs and profits, disregarding risk aversion themselves due to their ability to pool large numbers of independent risks, making their aggregate cash flows highly predictable. Consequently, the minimum price they can accept for covering user i's risk is simply the actuarially fair price: the expectation of user i.

Since user i's demand follows a Poisson distribution with parameter λ_i, $E(X_i) = \lambda_i$.

It is evident that a transaction can only occur when the highest price a user is willing to accept is higher than the lowest price a supplier is willing to accept.

Since $(e^k - 1)/k > 1$ always holds when $k > 0$, it follows that $PM_i > E(X_i)$ always holds as well. Under the aforementioned conditions, the transaction is always possible, and the actual transaction price is a value between $E(X_i)$ and PM_i [21].

4. Secondary Cloud Infrastructure

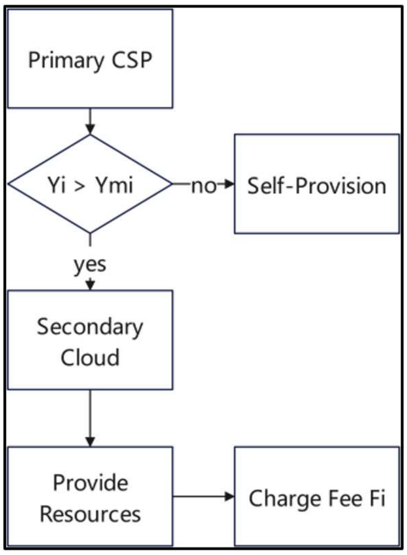
Reinsurance serves as a well-established financial mechanism within the insurance industry, enabling primary insurers to transfer portions of their underwritten risks to other specialized institutions. This process effectively functions as “insurance for insurance companies.” Its core operational objectives encompass helping direct insurers disperse exposure to catastrophic risks (such as earthquakes or hurricanes), balance financial fluctuations arising from claim volatility, and enhance overall underwriting capacity by freeing up capital reserves. Within the diverse landscape of reinsurance models, two primary categories exist: proportional reinsurance (which divides both premiums and claims according to a fixed ratio, exemplified by quota share arrangements) and nonproportional reinsurance (which triggers coverage based on specific claim amount thresholds, such as excess-of-loss reinsurance, where the reinsurer assumes liability for claims exceeding a predetermined dedu ctible level). This article primarily focuses on and draws inspiration from the structure of excess-of-loss (XoL) reinsurance due to its direct applicability to managing peak resource demands in computing.

By translating this mechanism, we propose the concept of a Secondary Cloud Infrastructure. Analogous to reinsurers in the insurance market, specialized secondary cloud companies can emerge to provide capacity relief to primary cloud service providers (CSPs) facing unpredictable demand surges. This model establishes a formalized framework for risk (demand) sharing among cloud infrastructure operators. It’s shown in Figure 3.

4.1. Theoretical model

Formal model construction: Consider a distributed ecosystem comprising m independent cloud computing markets. Within every market *i*, there exists one primary cloud service provider catering to *n* distinct users. For user *j* in market *i*, the demand for computing resources (e.g., CPU hours, storage GB-hours) during a specified time interval is modeled as a random variable X_{ij}.

Figure 3
Flowchart of secondary cloud



The total resource demand confronting the primary cloud provider in market *i* during that interval is consequently the aggregate random variable:

$$Y_i = \sum_{j=1}^{n} X_{ij} \tag{17}$$

Assuming that individual user demands X_{ij} follow independent Poisson distributions with parameters λ_{ij} (i.e., $X_{ij} \sim Poisson(\lambda_{ij})$), the aggregate demand Y_i for market i consequently follows a Poisson distribution. Its parameter Λ_i is the sum of the individual user demand rates:

$$\Lambda_i = \sum_{j=1}^{n} \lambda_{ij},\ hence\ Y_i \sim Poisson(\Lambda_i) \tag{18}$$

The value-at-risk (VaR) for this aggregate demand, representing the minimum resource capacity required to meet demand with probability p, is defined as

$$VaRY_i(p) = \inf\{y \in Supp(Y_i) | P(Y_i \leq y) \geq p\} \tag{19}$$

Critically, $VaRY_i(p)$ is a monotonically increasing function of p. Furthermore, due to the properties of the Poisson distribution, as p approaches 1 from below (i.e., $p \rightarrow 1^-$), $VaRY_i(p)$ tends toward infinity. This asymptotic behavior highlights a fundamental operational tension faced by primary CSPs operating in isolation: attempting to guarantee user demands with very high probabilities (e.g., 99.99% uptime/SLA) necessitates provisioning substantial excess capacity ($VaRY_i(p)$), much of which remains idle during normal operational periods, leading to significant capital inefficiency and underutilization costs. Conversely, reducing the provisioned resources to improve utilization rates inevitably increases the risk of failing to fully satisfy user demands during unexpected surges, potentially violating SLAs and damaging reputation.

The secondary cloud intervention: To resolve this dilemma, a secondary cloud company can establish contractual agreements with the m primary cloud providers. Under these agreements, analogous to XoL reinsurance treaties, each primary provider i pays a predetermined fee F_i to the secondary cloud company. In return, the secondary company commits to supplying additional resources whenever the primary provider i’s actual demand Y_i exceeds a pre-negotiated threshold Ym_i. This threshold Ym_i represents the “retention level” or “deductible” in insurance terms – the level of demand the primary provider agrees to handle using its own infrastructure [22].

Cost structure and optimization: The fee F_i charged by the secondary cloud company is actuarially fair, calculated as the expected value of the resources it must provide, conditional on the primary provider’s demand exceeding the threshold:

$$F_i = E(\max\{Y_i - Ym_i, 0\}) \tag{20}$$

Therefore, the total operational cost C_i for the primary cloud provider i comprises two components: the cost associated with maintaining its own infrastructure up to the retention level Ym_i, and the reinsurance-like fee paid to the secondary cloud [23]:

$$C_i = Ym_i + F_i \tag{21}$$

The primary provider’s strategic decision revolves around selecting the optimal retention threshold Ym_i^*. This optimization aims to minimize the total cost C_i, while ensuring sufficient capacity to meet its service level agreement (SLA) – specifically, guaranteeing that demand is fully met with probability p. Formally, the optimal threshold Ym_i^* is the solution to:

$$opt(Ym_i|p) = \{Ym_i|C_i = VaRY_i(p)\} \tag{22}$$

Solving this equation identifies the retention level Ym_i^*, where the total cost C_i (own infrastructure + secondary cloud fee) equals the capital cost of self-provisioning the full $VaRY_i(p)$ needed to meet the SLA independently. The optimal Ym_i^* strikes a balance: it is high enough to minimize the frequency of costly secondary resource calls (thus reducing F_i), but low enough to avoid excessive investment in rarely utilized peak capacity. This equilibrium point minimizes the primary provider's total expenditure while maintaining the required service reliability. The secondary cloud company, by pooling the excess demands from multiple independent primary providers, achieves its own efficiency through diversification. The likelihood of simultaneous, extreme demand surges across all primary providers is low, allowing the secondary company to maintain a smaller, more efficiently utilized resource pool than the sum of the peak capacities each primary provider would need alone, while still meeting its contractual obligations reliably. This structure effectively creates a risk-sharing consortium for cloud infrastructure, enhancing system-wide resilience and cost-effectiveness, particularly for handling bursty or unpredictable workloads.

4.2. Monte Carlo simulation

To better demonstrate the results of this article, we conducted a Monte Carlo simulation.

The simulation used the following parameters: 100 mutually independent cloud computing markets, each with a unique cloud computing provider and 1000 mutually independent users. The demand for cloud computing resources of every user in a specified time follows a mutually independent Poisson distribution, with parameters of 100 and 120, respectively.

We executed 100 simulation trials.

Tables 1 and 2 show some statistical results of the Monte Carlo simulation.

In Table 1, $DS(p)$ represents the amount of resources that the secondary cloud provider needs to provide when Ym_i is agreed as opt (p); $NC(p)$ indicates the number of times the actual demand exceeds $VaRY_i(p)$ in Monte Carlo simulation.

From Table 1, it can be found that if $\lambda = 100$, when the cloud computing provider prepares $VaRY_i(0.75)$ resources, on average, more than 20 out of 100 cloud computing providers cannot meet the customer's demand. That is to say, when $Ym_i = opt(Ym_i|0.75)$, on average, more than 20 out of 100 cloud computing providers need secondary cloud services. Similar results are also found for other parameters. This indicates that secondary cloud is effective.

Future simulations could incorporate privacy-preserving techniques (e.g., in-memory anonymization [24]) to handle sensitive user demand data.

4.3. Sustainability implications

In Table 2, URS (p) indicates the resource idle rate when Ym_i is set to opt (p); UR (p) indicates the resource idle rate when there is no secondary cloud agreement and the supplier's resources are $VaRY_i(p)$.

From Table 2, it can also be seen that for the simulation results of every parameter group, the URS (p) is lower than the UR (p).

Moreover, since the idle rate of resources is positively correlated with carbon emissions (although not linearly correlated [25]), reducing the idle rate of resources can help reduce carbon emissions.

From this perspective, secondary cloud is conducive to sustainable development.

Moreover, the lower the p, the greater the difference between URS (p) and UR (p). This indicates that reducing Ym_i which means

Table 1
Simulation results 1

	$\lambda = 100$		$\lambda = 120$	
	Aver	Max	Aver	Max
Y_i Aver	100,004	100,084	119,996	120,070
Y_i Max	100,784	101,172	120,899	121,595
DS(0.75)	6877.6	10,159	6584.1	11,072
DS(0.95)	692.12	2060	809.61	2482
NC(0.75)	25.72	35	23.79	35
NC(0.95)	5.3	10	4.69	9

Table 2
Simulation results 2

	$\lambda = 100$		$\lambda = 120$	
	Aver	Max	Aver	Max
URS(0.75)	0.21%	0.26%	0.21%	0.27%
URS(0.95)	0.51%	0.58%	0.48%	0.55%
UR(0.75)	0.26%	0.32%	0.25%	0.31%
UR(0.95)	0.52%	0.59%	0.48%	0.55%

transferring more resource demand to secondary clouds, can help reduce carbon emissions.

Secondary cloud protocols can also enhance transaction security in fintech scenarios, such as detecting fraudulent resource overuse patterns using hybrid ML models [26].

5. Conclusion

This study rigorously establishes the mathematical isomorphism between cloud computing resource scheduling and insurance risk transfer. Centralized resource/risk pools ($Y = \sum X_i$) under i.i.d. Poisson-distributed demands/losses ($X_i \sim Poisson(\lambda_i)$) consistently exhibit reduced buffer requirements due to risk aggregation: $VaR(p) < nVaR_i(p)$. This foundational similarity enables cross-domain methodology transfer, evidenced by the successful adaptation of insurance pricing frameworks to cloud services. When the user is risk-averse, cloud computing transactions are always feasible. Furthermore, the proposed secondary cloud infrastructure, mirroring excess-of-loss reinsurance, optimizes primary providers' costs ($C_i = Ym_i + F_i$) through threshold-based resource reallocation. These insights offer practical pathways for enhancing disaster recovery pools, burst-traffic management, and sustainable computing via risk-pooled resource optimization. Future work should explore non-Poisson distributions and blockchain-based automation of secondary cloud protocols [27, 28].

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

This work was supported by the Applied Basic Research Foundation of Yunnan Province [Grant Number: 202101BA070001-094], "Intelligent Analysis and Processing of Smart Tourism Big Data."

Author Contribution Statement

Xiaobo Cai: Conceptualization, methodology, validation, formal analysis, investigation, data curation, writing – original draft, writing – review & editing, visualization, supervision, project administration. **Haifei Qin:** Software, validation, investigation, resources, data curation, supervision, funding acquisition.

References

[1] Deutsch, L. P., Durham, L., & Lampson, B. W. (1968). *Time-sharing system reference manual (No. R-21). Advanced Research Projects Agency, Office of the Secretary of Defense*, Retrieved from: https://www.researchgate.net/publication/242636478_Time-sharing_system_reference_manual

[2] Parkhill, D. F. (1966). *The challenge of the computer utility*. Addison-Wesley.

[3] Forster, I., & Kesselman, C. (1999). *The GRID: Blueprint for a new computing infrastructure*. Morgan Kaufmann.

[4] Anderson, D. P., Cobb, J., Korpela, E., Lebofsky, M., & Werthimer, D. (2002). SETI@home: An experiment in public-resource computing. *Communications of the ACM*, *45*(11), 56–61. https://doi.org/10.1145/581571.581573

[5] Sugerman, J., Venkitachalam, G., & Lim, B. H. (2001). Virtualizing I/O devices on VMware workstation's hosted virtual machine monitor. In *Proceedings of the General Track: 2001 USENIX Annual Technical Conference*, 1–14.

[6] Amazon Elastic Compute Cloud. (2006). *In Wikipedia*. Retrieved from: https://en.m.wikipedia.org/wiki/Amazon_Elastic_Compute_Cloud

[7] Ahmed, M. I. (2024). Collaborative development in the cloud native. In M. I. Ahmed (Ed.), *Cloud-native DevOps: Building scalable and reliable applications* (pp. 265–306). Apress. https://doi.org/10.1007/979-8-8688-0407-6_7

[8] Gautam, B. P., Batajoo, A., & Shiratori, N. (2024). A proposal of JYAGUCHI computing platform to realize ClouEdge (Cloud-Edge) and serverless architecture. *Journal of Information Science and Engineering*, *40*(1), 89–105. https://doi.org/10.6688/JISE.202401_40(1).0006

[9] Hanson, N. (2002). *The great fire of London: In that apocalyptic year, 1666*. Wiley.

[10] Lemaire, J. (1987). Automobile insurance: Actuarial models. *Journal of Risk & Insurance*, *54*(3), 633–634. https://doi.org/10.2307/253374

[11] Ghani, M. A. N. U., She, K., Rauf, M. A., Khan, S., Alajmi, M., Ghadi, Y. Y., & Alkahtani, H. K. (2024). Toward robust and privacy-enhanced facial recognition: A decentralized blockchain-based approach with GANs and deep learning. *Mathematical Biosciences & Engineering*, *21*(3), 4165–4186. https://doi.org/10.3934/mbe.2024184

[12] Cao, Z., Zhao, Q., & Wang, J. H. Y. (2021). Edge-cloud collaborative architecture based multi-time scales rolling optimization of regional integrated electrical and natural gas energy system considering wind power uncertainty. *IET Generation, Transmission & Distribution*, *15*(19), 2684–2709. https://doi.org/10.1049/gtd2.12208

[13] Li, Y., Wang, H., Sun, J., Lv, H., Zheng, W., & Feng, G. (2024). Two-timescale joint service caching and resource allocation for task offloading with edge–cloud cooperation. *Computer Networks*, *254*, 110771. https://doi.org/10.1016/j.comnet.2024.110771

[14] Biswas, D., Jahan, S., Saha, S., & Samsuddoha, M. (2024). A succinct state-of-the-art survey on green cloud computing: Challenges, strategies, and future directions. *Sustainable Computing: Informatics and Systems*, *44*, 101036. https://doi.org/10.1016/j.suscom.2024.101036

[15] Liu, F., Zheng, Z., Shi, Y., Tong, Y., & Zhang, Y. (2024). A survey on federated learning: A perspective from multi-party computation. *Frontiers of Computer Science*, *18*(1), 181336. https://doi.org/10.1007/s11704-023-3282-7

[16] Dietz, S. (2021). Handbook of financial risk management. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(1), 402–403. https://doi.org/10.1111/rssa.12641

[17] Shaar, R., Kato, H., & Hikmat, R. F. (2025). Predicting value at risk in investment portfolio using Monte Carlo simulation: The case of the Syrian Internasional Islamic Bank. *El-Jizya:*

Jurnal Ekonomi Islam, *13*(1), 115–134. https://doi.org/10.24090/ej.v13i1.13803

[18] Hull, J. C. (2018). *Risk management and financial institutions* (5th ed.). Wiley.

[19] Nair, S. K., Porwal, S., Dimitrakos, T., Ferrer, A. J., Tordsson, J., & Sharif, T. (2010). Towards secure cloud bursting, brokerage and aggregation. In *2010 Eighth IEEE European Conference on Web Services*, 189–196. https://doi.org/10.1109/ECOWS.2010.33

[20] Hoseason, J. (2003). Pricing: Making profitable decisions. *Journal of Revenue and Pricing Management*, *2*(2), 175–177. https://doi.org/10.1057/palgrave.rpm.5170064

[21] Sun, X., Xu, R., Wu, L., & Guan, Z. (2021). A differentially private distributed data mining scheme with high efficiency for edge computing. *Journal of Cloud Computing*, *10*(1), 7. https://doi.org/10.1186/s13677-020-00225-3

[22] Aboagye, E., Asimit, V., Fung, T. C., Peng, L., & Wang, Q. (2025). A revisit of the optimal excess-of-loss contract. *European Journal of Operational Research*, *322*(1), 341–354. https://doi.org/10.1016/j.ejor.2024.11.027

[23] Tamma, C. R., Shamsheer, S., Venigalla, R., Akash, C., Begum, S. S. A., & Medishetti, S. K. (2024). Cost aware scheduling in hybrid Cloud-Fog environment using earthworm optimization algorithm. In *2024 International Conference on Advancement in Renewable Energy and Intelligent Systems*, 1–9. https://doi.org/10.1109/AREIS62559.2024.10893670

[24] Shamsinejad, E., Banirostam, H., BaniRostam, T., Pedram, M. M., & Rahmani, A. M. (2025). Providing and evaluating a model for big data anonymization streams by using in-memory processing. *Knowledge and Information Systems*, *67*(9), 7791–7824. https://doi.org/10.1007/s10115-025-02417-2

[25] R, D., J, U. U., Sharma, T., Singh, M. P., R, K., Selvan, S., & Krah, D. (2022). Energy-efficient resource allocation and migration in private cloud data centre. *Wireless Communications and Mobile Computing*, *2022*(1), 3174716. https://doi.org/10.1155/2022/3174716

[26] Vasant, M., Ganesan, S., & Kumar, G. (2025). Enhancing e-commerce security: A hybrid machine learning approach to fraud detection. *FinTech and Sustainable Innovation*, *1*, A7. https://doi.org/10.47852/bonviewFSI52024882

[27] Gu, K., Wang, Y., Qiu, J., Li, X., & Zhang, J. (2024). Blockchain-based data deduplication and distributed audit for shared data in cloud-fog computing-based VANETs. *IEEE Transactions on Network and Service Management*, *21*(5), 5548–5565. https://doi.org/10.1109/TNSM.2024.3424878

[28] Yıldırım Külekci, B., Korn, R., & Selcuk-Kestel, A. S. (2024). Ruin probability for heavy-tailed and dependent losses under reinsurance strategies. *Mathematics and Computers in Simulation*, *226*, 118–138. https://doi.org/10.1016/j.matcom.2024.06.018

How to Cite: Cai, X., & Qin, H. (2025). Similarities Between Cloud Computing and Insurance: Feasibility Study of Secondary Cloud Infrastructure. *FinTech and Sustainable Innovation*. https://doi.org/10.47852/bonviewFSI52026541