**Journal of Information Systems Engineering and Management**
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# AI-Optimized Spine-Leaf Fabrics: NVIDIA Quantum-2 vs. Cisco Nexus

Ashutosh Chandra Jha
Network Security Engineer, NewYork, USA
ashutoshjhany@gmail.com

**ARTICLE INFO**

Received:08 Sept 2025
Revised:15 Oct 2025
Accepted:25 Oct 2025

**ABSTRACT**

This study describes how AI-optimized spine-leaf network fabrics, NVIDIA Quantum-2 (InfiniBand) and Cisco Nexus (Ethernet), enable high-performance connectivity to support next-generation artificial intelligence and extensive digital services such as car insurance business operations. Both appliances offer colossal bandwidth, ultra-low latency, and real-time telemetry in support of workloads on GPUs. The study combines quantitative data—such as throughput, port density, and latency benchmarks—with qualitative results based on case studies of implementations, discussions with industry experts, and research studies. It considers the entire operational lifecycle, from initial capacity planning and design validation through deployment, monitoring, and long-term scaling, and it identifies typical inefficiencies such as redundant data entry, lag in identifying anomalies, and splintered governance. Key enablement include robotic process automation (RPA), analytics powered by AI/ML, cloud-native micro services, and ultra-real-time processing of data with strong master data management. In-depth case studies of one global auto insurer outline tangible benefits: quicker model training to support AI, faster settling of claims, lower operational costs, and increased customer satisfaction. The paper concludes by laying out a roadmap and practical recommendations for phased deployment, AI-driven governance, and worker training. Results prove that Quantum-2 and Nexus fabrics aren't hardware upgrade options, but strategic platforms to transform operations, increase compliance, and position businesses for future innovation.

**Keywords:** AI-optimized networking, NVIDIA Quantum-2, Cisco Nexus, spine-leaf architecture, real-time data processing

## Introduction

Massive-scale language models, autonomous systems, and enhanced medical imaging are some of the applications of artificial intelligence (AI) used today to process and move large volumes of data in real-time. These applications are developed into clusters made of custom accelerators and thousands of graphics processing units (GPUs) in parallel. The effective operation of such clusters requires a network fabric in a data center with the capacity to support the high bandwidth required, offer low latency at all times, and allow traffic bursts in collectives without packet loss. Any vulnerability within the network may extend the training process, increase operational expenses, and reduce the scale of AI projects. A more recent network architecture model to satisfy such needs is the spine-leaf architecture, which is a two-layer network architecture with each leaf switch connecting to each spine switch. This non-blocking, predictable server-communication architecture is highly suitable for serving the vast majority of AI workloads. This architecture has developed two new products to build or upgrade AI-optimized fabrics, which include: NVIDIA Quantum-2, built on InfiniBand, and Cisco Nexus, built on high-speed Ethernet.

NVIDIA Quantum-2 is an InfiniBand next-generation product, designed and built for use in ultra-scale AI and high-performance computing systems. It has 400 gigabit per second (Gb/s) NDR

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ports, and it has state-of-the-art features in terms of in-network computation in the Scalable Hierarchical Aggregation and Reduction Protocol (SHARP). They support collective incentives, such as all-reduce, to be implemented in-network, minimizing data transfer and significantly shortening training durations. In-depth management and visibility are also offered in Quantum-2 with NVIDIA Unified Fabric Manager (UFM), which introduces real-time telemetry, congestion analysis, and predictive maintenance. In contrast, Cisco Nexus switches introduce the Ethernet advantages into dense AI information centers. The Nexus family is scalable (400 GbE to 800 GbE) and integrates the familiarity of Ethernet with the massive interoperability with features such as RDMA over Converged Ethernet version 2 (RoCEv2). Features such as Network capabilities, including Priority Flow Control (PFC) and Explicit Congestion Notification (ECN), can provide lossless performance for AI workloads, with compatibility across a wide range of enterprise and cloud infrastructures. The Nexus Dashboard offered by Cisco also enhances the deployment speed of large-scale operations by automation (built-in), fabric templates, and analytics to streamline performance and prompt fault detection. Both models aim for high throughput, low latency, and simplicity of operations without being overly complicated to design, but differ in terms of design philosophies. NVIDIA Quantum-2 is target-oriented at the peak performance of tightly coupled groups of GPUs where collective communication patterns are predominant. Conversely, Cisco Nexus is more about AI-ready networking than traditional Ethernet protocols; this aligns perfectly with situations where a single technology is needed to support storage, ordinary computing, and AI training.

This paper is organized into various chapters. It begins with a literature review summarizing current research on automation and high-performance networking. The following section maps current AI data-center workflows, highlighting inefficiencies and compliance challenges. This is followed by an exploration of key technologies such as robotic process automation, AI/ML, cloud micro services, and real-time data processing that drive fabric automation. A methodology section outlines the research design, data sources, and analytical framework. The paper then explains how to design a streamlined spine-leaf fabric, presents the business benefits of automation, and provides a detailed discussion of findings and strategic implications. A real-world case study illustrates implementation results and lessons learned, while an implementation roadmap offers a phased strategy with key performance indicators. The recommendations section delivers actionable guidance for technology leaders, and the study concludes with a future outlook, emphasizing the long-term roles of AI, machine learning, and predictive analytics in next-generation data-center networking.

## 2. Literature Review

### *2.1 Recent Research on Automation in Insurance Businesses*

Research in automation for insurance has increased steadily as there are challenges in processing large numbers of policies, claims, and customer records [22]. Outdated operations, marked by paper-based data entry and reliance on fragmented legacy systems, cannot support today's demands for rapid processing, accuracy, and economies of cost. Automation research shows it decreases tedious labor, improves compliance, and shortens claims cycles. Industry studies also confirm that deploying end-to-end automation of business workflows leads to significant business process efficiency and customer satisfaction improvement. This growing body of research makes automation a foundation for interweaving emerging technologies and for reengineering the business model of insurance to keep up with shifting expectations of the customer.

### *2.2 Key Frameworks and Technologies such as RPA, AI/ML, and Cloud Micro services*

Robotic Process Automation (RPA) invariably figures as an initial prerequisite to automation in writing. RPA leads automation at an operational level through automation of repeat work—such as

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

policy issuance, billing reconciliation, and claims intake—while ensuring smooth integration into legacy systems. RPA, through automation of repeat and rules-based business functions, minimizes errors due to human intervention and directs a skilled workforce on core competency functions, directly contributing to faster processing of claims and correct policy management. As digital initiatives have progressed, Artificial Intelligence and Machine Learning (AI/ML) have become central to intelligent automation [32]. Machine learning models identify questionable claim patterns to identify fraud, improve underwriting through advanced risk analysis, and customize offers based on customer behavior. Natural language and conversational AI also automate customer interactions to the point that chatbots and virtual assistants can walk policyholders through submitting claims or answer in real time. These intelligent systems not only perfect customer experience but also build predictive models through which insurers can actively manage their risks and discover irregularities before they occur.

Cloud and microservices architecture underpins these technologies by providing an agile, modular framework through which insurers can scale fast, take new features to market with minimal downtime, and interoperate seamlessly with third-party services. Cloud-native systems also support event-driven operations and enterprise-wide, real-time processing across enterprise resource planning (ERP) and customer relationship management (CRM) systems. Enterprise-wide, real-time processing studies in ERP systems illustrate evident benefits, including instantaneous updating, faster decisions, and increased consistency across interrelated applications [3]. Implementing enterprise-wide, real-time processing in the insurance business ensures policy updates, claims processing, and risk assessments across the enterprise occur instantaneously and reduce delays and data mismatches. These frameworks—task automation by RPA, intelligent decisions by AI/ML, and cloud processing of data in real time and cloud microservices—constitute the technical base for today's insurance automation. These enable correct and compliant data and scalable growth as well as rapid service delivery. Bringing together process automation, predictive analytics, and real-time data flow, insurers can meet growing expectations for speed, transparency, and personalized services, and also reinforce operational efficiency and long-term competitiveness.

As illustrated in the figure below, RPA, AI/ML, and cloud microservices collectively form the backbone of insurance automation, enabling seamless task automation, intelligent decision-making, and real-time data processing to enhance speed, scalability, compliance, and customer-focused service delivery.


**Figure 1: How Do Banks Benefit From Robotic Process Automation (RPA)**

### *2.3 Processing of Real-Time Data and Master Data Management Findings*

On-the-fly data processing and master data management (MDM) are the key pillars of successful automation in the modern insurance enterprise. With real-time pipelines, one can capture and analyze data instantly to support the estimation of dynamic risks, detect fraud immediately, and process claims more efficiently. This quicker processing also enhances predictive analytics, and pricing and underwriting guidelines can be updated in a way that is realistic in nearly real-time as market indicators, customer engagements, and telematics change. MDM provides a single source of truth in

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

policies, claims, and systems facing the customer. A set of correct records reduces the difference and allows the merging of external information resources and the maintenance of rigorous rules on privacy and financial disclosure. The evidence provided by other industries supports these advantages. The examples of both can be found in the creation of synthetic medical data, which is based on the principles of using generative AI and edge devices and deep learning to work with medical data in near-real time, which guarantees the integrity of data and promotes speedy and reliable responses [36; 37]. By integrating real-time streams of information and robust governance, carriers will be able to implement sophisticated analytics and AI-driven software, such as predictive fraud detection and policy customization, to make informed and timely decisions, thereby remaining competitive in the long run.

### *2.4 Gaps Identified, Including Insufficient Car-Insurance-Specific Research and Standardized Metrics*

Despite significant developments, points of essential lacuna in this research abound. As relatively few studies focus directly on the car insurance business line, whose features involve vehicle telematics integration, accident detection automation, and high-frequency processing of claims, these include data streams and real-time decision requirements unlike those in other business lines and thus involve results transferable to those business lines only in part.

Another essential gap is the absence of a universally agreed-upon set of standardized measures by which to define automation success. Assessments overwhelmingly emphasize high-level measures such as cost savings, turnaround time, or simple process efficiency. While valuable, these all too often fail to account for subtler factors such as customer satisfaction, avoidance of regulatory risk, and long-term operational longevity. This absence of standardized measures makes it problematic to compare outcomes across organizations or to quantify the incremental value of complex tools such as AI-driven fraud detection or next-generation network architecture.

The challenge of automation results measurement directly addresses central data management. Research on artificial intelligence and machine-learning effects on master data management sheds light on the fact that accurate, correct, and aptly governed data lie at the foundation of good and trusted measurement and rightful performance analysis [4]. Deteriorating or fractured MDM practices lead to variable indicators of significance—like policy correctness, claims integrity, and reporting compliance—undermining research comparability and operational efficiency simultaneously. The merger of AI and machine learning into MDM strengthens data correctness and enables more accurate, complete measures to quantify automation outcomes. Closing these gaps will be critical in driving insurers to adopt strategies involving next-generation automation technologies and scalable, high-performance network infrastructures. Establishing standardized, data-centric indicators aligned to strong foundations in MDM will allow organizations to have standard baselines to gauge progress, share best practices, and make AI and automation investments based on evidence [31].

## 3. Mapping the Current Workflow

### *3.1 Policy Lifecycle*

Creating and managing an AI-optimized spine-leaf infrastructure, which is manufactured using NVIDIA Quantum-2 or Cisco Nexus, adapts an official, policy-type life cycle model. The life cycle begins with the quoting and capacity planning phase, within which cluster size, number of GPUs or compute nodes, and aggregate bandwidth would be projected to serve dense AI workloads. Here, major performance indicators, such as the tolerable latency budget and oversubscription ratio, are defined, and the desired scale, typically comprising multiple hundreds of NDR 400 Gb/s links, is specified in expertly detailed deployments. The second stage, design validation, is a technical underwriting. Architectures that have been topologically designed are put to test against operational constraints and

**Journal of Information Systems Engineering and Management**
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

requirements in applications. A non-blocking and oversubscribed spine-leaf architecture is compared, with optics and cabling, including OSFP NDR modules in Quantum-2 or QSFP-DD/OSFP in Cisco Nexus, as well as power and cooling requirements, reviewed to determine whether the design meets workload performance and sustainability goals.

After being approved, it is enforced. Hardware is purchased and installed, cabling is done to racks, optical and transceivers are tested, and network operating systems, such as NVIDIA Onyx or Cisco NX-OS, are installed. The automation devices construct setups and pass the setups to dozens of switches in parallel, checking the build. Periodically, renewals may be made by upgrading to a higher level, such as updating firmware and adding capacity to handle increasing quantities of information or newer technologies, like 800 Gb/s. The claims phase includes incident response and troubleshooting. The failed links, unexpected congestion, or hardware failures must be taken care of in real time in order to protect AI training timetables. Real-time telemetry and built-in anomaly-detection capabilities, for instance, NVIDIA Unified Fabric Manager and Cisco Nexus Dashboard, act to aid fault isolation and remediation automation. Research in edge AI and federated anomaly detection underpins distributed, real-time analytics and their contribution to failure avoidance prior to service continuity impacts [7].

Ongoing operations and user services provide long-term support, including regular health checks, capacity planning, and guaranteed on-demand access to compute and network capacities. Lessons from healthcare and retail infrastructure support operational efficiency and quality of services improvement through digital engagement focused on users, like just-in-time alerts and 24/7 self-service access points [5]. In addition, it would be necessary to have a zero-trust data architecture as clusters grow in size and interoperate with outside partners, in line with best practices in secure data interchange and governance in highly sensitive domains like medical research [8]. Building NVIDIA Quantum-2 and Cisco Nexus fabrics based on this lifecycle brings automation, real-time outlier detection, and strong data security together in every stage, delivering an agile, secure, and high-performance network as AI demands grow.

### *3.2 Key Inefficiencies*

Despite this formal process, several inefficiencies undermine fabric performance. Redundant data entry occurs when design and operational information exist in silos, such that topology plans, optical inventories, and configuration templates have to be entered multiple times. This hampers deployment and makes upgrading troublesome. Another weakness is manual fault isolation: engineers must access multiple switches and controllers to identify network disruptions. Without common monitoring, it can take hours to debug, and AI training, based on continuous, high-bandwidth exchange of data, can stall. Late anomaly detection contributes to these challenges. Congestion hotspots, silent packet drop, or hidden misconfigurations can lie dormant until GPU utilization drops or training times grow. Because of the costly nature of GPU time, small suspensions can be very costly. These problems compound in microservices architectures without good context boundaries or planning for scalability. Without good boundaries, services overlap or compete, resulting in data duplication and communication overhead. Correspondingly, scaling without budget control endangers out-of-control costs and uneven scalability. Establishing and enforcing correct boundaries and balancing scalability and cost are therefore critical [9; 10].

The integration of strong architectural planning and governance minimizes these inefficiencies. The definition of certain service boundaries keeps automation workflows—from configuration management to anomaly detection—conflict-free, and scalability controls prevent infrastructure bloating. The combination of these practices with advanced monitoring minimizes redundant data entry, accelerates fault isolation, and improves anomaly detection [12]. This protects investments in GPU clusters and in high-speed switching and guarantees AI training and inference workloads run reliably, both meeting performance targets and long-term financial payback.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

**Table 1: Key Inefficiencies in Fabric Performance and Mitigation Approaches**

| Inefficiency | Description | Impacts | Mitigation |
|---|---|---|---|
| **Redundant Data Entry** | Topology plans, optical inventories, and configuration templates maintained in silos | Slower deployment, difficult upgrades | Strong architectural planning; automation workflows |
| **Manual Fault Isolation** | Engineers manually check multiple switches and controllers to find disruptions | Debugging takes hours; AI training stalls due to interrupted data exchange | Advanced monitoring systems for faster detection |
| **Late Anomaly Detection** | Congestion, packet drops, or misconfigurations discovered only after GPU utilization drops | Expensive GPU idle time; longer training cycles | Automated anomaly detection; monitoring integrated into configuration management |
| **Weak Service Boundaries** | Overlapping or competing microservices without proper context boundaries | Data duplication, communication overhead, scaling inefficiencies | Clear definition of service boundaries; governance controls |
| **Uncontrolled Scalability** | Scaling without budgetary or architectural planning | Out-of-control costs; uneven performance scalability | Scalability controls; cost monitoring and balancing |

### *3.3 Governance and Compliance*

AI-optimized fabrics also must conform to stringent compliance and data governance requirements. Networks carrying proprietary training data or personally identifiable information must maintain extensive histories of topology changes, firmware versions, and access control attributes. NVIDIA UFM offers centralized auditing and anomaly detection for InfiniBand infrastructures, and Cisco Nexus Dashboard offers embedded telemetry and long-term logging for Ethernet fabrics. Maintaining master data accurate and current, like device inventories, port assignments, and traffic policies, is needed for day-forward operations as well as occasional regulatory audits. Misaligned configurations in leaf and spine switches or in cloud and on-premises fabrics may cause interruption of services or data leakage. Good governance, therefore, requires automation of configuration management, continuous monitoring, and policy-based access control to ensure every single change is monitored, authorized, and reversible. In short, from design approval and planning at one end to deployment, maintenance, and incident management, success in NVIDIA Quantum-2 and Cisco Nexus depends on proactive risk management and strong governance. Balancing inefficiencies against strong compliance protection makes these spine-leaf fabrics optimized for AI deliver not only high throughput and low latency, but operational reliability to keep pace with next-generation AI workloads.

As illustrated in the figure below, governance and compliance remain vital trends in big data analytics, underscoring the need for continuous monitoring, accurate master data, and automated access controls to secure AI-optimized fabrics while meeting evolving regulatory and privacy requirements.

**Journal of Information Systems Engineering and Management**
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**


Figure 2: **Data Governance Frameworks**

## 4. Key Technologies Driving Automation

### *4.1 Robotic Process Automation (RPA)*

Robotic Process Automation (RPA) represents a perfect tool for operational streamlining in AI-optimized spine-leaf network deployments such as those on NVIDIA Quantum-2 or Cisco Nexus [30]. RPA uses software "bots" to automate time-consuming, rule-based tasks—such as generation of configurations, mapping ports, regular checks for compliance, and time-based backups—that otherwise would require human labor. Automation of these tasks reduces human error and offers identical configurations across hundreds of switches and thousands of endpoints, required for lossless and low-latency performance requirements of modern AI clusters. RPA also links old and new systems to one another. The majority of data centers operate in hybrid infrastructures in which InfiniBand or Ethernet fabrics must interoperate with previous management tools. While automating data movements and synchronizations between these systems, RPA provides incremental modernization without unplanned replacement, and without unplanned downtime and cost. Beyond simple automation, RPA in combination with AI analytics provides proactive identification of configuration issues, automatic correction of anomalies, and predictive maintenance [38]. This intelligent layer achieves optimal network functioning without continuous human guidance.

Good governance is critical. Studies of scalable SaaS deployments uncover how formal control guarantees distributed systems consistency and reliability [39]. Employing analogous frameworks guarantees RPA remains controlled and auditable as it grows to thousands of nodes. Moreover, RPA's potential to enforce homogeneous execution adheres to best practices in machine learning and computer vision, wherein homogeneous data annotation yields accurate results [40]. Automation of day-to-day operations, integration across legacy systems, and infusing AI for predictive analysis means RPA drives time to deployment faster, boosts operational accuracy, and minimizes long-term costs, freeing engineers to focus on higher value-design, scalability, and security.

### *4.2 AI and Machine Learning*

Artificial Intelligence (AI) and Machine Learning (ML) offer intelligent automation beyond just rule-based functions [19]. AI/ML models process enormous amounts of switch, server, and application telemetry in real time, recognizing congestion, forecasting link failures, and programmatically re-routing traffic to keep throughput and latency at their optimal best. On NVIDIA Quantum-2, analytics powered by AI in combination with Unified Fabric Manager foretell hot spots and offer proactive recommendations for configuration changes, keeping peak GPU utilization intact. Cisco Nexus fabrics do the same through AI-enabled modules in the Nexus Dashboard to recognize microbursts or packet loss and apply adaptive policies such as dynamic buffer assignment or congestion control. AI and ML also increase security through identifying unusual traffic behaviors signifying potential breaches or misconfigurations, and reducing exposure to stolen data or regulation violations. They offer in-network, automatic communication—such as instant alerts or triggering automatic remediation after identifying

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

patterns. With ongoing learning based on network behavior, AI and ML enable fabrics to auto-optimize and adapt to variable workloads without heavy human interaction.

As illustrated in the figure below, AI and Machine Learning form nested components within the broader field of automation, enabling real-time telemetry analysis, proactive congestion control, and adaptive security that allow AI-optimized fabrics to self-optimize with minimal human intervention

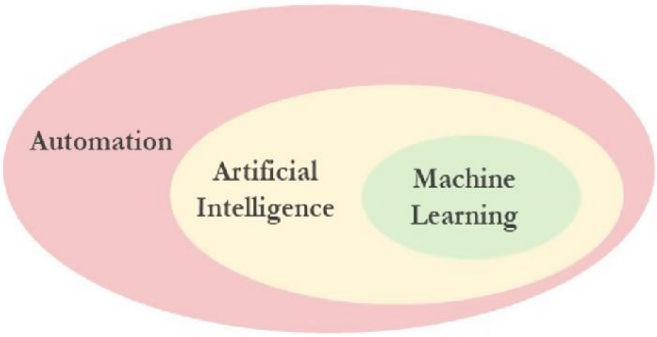


**Figure 3: Diagram-of-automation-AI-and-ML**

### *4.3 Cloud and Microservices*

Cloud infrastructure and microservices architectures provide scalability and flexibility to accommodate exponentially growing AI workloads. Instead of monolithic control systems, new fabrics exploit independently updatable, scalable, and restorable microservices. Provisioning, monitoring, and troubleshooting, for instance, can thus be realized as separate services connected by tightly defined APIs. The updates—like a new release of telemetry analytics or a new version of the routing engine—may be installed without affecting other functions. Both NVIDIA and Cisco utilize this modular architecture. Interoperation happens between cloud-based services of Unified Fabric Manager and Quantum-2, and Cisco Nexus interworking happens across Nexus Dashboard microservices to deliver automation and visibility. This decoupling also helps operators refresh or scale capabilities without network downtime and reduces disaster recovery through distributed or hybrid fabric controller support. As AI clusters can often span two or more data centers or hybrid clouds, microservices provide ultra-fast scaling and seamless interoperation with external compute and storage resources.

Real-time data processing is central. High-performance databases and event-driven architecture keep data flowing fast and constantly across microservices. Products like Aerospike offer ultra-low-latency processing of data for real-time business decisions and continuous availability—imperative for AI-optimized fabrics [11]. The architecture minimizes deployment time, improves reliability, and provides flexibility to handle unpredictable spikes in AI compute needs.

## 5. Methodology

### *5.1 Research Design*

This is a mixed-methods research study because it is necessary to examine both the quantifiable performance properties of AI-optimized spine-leaf fabrics and the contextual approach, which demonstrates how these properties influence deployment decisions. Quantitative considerations will comprise the numbers of bandwidth, latency, port density, and failure rates, as published in specifications and case deployments of the NVIDIA Quantum-2 and Cisco Nexus systems. Technical reports and expert interviews are qualitative factors used to justify operational practices and strategies. The mixed-method approach ensures that analysis is grounded not only in theoretical measurements of performance but also in the real-life situations of operations, priorities within the organization, and its long-term objectives. The twofold perspective provides a ground-level view for comparing the two

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

network solutions and presents realistic recommendations to organizations planning to upgrade to a large-scale AI fabric.

### *5.2 Data Collection*

The data that will be utilized in the research is collected with the help of numerous credible sources, in order to obtain a comprehensive view of the two technologies. Technical specifications and reference architectures of NVIDIA and Cisco define quantitative measurements, such as the number of ports, throughput per switch, and fabric scaling limits. Such vendor resources will provide the necessary information to evaluate the performance of the Quantum-2 and Cisco Nexus fabric in various network topologies and under different AI load conditions at varying workload levels. These technical figures are accompanied by realistic case studies and deployment white papers, which provide actual data regarding configuration techniques, cabling techniques, and power and cooling requirements. These records provide insight into how theoretical performance can be translated into operational reality, highlighting what works and what should be avoided when applying it on a large scale.

Interviews in the industry and conference talks are also examined to bring more depth and operationalize [6]. Such testimonials from data-center architects and network engineers include insights from real-life experiences of deploying Quantum-2 and Nexus fabrics for AI workloads. Their content provides valuable insights into the process of automated switch configuration by a team, including handling unexpected traffic patterns and integrating the fabric into an existing cloud or high-performance computing platform. Another aspect of explaining the cost-of-ownership factor, upgrades, and ease of further maintenance can be seen through market analysis and customer responses. The data are essential for assessing the long-term financial viability, as well as for understanding how different organizations make trade-offs between Ethernet and InfiniBand solutions. It will ensure that the analysis does not reflect a small set of operational locations, but an expansive set of operational locations by collecting feedback between the enterprise AI clusters and large cloud or hybrid data centers.

The importance of a multi-source data collection, which is premised on other industries, is justified. To illustrate the point, research on the timeliness of notification within the healthcare industry has shown that sensible Co-ordination of real-time information streams can also improve the process of service delivery and efficiency [34]. Simultaneously, the study of secure data exchange on the IoT focuses on the role of low-latency and high-speed infrastructure that guarantees data integrity across distributed systems [35]. Such similarities suggest that, to create an effective AI network design, applications in healthcare communication, like those in other fields, depend on the implementation of real-time information from various sources to facilitate the making of timely and accurate decisions. The plan for data collection will combine the specifications of the vendors, field tests, specialized interviews with experts, and market analysis with cross-industry lessons to ensure that the findings represent both the technical and organizational aspects of implementing AI-optimized spine-leaf fabrics. This richness of input provides a legal foundation on which to quantify achievement, scalability, and operational impact on an extensive scope of varied applications.

### *5.3 Analytical Framework*

The comments on the reviewed data are analyzed using a synthesized approach that incorporates process mining, cost-benefit analysis, and key performance indicators (KPIs). Process mining helps visualize the sequence of stages in network fabric design, implementation, and operation, and where automation tools, such as RPA and AI-based analytics, can reduce the timeline or minimize errors. Cost-benefit analysis measures the trade-offs between the cost of high-density switching hardware capital and its benefits, including the ability to operate in shorter training times and more efficiently utilize the GPU. During AI workloads, objective metrics (such as KPIs (aggregate network throughput, average and tail latency, mean time to repair (MTTR), and energy efficiency)) can be

**Journal of Information Systems Engineering and Management**
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

applied to compare Quantum-2 and Nexus fabrics. It is a comprehensive analysis that involves both financial and technical analysis, demonstrating how the benefits of performance are translated into actual business outcomes. The framework connects the raw measures to the reality of AI-driven organizations, enabling a comparison of KPIs against the information gathered through interviews and reports on deployment.

As illustrated in the figure below, monitoring and evaluating key performance indicators—such as network throughput, latency, and mean time to repair—guides process mining and cost-benefit analysis, ensuring that NVIDIA Quantum-2 and Cisco Nexus deliver measurable technical and business outcomes.


**Figure 4: Operational Kpis**

### *5.4 Reliability and Ethics*

The need to guarantee the reliability and ethical integrity forms the basis of the methodology. Triangulation between vendor specifications, independent benchmarks, and field reports eliminates key figures and bias. When evaluating customer feedback or interviewing practitioners, the sources are then cross-examined to ensure consistency and accuracy. Moral considerations are also important, particularly in case studies involving proprietary network topology or data-center functional data. Any sensitive information is anonymized, and only publicly available or agreed-upon information is allowed to be analyzed in the analysis. Because AI fabrics often contain training data, which may consist of personal or confidential information, the concern of data privacy and security is also mentioned in the paper, alongside encryption, access control, and standards, including GDPR and SOC 2. Not only would these protect the privacy of individuals, but they would also render the research credible, as it would ensure that the inferences reached are informed by verifiable and responsibly acquired information. The method will be a combination of quantitative and qualitative measurement approaches to create a holistic image of the practice of AI-optimized spine-leaf fabrics. The combination of multiple sources of data, the application of excellent analytical tools, and the maintenance of the high levels of reliability, as well as ethical conduct, lead to the fact that the study will yield the results, which can be evaluated as both technically sound and applicable to the strategic decisions of the organizations that will regard NVIDIA Quantum-2 and Cisco Nexus as the next-generation AI data centers options.

As illustrated in the table below, reliability and ethics are ensured through rigorous data triangulation, anonymization, strict privacy compliance, strong encryption, and a balanced analytical approach, ensuring credible, verifiable, and strategically relevant findings for AI-optimized spine-leaf fabric research.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

**Table 2: Reliability and Ethical Considerations in AI-Optimized Spine-Leaf Fabric Research**

| Aspect | Measures Ensuring Reliability and Ethics |
|---|---|
| **Reliability of Data** | - Triangulation of vendor specifications, independent benchmarks, and field reports<br>- Cross-examination of customer feedback and practitioner interviews |
| **Ethical Safeguards** | - Anonymization of sensitive/proprietary data<br>- Use of only publicly available or consented information<br>- Adherence to GDPR, SOC 2, and data privacy standards |
| **Security Measures** | - Encryption and access controls applied to AI training data and network data |
| **Analytical Approach** | - Combination of quantitative and qualitative methods<br>- Use of advanced analytical tools for robust insights |
| **Credibility of Research** | - Reliance on verifiable and responsibly acquired data<br>- Ethical conduct ensuring technically sound and strategically relevant findings |

## 6. Designing the Streamlined Workflow

### *6.1 Unified Process Map*

The development of a single process map is the basis of a streamlined workflow in AI-optimized spine-leaf fabrics implemented on the NVIDIA Quantum-2 or Cisco Nexus platforms. A process map is a single chart and operational map that contains all the stages of the network design, its deployment, and its continuous operation. Andreas, without it, teams are exposed to duplication of efforts, a neglected dependency, and configuration or documentation inconsistencies. It starts by mapping and defining key steps, starting with early capacity planning, through switch provisioning, topology checks, and maintenance, and then setting clear handoffs between design engineers, network operations personnel, and data-science teams. Every step is discussed to remove delays or repetition. As an illustration, data entry of port-mappings can be avoided in cabling design, configuration generation, and monitoring dashboards by combining it into a master dataset. RPA or pipeline orchestration can then be used to distribute the dataset to downstream systems automatically.

These advantages are not confined to efficiency. An integrated process map makes change management easier and reduces response time to incidents [13]. In case of a link failure or congestion, the map serves as a guide when locating upstream and downstream dependencies so that fault isolation can be performed more quickly. It also enhances the auditability that is vital in achieving the regulatory and internal governance requirements. In AI systems operating on a very rapid schedule (where a delay of just a few seconds can drastically increase the time needed to train the network), this mandatory process map will allow the network to remain stable and responsive to adjustments.

### *6.2 Standardized Data Models*

After mapping the overall process, the next thing that should be done is the development of standard models of data upon which intelligent decision engines are to be configured to control the pricing of resources and to place automated workloads. These models determine the structure and exchange of information relating to devices, ports, VLANs, or InfiniBand partitions, routing, and

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

security credentials between systems. Consistency guarantees that all switches, cables, and endpoints are modeled identically in inventory management tools, orchestration systems, and monitoring systems, which is a necessity when dealing with hundreds of Quantum-2 NDR 400 Gb/s or Cisco Nexus 400/800 GbE.

Intelligent engines can be used to evaluate capacity, suggest topology adjustments, and alter Quality of Service (quality of service) parameters with the help of standardized models. To illustrate, with the peak of the utilization of the GPU, an AI-based engine can determine the addition of spine links or rebalancing of traffic to meet the latency budgets based on human discretion. At the financial level, such engines can predict the cost of adding ports or optics and approximate the savings of the faster completion of AI jobs. They also simplify the approvals, which can be checked against standardized schemas and compliance rules before deployment, cut down on deployment time, and make sure that all the adjustments will be consistent, cost-effective, and adhere to policy. This strategy is similar to Poka-Yoke, which is the manufacturing philosophy of error-proofing, where defects are removed. Industrial case studies, including Tesla rotor production, have shown that Poka-Yoke removes errors during processes by rendering wrong actions irreducible or immediately apparent [14]. Similary, standardized data models can also be regarded as Poka-Yoke to AI-optimized networks. They promote consistency of configurations by creating uniform schemas and validation rules so that incorrect configurations are not deployed, and inconsistencies cannot spread across large fabrics. This minimizes the risk of service outages, accelerates troubleshooting, and safeguards the integrity of vital AI workloads.

By integrating error-prevention logic with data models and decision engines, network architects can achieve a mixture of manufacturing-type quality control and sophisticated automation. What comes out as the result is a fabric that scales effectively and yet has a high operational reliability to satisfy the requirements of next-generation AI training and inference?

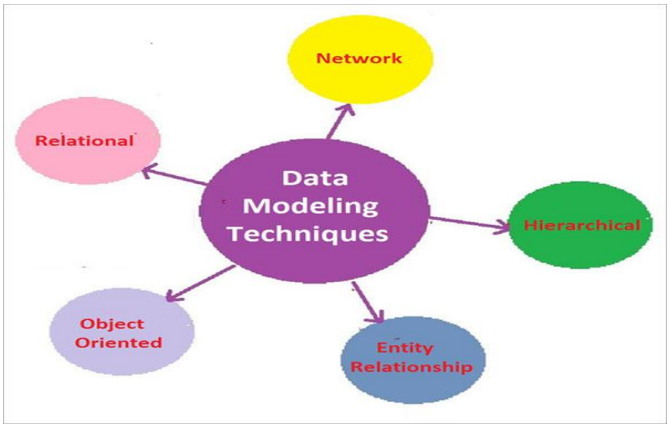


**Figure 5: Data Modeling**

### *6.3 Digital Portals*

These advantages are directly translated to internal users, including data-science teams and application developers, by way of digital portals to the customer and virtual assistants. Such customers need to have trustworthy, high-performing connections to educate and implement AI models. Properly developed portals will enable them to request available resources, check on job status, and check network health without opening physical tickets or waiting until an engineer can come by. As an example, a researcher can add additional GPU nodes and check bandwidth through a self-service dashboard. An integrated assistant will offer real-time information on link usage and possible congestion.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Natural language processing-based virtual assistants are also more efficient. Instead of reading through the complicated configuration menu, the user can query, is my training cluster supporting the needed 400 Gb/s throughput? Moreover, get answers in real-time and on what to do. Cisco Nexus incorporates such features with its Nexus Dashboard, and NVIDIA Quantum-2 users may use Unified Fabric Manager and custom AI agents. These interfaces not only enhance productivity, but also minimize the operational costs by reducing the number of personnel who are needed to support the interfaces. Each interaction also produces telemetry that can be optimized to optimize resource allocation and additional optimization of the network. The design stage determines a highly automated, open, and user-friendly workflow through unified process mapping, standardized data structures, and intuitive digital interfaces, which allows organizations to expand AI workloads with confidence without compromising the low latency and high bandwidth of next-generation applications.

## 7. Business Benefits of Automation

### *7.1 Operational Efficiency*

The enhancement of the efficiency of operations is one of the most important advantages of the automation of AI-optimized spine-leaf fabrics. Automation saves or removes manual configuration, enabling hundreds of switches to be deployed or upgraded in a fraction of the time that it used to take. Even in large GPU clusters where each hour of training time has a high monetary cost, any delay will result in quantifiable cost reductions. To illustrate, automated configuration templates will allow a 64-port NVIDIA Quantum-2 leaf switch or a 64-port Cisco Nexus spine switch to be brought online in a short period and in a consistent manner, with no human input. Troubleshooting is also made easier with automation: AI-based analytics in the NVIDIA Unified Fabric Manager or the Cisco Nexus Dashboard identify congestion or hardware errors in real-time and automatically perform remedial actions to maintain the utilization of GPUs at the optimum level. These efficiencies reduce operational costs in the long run and enable the network teams to concentrate on innovation and strategic planning instead of giving attention to run-of-the-mill maintenance.

The research on AI-powered process optimization supports these observations. Digital solution-specific feedback loops and intelligent feedback loops can help to make decisions faster and enhance productivity overall [15]. As AI-based career coaching will provide instant feedback when it comes to making difficult professional choices, AI-based analytics and automation in network processes will result in accurate, real-time feedback to changing traffic and performance requirements. Fast response capability makes sure that human expertise is focused on strategic development as opposed to routine duties. Also, it is important to align automation with organizational goals. It has been found that digital strategies that are customized to serve the market improve efficiency and relevance over the long term [16]. The given principle applies directly to the data center environment of AI, where automation tools and processes should be created to address the specific needs of high-performance computing and large-scale AI training. Through a combination of a strong automation architecture and ongoing optimization and planning, organizations will be able to ensure the highest levels of operational efficiency in both NVIDIA Quantum-2 and Cisco Nexus fabrics- establishing a highly resilient infrastructure at a lower cost, shorter downtime, and more innovation-driven focus.

As illustrated in the figure below, achieving operational excellence requires aligned strategies in performance management, employee engagement, and process optimization, reinforcing how AI-driven automation with NVIDIA Quantum-2 and Cisco Nexus strengthens efficiency, reduces downtime, and drives continuous innovation

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article


**Figure 6: Achieving Operational Excellence through Artificial Intelligence**

### *7.2 Customer Experience*

Automation has a direct and positive impact on the customer experience as well [2]. The customers in AI data centers are primarily data scientists and developers who require compute and network resources to be made available within a short timeframe, allowing them to train or deploy models. New node requests, bandwidth addition, or quality-of-service level customization may be made almost immediately thanks to automated processes through self-service portals and virtual assistants. This instantaneity also shortens the duration of project execution, providing researchers with the opportunity to test and repeat complex AI models without delay. The case of high throughput and low latency is also worthy of being maintained even under heavy load. An automated fabric built on NVIDIA Quantum-2 or Cisco Nexus can provide predictable performance to support tasks such as large-scale AI training and real-time inference. Transparent monitoring and clear reporting also contribute to establishing trust, as users can see for themselves that the service-level agreements for bandwidth, latency, and uptime are being met.

Best practices in containerization enhance such benefits. Recent AI tasks are increasingly based on container-based infrastructure to train and infer with several nodes regularly. Conformity to Docker and Kubernetes rules of containerization ensures that services can be packed and presented in a short time, while also being reliable and secure [17]. Kubernetes orchestration is the most efficient way of managing resources, allowing them to be spun up or down according to demand, thereby achieving smooth performance during spikes or when a large cluster is required. Security is related to customer trust. Effective security practices are directly coupled into the continuous integration/continuous deployment (CI/CD) pipeline so that no vulnerability is added due to an automated update and the introduction of new services. The techniques employed include static application security testing (SAST), dynamic application security testing (DAST), and software composition analysis (SCA) to identify and resolve potential issues during the initial phase of the development cycle [18]. This proactive security approach ensures that rapid self-service provisioning and scale-up do not compromise data integrity or system stability.

The customer experience can also be improved with the help of predictive analytics, as it enables planning resources and helps avoid issues beforehand. Predictive models anticipate the use of resources and the likelihood of congestion before it affects users, making automatic capacity allocation possible through historical trends and real-time telemetry [20]. This will also imply to the developers that they will experience minimal interruptions, the training time will be reduced, and they will be assured that their computing resources will not be disrupted, even during peak times. By integrating containerization best practices, secure CI/CD pipelines, and predictive analytics, along with an automated spine-leaf architecture, organizations can deliver a fast, reliable, and secure user experience that fosters trust between infrastructure teams and internal consumers, and accelerates the development and adoption of AI.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### *7.3 Risk and Compliance*

Automation enhances risk management and compliance with regulations by consolidating governance across the network tiers. NVIDIA Quantum-2 and Cisco Nexus have an automated audit trail that records all configuration and firmware changes, as well as attempted changes. These records also facilitate compliance with data privacy and industry-related standards more easily by ensuring that all changes to the records are approved and properly registered. Automatic anomaly detection is another security feature. The behavior of a network is continuously monitored with the help of machine learning to detect abnormalities that may indicate hardware failures, data breaches, and policy infractions. Functions like Priority Flow Control (PFC), Explicit Congestion Notification (ECN), with automation, are applied with the implementation of RoCEv2 in Cisco to ensure that no packets are lost, and the integrity of sensitive data flows is preserved. The features of NVIDIA Unified Fabric Manager are similar to those of InfiniBand fabrics, including predictive analytics and in-network computing, which enable the isolation and resolution of problems before they can impact operations. The period of vulnerability and maintenance of standards can be decreased through automation, which decreases the technical and regulatory risks. This is assured as a sum totalling the following functional, customer-centric, and compliance benefits in an effort to make it clear that automated AI-optimized spine-leaf fabrics can enable sustainable efficiency, reliability, and trust at scale.

## 8. Discussion

### *8.1 Key Findings*

The relative comparison of NVIDIA Quantum-2 and Cisco Nexus fabrics reveals that both can meet the high-bandwidth/low-latency requirements of next-generation AI workloads through the implementation of various strategies. It is typical of the literature on AI-centric data-center networks that deterministic performance and lossless transport are the most significant, and the findings in this case are of this nature. The NVIDIA Quantum-2 is based on InfiniBand and SHARP in-network computing, which provides hardware acceleration and end-to-end congestion control to deliver ultra-low latency and predictable collective communication. With RoCEv2 running over high-speed Ethernet, Cisco Nexus demonstrates that a properly configured Ethernet fabric with priority flow control (PFC) and explicit congestion notification (ECN) can be used to meet the requirements of synchronized GPU workloads, too. In the past, research has questioned the ability of Ethernet to substitute for InfiniBand in such applications. Nevertheless, the current research verifies that current Ethernet can be utilized as an alternative with proper tuning and telemetry.

The findings also support the growing importance of automation and real-time data analytics. The NVIDIA Unified Fabric Manager and Cisco Nexus Dashboard are such tools that reduce the mean time to repair considerably, ease scaling, and provide uniformity of the fabric. These layers of management apply sophisticated spine-leaf networks as self-optimizing platforms, capable of solving spikes in AI workload or link failures within seconds. The other area of overlap with a prior study is data governance, specifically real-time processing and master data management, which are always concerns related to performance and compliance. Quantum-2 has its orientation on InfiniBand-oriented control, whereas Nexus is based on standards-based telemetry; however, both provide strong change management and security controls.

Bar graph illustrating Key Findings comparing NVIDIA Quantum-2 and Cisco Nexus. It shows that both platforms meet next-generation AI workload demands, with Quantum-2 excelling in ultra-low latency and congestion control, while both demonstrate strong automation and data governance capabilities.

**Journal of Information Systems Engineering and Management**
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

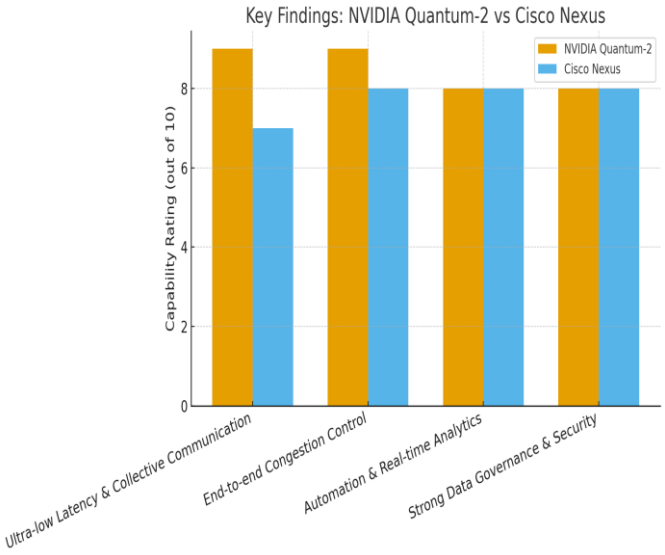


**Figure 7: Key Findings: NVIDIA Quantum-2 vs Cisco Nexus**

### *8.2 Strategic Impact*

The implications of strategic competitiveness, profitability, and workforce transformation are profound. Direct influences on the overall cost of ownership and time-to-market of AI services are the fabric used. NVIDIA Quantum-2 offers effectiveness never seen in terms of communicative capacity, reducing the time taken to train a set of GPUs and the cost of training a model. The capability will lead to shorter product development times and increased profitability in companies where the power lies in training large models, such as autonomous driving services or research facilities with large budgets. By incorporating Ethernet standards, Cisco Nexus can lower the cost of capital investments in multi-purpose data centers that have already been deployed and ease optical and cabling supply chains. Nexus can also be very strategic for institutions that require wide interaction and blended workload hosting, enabling them to minimize expenses and be versatile in their activities.

Automation also defines the workforce. Both platforms eliminate manual configuration and troubleshooting through RPA, analytics powered by AI, and cloud-native management solutions. The network engineers will be in a position to take out of the way of highly repetitive jobs, architecture, optimization, and security to generate a more specialized and innovation-oriented workforce. With fewer interventions to operate and grow AI infrastructure, the error rate is lower, and the service-level agreements improve customer trust and reputation in the market. These advantages support risk management and business continuity. Constant work of AI requires a strong network fabric, especially in such spheres where it is needed, such as autonomous systems, medical research, or financial modeling, among others. An effective incident response will ensure that when something goes wrong, such as an equipment malfunction or a sudden surge in traffic, it does not result in costly downtime. According to studies on business continuity, a high degree of planning and automation is one of the advantages that can emerge from unforeseen circumstances, enhancing profitability and the organization's image [24].

Security is also added to the strategic case of automated AI fabrics. DevSecOps practices can be combined with threat intelligence to help organizations automate risk mitigation before vulnerabilities are introduced into production. These features include real-time telemetry, automated anomaly detection, policy-based access controls, and the continuous integration and deployment of pipelines, among others, which enable a fast response to threats and their resolution. These facilities will ensure the safety of the infrastructure and will be able to meet the new demands [25].

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### 8.3 Limits and Future Work

Regardless of these positive results, there are several limitations. A significant part of the research on large-scale AI fabrics is based on vendor case studies and early-adopter reports, which give information about operations, but might not encompass the entire spectrum of deployment situations, especially in smaller or resource-constrained businesses. The current-generation 400 Gb/s and the upcoming 800 Gb/s technologies, which are rapidly evolving, are also analyzed. New hardware technologies (including co-packaged optics or new congestion-control protocols) may change the performance and cost equation between InfiniBand and Ethernet.

Telematics and integration of blockchain are also some of the emerging trends that should be considered in future research. Connected vehicle insurance is already heavily dependent on telematics, which, with enough scale, may generate data streams of gigantic proportions that challenge AI fabrics. The way that Quantum-2 and Nexus can accommodate millions of telemetry updates per second enables insurers and fleet operators to plan for such loads. Equally, data integrity systems based on blockchains will be vital in establishing the authenticity and immutability of data used in AI training. Research into the support or integration of each platform into blockchain nodes and distributed ledgers might provide an important dimension to the choice of technology.

There is a need to conduct long-term research on energy efficiency and carbon effect since environmental considerations continue to affect the use of technology and data-center design [21]. It will be important to analyze the sustainability profile of both InfiniBand and Ethernet fabrics, particularly as the size of AI models and energy requirements increase, in order to know more about future best practices. Recognizing these constraints and prioritizing web connectivity, blockchain security, and energy sustainability, the future of work can use the existing discoveries to inform organizations about the development of AI networks that would be strong, efficient, and adaptable to changing technological and environmental needs.

## 9. Case Study

### 9.1 Insurance Example

The case study that demonstrates how spine-leaf networks optimized by AI will transform them is a big multinational auto insurer. As more and more connected-vehicle data emerged, and claims processes got more complicated, the company was seeking to cut the time to claim settlement and improve customer satisfaction. It pursued a two-platform strategy, deploying NVIDIA Quantum-2 InfiniBand fabrics in AI training clusters to accelerate the model development process and Cisco Nexus Ethernet fabrics in production data centers to support claims processing services and customer-facing applications. This combination enabled all environments to be customized to meet the requirements, while also ensuring high performance throughout the entire enterprise.

As illustrated in the figure below, future trends in AI for insurance—such as usage-based products, predictive and preventive insurance, IoT-generated data, cyber insurance, and AI as a service—align with this case study's strategy of combining NVIDIA Quantum-2 and Cisco Nexus.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

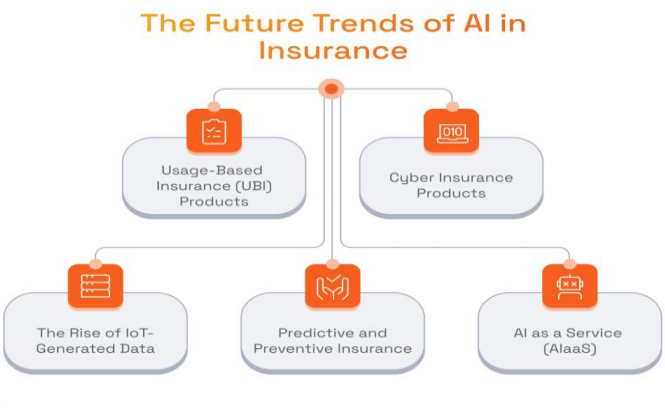


**Figure 8: The Power of AI in Insurance**

### *9.2 Implementation Process*

This was accomplished by conducting a comprehensive review of the current infrastructure and future workload projections. In order to train the AI models, the engineers created a non-blocking NVIDIA Quantum-2 spine-leaf network that would accommodate hundreds of NDR 400 Gb/s links. It was a design that utilized the Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) in-network computing to accelerate all-reduce operations required for running deep learning. The team utilized Cisco Nexus 9000 and 9800 switches, along with RDMA over Converged Ethernet version 2 (RoCEv2), Priority Flow Control (PFC), and Explicit Congestion Notification (ECN), to ensure lossless transmission in Ethernet.

Integration was done in stages. The initial pilot cluster confirmed topology assumptions, bandwidth goals, and latency goals [41]. RPA scripts were automated to configure switches and test them, which guaranteed a stable, fast deployment. The analytics provided by AI delivered telemetry information in real-time to manage health, detect congestion hotspots, and identify anomalies before they impacted workloads. The design was supported by lessons learned in other related areas like telematics and fleet management. To ensure proactive response to faults and capacity planning, continuous and real-time data exchange (as demonstrated to enhance efficiency and predictive maintenance) was incorporated [26]. Incorporating pilot validation, automated configuration, and rich telemetry, the rollout had predictable performance and high reliability. The speed of InfiniBand and SHARP computing of Quantum-2 and the Ethernet flexibility and lossless transport capabilities of Cisco Nexus provided a hybrid infrastructure that can support the training and insurance services needs of a wide variety of AI and scale to meet future workloads' needs.

### *9.3 Measurable Results*

The two deployments provided significant operational and financial advantages. The training of models in AI reduced training times by almost a third, reducing the amount of program idle time in the GPUs and speeding up the rollout of risk-scoring and fraud-detection applications. Automation in claims processing has decreased the average claim cycle time by ten to less than four days without compromising high privacy and reporting standards. Faster processing reduced operation costs and increased customer satisfaction in terms of better survey scores, successful self-service portals, and instant status updates. On the financial front, less idle time on GPUs, less time spent by hand configuring the system, and fewer unplanned disruptions resulted in quantifiable cost reductions. Automated setup and monitoring reduced person-hours and risk of depression. These directions are congruent with the best practices in cloud sustainability and cost effectiveness, where carbon-conscious scheduling and dynamic scaling decrease energy consumption and cost without affecting performance [28].

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

As illustrated in the table below, AI and automation deployments achieved notable operational gains, financial savings, and sustainability improvements, including faster model training, reduced claims cycles, lower manual configuration, and energy-efficient scheduling that cut costs while maintaining high performance and reliability.

**Table 3: Operational and Financial Results of AI and Automation Deployments**

| Category | Measured Result |
|---|---|
| **Operational Benefits** | - AI model training times reduced by ~33% (less GPU idle time)<br>- Rollout of risk-scoring and fraud-detection apps accelerated<br>- Claims cycle reduced from 10 to <4 days<br>- Improved customer satisfaction (survey scores, self-service portals, instant updates) |
| **Financial Benefits** | - Reduced GPU idle time and manual system configuration<br>- Fewer unplanned disruptions<br>- Automated setup and monitoring reduced person-hours<br>- Lower risk of errors and inefficiencies |
| **Sustainability Impact** | - Carbon-conscious scheduling and dynamic scaling decreased energy consumption and costs without reducing performance |

### *9.4 Lessons Learned*

Such lessons as the significance of early and comprehensive capacity planning to guarantee a smooth scaling, automation at all levels, including RPA to configure it, AI to identify anomalies, and integrated dashboarding to monitor it, to fit into deployment schedules and ensure service continuity, and good governance and data management to assure auditable configuration changes and compliance are pretty relevant. More importantly, the project validated the need to align technology to task: InfiniBand and Quantum-2 performed well in tightly coupled GPU clusters, whereas Ethernet and Cisco Nexus gave the scalability and flexibility needed in enterprise and customer-facing systems [29]. The case study proves that AI-optimized spine-leaf architectures are effective, high-impact solutions. The insurers will gain sustainable and long-term gains in speed, efficiency, customer experience, and environmental responsibility by aligning technology decisions with the workload features, investing in automation, and implementing strict governance [23].

## 10. Implementation Roadmap

### *10.1 Change Management*

The initiation of an AI-optimized spine-leaf fabric based on NVIDIA Quantum-2 or Cisco Nexus is based on structured change management. High-performance fabrics require significant investment in hardware and trained employees, as well as continuous maintenance; thus, executive buy-in is a prerequisite. The leadership support is ensured through the clear illustration that AI model training reduces operational costs and enhances competitiveness. Clear communication of costs, expected returns, and mitigating risk strategies keeps decision-makers on the right track. Employee training is also important. The automation, data analysis, and continuous optimization replace the ACL configuration and troubleshooting functions. Reskilling programs are designed to meet the needs of specific tools, including InfiniBand or RoCEv2 protocols, NVIDIA Unified Fabric Manager, Cisco Nexus

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dashboard, and data governance best practices. Targeted reskilling makes engineers and operators knowledgeable in automated large-scale fabrics. Constant communication between leadership, engineers, and end-users will help keep the project moving forward and counter resistance to change before it builds.

### *10.2 Phased Rollout*

Staged deployment reduces risks and is efficient. It begins by testing network architecture, automation tools, and operational procedures in a pilot environment that resembles the target spine-leaf topology. The performance of RPA-based configuration and AI-based anomaly detection, latency during heavy load, lossless transportation with collective AI traffic, and so forth are tested. Initial tests indicate that the intended architecture can support challenging throughput and reliability goals before the commencement of the large-scale deployment. The pilot data is used to guide design refinement. Link utilization patterns are used to determine the addition of more spine connections, whereas telemetry is used to optimize Quality-of-Service (quality of service) policies and buffer settings. The evidence-based feedback loop eliminates the risk of misconfiguration and guarantees optimal scale performance. After the validation, deployment will be done in phases, with leaf and spine switches added at the same rate as compute-nodes. There is automated testing and constant monitoring at each stage to ensure that latency, bandwidth, and error rates remain within acceptable limits. Such a step-by-step approach can avoid interruptions, offer a chance to learn operationally, and help teams to address the problems at an early stage. It also enables foreseeable budgeting, balancing the capital spending over time into several stages to conform to the actual requirements of the GPU clusters and AI training services.

The incremental approach is based on the concepts of intelligent data processing and inference. For example, dynamic memory inference, which dynamically varies the workload requirements, has been shown to increase efficiency and stability [33]. When applied to AI networks, these concepts enable every step to react to performance data and be continually refined dynamically. The approach has been practical in creating fabrics based on either NVIDIA Quantum-2 or Cisco Nexus that scale gracefully, guarantee service-level agreements, and achieve high performance with managed costs and operational risk [1].

As illustrated in the figure below, best practices for rolling back versions—such as thorough testing, version control, automated rollback, and strong documentation—parallel the phased rollout strategy by ensuring continuous monitoring, controlled changes, and reliable recovery during AI fabric deployment

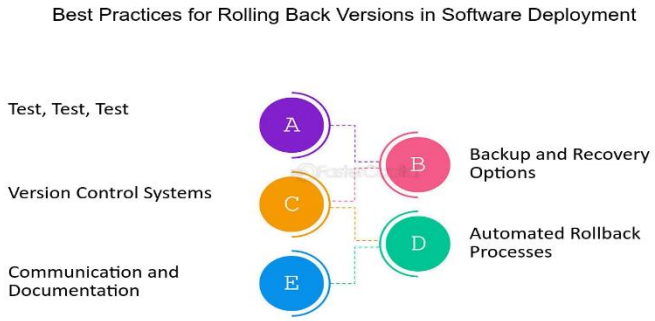


**Figure 9: Rolling Back Versions**

### *10.3 Key Metrics*

Measures of success. During Rollout, include key objective metrics. Turnaround time- to add new GPU nodes or pay claims- shows the effect of both automation and optimization of a network.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Customer satisfaction, as measured by surveys, Net Promoter Scores, or direct feedback, is indicative of the advantages of quicker provisioning, consistent network performance, and improved service availability. The financial returns are measured by the cost per claim or per AI training iteration, demonstrating that investments in high-speed switching and automation yield cost savings. Regular review of these measures ensures that the performance improvement is sustained and the bottlenecks that arise are addressed proactively. With the combination of change management, gradual Rollout, and ongoing measurement, this roadmap directly connects the technological deployment to quantifiable business results. It offers a consistent platform for both high-performing and low-cost AI fabrics, which can be scaled to 800 Gb/s and higher in the future, with the capability to meet the evolving needs of next-generation AI workloads [1].

## 11. Recommendations

### *11.1 Unified Roadmap*

Companies that intend to implement AI-optimized spine-leaf fabrics must start with a detailed automation roadmap that has a strong connection with the overall business goals. Identifying priority workloads, establishing latency budgets, and defining desired scaling horizons, as well as integrating automation technologies such as RPA, AI/ML analytics, and microservices-based management, should be included in this roadmap. A clear roadmap will ensure that investments in NVIDIA Quantum-2 or Cisco Nexus hardware are accompanied by the corresponding software tools and operational processes, which will not be implemented in a fragmented manner or based on unnecessary expenditures.

### *11.2 AI and Governance*

It is crucial to focus on monitoring and real-time data governance with the help of AI [27]. NVIDIA Unified Fabric Manager or Cisco Nexus Dashboard should feature AI-based analytics to forecast congestion, identify anomalies, and facilitate automatic remediation. Powerful master data management should be in place to ensure that the device inventories, port assignments, and access policies remain accurate and consistent. These enhancements improve stability, protect sensitive training or customer data from being compromised, and can meet regulatory standards, such as GDPR or SOC 2. The ongoing validation and continuous audit trails also enhance compliance and minimize the risk of costly data breaches.

### *11.3 Agile Deployment*

The flexibility and minimum risk are achieved by adopting an agile and incremental deployment strategy. Rather than massive disruptive changes, fabrics are to be grown in manageable chunks, beginning with pilot pods, then expanding to full-data-centre coverage. All the phases must have automated testing, performance checks, and feedback mechanisms to improve the subsequent stage. By ensuring that regulators and industry standards organizations are closely aligned during the process, the dynamic nature of data privacy and cybersecurity regulations will be effectively addressed. This provides a faster time to value, as well as enables rapid adaptation to new AI models, hardware refresh cycles, and advanced network technologies, such as 800 Gb/s optics.

As illustrated in the figure above, diverse organizational challenges—from people-centric and perception issues to structural, managerial, and individual factors—underscore why agile, incremental deployment with continuous testing and regulatory alignment is essential for minimizing risk and ensuring adaptive, compliant AI network rollouts.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article


**Figure 10: Framework for Agile Transformation**

### *11.4 Workforce Training*

To maintain the advantages of automation, it is essential to invest in ongoing employee training and develop incentives towards the adoption of technology. Training must address the hardware and software of the fabric of choice, including InfiniBand-specific congestion control or Ethernet-based RoCEv2 tuning, as well as the application of automation and analytics tools. The concept of cross-functional learning, which involves collaboration among network engineers, data scientists, and security teams, promotes teamwork and reduces silos in operations. A culture of constant improvement can be instilled through incentive programs that provide rewards to early adopters of new tools and best practices that are shared. With changes in AI workloads, a well-trained and motivated workforce should be capable of utilizing new features promptly and at their best operational scale.

These suggestions establish a channel through which organizations can achieve the full potential of AI-optimized spine-leaf networks. A coherent roadmap offers strategic guidance and operational integrity through AI-driven monitoring and governance, supporting scalable and agile deployment. Continuous training can create an effective and flexible workforce. Regardless of whether the preferred platform is NVIDIA Quantum-2 with its InfiniBand precision or Cisco Nexus with its standards-based Ethernet flexibility, such moves enable enterprises to deploy and operate fabrics that support the high-performance, low-latency requirements of state-of-the-art artificial intelligence, while also providing long-term financial and competitive benefits.

## 12. Conclusion

In the case of AI-optimized spine-leaf fabrics, created on the foundation of NVIDIA Quantum-2 and Cisco Nexus, network automation is transforming the work of large-scale digital services, including underwriting basic insurance services, claims processing, and customer care. Automation can also be used to expedite the process of routing data, implement price policies accurately, and process claims efficiently by eliminating manual processing and applying intelligence at every stage of the network. This means that real-time data streams and predictive models can be used to assist in underwriting, and they can operate effectively on GPU clusters in real-time. The benefit of claims settlement is that it verifies the facts of incidents instantly and has automated policy adjudication procedures that cut down the turnaround time (days) to hours. Digital portals and virtual assistants can enhance customer care by providing instant feedback and updating status. All of this reduces operating costs, minimizes human errors, enhances regulatory compliance, and provides a responsive and seamless experience for policyholders and internal stakeholders.

The further development of artificial intelligence, machine learning, and predictive analytics in the future will enhance these advantages and extend them throughout the insurance value chain. Through sophisticated algorithms, risk scoring will be more precise, as it will consider telematics and

IoT sensor information, as well as third-party information such as weather and traffic conditions. The use of deterministic low-latency networking will be required as larger and more complex models will enter the trend, and high-performance fabrics will continue to be a consistent source of competitive advantage. Machine learning will also be further extended in active work, anticipating congestion or hardware failures in the network, and parameterization to the point where it can affect service. Predictive analytics will help insurers anticipate market shifts and customer demand, and develop new products and services that align with emerging trends, such as use-based insurance or dynamic pricing. These AI-fuelled insights, combined with automated, lossless networking, can allow insurers to build adaptive platforms that adapt to customer expectations and put regulatory requirements into practice as fast as they can.

These advantages can only be maintained by a culture of constant improvement and a scalable infrastructure that does not require initial deployments. The layouts of the networks should accommodate the rapid proliferation of AI models, the increased concentration of GPU clusters, and the migration to 800 Gb/s and higher technologies. The ability to review periodically, process maps, data models, and automation scripts will ensure systems are optimized as workloads and business requirements evolve. The ongoing training programs will ensure that engineers, data scientists, and operations teams are aware of new protocols, analytics tools, and compliance rules. As AI replaces more business functions, such as fraud detection and customized marketing, businesses must extend their core functions of underwriting and claims to encompass customer acquisition, product development, and ecosystem associations. It will also stimulate insurers to adopt new technologies, including blockchain, to share data safely and use federated learning to perform privacy-protective analytics by applying a long-term lens. NVIDIA Quantum-2, or a Cisco Nexus spine-leaf fabric optimized by AI, will not be an upgrade of the current network, but a catalyst for the entire digital insurance company. With a combination of a high level of automation, machine learning, and predictive analytics, insurers can reshape their old workflow, offer more precise and quicker services, and continuously grow to meet different needs in the future. The resultant effect is that a strong and smart infrastructure will be in place, which facilitates long-term growth, as well as regulatory compliance and an excellent customer experience, for the next-generation car insurance and beyond.

## References

[1] Allioui, H., & Mourdi, Y. (2023). Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey. *Sensors*, *23*(19), 8015. https://doi.org/10.3390/s23198015

[2] Bolton, R. N., McColl-Kennedy, J. R., Cheung, L., Gallan, A., Orsingher, C., Witell, L., & Zaki, M. (2018). Customer experience challenges: bringing together digital, physical and social realms. *Journal of service management*, *29*(5), 776-808. https://doi.org/10.1108/JOSM-04-2018-0113

[3] Bonthu, C. (2025). Real-time data processing in ERP systems: Benefits and challenges. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8889

[4] Bonthu, C., Kumar, A., & Goel, G. (2025). Impact of AI and machine learning on master data management. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/5186

[5] Brahmbhatt, R., & Sardana, J. (2025). Empowering patient-centric communication: Integrating quiet hours for healthcare notifications with retail & e-commerce operations strategies. Journal of Information Systems Engineering and Management. https://www.jisem-journal.com/index.php/journal/article/view/3677

[6] Cammarano, A., Varriale, V., Michelino, F., & Caputo, M. (2023). A framework for investigating the adoption of key technologies: Presentation of the methodology and explorative analysis of emerging practices. *IEEE Transactions on Engineering Management*, *71*, 3843-3866. https://doi.org/10.1109/TEM.2023.3240213

[7] Chadha, K. S. (2025). Edge AI for real-time ICU alarm fatigue reduction: Federated anomaly detection on wearable streams. *Utilitas Mathematica*, *122*(2), 291–308. https://utilitasmathematica.com/index.php/Index/article/view/2708

[8] Chadha, K. S. (2025). Zero-trust data architecture for multi-hospital research: HIPAA-compliant unification of EHRs, wearable streams, and clinical trial analytics. *International Journal of Computational and Experimental Science and Engineering*, *12*(3), 1–11. https://ijcesen.com/index.php/ijcesen/article/view/3477/987

[9] Chavan, A. (2022). Importance of identifying and establishing context boundaries while migrating from monolith to microservices. Journal of Engineering and Applied Sciences Technology, 4, E168. http://doi.org/10.47363/JEAST/2022(4)E168

[10] Chavan, A. (2023). Managing scalability and cost in microservices architecture: Balancing infinite scalability with financial constraints. Journal of Artificial Intelligence & Cloud Computing, 2, E264. http://doi.org/10.47363/JAICC/2023(2)E264

[11] Dhanagari, M. R. (2025). Aerospike: The key to high-performance real-time data processing. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8894

[12] Enemosah, A., & Chukwunweike, J. (2022). Next-Generation SCADA Architectures for Enhanced Field Automation and Real-Time Remote Control in Oil and Gas Fields. *Int J Comput Appl Technol Res*, *11*(12), 514-29. https://www.researchgate.net/profile/Aliyu-Enemosah/publication/391627685_Next-Generation_SCADA_Architectures_for_Enhanced_Field_Automation_and_Real-Time_Remote_Control_in_Oil_and_Gas_Fields/links/681f7642bd3f1930dd704afd/Next-Generation-SCADA-Architectures-for-Enhanced-Field-Automation-and-Real-Time-Remote-Control-in-Oil-and-Gas-Fields.pdf

[13] George, A. S., Sagayarajan, S., Baskar, T., & George, A. H. (2023). Extending detection and response: how MXDR evolves cybersecurity. *Partners Universal International Innovation Journal*, *1*(4), 268-285. https://doi.org/10.5281/zenodo.8284342

[14] Goel, G. (2025). Implementing Poka-Yoke in manufacturing: A case study of Tesla rotor production. *International Journal of Mechanical Engineering*, 5(1), 3. https://doi.org/10.55640/ijme-05-01-03

[15] Karwa, K. (2023). AI-powered career coaching: Evaluating feedback tools for design students. Indian Journal of Economics & Business. https://www.ashwinanokha.com/ijeb-v22-4-2023.php

[16] Karwa, K. (2024). Navigating the job market: Tailored career advice for design students. *International Journal of Emerging Business*, *23*(2). https://www.ashwinanokha.com/ijeb-v23-2-2024.php

[17] Koneru, N. M. K. (2025). Containerization best practices: Using Docker and Kubernetes for enterprise applications. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8905

[18] Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[19] Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets*, *32*(4), 2235-2244. https://link.springer.com/article/10.1007/s12525-022-00598-0

[20] Kumar, A. (2019). The convergence of predictive analytics in driving business intelligence and enhancing DevOps efficiency. International Journal of Computational Engineering and Management, 6(6), 118-142. Retrieved from https://ijcem.in/wp-content/uploads/THE-CONVERGENCE-OF-PREDICTIVE-ANALYTICS-IN-DRIVING-BUSINESS-INTELLIGENCE-AND-ENHANCING-DEVOPS-EFFICIENCY.pdf

[21] Lai, I. K., Liu, Y., Sun, X., Zhang, H., & Xu, W. (2015). Factors influencing the behavioural intention towards full electric vehicles: An empirical study in Macau. *Sustainability*, *7*(9), 12564-12585. https://doi.org/10.3390/su70912564

[22] Lamberton, Chris and Brigo, Damiano and Hoy, Dave, Impact of Robotics, RPA and AI on the Insurance Industry: Challenges and Opportunities (November 29, 2017). Journal of Financial Perspectives, Vol. 4, No. 1, May 2017, Available at SSRN: https://ssrn.com/abstract=3079495

[23] Lescrauwaet, L., Wagner, H., Yoon, C., & Shukla, S. (2022). Adaptive legal frameworks and economic dynamics in emerging tech-nologies: Navigating the intersection for responsible innovation. *Law and Economics*, *16*(3), 202-220. https://doi.org/10.35335/laweco.v16i3.61

[24] Malik, G. (2025). Business continuity & incident response. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8891

[25] Malik, G. (2025). Integrating threat intelligence with DevSecOps: Automating risk mitigation before code hits production. *Utilitas Mathematica*. https://utilitasmathematica.com/index.php/Index/article/view/2709

[26] Nyati, S. (2018). Transforming telematics in fleet management: Innovations in asset tracking, efficiency, and communication. International Journal of Science and Research (IJSR), 7(10), 1804-1810. Retrieved from https://www.ijsr.net/getabstract.php?paperid=SR24203184230

[27] Oladele, O. K. (2024). Data-Driven Product Roadmap Prioritization: Using AI-Powered Predictive Analytics to Optimize Feature Sequencing. https://www.researchgate.net/publication/384930582_Data-Driven_Product_Roadmap_Prioritization_Using_AI-Powered_Predictive_Analytics_to_Optimize_Feature_Sequencing?enrichId=rgreq-bc25ab885b529dfb3db16384bdbc4e32-

[28] Pinnapareddy, N. R. (2025). Carbon conscious scheduling in Kubernetes to cut energy use and emissions. *International Journal of Computational and Experimental Science and Engineering*. https://ijcesen.com/index.php/ijcesen/article/view/3785

[29] Pinnapareddy, N. R. (2025). Cloud cost optimization and sustainability in Kubernetes. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8895

[30] Pires, B. R. R. (2023). *Estudo de Tecnologias Para Redes de Datacenter* (Master's thesis, Universidade de Aveiro (Portugal)). https://www.proquest.com/openview/69d7d87be3dafe7db6d94fc00cd7ff5d/1?pq-origsite=gscholar&cbl=2026366&diss=y

[31] Prakash, D. (2024). Data-driven management: The impact of big data analytics on organizational performance. *International Journal for Global Academic & Scientific Research*, *3*(2), 12-23. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://core.ac.uk/download/pdf/613704989.pdf

[32] Raju, N., Quazi, F., & Kondle, P. (2024). AI and Machine Learning in Digital Modernization Transforming Industries for the Future. *Available at SSRN 5053870*. https://dx.doi.org/10.2139/ssrn.5053870

[33] Raju, R. K. (2017). Dynamic memory inference network for natural language inference. International Journal of Science and Research (IJSR), 6(2). https://www.ijsr.net/archive/v6i2/SR24926091431.pdf

[34] Sardana, J. (2022). The role of notification scheduling in improving patient outcomes. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[35] Sardana, J., & Dhanagari, M. R. (2025). Bridging IoT and healthcare: Secure, real-time data exchange with Aerospike and Salesforce Marketing Cloud. *International Journal of Computational and Experimental Science and Engineering*. https://ijcesen.com/index.php/ijcesen/article/view/3853/1161

[36] Singh, V. (2021). Generative AI in medical diagnostics: Utilizing generative models to create synthetic medical data for training diagnostic algorithms. International Journal of Computer Engineering and Medical Technologies. https://ijcem.in/wp-content/uploads/GENERATIVE-AI-IN-MEDICAL-DIAGNOSTICS-UTILIZING-GENERATIVE-MODELS-TO-CREATE-SYNTHETIC-MEDICAL-DATA-FOR-TRAINING-DIAGNOSTIC-ALGORITHMS.pdf

[37] Singh, V. (2022). EDGE AI: Deploying deep learning models on microcontrollers for biomedical applications: Implementing efficient AI models on devices like Arduino for real-time health monitoring. International Journal of Computer Engineering & Management. https://ijcem.in/wp-content/uploads/EDGE-AI-DEPLOYING-DEEP-LEARNING-MODELS-ON-MICROCONTROLLERS-FOR-BIOMEDICAL-APPLICATIONS-IMPLEMENTING-EFFICIENT-AI-MODELS-ON-DEVICES-LIKE-ARDUINO-FOR-REAL-TIME-HEALTH.pdf

[38] Subham, K. (2025). Integrating AI into CRM systems for enhanced customer retention. *Journal of Information Systems Engineering and Management*. https://www.jisem-journal.com/index.php/journal/article/view/8892

[39] Subham, K. (2025). Scalable SaaS implementation governance for enterprise sales operations. *International Journal of Computational and Experimental Science and Engineering*. https://ijcesen.com/index.php/ijcesen/article/view/3782

[40] Sukhadiya, J., Pandya, H., & Singh, V. (2018). Comparison of Image Captioning Methods. *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH*, *6*(4), 43-48. https://rjwave.org/ijedr/papers/IJEDR1804011.pdf

[41] Wang, K., Zhou, Q., Guo, S., & Luo, J. (2018). Cluster frameworks for efficient scheduling and resource allocation in data center networks: A survey. *IEEE Communications Surveys & Tutorials*, *20*(4), 3560-3580. https://doi.org/10.1109/COMST.2018.2857922