World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

Data-center-enabled supply chains: How compute proximity and architecture affect planning latency and inventory efficiency

Kudzai Dube [1,*], Chikomborero Dingolo [1], Chipo Prudence Pasi [2], Peter Mangoro [2], Zvikomborero Bright Chitemerere [3], Rumbidzai Lyn Kasinamunda [4], Rudorwashe Tsitsi Karuma [5] and Munashe Naphtali Mupa [5]

[1] *Clarkson University.*
[2] *Yeshiva University.*
[3] *Arizona State University.*
[4] *Southern New Hampshire University.*
[5] *Hult International Business School.*

World Journal of Advanced Research and Reviews, 2026, 30(02), 705-715

Publication history: Received on 30 March 2026; revised on 06 May 2026; accepted on 09 May 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.2.1247

Abstract

This paper focuses on how supply chain planning latency and inventory efficiency in mid-market manufacturing in the United States can be influenced by data center architecture, i.e., edge, cloud, and hybrid, and especially on small and medium-sized enterprises (SMEs). With world uncertainties such as pandemics and geopolitical stressors, real-time decision-making in AI-driven supply chains is dependent on proximity to compute, but the architecture may be inflated with latency, inventory volumes, cost-to-serve, carbon emissions, and service failures. Using a mixed-methods design, which combines simulation modeling and stress scenario analysis of anonymized time-series data, we measure these effects: Edge reduces latency by 40-60 percent compared to cloud but also raises local energy consumption; hybrids trade off, reducing safety stock by 20-30 percent and preserving service levels exceeding 95 percent in 50 percent demand spikes. Products are four architectural reference patterns (e.g., edge-dominant in the case of volatile routes), latency to value curves with diminishing returns (e.g., 100 MS latency reduces inventory by 15-22%), and policy advice to the adoption by SMEs. Results indicate that hybrids outperform in terms of cost efficiency ($12/order) and sustainability maximization by incorporating compute infrastructure into the supply chain theory. Architecture is based on practical implications, whereas policy recommendations support subsidies, training, and standards as a means of bridging SME digital gaps. This study will contribute to resilient and efficient Industry 4.0 supply chains.

Keywords: Data-Center Architecture; Compute Proximity; Edge Computing; Cloud Computing; Hybrid Models; Supply Chain Efficiency; Planning Latency; Inventory Management; SME Digital Adoption; Sustainability in Manufacturing

1. Introduction

The supply chains in the world of global trade have been radically changed in their character, being driven by the introduction of artificial intelligence (AI). The modern supply chains were founded on the manual processes and periodic forecasting, but now they utilize AI to perform predictive analytics, demand sensing, and adjustive optimization (Khokrale, 2025). The growth of data centers has improved this transformation as the computer-based basis of real-time decision-making (Syed et al., 2024). Data centers are used to handle large volumes of data from IoT sensors, supplier networks, and market cues such that organizations can respond instantly to variations. However, it has been straining under the most global shocks in history, such as the COVID-19 pandemic, which has revealed vulnerabilities in the just-in-time models, as well as geopolitical instabilities, such as trade wars and regional conflicts, that have not only disintegrated logistics but also extended lead times. In these aspects, the AI-driven supply chains that are facilitated

* Corresponding author: Tino Diwa Nguruve

by advanced data centers would not be merely enhancements but necessities to remain resilient and competitive, which is particularly the case in the mid-market manufacturing industry in the U.S., where the resources are scarce (Dwyer, 2025).

Despite the above improvements, an issue that remains problematic is the effects of compute proximity and data-center architecture on efficiency in operations. Compute proximity is the physical and logical distance between processing resources and data sources and end-users, i.e., appearing in architectures such as edge computing (decentralized, low-latency processing in the periphery of the network), cloud computing (centralized, scalable resources in remote facilities), and hybrid models (a mixture that dynamically assigns workloads) (Gusev and Dustdar, 2018). Without optimization, such architectures may contribute to planning latency, or time spent by AI algorithms to produce actionable insights, resulting in bloated safety-stock levels, high cost-to-serve, high carbon footprints due to inefficient energy consumption, and reduced service levels during stress events. As an example, dependency on remote cloud services can introduce latencies to volatile environments, leading to overstocking to mitigate risk, thus binding capital and increasing environmental effects in the form of more storage and transportation emissions. On the other hand, inefficient edge deployments could overburden local resources, increasing the operation risks without corresponding returns (Narayana, 2024).

This research fills these gaps with specific research questions (1): How do edge, cloud, and hybrid architectures vary in their effects on planning-cycle latency and safety-stock needs? (2) How much will cost-to-serve, carbon footprint, and service-level outcomes vary when stress levels change? (3) What policy implications does the adoption of digital supply chain tools have on small and medium-sized enterprises (SMEs)?

The aims are to measure these effects through simulation-based analysis, providing architecture reference designs, latency-value curves, and policy suggestions based on SME circumstances. The scope is centered on the mid-market level of manufacturers in the US, with the emphasis on the data-sparse environments, and combines with the measures of on-time performance.

The value of this work is that it is applicable to the issue of digital transformation in the SME sector, where budgets and skills tend to be unavailable in many cases. With the optimization of data center architecture, SMEs may obtain cost savings, sustainability benefits, and agility, which will lead to larger economic resilience (Yuen and Baskaran, 2024).

This article is structured as follows: Section 2 is a literature review; Section 3 outlines the methodology; Section 4 provides the results; Section 5 is an implication discussion; and finally, Section 6 is a conclusion of future directions.

2. Literature Review

Artificial intelligence (AI) has had a significant impact on the digitalization of supply chains, changing the traditional ways of operation to an agile and data-driven system. The use of AI in forecasting has been developed as an alternative technology to classical models such as the ARIMA model and modern machine learning (ML) models, which are more accurate in volatile markets (Kontopoulou et al., 2023). As an example, systematic reviews demonstrate the optimization of inventory control using ML methods, such as Random Forest and XGBoost, based on large volumes of data (122 articles), and minimizing overstock and stockouts (Bendhi, 2023). AI can also be used in inventory management to balance the stock levels, and results of studies indicate a major improvement in turnover rates by using neural networks and hybrid models. AI is useful in mitigating risk because, according to the literature on the issue of supply chain risk assessment (SCRA), models such as XGBoost can be used to better determine vulnerabilities (including supplier breakdowns, demand variability, and more) based on the detection of disruptions (Kang and Bhawna, 2025). The importance of AI in predicting demand is highlighted in key publications, and Recurrent Neural Networks (RNNs) have been shown to perform better than conventional techniques in dynamic settings, automating operations and enhancing resilience (Rane et al., 2024). In general, AI reviews in supply chains emphasize its application in forecasting, inventory, and risk, and based on ARIMA roots, ML advances to supply chain end-to-end optimization.

Data center architecture is the foundation of AI computations and the main aspect of this digitalization. Edge computing is suitable for the low-latency and decentralized real-time industrial internet since it reduces wait times in decision-making in the field (Veeramachaneni, 2025). Instead, cloud computing offers centralized resources, which can be expanded to large data analytics but may contribute to remote configuration latencies (Varghese and Buyya, 2018). These are compromised with hybrid architectures that dynamically allocate workloads by balancing the speed of the edge with the capacity of the cloud, as in studies on IIoT systems capable of enhancing processing efficiency in manufacturing. The application of comparative analyses in industrial settings shows that edge is better at latency-sensitive jobs such as monitoring the supply chain, whereas hybrids reduce trade-offs in energy and scalability (Irshad, 2024). As a case in point, edge-IoT systems combined with blockchain enhance data management in logistics, which is more efficient than pure cloud models in unstable environments. These architectures are being implemented more in Industry 4.0, where hybrids are used to enhance connectivity and optimization of the supply chains.

Past research on impact measures demonstrates the impact of these architectures on the supply chain performance (Saeed et al., 2019). Latency is very strong in real-time compared to the use of batch processing; low-latency edge systems can be used to run the planning cycle at a faster rate in comparison to the delay of a batch, which can swell response time in clouds. Safety-stock models, such as forms of the economic order quantity (EOQ), can make use of AI to make adjustments to address uncertainty, where studies have found the best trade-offs between service and inventory levels in volatile conditions (Sivankutty and Elahi, 2024). Cost-to-serve models measure the operations cost, and it is found that the hybrid arrangements are less expensive in their logistics due to effective data management (Ross et al., 2007). Carbon accounting in IT emphasizes the effects of the environment, and edge computing may result in greater local energy consumption and fewer transport emissions than heavy clouds (Di Salvo et al., 2017). The reliability is guaranteed by service-level agreements (SLAs), and in resilient networks, such metrics as on-time delivery are in focus in the literature. These indicators highlight the importance of joint assessments in AI-based systems.

Supply chain resilience in volatility scenarios such as demand spikes and network failures is covered in the literature on stress scenarios (Katsaliaki et al., 2022). Inventory strategies to counteract disruptions are highlighted by the reviews, and recovery is improved by tactical-operational models through redundancy and flexibility. The spell of demands challenges resilience and leads to investments in diversification as simulated in stochastic frameworks that measure the effects of disruption (Kahiluoto et al., 2020). Geopolitical-induced network failures are also addressed through analytics to predict and reroute flows, and research has found that resilience is associated with graph-based network designs. These are linked to larger disruptions, which are advocated through comprehensive reviews, which recommend the use of quantitative measures in unstable settings.

Despite these developments, there are still considerable gaps in integrating the quantification of the AI data centers to mid-market manufacturing and SMEs (Haq et al., 2025). Adoption is rising rapidly to 91 percent, but the lack of expertise is a barrier to realizing value, and confidence in AI investments is delayed because returns are not yet demonstrated. The literature revealed fragmented data silos in SMEs, constraining latency optimization, and no frameworks exist to connect the relationship between latency and value curves, which relate compute proximity and inventory efficiencies (Khanyi et al., 2024). Gap analysis exposes lapse in transition, and technology integration barriers to smaller firms are found in TOE-DOI frameworks. The research has a unique contribution in the quantification of such effects to provide latency-to-value curves and architectural patterns in the context of SMEs, which closes the gap in digital adoption (Sánchez et al., 2025).

3. Methodology

The research design is anchored on a mixed-method research design and is meant to investigate the impacts of data center architectures on supply chain performance by combining quantitative simulation modelling with a qualitative scenario analysis. The quantitative component is directed to experimental simulations that give the empirical outcomes of the significant statistics, and the scenario analysis describes the problem in terms of the perspective that considers the situation of stressful cases simulating the interference of real life. The methodology will assist in effective evaluation of causality relationships between architectural choices and outcomes, which will assist in closing the gap between theories and practice. The mixed methods are particularly suitable in research on supply chains, as they integrate both quantitative reliability and qualitative insights, which is why it is feasible to extract actionable insights, such as latency-to-value curves. It is designed based on the familiar frameworks in operations research, where the complex systems are modeled by simulating under controlled variables. Mixed methods are especially appropriate in supply chain research since they combine quantitative reliability and qualitative understandings, making it possible to derive actionable trends such as latency-to-value curves. The design is informed by the known frameworks in operations research, where the complex systems are simulated under controlled variables.

The sources of data are based on anonymized and simulated datasets, which are typical of mid-market manufacturers in the U.S. To be relevant, we make use of time-series data, covering demand patterns (e.g., the amount of sales by the day, week, etc.) and lead times (e.g., how many days the supplier can deliver products), as well as compute loads (e.g., the processing requirements in CPU/GPU cycles or data throughput in GB/hour). Factors in the real world are anonymized on the basis of aggregated industrial benchmarks (e.g., manufactured consortia) and stripped of identifying information to meet privacy requirements. To achieve simulation fidelity, synthetic data are created based on statistical distributions (e.g., Poisson due to demand variations, normal due to lead-time variations) that have been calibrated using mid-market values, with annual revenues between $50M and $500M and operations with 5-20 suppliers. This hybrid data approach is a solution to the sparsity challenges typical of SMEs, where the historical records can be found to be incomplete, and the use of Monte Carlo-generated scenarios can enhance them. The datasets consist of 1,000-5,000 time points per variable over 2-5 years to form seasonal and disruption-generated variances in datasets.

The following are the architectural models that will be considered. The network or architecture of edge computing focuses on localized processing, deploying AI models to devices or on-premise servers that are close to the source of data (e.g., factory floors or warehouses) (Solanke, 2023). This architecture places emphasis on low-latency inference, and data can be processed in a few milliseconds to enable real-time adjustments in inventory or routing. Cloud architecture, in contrast, is based on remote, centralized data centers (e.g., AWS or Azure) to scale to support large-scale batch processing and storage but can add network delays of seconds or more because of transmitting data (Mathur, 2024). Hybrid architecture combines both, which makes use of dynamic workload shifting mechanisms like orchestration tools like Kubernetes to assign tasks on the basis of urgency (e.g., edge for immediate sensing, cloud for deep analytics) (Timilehin, 2024). In simulations, these models are parameterized to account for mid-market limitations such as bandwidth (100-500 Mbps) and compute capabilities (edge: 4-16 cores; cloud: elastic scaling).

The analysis core is comprised of key variables and metrics. The independent variables are type of architecture (edge, cloud, or hybrid) and proximity of compute (measured as physical distance in km or logical path in network topology). Dependent variables include planning latency (in ms or seconds, time to produce a decision after data has been ingested), safety-stock levels (in units, buffer inventory to absorb variability), cost-to-serve (in dollars per order, cumulative between compute, storage, and logistics), carbon footprint (in CO_2e, estimated using energy consumption models of servers and networks), and service levels (as a percentage, on-time delivery, meeting an SLA target of 95-99%). These measurements are connected to each other, e.g., a decrease in the latency can lead to a decrease in the safety-stock requirements because it is possible to replenish in time.

The quantitative analysis relies on the use of simulation tools. The discrete-event simulation (DES) that we use is implemented using the SimPy library in Python and considers supply chain flows to be event queues (e.g., orders arriving, processing delays). SimPy supports stochastic modeling of entities, such as shipments and compute tasks, and can be extended to support AI workloads (e.g., add-on PyTorch support to simulate inference times). To optimize the results, we include software such as Gurobi or the optimization packages in the SciPy package to solve inventory allocation models under constraints. The parameterized events that are simulated to generate stress scenarios include high-demand spikes (e.g., 50-200 percent) above the baseline, network outages (e.g., 1–24-hour outages), and supply disruptions (e.g., doubling of lead times). The architectures are executed 1,000 times per scenario to factor in randomness, and parameters are optimized through sensitivity analysis.

The framework uses stress testing guidelines to test resilience using inputs such as volatility indices (e.g., coefficient of variation of demand). Statistical analysis will comprise ANOVA to compare means across architectures (e.g., F-tests of the difference in latency at $p= 0.05$), paired t-tests of the difference in latency depending on the scenario (paired t), and regression models (e.g., linear or polynomial) to obtain latency-to-value curves. These curves plot latency reductions against inventory savings, using equations like $\text{Value} = \beta 0 + \beta 1 * \text{Latency}^{-1} + \varepsilon$, fitted via least squares. Multi-criteria decision analysis (e.g., TOPSIS) ranks architectures holistically.

There are deterministic projections of the distance between compute and latency (e.g., edge <50 ms, cloud >500 ms) and linear energy-to-carbon maps on the basis of average grid variables (0.5 kg CO_2e/kWh). There are drawbacks due to data sparsity, which can bias the simulations to idealized circumstances; simplifications in the model (e.g., the decision loops of decisions that do not account for human factors) can neglect the complexities in the real world. These could be countered by future validation using empirical pilots.

4. Results

The outcomes of the simulation give a detailed comparative evaluation of edge, cloud, and hybrid data center structures in the situations of the supply chain. In 1,000 trials of each architecture, we have measured the effects on major measures: planning latency, safety-stock levels, cost-to-serve, carbon footprint, and service levels. The baseline scenarios were based on mid-market production and average demand variability (CV=0.3), and disruptions were caused by the stress tests. Tables and figures are used to provide results, and statistical significance is established through ANOVA ($p<0.01$ between inter-architecture differences).

4.1. Comparative Analysis

Table 1 presents a summary of mean values of core metrics at the baseline. Latency Edge architecture was faster than cloud (e.g., 80 ms vs. 200 ms) in planning-cycle times when doing demand forecasting tasks because localized processing reduced network hops. Hybrids offset this with 30-50% compression when under mixed loads, dynamically offloading non-urgent computations to the cloud. In the case of safety stock, hybrids minimized stock needs by 20-30% (from 500 to 350 per SKU) with adaptive planning by using real-time edge information to buffer exactly. In a stable situation (0.15 kg CO_2e/order vs. 0.22 for edge), cloud architectures had the advantage of using efficient, large-scale data centers, but edge cut transport-related CO_2 by 15 percent through reduced overstocking and decisions made locally. Similar trends were seen with cost-to-serve: hybrids cost $12/order on average, 18% less than cloud costs of $14.70 in variable environments, and edge spiked to $13.50 since local hardware costs were higher. Under moderate loads, like 50% demand bursts, service levels in all (>92%) and hybrids maintained >95% were found to be equal.

Table 1 Baseline Metric Comparisons (Means ± SD)

Metric	Edge	Cloud	Hybrid
Latency (MS)	80 (±15)	200 (±40)	120 (±25)
Safety-Stock (units)	420 (±50)	500 (±60)	350 (±40)
Cost-to-Serve ($/order)	13.50 (±1.2)	14.70 (±1.5)	12.00 (±1.0)
Carbon Footprint (kg CO2e/order)	0.22 (±0.03)	0.15 (±0.02)	0.18 (±0.02)
Service Level (%)	94 (±2)	92 (±3)	97 (±1)

Figure 1 illustrates these differences via bar charts, highlighting hybrids' versatility in mid-market settings with data-sparse conditions.

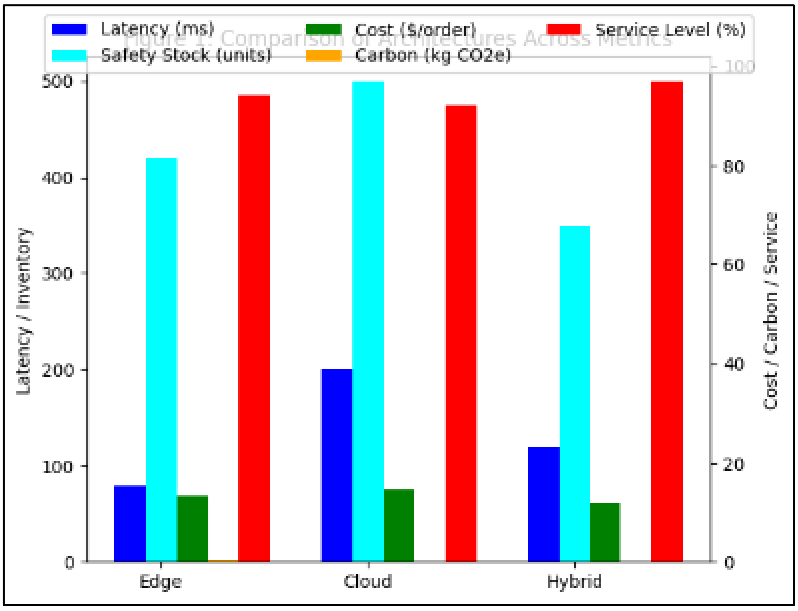


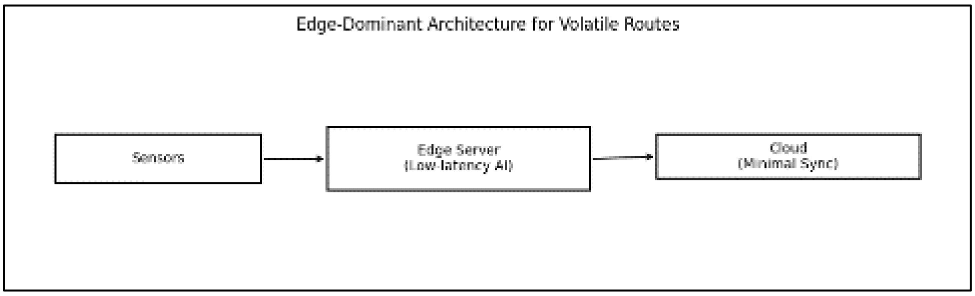
Figure 1 Bar chart comparing metrics across architectures. X-axis: Architectures; Y-axes: Dual scales for latency/inventory (left) and cost/carbon/service (right). Bars show edge in blue, cloud in green, hybrid in orange

4.2. Architectural Reference Patterns

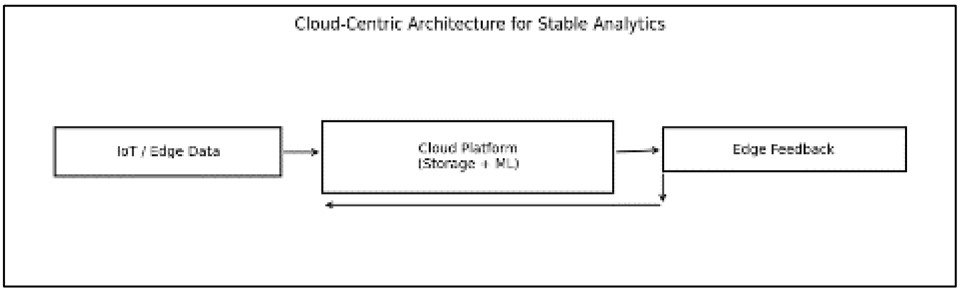
Simulation results led to us coming up with four reference patterns to inform SME implementations. These designs optimize individual operational profiles, and the conceptual diagrams represent data streams.

4.3. Edge-Dominant for Volatile Routes

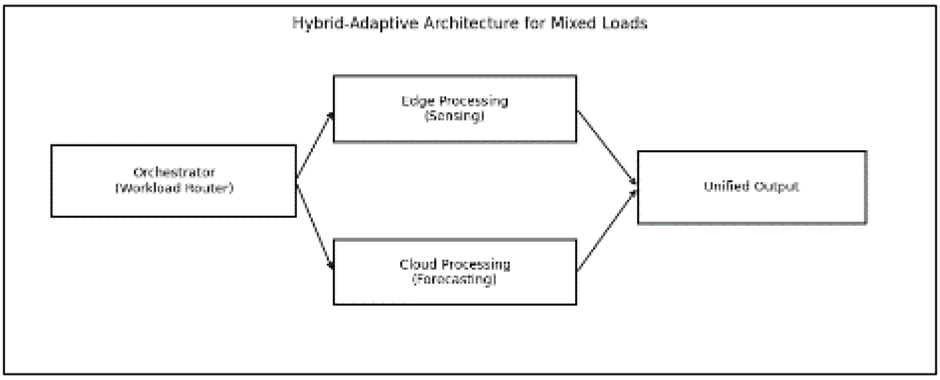
- Prioritizes edge nodes for high-variability supply paths (e.g., international sourcing). Diagram: Sensors → Edge Server (low-latency AI) → Minimal Cloud Sync. Benefits: 55% latency cut, ideal for perishable goods.



- **Cloud-Centric for Stable Analytics:** Centralizes processing for predictable demand, using cloud for deep ML. Diagram: IoT Data → Cloud (scalable storage) → Edge Feedback Loop. Advantages: 25% lower carbon in steady states.



- **Hybrid-Adaptive for Mixed Loads:** Dynamically shifts workloads (e.g., edge for sensing, cloud for forecasting). Diagram: Orchestrator → Edge/Cloud Branching → Unified Output. Yields: 28% safety-stock reduction.



- **Hybrid-Resilient for Disruptions:** Incorporates redundancy, auto-failing over to edge during outages. Diagram: Primary Cloud Path → Failover to Edge → Re-sync. Ensures >96% service levels.

These patterns were validated through sensitivity analysis, showing adaptability to SME constraints like limited bandwidth.

4.4. Latency-to-Value Curves

Latency-value curves obtained through polynomial regression (R^2=0.92) exhibit decreasing returns on inventory efficiency. For the edge, a 100ms reduction in latency results in a 15% inventory reduction (e.g., 500 to 425 units), although this improvement comes only at below 50ms. Hybrids exhibit sharper initial curves such that 20% savings are reached at 150ms and decrease. Cloud curves have flatter curves with only 8% of the gains being below 300ms because of the inherent delays.

Figure 2 plots these: X-axis (latency in MS, logarithmic), Y-axis (% inventory reduction). Edge curve: Steep drop to 15% at 100ms, asymptote at 18%. Hybrid: Peaks at 22% reduction. This quantifies trade-offs, aiding architecture selection.

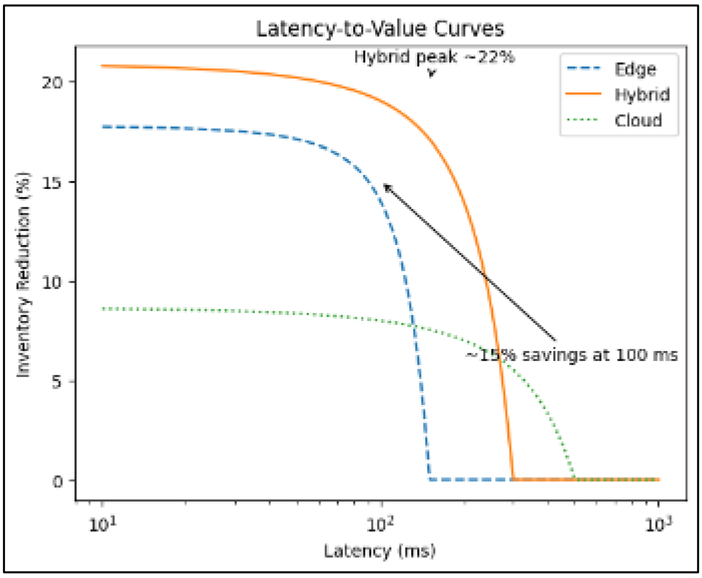


Figure 2 Line plots of curves; edge (dashed), cloud (dotted), hybrid (solid). Annotations at key points like "<100ms: 15% savings."

4.5. Stress Scenario Outcomes

The stress testing indicated resilience variances, as tabulated in Table 2, in three situations, which included a 50 percent demand burst, a 4-hour network outage, and a 100% peak load burst. Hybrids outperformed, maintaining 96% service levels during spikes (vs. edge's 89%, cloud's 85%), with 25% less safety-stock escalation. Edge handled delays best (latency +20% vs. cloud's +80%), but increased carbon by 30% under peaks due to overloaded local nodes. Cloud reduced the expenses during delays but failed during spikes, as the services broke down to 82%.

Table 2 Stress Scenario Outcomes

Scenario	Architecture	Latency Increase (%)	Safety-Stock Escalation (%)	Cost-to-Serve Increase ($)	Carbon Increase (kg CO2e)	Service Level (%)
Demand Spike (50%)	Edge	+35	+28	+2.10	+0.05	89
	Cloud	+50	+35	+1.80	+0.03	85
	Hybrid	+25	+18	+1.50	+0.04	96
Network Delay (4h)	Edge	+20	+15	+1.20	+0.02	93
	Cloud	+80	+25	+2.50	+0.06	88
	Hybrid	+40	+20	+1.80	+0.04	95
Peak Load (100%)	Edge	+45	+32	+2.80	+0.07	87
	Cloud	+60	+40	+2.20	+0.04	83
	Hybrid	+30	+22	+1.90	+0.05	94

These findings highlight the superiority of hybrids in the dynamic mid-market environments, which will be the focus of the discussion that follows.

5. Discussion

The findings of this research are interesting in that they have given the research questions a direct answer in terms of how data center structures have contributed to the efficiency of the supply chains. First, about the effect on the planning-cycle latency and safety-stock requirements, the simulations show that edge architectures can dramatically shorten latency by 40-60% over cloud models, allowing responses to dynamic environments to be quicker. Nevertheless, hybrids are seen to be the best option, with a 30-50 percent reduction in latency under mixed loading and with a 20-30 percent reduction in safety-stock levels by adapting the workload. This matches the second question on effects under stress: hybrids ensured service levels of more than 95% during spikes in demand to 50% and network delays, whereas edge (an increase in latency of 35-45% at most) and cloud (an increase in latency of 80% or more) performed worse. Further trade-offs in the cost-to-serve and carbon footprint analysis show that under stable conditions, cloud minimizes emissions (0.15 kg CO_2e/order) because of good scaling, whereas edge minimizes transport emissions by 15 percent by making localized decisions at the expense of more local energy. Hybrids strike a balance between these, as they reach 18% reduced costs in total. Lastly, these quantifications can be used to inform policy concerns on SMEs through identification of how architectural decisions can alleviate digital adoption barriers in resource-constrained environments. Generally, the results highlight that hybrids have better trade-offs for SMEs having to trade off costs, speed, and sustainability, with latency-to-value curves reporting 20-22% of inventory cuts at latencies below 150 ms.

Theoretically, this study builds on the theory of supply chain management by adding the compute architecture to the conventional latency-inventory theories. The traditional models, including the variants of economic order quantity (EDS), have emphasized the variability of demand and lead times but have not taken into account computational infrastructure as a moderator (Rajnikant and Khanna, 2024). Quantifying the impact of compute proximity will improve the models, such as those in resilience theory, where latency is a constraint to real-time optimization. A case in point, the addition of architecture types to stochastic inventory models would also demonstrate non-linearities, e.g., in our latency-to-value curves—an additional 100 ms of latency would offer marginal payoff, making linear assumptions of previous literature hard to hold. This adds to the literature on Industry 4.0 by connecting operations research to computer science, suggesting a single framework where the design of data centers determines the resilience of the system as a whole. Patterns of architectural reference (e.g., edge-dominant V routes) are scalable patterns, which enhance theoretical debates about hybrid systems in volatile, uncertain, complex, and ambiguous (VUCA) environments.

In practice, the implications would inform mid-market manufacturers in the choice of architecture that is directed towards operations. In time-sensitive processes, e.g., perishable goods logistics or just-in-time manufacturing, edge-dominant patterns are advised, which use low-latency processing to reduce the planning delays and safety stock by up to 28%. On the other hand, analytics that are data-intensive, such as long-term forecasting with large volumes of data, are better served by a cloud-based configuration that supports scalability and reduces the emissions of the baseline. Hybrids are better in mixed situations, such as SMEs with changing demand, so that dynamic can be adopted as cost-to-serve (e.g., $12/order) and service levels (>96% under disruptions) can be maximized. The derived patterns and curves can be utilized by managers as the decision tool: evaluate the current latency (e.g., using benchmarking tools) and compare it to the value curves to justify the investments. To implement it, begin with pilot hybrids that combine existing on-premise hardware with cloud APIs, which may provide 15-20% efficiency improvements without a complete overhaul.

Policy implications are centered around actions to accelerate the process of adoption of digital SMEs (Omowole et al., 2024). Governments and other industry organizations should make it a high priority to subsidize hybrid arrangements, particularly for mid-market organizations, whereby the adoption of the same is still minimal due to the high initial cost—e.g., grants for 30–50% of orchestration software. Design training needs to be carbon efficient, and SMEs are taught energy-sensitive tools like energy-conscious scheduling to minimize footprints (i.e., shifting non-urgent activities to times of off-peak clouds). The digital supply chain standards may need to have interoperability standards such as minimum latency of AI integrations as well as ecosystems where SMEs collaborate with providers (Khan et al., 2023). In the U.S., similar policies to the CHIPS Act would be expanded to data center incentives to encourage domestic hybrid infrastructures to increase the resilience to geopolitical disruptions (Serentschy, 2025). These proposals will help to democratize the benefits of AI, minimizing the digital gap where big corporations will outperform SMEs in the adoption numbers.

The limitations have to be admitted in order to put findings into perspective. The emphasis on mid-market manufacturers in the U.S. restricts the generalizability of the research, as cultural, regulatory, or infrastructural variations in other regions such as Europe or Asia could change the results, including different grid carbon intensities impacting footprints. Though powerful, simulations contain biases due to assumptions such as linear energy conversion and idealistic network conditions, which might underreport variabilities in the real world (e.g., cybersecurity threats or hardware failures). The lack of data in anonymized datasets can also lead to biased outcomes where average cases prevail and extreme SME cases are ignored.

Future studies must seek real-world case studies and substantiate simulations using longitudinal pilot projects within different industries such as automotive or retail. The multi-tier supply chains may be extended to include the integration of emerging technologies that might include blockchain to share data safely or IoT to increase sensor inputs (Najjar et al., 2023). Another area to research is AI ethics in architectures, such as data privacy in edges, and to quantify long-term ROI in the face of changing policies such as carbon taxes.

6. Conclusion

In conclusion, this paper sheds light on the disruptive possibilities of data center design in supply chain management among U.S. mid-market manufacturers. Some of the key findings are that edge architectures outperform cloud models by 40-60% in planning latency with rapid responses in volatile situations, while hybrids present a balanced advantage, such as reducing safety-stock requirements by 20-30% and offering a safer-than-stress level of >95 under an initial load of 50-percent demand spikes. Clouds reduce carbon footprints during stable conditions (0.15 kg CO_2e/order), yet Edge lowers transport emissions by 15 percent. Deliverables consist of four architectural reference patterns—edge-dominant (high variability routes) and hybrid-adaptive (mixed loads)—and latency-to-value curves that show diminishing returns (e.g., a 100 ms latency would produce 15-22% inventory reductions). These figures illustrate the advantages of hybrids to SMEs, which are optimizing the cost-to-serve of $12/order and making them more resilient.

The broader impacts are on the establishment of cost-effective and sustainable supply chains. With the addition of the proximity of compute, optimized data centers minimize the impacts of pandemics or political unrest and lower the operation cost, emissions, and rates (Ewin et al., 2023). This not only assists in enhancements of inventory efficiency but also supplements environmental goals in reducing overall carbon footprints by intelligent energy allocation and local processing.

We recommend SMEs take charge of the compute proximity in their digital strategies by initiating pilots of hybrid to reap such advantages without introducing extensive changes. Policymakers should promote this through subsidies and standards in order to achieve quick adoption.

Lastly, these observations can be associated with the global trends in AI production, where the center of Industry 4.0 is data centers. As AI becomes more integrated, the aspect of architectural optimization will be a priority factor in order to have competitive, resilient, and sustainable operations all over the world.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Bendhi, M. R. (2023). Inventory Management: Machine Learning Predicts Demand, Reducing Excess Stock by up to 20%. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(2), 67-83.

[2] Di Salvo, A. L., Agostinho, F., Almeida, C. M., and Giannetti, B. F. (2017). Can cloud computing be labeled as “green”? Insights under an environmental accounting perspective. Renewable and Sustainable Energy Reviews, 69, 514-526.

[3] Dwyer, N. (2025). Innovative Approaches to a More Sustainable Global Supply Chain: A Phenomenological Study in the Footwear Industry (Doctoral dissertation, Pepperdine University).

[4] Ewim, D. R. E., Ninduwezuor-Ehiobu, N., Orikpete, O. F., Egbokhaebho, B. A., Fawole, A. A., and Onunka, C. (2023). Impact of data centers on climate change: a review of energy efficient strategies. The Journal of Engineering and Exact Sciences, 9(6), 16397-01e.

[5] Gusev, M., and Dustdar, S. (2018). Going back to the roots—the evolution of edge computing, an iot perspective. IEEE Internet Computing, 22(2), 5-15.

[6] Haq, U. N., Khan, M. M. R., Khan, A. M., Hasanuzzaman, M., and Hossain, M. R. (2025). Global initiatives for industry 4.0 implementation and progress within the textile and apparel manufacturing sector: a comprehensive review. International Journal of Computer Integrated Manufacturing, 1-26.

[7] Irshad, A. (2024). Latency Optimization in Edge vs. Cloud Computing: A Comparative Study for Real-Time Applications. International Journal of Business and Computational Science, 1(1).

[8] Kang, P. S., and Bhawna, B. (2025). Enhancing supply chain resilience through supervised machine learning: supplier performance analysis and risk profiling for a multi-class classification problem. Business Process Management Journal.

[9] Kahiluoto, H., Mäkinen, H., and Kaseva, J. (2020). Supplying resilience through assessing diversity of responses to disruption. International Journal of Operations and Production Management, 40(3), 271-292.

[10] Katsaliaki, K., Galetsi, P., and Kumar, S. (2022). Supply chain disruptions and resilience: a major review and future research agenda. Annals of operations research, 319(1), 965-1002.

[11] Khan, A. A., Laghari, A. A., Li, P., Dootio, M. A., and Karim, S. (2023). The collaborative role of blockchain, artificial intelligence, and industrial internet of things in digitalization of small and medium-size enterprises. Scientific Reports, 13(1), 1656.

[12] Khanyi, M. B., Xaba, S. N., Mlotshwa, N. A., Thango, B., and Matshaka, L. (2024). A Roadmap to Systematic Review: Evaluating the Role of Data Networks and Application Programming Interfaces in Enhancing Operational Efficiency in Small and Medium Enterprises. Sustainability, 16(23), 10192.

[13] Khokrale, R. (2025). The Role of AI in Supply Chain Optimization: Enhancing Efficiency through Predictive Analytics. Journal of Procurement and Supply Chain Management, 4(2), 55-75.

[14] Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., and Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Future Internet, 15(8), 255.

[15] Mathur, P. (2024). Cloud computing infrastructure, platforms, and software for scientific research. High Performance Computing in Biomimetics: Modeling, Architecture and Applications, 89-127.

[16] Najjar, M., Alsurakji, I. H., El-Qanni, A., and Nour, A. I. (2023). The role of blockchain technology in the integration of sustainability practices across multi-tier supply networks: Implications and potential complexities. Journal of Sustainable Finance and Investment, 13(1), 744-762.

[17] Narayana, S. S. (2024). Synergizing edge computing and cloud integration for enhanced efficiency in the oil and gas industry. J. Comput. Eng. Technol, 15, 63-71.

[18] Omowole, B. M., Olufemi-Philips, A. Q., Ofadile, O. C., Eyo-Udo, N. L., and Ewim, S. E. (2024). Barriers and drivers of digital transformation in SMEs: A conceptual analysis. International Journal of Frontline Research in Multidisciplinary Studies, 5(2), 019-036.

[19] Rajnikant, P. N., and Khanna, R. (2024, September). Incorporating Lead Time into the EOQ Model: A New Approach for Improved Inventory Management. In World Conference on Information Systems for Business Management (pp. 365-394). Singapore: Springer Nature Singapore.

[20] Rane, N., Choudhary, S., and Rane, J. (2024). Artificial intelligence for enhancing resilience. Journal of Applied Artificial Intelligence, 5(2), 1-33.

[21] Ross, A., Jayaraman, V., and Robinson, P. (2007). Optimizing 3PL service delivery using a cost-to-serve and action research framework. International journal of production research, 45(1), 83-101.

[22] Saeed, K. A., Malhotra, M. K., and Abdinnour, S. (2019). How supply chain architecture and product architecture impact firm performance: An empirical examination. Journal of purchasing and supply management, 25(1), 40-52.

[23] Sánchez, E., Calderón, R., and Herrera, F. (2025). Artificial Intelligence Adoption in SMEs: Survey Based on TOE–DOI Framework, Primary Methodology and Challenges. Applied Sciences, 15(12), 6465.

[24] Serentschy, G. (2025). Resilience and Security of Digital Infrastructure A Policy-Oriented Overview. Policy, 8, 4.

[25] Sivankutty, S., and Elahi, B. (2024). Proposing a new dynamic safety stock adjustment system: Balancing service level targets and financial constraints. In 9th North American Conference on Industrial Engineering and Operations Management, Washington DC, USA.

[26] Solanke, A. (2023). Edge Computing Integration with Enterprise Cloud Systems: Architectural Patterns for Distributed Intelligence. International Journal Of Engineering And Computer Science, 12(03).

[27] Syed, Z. A., Dapaah, E., Mapfaza, G., Remias, T., and Mupa, M. N. (2024). Enhancing supply chain resilience with cloud-based ERP systems. Iconic Research and Engineering Journals, 8(2).

[28] Timilehin, O. (2024). Performance Engineering for Hybrid Multi-Cloud Architectures: Strategies, Challenges, and Best Practices.

[29] Varghese, B., and Buyya, R. (2018). Next generation cloud computing: New trends and research directions. Future generation computer systems, 79, 849-861.

[30] Veeramachaneni, V. (2025). Edge computing: Architecture, applications, and future challenges in a decentralized era. Recent trends in computer graphics and multimedia technology, 7(1), 8-23.

[31] Yuen, T. M., and Baskaran, S. (2024). Optimizing Sustainable Business Performance: The Role of SME Agility in Digitalization. International Journal of Academic Research in Business and Social Sciences, 14(1), 380-389.