RAIS RESEARCH ASSOCIATION for INTERDISCIPLINARY STUDIES

DOI:10.5281/zenodo.22801775

The Network-to-Model Continuum: A New Framework for Enterprise Generative AI Success

Kathleen M. McMackin[1], Atif Farid Mohammad[2]

[1]*Capitol Technology University, Laurel, MD, USA, kmcmackin@captechu.edu*
[2]*Capitol Technology University, Laurel, MD, USA, afmohammad@captechu.edu*

Abstract: As enterprises transition Generative AI (GenAI) from experimental pilots to production-scale services, a significant performance bottleneck has emerged. This bottleneck is commonly termed the "Network Wall." While existing research focuses primarily on model optimization and data quality, this paper introduces the **Network-to-Model Continuum**. This concept provides a holistic framework that treats the end-to-end data path as an integral component of the GenAI's functional architecture. The continuum encompasses three specific segments of enterprise connectivity, which include the user's edge device, the wide-area network, and the high-performance compute fabric. Through a pilot study of four exemplar enterprise cases, this research analyzes how performance variances across this continuum dictate the success or failure of GenAI initiatives. The findings demonstrate that end-to-end network maturity is a non-compensatory necessary condition for success. The results indicate that even high-quality data fails to deliver value when the continuum is compromised. Furthermore, the study identifies a potential failure mode termed **Network-Induced Hallucination**. In this phenomenon, latency and jitter in the data path result in model inaccuracies. This paper provides a preliminary diagnostic framework for researchers and practitioners to identify and mitigate infrastructure-based bottlenecks. These insights ensure more reliable and scalable enterprise GenAI deployments.

Keywords: Generative AI, Network-to-Model Continuum, TOE Framework, Network Maturity, Network-Induced Hallucinations, Enterprise Infrastructure, Retrieval-Augmented Generation (RAG)

Introduction

The rapid integration of Generative Artificial Intelligence (GenAI) into the enterprise ecosystem has created a new frontier for organizational productivity. However, as these systems move from isolated pilots to global production, a "Great Decoupling" has occurred. While AI compute power has grown exponentially, the network infrastructure required to transport the necessary data has only evolved linearly (Brynjolfsson & McAfee, 2014). This disparity has given rise to the "Network Wall," a threshold where system performance is throttled not by the model's intelligence, but by the network's inability to sustain the required data flow (IEEE, 2025).

Despite the critical nature of this infrastructure, many organizations suffer from "Infrastructure Blindness." This is a phenomenon where strategic focus remains on model selection while neglecting the end-to-end (E2E) connectivity that serves as the AI's nervous system. According to the *Cisco AI Readiness Index* (2025), only 5% of organizations possess the network maturity required to sustain these workloads. Furthermore, Gartner (2024) predicts that 30% of GenAI projects will be abandoned after the pilot phase due to unforeseen technical complexities. This paper argues that these complexities are failures within the Network-to-Model Continuum. This study defines the continuum as the holistic, bi-directional data path from the user's edge device to the AI model's computational weights.

Theoretical Foundation: Expanding the "T" in TOE

To analyze the complexities of enterprise GenAI adoption, this study integrates two robust theoretical lenses: the Technology-Organization-Environment (TOE) framework and Information Processing Theory (IPT).

The TOE Framework

The TOE framework, first articulated by Tornatzky and Fleischer (1990), posits that the adoption and implementation of technological innovations are influenced by three distinct contexts. The Technological Context describes the internal and external technologies relevant to the firm; the Organizational Context refers to measures such as firm size and the quality of human resources; and the Environmental Context encompasses the industry and regulatory arena. While TOE is a standard for studying cloud computing (Baker, 2012), it has historically suffered from "Infrastructure Blindness." This research extends TOE by asserting that the network is a dynamic and critical determinant of the technology's actual performance.

Information Processing Theory (IPT)

To understand the mechanics of how the network influences GenAI, the study employs IPT. Originally developed by Galbraith (1974), IPT suggests that success is achieved when an organization's Information Processing Capacity (IPC) meets the Information Processing Requirements (IPR) of its tasks. In the GenAI era, IPC is the network's ability to move data with near-zero latency. As Mithas et al. (2024) argue, the modern GenAI-driven firm must treat its infrastructure as a core component of its processing capacity. This research therefore proposes expanding the "T" in TOE to encompass the entire Network-to-Model Continuum, viewing the network as an active participant in the AI's cognitive process rather than a passive utility.

The Network-to-Model Continuum

In traditional Information Technology (IT) research, the network and the application have historically been treated as discrete, independent entities. The network was viewed as a passive utility—a "pipe" whose primary requirement was to remain operational (Kurose & Ross, 2021). However, the architectural demands of Generative AI (GenAI) have rendered this siloed perspective obsolete. This paper proposes the **Network-to-Model Continuum** as a holistic unit of analysis. This continuum represents the integrated, bi-directional data path that connects the user's intent at the edge to the GenAI model in the cloud or data center.

The "Continuum" concept is rooted in the realization that in a GenAI environment, the network functions as an extension of the model's cognitive process. Unlike traditional web applications where data is requested and received in linear bursts, GenAI requires a constant, high-speed "chatter" between the model, the user, and various distributed databases. If any segment of this path experiences high latency or packet loss, the model's ability to process and distribute information is compromised. Therefore, the network is not merely a support structure; it is an integral component of the AI's functional architecture. This study suggests that the "intelligence" of the system is not located solely in the model but is distributed across the entire path. A failure at any point in the continuum results in a failure of the output.

To provide a granular framework for analysis, the Network-to-Model Continuum is divided into three distinct segments, each with unique technical requirements and potential failure points.

The First Mile: The Edge and Campus

The "First Mile" is the user's point of entry, typically involving Wi-Fi 7, 5G, or high-speed Ethernet. In this segment, the primary challenge is Jitter (variance in latency). For voice-based GenAI or real-time vision systems, even minor jitter can corrupt the data packet before it reaches the model, leading to inaccurate outputs. Instability in the First Mile leads to "fragmented prompts," where the model receives a distorted representation of the user's intent, resulting in erroneous outputs. In high-stakes environments like healthcare, Shah et al. (2024) demonstrate that flawless edge connectivity is required to ensure that ambient data reaches the model without corruption or delay.

The Middle Mile: The Transport and WAN

The "Middle Mile" is the bridge between the enterprise perimeter and the computational core. This segment relies on Software-Defined Wide Area Networks (SD-WAN) and dedicated Cloud On-ramps (e.g., Azure ExpressRoute). The critical variable here is **Path Determinism**. Because GenAI traffic is "bursty" and high-volume, it often competes with standard business traffic. Without high-maturity transport layers, the Middle Mile becomes the primary site of "Inference Lag," where the delay in data transit results in a degraded user experience and low adoption rates. Managing this path is essential for maintaining the life cycle of enterprise copilots (Zhu et al., 2024). Without a deterministic transport layer, the continuous feedback loops required for AI performance and model evolution collapse.

The Last Mile: The Compute and Fabric

The "Last Mile" occurs within the data center or cloud environment where the AI model is hosted. This is the site of the **"Network Wall,"** where the exponential growth of GPU performance has outpaced the linear growth of network bandwidth (IEEE, 2025). In this segment, even a 1% packet loss can be catastrophic. High-performance fabrics, such as RDMA over Converged Ethernet (RoCE v2), are required to ensure "Lossless" delivery of data between the GPU cluster and the vector databases (NVIDIA, 2025). As demonstrated by the "AI PB" system (Park et al., 2024), "grounding" the model in real-time data is a high-bandwidth task. Furthermore, the integration of Knowledge Graphs (Xu et al., 2024) increases the "chatter" across the network, making a lossless fabric (IEEE, 2025) a prerequisite for accuracy.

The Interdependence of the Continuum

The Network-to-Model Continuum operates on a "Weakest Link" principle. An organization may possess "AI-Native" maturity in its data center (Last Mile) but remain "Legacy" at the branch (First Mile). Because the AI application relies on the entire path, the resulting system performance will be capped by the least mature segment. This holistic view allows researchers to identify why "Data-Rich" organizations often fail; they have focused on the data at the end of the pipe while ignoring the instabilities of the pipe itself. By managing the continuum as a single system, enterprises can mitigate the risk of Network-Induced Hallucinations and achieve the "Zero Friction" state.

Methodology

This study employs a qualitative multi-case study design (Yin, 2018). This paper presents the results of an initial pilot study consisting of four "polar type" cases (S1, S2, F1, F2) selected from a planned larger study of 20 organizations (Eisenhardt, 1989). The pilot cases consist of two "Pacesetters" (Successes) and two "Laggards" (Failures). Data was extracted from publicly available primary artifacts, including 10-K filings, technical white papers, and legal rulings. The analysis utilizes Cross-Case Pattern Matching to identify where "breaks" occurred along the Network-to-Model Continuum.

To evaluate the cases, the researcher categorized the end-to-end network of each organization based on its technical capabilities. Networks were classified as **"Legacy"** if they relied on best-effort connectivity and lacked real-time observability; **"Reactive"** if they possessed basic cloud-integration but lacked optimization for AI; and **"AI-Native"** if they utilized dedicated, high-speed interconnects, edge computing, and automated telemetry to support the specific demands of the AI workload.

Weakest Link System Performance Model

The Network-to-Model Continuum operates on a "Weakest Link" principle, a concept rooted in reliability engineering (Weibull, 1951) and the Theory of Constraints (Goldratt, 1984). In an enterprise environment, an organization may achieve GenAI-Native maturity in its data center yet remain at a "Legacy" level at the branch office. Because every GenAI transaction must traverse the entire continuum, the total system performance is capped by the least capable segment.

.

Formal Definition

Let each segment performance metric $P_i(t) \in [0, 1]$, where $\in \{FM, MM, LM\}$. The system performance S(t) is defined as:

$$S(t) = \min \{P_{FM}(t),\ P_{MM}(t),\ P_{LM}(t)\}$$

where $S(t) = 1$ represents the Zero-Friction state; $S(t) = 0$ represents total system failure.

Corollary — Bottleneck Identification

The bottleneck segment at time *t* is the unique element:

$$i^*(t) = \arg\min \{P_i(t)\}, \quad i \in \{FM, MM, LM\}$$

All remediation effort should be directed at segment $i^*(t)$ to maximize marginal improvement in $S(t)$.

First Mile Prompt Fidelity Function

The First Mile (edge, Wi-Fi 7, 5G) is primarily impaired by jitter, the variance in latency. Jitter corrupts the data packet before it reaches the model, producing “fragmented prompts”; a phenomenon where network-level instabilities cause a distorted representation of user intent (Zhang et al, 2024). The Prompt Fidelity function Φ(J) quantifies the quality of the signal arriving at the model gateway.

Formal Definition

$$\Phi(J) = \max\{0, 1 - (J / J_{max})^{\alpha}\}$$

where J is observed jitter (ms), J_{max} is the maximum tolerable jitter, $\alpha > 0$ is the sensitivity exponent.

For $J = 0$: $\Phi(0) = 1$ (perfect fidelity); For $J \geq J_{max}$: $\Phi = 0$ (total prompt corruption).

Fragmented Prompt Condition

A Fragmented Prompt event FP occurs when fidelity falls below a minimum acceptable threshold Φmin:

$$FP = \{\Phi(J) < \Phi_{min}\}$$

Model Accuracy Degradation

Drawing on Information Processing Theory (Galbraith, 1974), this model posits that the "Information Processing Capacity" of the network directly dictates the accuracy of the output. Given that prompt fidelity directly affects model output accuracy A, the following monotone relationship holds:

$$A(\Phi) = A_{max} \cdot \Phi^{\beta}$$

where $Amax$ is the accuracy under perfect connectivity and $\beta > 0$ controls degradation steepness.

When $\beta = 1$, accuracy decays linearly with fidelity. Higher β implies sharper cliff degradation.

Results and Discussions: The Continuum in Practice

The analysis of the four pilot cases demonstrates that the success of a GenAI implementation is contingent upon the stability of the entire data path. A summary of the results is shown in Table 1.

Table 1: Comparative Analysis of the Network-to-Model Continuum

Case Study	Primary Segment Focus	Network Infrastructure Status	Implementation Outcome
S1: Morgan Stanley	Middle Mile (WAN/Cloud)	**AI-Native** (Private Fabric)	**Success**: 98% Adoption
S2: Walmart	First Mile (Edge)	**AI-Native** (Edge Compute)	**Success**: Global Scale
F1: McDonald's	First Mile (Edge)	**Legacy** (Standard Wi-Fi)	**Failure**: Project Terminated
F2: Air Canada	Last Mile (Sync)	**Reactive** (Passive Pipe)	**Failure**: Legal Liability

Successes: Achieving "Zero Friction"

Morgan Stanley (S1) achieved 98% adoption by ensuring a deterministic "Middle Mile." By utilizing private interconnects, the organization bypassed the public internet, ensuring that complex RAG queries did not time out (OpenAI, 2024). Walmart (S2) bypassed WAN bottlenecks by processing AI at the "First Mile" edge in over 10,000 stores (Walmart Global Tech, 2021), effectively shortening the continuum and ensuring real-time responsiveness.

Failures: The Cost of a Broken Continuum

McDonald's (F1) failed because the "First Mile" was unstable, leading to corrupted inputs (AIID, 2024). High jitter at the drive-thru edge resulted in the model providing inaccurate responses. Air Canada (F2) suffered a "Last Mile" failure where the network failed to sync the chatbot with the "Source of Truth" in real-time. The resulting "hallucinated" policy led to a court ruling of "negligent misrepresentation" (*Moffatt v. Air Canada*, 2024).

Discussion: The Continuum and Network-Induced Hallucinations

The evidence suggests the continuum follows a "Weakest Link" logic. A failure in any one segment renders the entire system ineffective. Furthermore, this research identifies an inherent failure mode termed **Network-Induced Hallucination**. Unlike traditional stochastic hallucinations, these errors occur when the network fails to deliver "grounding" data within the model's required temporal window (Zhang et al., 2024). This suggests that "Data Quality" is a secondary concern to "Data Delivery." If the network cannot guarantee the timely delivery of data, the quality of that data becomes irrelevant to the final output. This finding suggests that the "Technological Context" in the TOE framework should be expanded to include the end-to-end health of the data path.

Limitations and Future Work

A significant hurdle in the study of enterprise GenAI infrastructure is the proprietary nature of specific network performance metrics. Corporations rarely publish granular data regarding internal latency, jitter, or packet loss ratios. Consequently, this study relies on "Technical Proxies." Much of the available data is triangulated from multiple sources and direct access may provide additional data points.

There is a natural tendency for organizations to publicize successful projects while obscuring or minimizing failures. This "survivorship bias" in available literature can make it more difficult to find granular technical data for the "Laggard" cases compared to the "Pacesetter" cases. This study sought to mitigate this by searching legal databases (Nexis Uni) and the AI Incident Database to find objective, third-party accounts of failures that might not be found in traditional corporate reporting.

While this paper establishes the theoretical foundation of the continuum, further research is required to provide a quantitative diagnostic tool. The next phase of this research will expand this study to a full 20-case sample and introduce the T-GOME Maturity Scale to provide a diagnostic roadmap for the Network-to-Model Continuum.

Conclusions

This research has established the Network-to-Model Continuum as a vital framework for understanding the success and failure of enterprise GenAI. By analyzing four exemplar cases, this study has demonstrated that the network is not merely a background utility but a primary determinant of AI efficacy. The findings suggest that "Infrastructure Blindness" is a significant risk factor that leads to the "Network Wall," where even the most advanced models fail due to transport-layer instabilities. The identification of Network-Induced Hallucinations provides a new technical lens through which to view model inaccuracies, shifting the responsibility for groundedness from the data alone to the entire end-to-end data path.

To apply this framework, practitioners and organizations must transition from a model-centric to a continuum-centric deployment strategy. Organizations can utilize the Network-to-Model Continuum as a diagnostic roadmap by performing systematic audits across the three miles of connectivity. In the First Mile, IT leaders must ensure that edge environments provide the low-jitter stability required for high-fidelity data capture. Along the Middle Mile, organizations should prioritize the use of software-defined wide-area networks and dedicated cloud on-ramps to maintain a deterministic path for AI traffic. Finally, in the Last Mile, architects must deploy lossless fabrics to prevent the performance collapse known as the Network Wall. By implementing end-to-end observability tools, organizations can gain the necessary visibility to monitor the health of this continuum in real-time, thereby mitigating the risk of network-induced errors and ensuring a sustainable return on investment.

References

Artificial Intelligence Incident Database. (2024). *Incident 475: McDonald's IBM AI drive-thru errors*. https://incidentdatabase.ai/cite/475

Baker, J. (2012). The technology-organization-environment framework. In Y. K. Dwivedi, M. R. Wade, & S. L. Schneberger (Eds.), *Information Systems Theory* (pp. 231–245). Springer. https://doi.org/10.1007/978-1-4419-6108-2_12

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age*: Work, progress, and prosperity in a time of brilliant technologies. W. W. Norton & Company.

Cisco Systems. (2025). *Cisco AI readiness index 2025: Realizing the value of AI*. https://www.cisco.com/c/m/en_us/solutions/ai/readiness-index.html

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Galbraith, J.R. (1974). Organization design: An information processing view. *Interfaces, 4*(3), 28-36. https://doi.org/10.1287/inte.4.3.28

Gartner. (2024, July 29). *Gartner predicts at least 30% of generative AI projects will be abandoned*. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

Goldratt, E.M., & Cox, J. (1984). *The goal: A process of ongoing improvement.* North River Press.

IEEE Standards Association. (2025). *Standard for high-performance interconnects in distributed AI systems (IEEE P3185)*.

Mithas, S., Chen, Z., & Tafti, A. (2024). Information processing and the AI-driven firm. *MIS Quarterly*, *48*(1), 125–144.

Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII). https://canlii.ca/t/k2v9z

OpenAI. (2024). *Morgan Stanley uses AI evals to shape the future of financial services*. https://openai.com/index/morgan-stanley/

NVIDIA. (2025, December). *NVIDIA Spectrum-X white paper*. https://resources.nvidia.com/en-us-networking-ai/nvidia-spectrum-x

Park, D., Park, S., Hong, I., Lee, H., Park, J., Lee, S., ... & Loh, H. (2024). AI PB: A grounded generative agent for personalized investment insights. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Shah, N. H., Ambers, N., Pandya, A., Keyes, T., Banda, J. M., Nallan, S., ... & Pfeffer, M. A. (2024). Adoption and use of large language models at an academic medical center. *NEJM AI*, *1*(1).

Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington Books.

Walmart Global Tech. (2021, March 11). *Inside the network powering Walmart.* Walmart Global Tech Blog.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 18*(3), 293-297.

Yin, R. K. (2018). *Case study research and applications* (6th ed.). Sage Publications.

Zhang, Y., Li, X., & Wang, J. (2024). The impact of network jitter on LLM performance. *Journal of Network and Computer Applications*, *228*.

Zhu, Y., Demarne, M., Deng, K., Wang, W., Sahoo, N., Vermareddy, D., ... & Krishnan, S. (2024). ENCO: Life-cycle management of enterprise-grade copilots. *Proc. 46th Int. Conf. on Software Engineering*. https://doi.org/10.48550/arXiv.2412.06099