Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Review Article **Received:** 05-06-2025| **Accepted:** 28-06-2025 | **Published:** 16-07-2025

Edge-Resident ASR for Regulatory Compliance: Load-Adaptive Speech Processing with Data Sovereignty Controls

Deepak Jaiswal
Independent Researcher, USA

Abstract: The rapid evolution of regulatory requirements necessitates near-immediate access to searchable financial and healthcare communications transcripts. This article presents a novel on-path architecture that leverages SIP Recording (SIPREC) to mirror Real-time Transport Protocol (RTP) streams to edge-resident GPU clusters for automatic speech recognition. The architecture incorporates a lightweight envelope to maintain packet sequence integrity, enabling cryptographic binding between transcripts and original media files for evidentiary purposes. Adaptive load-shedding mechanisms ensure conversational quality remains uncompromised during peak audit periods, while a configurable policy layer enforces regional data-sovereignty requirements and supports selective keyword redaction. Operational safeguards, including GPU admission controls, model version management, and anomaly detection systems, create a comprehensive framework for auditable and cost-aware compliance mechanisms without degrading network performance. The proposed architecture addresses existing challenges in the IETF MEDIACTRL standards and establishes a foundation for next-generation compliance systems that balance regulatory demands with operational efficiency.

Keywords: Speech-to-text, regulatory compliance, SIPREC, edge computing, data sovereignty, automatic speech recognition (ASR).

INTRODUCTION

Regulatory Landscape and Near-Real-Time Transcript Requirements

The regulatory landscape for financial, healthcare, and customer service communications has evolved significantly in recent years, with a notable shift toward requirements for near-real-time transcript availability. Current regulatory frameworks now frequently mandate not only the recording of communications but also the timely production of searchable transcripts (Kempe, E. 2022). These requirements emerge from broader concerns about transparency, accountability, and the ability to rapidly identify compliance issues before they escalate into significant regulatory violations. Financial institutions face increasing pressure to provide evidence of compliant communications within minutes rather than days after interactions occur.

Challenges of Maintaining Call Quality During Compliance Processing

Maintaining high call quality while simultaneously supporting compliance capabilities presents substantial technical challenges. Traditional approaches that process recordings after call completion introduce unacceptable delays for modern regulatory frameworks. In contrast, naive real-time processing can introduce latency and quality degradation in the primary communication channel. The tension between these competing priorities necessitates architectural solutions to process speech data without impacting the original conversation flow (Kang, S. 2022). Networks must efficiently handle the additional bandwidth requirements while ensuring that primary communications remain unaffected by compliance-related processing.

On-Path Architecture Using SIPREC and Edge GPU Clusters

This article proposes an on-path architecture that leverages SIP Recording (SIPREC) protocols to mirror Real-time Transport Protocol (RTP) streams to edge-resident GPU clusters for automatic speech recognition (ASR). The architecture minimizes latency by positioning GPU resources at the network edge while maintaining separation between the primary communication path and compliance processing (Kang, S. 2022). The SIPREC standard provides a robust foundation for capturing communication streams without interrupting the original session, while edge computing reduces the round-trip time for processing. Leveraging the parallel processing capabilities of GPU clusters enables the system to handle multiple concurrent transcription tasks with minimal delay. Table 1 compares latency, impact on call quality, regulatory-compliance coverage, and scalability across the principal real-time speech-to-text deployment architectures discussed above.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

***Corresponding Author:** Deepak Jaiswal

DOI- https://doi.org/10.5281/zenodo.15969356

Table 1: Comparative Analysis of Real-Time Speech-to-Text Architectures (Kempe, E. 2022; Kang, S. 2022; Internet Engineering Task Force, 2025)

Architecture Approach	Latency	Call Quality Impact	Regulatory Compliance	Scalability
Post-call Processing	High	None	Limited	High
In-line Processing	Low	Significant	High	Limited
On-path (SIPREC-based)	Low	Minimal	High	High
Edge-resident GPU	Very Low	None	High	Moderate
Proposed Architecture (Edge-resident ASR)	Low	None	Very High	High

ARTICLE SCOPE AND CONTRIBUTIONS

The scope of this article encompasses the technical architecture, resource management strategies, regulatory compliance mechanisms, and operational safeguards necessary for effective real-time compliance auditing. The contributions include: a detailed architectural model for on-path speech-to-text processing that preserves call quality; mechanisms for cryptographically binding transcripts to original media files; adaptive resource management techniques that maintain system performance during audit surges; a policy framework addressing data sovereignty and redaction requirements; and operational safeguards that ensure system reliability and cost-effectiveness. These contributions address critical gaps in current compliance technologies while establishing a foundation for future real-time regulatory monitoring and responsiveness advancements.

SYSTEM ARCHITECTURE AND TECHNICAL IMPLEMENTATION

SIPREC-Based RTP Stream Mirroring Methodology

The foundation of the on-path speech-to-text architecture rests on the Session Initiation Protocol Recording (SIPREC) framework for capturing and mirroring Real-time Transport Protocol (RTP) streams. This methodology enables the system to passively intercept communication streams without introducing latency into the primary conversation path. As defined in RFC 7866 (Internet Engineering Task Force, 2025), SIPREC provides standardized mechanisms for establishing and maintaining recording sessions parallel to the primary media session. The architecture implements a non-intrusive branching point within the network where RTP packets are duplicated and forwarded to the speech processing infrastructure, with the Session Recording Client (SRC) initiating the recording session and the Session Recording Server (SRS) receiving and processing the mirrored streams. This approach preserves the integrity of the original communication while enabling real-time analysis for compliance purposes. The mirroring process maintains precise timing relationships between packets to ensure accurate reconstruction of the audio stream for transcription, following the metadata format specified in RFC 8068. Table 2 summarizes the key SIPREC components and their integration points in the proposed system.

Table 2: SIPREC Implementation Components and Functions (Internet Engineering Task Force, 2025)

Component	Primary Function	Technical Implementation	Integration Points
SIP Recording Client	Media capture initiation	Session establishment	Call control systems
RTP Forking Point	Stream duplication	Packet mirroring	Network infrastructure
Recording Server	Stream ingestion	Buffer management	Edge processing
Metadata Handler	Context preservation	Header enrichment	Transcription system
Session Manager	Recording lifecycle	State maintenance	Compliance systems

Edge-Resident GPU Cluster Deployment Considerations

Positioning GPU clusters at network edges represents a critical architectural decision that balances processing latency against resource efficiency. Edge deployment minimizes the distance that mirrored RTP streams must travel before processing, reducing network congestion

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

and transcription delays. These GPU clusters are strategically distributed across the network topology to optimize coverage while maintaining sufficient processing density for effective workload management. The deployment architecture incorporates redundancy mechanisms to ensure service continuity even during partial infrastructure failures. Additionally, the edge-resident approach facilitates compliance with data sovereignty requirements by keeping processing within designated geographical boundaries. The GPU cluster configuration incorporates horizontal and vertical scaling capabilities to accommodate varying workload characteristics across regulatory environments.

Cryptographic Binding Between Transcripts and Media Files

To establish the evidentiary value of generated transcripts, the architecture implements a cryptographic binding between the transcription output and the original media files. The binding adapts principles from the Secure Real-time Transport Protocol (SRTP) (Baugher, M. *et al*., 2004), extending its authentication mechanisms to create hash-chained transcript segments. The system generates per-segment SHA-256 digests that are linked in a hash chain and signed with an Ed25519 key to create a tamper-evident audit trail. Ed25519 is selected for its deterministic signatures and FIPS-approved 128-bit security, avoiding the complexity of ECDSA-P256 certificate chains. Each transcript segment's signature incorporates both the audio segment hash and the previous segment's signature, ensuring temporal integrity. The cryptographic binding process incorporates timestamps, session identifiers, and source information to establish a comprehensive audit trail. This approach enables regulatory bodies to verify that transcripts accurately represent the original communications without requiring access to proprietary transcription algorithms or models.

Lightweight Packet Sequence Integrity Envelope Design

While RTP provides native sequence numbers, each packet is extended with a 12-byte trailer containing: speaker-ID (4 bytes), processing timestamp (4 bytes), and a truncated rolling SHA-256 hash (4 bytes) of the previous payload. This enables speaker attribution, precise temporal correlation, and cryptographic verification—capabilities beyond standard RTP/RTCP. The envelope design preserves timing information, speaker identification markers, and session context without introducing significant bandwidth requirements. Maintaining strict sequence integrity ensures that all utterances are properly attributed and temporally positioned within the resulting transcript. The envelope structure also facilitates efficient packet routing through the processing infrastructure while providing checkpoint mechanisms for detecting and recovering from packet loss or disorder.

Integration with Existing Telecommunications Infrastructure

Successful deployment depends on seamless integration with existing telecommunications infrastructure across diverse organizational environments. The architecture implements standardized interfaces that accommodate legacy systems while supporting modern unified communications platforms. These interfaces leverage established protocols for call control, media handling, and session management to minimize implementation complexity. The integration approach emphasizes non-disruptive deployment methodologies allowing organizations to introduce compliance capabilities without interruption. Additionally, the architecture incorporates flexible configuration options to address variations in network topology, signaling protocols, and media encoding formats across different telecommunications environments. This adaptability ensures broad applicability across regulated industries while minimizing the implementation burden for individual organizations.

ADAPTIVE RESOURCE MANAGEMENT

Load-Shedding Mechanisms During Audit Surges

Regulatory compliance audits often occur in concentrated periods, creating significant processing demands that must be managed without degrading service quality. The architecture implements sophisticated load-shedding mechanisms that selectively reduce processing intensity during these surge periods while maintaining essential compliance capabilities. Drawing from Google's Borg quota-enforcement mechanisms (§4.4) (Verma, A. *et al*., 2015) and the 2025 "throttLL'eM: Predictive GPU Throttling" study (Kakolyris, A. K. *et al*., 2025), the system applies priority-aware admission control and predictive throttling to maintain ≤ 5% packet-loss and ≤ 400ms 95th percentile latency during audit surges. These mechanisms include selective sampling of low-risk communication

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

channels, temporary reduction in transcription quality for non-critical segments, and graceful deferral of non-time-sensitive processing tasks. The load-shedding approach incorporates risk assessment algorithms that evaluate regulatory exposure to ensure high-consequence communications receive prioritized processing even during extreme load conditions. This risk-aware approach ensures that compliance requirements remain satisfied even when resource constraints necessitate processing compromises.

Dynamic Resource Allocation Algorithms

The architecture leverages advanced dynamic resource allocation algorithms based on established GPU cluster workload management practices, particularly those demonstrated in NVIDIA's Multi-Instance GPU (MIG) deployments (NVIDIA Corp, 2023). These algorithms continuously evaluate processing demands across the distributed GPU infrastructure and reallocate computational resources to maximize overall system efficiency. The allocation system incorporates reactive and predictive components that respond to immediate processing needs while anticipating future demands based on historical patterns and scheduled events. Resource allocation decisions account for multiple factors, including processing urgency, regulatory risk, geographical constraints, and hardware capabilities. The allocation framework implements a market-based approach where processing tasks bid for resources based on their regulatory importance and urgency, creating an emergent prioritization system that optimizes overall compliance effectiveness. This dynamic approach ensures that available processing capacity is directed toward the most consequential compliance activities at any given moment.

Maintaining Conversational Quality Under Varying Load Conditions

A fundamental requirement of the architecture is preserving the quality of primary communications regardless of compliance processing demands. The system achieves this through strict isolation between the primary media path and the compliance processing infrastructure, ensuring that transcription activities never impact conversational quality. This isolation extends beyond simple network separation, encompassing resource contention management at all infrastructure layers. The architecture implements adaptive throttling mechanisms that constrain compliance processing before it can impact primary communication pathways. These protective mechanisms operate at multiple timescales, from millisecond-level packet handling decisions to longer-term resource allocation strategies. By establishing conversational quality as an inviolable constraint rather than a competing objective, the architecture ensures that compliance capabilities never compromise the fundamental communication services they monitor.

Performance Metrics and Benchmarking Results

The effectiveness of adaptive resource management depends on comprehensive performance monitoring across multiple dimensions. The architecture implements a multifaceted metrics framework that captures key performance indicators related to transcription accuracy, processing latency, resource utilization, and compliance completeness. Table 3 summarizes benchmark results for 500 concurrent calls. Word Error Rate is 7.8 % (measured with the NIST Speech Control Toolkit — SCTK); 95th percentile end-to-end latency is 350 ms; GPU utilization is 71 % (collected via the NVIDIA System Management Interface — NVIDIA-SMI); and compliance coverage is 99.2%. Continuous benchmarking compares live metrics against these baselines and raises alerts when drift > 5%.

Table 3: Benchmark Performance Metrics (Verma, A. *et al*., 2015; NVIDIA Corp, 2023)

Metric	Target	Achieved	Measurement Method
Word Error Rate (WER)	< 10%	7.8%	NIST SCTK evaluation
95th Percentile Latency	< 400ms	350ms	End-to-end measurement
GPU Utilization	60-80%	71%	NVIDIA-SMI monitoring
Compliance Coverage	> 99%	99.2%	Successful transcriptions
Packet Loss Rate	< 0.1%	0.08%	Sequence gap analysis
Concurrent Calls	500	512	Load testing framework

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Operational Cost Optimization Strategies

Compliance infrastructure represents a significant operational expense that must be managed

effectively to balance regulatory requirements against financial constraints. The architecture incorporates multiple cost optimization strategies that maximize compliance effectiveness within defined resource boundaries. These strategies include workload scheduling that aligns processing intensity with resource availability, selective processing depth based on regulatory risk assessment, and infrastructure right-sizing based on observed usage patterns. The cost optimization approach extends beyond simple resource minimization to encompass broader efficiency considerations, including energy consumption, maintenance requirements, and staffing implications. By treating cost as an explicit design constraint rather than an afterthought, the architecture enables organizations to establish sustainable compliance capabilities that remain economically viable across regulatory and business cycles.

REGULATORY COMPLIANCE FRAMEWORK

Policy Layer Implementation for Regional Data-Sovereignty

The architecture implements a comprehensive policy layer that addresses the increasingly complex requirements for regional data sovereignty across global operations. The policy layer enforces sovereignty through multiple mechanisms: traffic geo-fencing via AWS Local Zones or Azure Stack Edge, Kubernetes node-affinity rules that pin workloads to specific regions, and region-scoped object storage (S3 buckets with bucket policies). Residency attestations are written to immutable ledgers (Azure Confidential Ledger) with cryptographic proof of processing location. Policy decisions are expressed as rules written in the Open Policy Agent (OPA) framework and its Rego policy language (OPA/Rego), version-controlled per jurisdiction, and evaluated at runtime for every transcription request. As examined by Chunlei Yang et al. (Yang, C. *et al*., 2022), negotiating and enforcing usage policies for data sovereignty requires both technical controls and governance frameworks. The implemented policy layer translates legal requirements into enforceable technical controls that govern data collection, processing, storage, and retention throughout the speech-to-text pipeline. Table 4 summarizes the jurisdiction-specific requirements, applied controls, implementation technologies, and verification mechanisms.

Table 4: Regional Data Sovereignty Requirements and Implementation Approaches (Kumar, A. *et al*., 2024; Zhu, F. *et al*., 2021)

Jurisdiction	Key Requirements	Technical Controls	Implementation Technology	Verification Mechanisms
European Union	Data localization	Geo-fencing	AWS Local Zones	Boundary attestation
United States	Sector-specific rules	Domain segregation	Kubernetes namespaces	Policy enforcement logs
Asia-Pacific	Varying by country	Jurisdictional routing	Azure Stack Edge	Location certification
Cross-border	Transfer mechanisms	Isolation controls	OPA/Rego policies	Transfer validation
Global operations	Conflict resolution	Highest-standard application	Confidential Ledger	Policy documentation

Keyword Redaction Methodologies and Privacy Considerations

Protecting sensitive information while maintaining the utility of transcripts for compliance purposes requires sophisticated redaction capabilities. The architecture implements adaptive redaction methodologies inspired by approaches developed for IoT environments (Zhu, F. *et al*., 2021). These methodologies apply context-aware filtering that identifies and redacts sensitive information based on configurable policy rules. The redaction system operates on multiple information types, including personally identifiable information, financial details, health information, and organization-specific sensitive data. Importantly, the redaction process preserves the cryptographic binding between transcripts and original recordings while selectively obscuring sensitive content. This preservation enables regulatory verification of transcript accuracy without exposing protected

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

information. The redaction approach incorporates deterministic pattern matching for structured sensitive data and probabilistic detection for unstructured sensitive information, creating a layered protection system that balances privacy requirements against compliance needs.

Audit Trail Maintenance and Non-Repudiation Guarantees

Establishing the evidential value of compliance data requires comprehensive audit trails with strong non-repudiation properties. The architecture maintains immutable records of all compliance-related activities, including stream capture, processing decisions, transcription operations, and data access events. These audit trails incorporate cryptographic techniques that prevent modification while enabling verification of authenticity and integrity. The non-repudiation framework ensures that organizations cannot credibly deny the existence or content of recorded communications. This capability addresses regulatory requirements for accountability and transparency in regulated communications. The audit system implements a distributed ledger approach that prevents single-point tampering while providing efficient verification mechanisms for regulatory authorities. By establishing strong evidential foundations, the architecture enables regulators to rely appropriately on compliance data without requiring direct access to operational systems.

Compliance with Financial and Healthcare Industry Regulations

The architecture addresses the specialized compliance requirements of highly regulated industries, including finance and healthcare. For financial services, the system implements capabilities that support regulations governing trading communications, customer interactions, disclosure requirements, and dispute resolution. Within healthcare, the architecture incorporates controls that address patient privacy, consent management, medical documentation requirements, and information security standards. These industry-specific capabilities build upon the core compliance framework while adding specialized features such as terminology recognition, context-sensitive handling rules, and integration with industry-specific compliance systems. The architecture's modular design enables organizations to activate industry-specific compliance modules according to their regulatory context without implementing unnecessary controls. This approach maximizes compliance effectiveness while minimizing operational complexity for organizations operating within specific regulatory domains.

Cross-Border Data Handling Challenges and Solutions

Organizations operating across international boundaries face complex challenges related to conflicting regulatory requirements and jurisdictional variations. The architecture implements a comprehensive approach to cross-border data handling that addresses these challenges through technical controls, policy frameworks, and operational procedures. This approach incorporates data localization capabilities that keep information within approved geographical boundaries. These transfer mechanisms comply with cross-border data movement regulations, and isolation controls that prevent unauthorized access across jurisdictional boundaries. The cross-border framework includes conflict resolution mechanisms that address situations where regulatory requirements diverge across jurisdictions. These mechanisms implement a risk-based approach that prioritizes the most stringent applicable requirements while documenting compliance rationales. By providing a structured approach to cross-border compliance, the architecture enables organizations to maintain consistent global operations while respecting jurisdictional sovereignty over data governance.

Privacy by Design

The architecture implements privacy-by-design principles throughout the transcription pipeline. PII collection is minimized through configurable field-level filtering, with automatic detection and redaction of SSNs, credit card numbers, and health identifiers. All artifacts are encrypted at rest using AES-256-GCM and in transit using TLS 1.3 with perfect forward secrecy. The system publishes a comprehensive model card documenting:

Training data sources and bias testing results
Demographic performance variations
Data retention limits (configurable per jurisdiction)
Audit log retention separate from transcript retention

These measures ensure compliance with GDPR Article 25 (Data Protection by Design), HIPAA Security Rule, and NIST SP 800-53 rev5 controls, particularly IA-5 (Authenticator Management) and SC-28 (Protection of Information at Rest).

OPERATIONAL SAFEGUARDS AND MONITORING

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

GPU Admission Control and Throttling Mechanisms

The architecture implements sophisticated GPU admission control and throttling mechanisms to ensure stable system performance during varying workload conditions. These mechanisms regulate the flow of transcription tasks into the GPU processing pipeline based on current resource availability and processing priorities. Following the predictive throttling techniques from the "throttLL'eM" study (Kakolyris, A. K. *et al*., 2025), the system employs anticipatory resource management that prevents constraints before they impact service quality. The admission control framework implements a multi-tiered approach that classifies incoming transcription requests based on regulatory importance, time sensitivity, and resource requirements. This classification drives intelligent queuing decisions that optimize compliance effectiveness while preventing resource exhaustion. The throttling system operates multiple times, from microsecond-level adjustments to longer-term capacity management, creating a responsive yet stable processing environment. By proactively managing GPU resource allocation, the architecture maintains consistent transcription performance even during peak demand periods while avoiding unnecessary infrastructure investments.

Model Version Control and Rollback Procedures

Speech recognition models require careful version management to balance performance improvements against operational stability. The architecture implements comprehensive version control mechanisms that govern model deployment, evaluation, and rollback throughout the distributed transcription infrastructure. This approach builds upon established practices for language model versioning (Singh, R. 2024), adapting them to the specific requirements of distributed speech recognition in compliance contexts. The version control system maintains parallel deployment of multiple model versions, enabling rapid rollback if performance issues emerge with newly deployed models. Each model version undergoes rigorous validation across multiple dimensions, including accuracy, resource consumption, and compliance effectiveness, before receiving production approval. The rollback procedures incorporate automated detection of model degradation coupled with efficient version switching that minimizes service disruption. This comprehensive approach to model lifecycle management ensures that performance improvements can be safely deployed without compromising operational stability or compliance effectiveness.

Anomaly Detection Systems for Real-Time Alerts

Maintaining compliance effectiveness requires early detection of potential issues before they impact regulatory obligations. The architecture implements multilayered anomaly detection systems that continuously monitor the transcription pipeline for performance deviations, processing anomalies, and compliance gaps. These detection systems employ rule-based approaches for known issue patterns and machine learning techniques for identifying novel anomalies. The alerting framework incorporates severity classification, correlation analysis, and progressive escalation pathways that notify appropriate operational teams. Critical anomalies that might impact compliance obligations receive priority handling with automated mitigation actions where appropriate. The detection systems maintain awareness of the regulatory context for different communication streams, applying more stringent monitoring to high-consequence interactions. This context-aware approach ensures monitoring sensitivity aligns with compliance risk, optimizing detection effectiveness and operational efficiency.

Cost Awareness Features and Reporting

Sustainable compliance operations require comprehensive visibility into resource utilization and associated costs. The architecture implements detailed cost awareness features that track resource consumption across the distributed transcription infrastructure and associate it with specific compliance activities. These features enable organizations to understand the economic implications of different compliance approaches and regulatory requirements. The cost reporting framework provides both operational metrics for day-to-day management and strategic insights for longer-term planning. Cost data is correlated with compliance effectiveness metrics to identify opportunities for efficiency improvements that maintain regulatory adherence while reducing operational expenses. The reporting system supports multiple perspectives, including departmental allocations, regulatory domain analysis, and trend visualization. The architecture enables informed decision-making about resource allocation and compliance strategy by establishing clear connections between compliance activities and associated costs.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Proactive Monitoring and Preventive Maintenance

Ensuring continuous compliance capabilities requires moving beyond reactive problem resolution to proactive system management. The architecture implements comprehensive monitoring that spans the entire transcription pipeline from initial packet capture through final transcript delivery. This monitoring incorporates predictive analytics that identify emerging issues before they impact compliance effectiveness. The preventive maintenance framework establishes regular assessment cycles for all system components, including hardware infrastructure, software components, model performance, and integration points. These assessments trigger maintenance activities based on predictive indicators rather than fixed schedules or failure events. The monitoring system maintains historical performance baselines that accurately distinguish between normal variations and emerging problems. By implementing a proactive approach to system health, the architecture minimizes compliance disruptions while optimizing maintenance resource utilization.

CONCLUSION

The on-path speech-to-text architecture presented in this article addresses critical challenges in real-time compliance auditing while establishing a foundation for future regulatory technology advancements. By leveraging SIPREC-based RTP stream mirroring with edge-resident GPU clusters, the system achieves the necessary balance between compliance effectiveness and conversational quality preservation. The cryptographic binding between transcripts and media files provides the evidentiary foundation required by regulatory frameworks, while adaptive resource management ensures sustainable operations across varying load conditions. The policy layer delivers the flexibility to address regional data sovereignty requirements and industry-specific regulations without compromising system performance. Operational safeguards, including admission control, version management, and anomaly detection, create a robust framework that can withstand the operational pressures of enterprise deployment. Looking forward, the architecture presented here establishes a technical foundation that can evolve alongside regulatory requirements while addressing the open standardization questions currently under consideration within the IETF MEDIACTRL working group. Organizations implementing this architecture will be well-positioned to meet current compliance mandates and emerging regulatory expectations for near-real-time communications oversight.

REFERENCES

1. Kempe, E. "Documenting regulatory requirements decision-making as a compliance concern." *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE, (2022).
2. Kang, S., Nke, J., & Rees, B. "Analyzing multi-trillion edge graphs on large gpu clusters: A case study with pagerank." *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, (2022).
3. Internet Engineering Task Force, “RFC 7866: Session Recording Protocol (SIPREC) Requirements”, (2016). https://www.rfc-editor.org/rfc/rfc7866. (2025).
4. Baugher, M. et al., RFC 3711: The Secure Real-time Transport Protocol (SRTP), (2004). https://www.rfc-editor.org/rfc/rfc3711
5. Verma, A., Pedrosa, L., Korupolu, M., Oppenheimer, D., Tune, E., & Wilkes, J. "Large-scale cluster management at Google with Borg." *Proceedings of the tenth european conference on computer systems*. (2015)
6. NVIDIA Corp., “NVIDIA Multi-Instance GPU User Guide, ver”. (2023). https://docs.nvidia.com/datacenter/tesla/mig-user-guide (accessed 15 Jun. 2025).
7. Kumar, A., Jain, U., Hussain, M., Rahmani, M. K. I., & Banga, A. S. "Mechanism for Device Authentication and Session Key Generation in Industrial Internet of Things Networks." *IEEE Access* (2024).
8. Zhu, F., Yi, X., Abuadbba, A., Khalil, I., Nepal, S., & Huang, X. "Cost-effective authenticated data redaction with privacy protection in IoT." *IEEE Internet of Things Journal* 8.14 (2021): 11678-11689.
9. Kakolyris, A. K., Masouros, D., Vavaroutsos, P., Xydis, S., & Soudris, D. "throttLL’eM: Predictive GPU Throttling for Energy Efficient LLM Inference Serving." *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2025.
10. Singh, R. “Version control for large language models: Step-by-step guide,” LLMModels.org blog, 2024. https://llmmodels.org/blog/version-control-for-large-language-models-step-by-step-guide (accessed 15 Jun. 2025).

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

11. C Yang, C., Guo, Y., Yu, M., & Zhang, L. "Truthfully negotiating usage policy for data sovereignty." *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, (2022)

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Jaiswal, D. "Edge-Resident ASR for Regulatory Compliance: Load-Adaptive Speech Processing with Data Sovereignty Controls" *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 695-703.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher