Authority Governance for IT/OT Bridge Operations: A Reference Implementation for the Five Eyes Operational Technology Principles

Working Paper - BLADE-INFRA-OT

Burak Oktenli
Independent Researcher · Washington, DC, USA
ORCID: 0009-0001-8573-1667 · burak@authrex.systems
Version 1.0 · May 2026 · License: CC BY 4.0

Abstract

The IT/OT segmentation boundary is the most consequential interface in critical-infrastructure cybersecurity. Joint guidance from the Cybersecurity and Infrastructure Security Agency, the Australian Signals Directorate, and the National Security Agency, issued in December 2025 under the title *Principles for the Secure Integration of AI in Operational Technology*, called explicitly for hardware-enforced boundary controls and bounded authority delegation at this interface. The subsequent Five Eyes advisory of 1 May 2026, *Careful Adoption of Agentic AI Services*, extended this concern to autonomous agents acting on operator behalf. Within weeks, the Dragos threat-intelligence report on the Monterrey municipal water utility documented an adversary attempt to use AI-assisted tooling to pivot laterally from a compromised IT environment toward operational-technology control gateways - an attempt that failed at the segmentation boundary but that confirmed the threat-model the federal guidance had been warning about. This paper presents BLADE-INFRA-OT, a reference implementation of the joint principles: a bump-in-the-wire governance appliance that applies the AUTHREX eight-stage authority pipeline to every cross-boundary message at protocol-aware granularity. The paper specifies the architecture of the appliance, the allocation of the AUTHREX pipeline between governance and network planes, the audit-ledger schema designed for NERC CIP-008 evidentiary use, and the simulation methodology against the Monterrey threat pattern. The contribution is methodological and reproducible: a documented design, an open simulation, and an open Interface Control Document, released under CC BY 4.0.

Keywords: operational technology, IT/OT bridge, critical infrastructure protection, NERC CIP, IEC 62443, agentic AI, authority governance, AUTHREX, BLADE-INFRA-OT.

Status: Fundamental research, openly published. Hardware TRL 2–3; simulation TRL 3–4. No penetration testing performed; no operational utility is targeted; the simulation is scripted against documented threat patterns.

1 Introduction

The Purdue Reference Model for industrial control systems organized manufacturing, utility, and process-control networks into vertically stratified layers, with the corporate IT zone at the top and the field instrumentation zone at the bottom. Implicit in the model was the assumption that an attacker reaching one layer would be constrained in their ability to act on the next. This assumption was historically defended by network segmentation alone: an unannounced traffic rule between Levels 4 and 3, often instantiated as a unidirectional gateway or a firewall pair, was considered sufficient to keep IT compromises from translating into OT consequences.

Two developments have changed the operational reality of that assumption. First, the integration of operational technology with corporate IT systems has accelerated; the use of OPC UA gateways to expose

process variables to enterprise dashboards, of remote-access mechanisms for vendor maintenance, and of cloud connectivity for predictive analytics has multiplied the number of legitimate cross-boundary flows. Second, the threat actors operating against utilities and process industries have become more capable. The May 2026 Dragos report on the Monterrey municipal water utility documented an adversary using AI-assisted tooling to script and adapt lateral-movement techniques in real time; the pivot attempt from the compromised IT environment toward the OT control gateways was blocked at the segmentation boundary, but the attempt itself was new. It demonstrated that the threat model the federal guidance had been warning about had been instantiated in the field [R1, R3].

This paper proposes that boundary segmentation alone is no longer sufficient and that the boundary itself must become an active governance device. We present BLADE-INFRA-OT, a reference appliance that places the AUTHREX eight-stage authority pipeline on the IT/OT segmentation line as a bump-in-the-wire inspection element. Every cross-boundary message is parsed at the protocol layer, scored for provenance, evaluated for anomalous AI-generated patterns, mapped to an operator authority tier, checked for consensus across redundant inspection nodes, optionally held in a bounded deliberation window, gated by an aggregate risk model, and finally either propagated, deliberated, or isolated. The decision is committed to a TPM-signed append-only audit ledger before any action is taken.

The paper's contributions are: (i) a reference architecture for a hardware-enforced IT/OT governance appliance derived from the joint Five Eyes principles [R1, R2]; (ii) a pipeline-allocation scheme that meets industrial latency budgets while supporting multi-protocol OT environments; (iii) a simulation methodology that exercises the appliance against four scenarios, including the Monterrey threat pattern; and (iv) an open ICD, an open simulation, and an open working paper, all released under CC BY 4.0.

2 Threat context

Three documentary anchors frame the threat context.

2.1 The December 2025 joint principles

In December 2025, CISA, the Australian Signals Directorate, and the National Security Agency published *Principles for the Secure Integration of AI in Operational Technology* [R1]. The document articulates governance principles for any deployment of AI-derived or AI-informed automation in OT environments. Of the principles, four are directly load-bearing for the bridge governance problem: provenance verification of inputs and instruction sources; segmentation and minimum-necessary access between IT and OT; bounded authority delegation to AI-derived components; and forensic auditability of every cross-boundary action. BLADE-INFRA-OT is structured to satisfy each of these.

2.2 The May 2026 agentic AI advisory

On 1 May 2026, CISA, NSA, and Five Eyes partners issued *Careful Adoption of Agentic AI Services* [R2]. The advisory warned that autonomous-agent components may issue commands not intended by their principals, particularly when operating against composite tool surfaces. For OT environments, this concern has a sharp interpretation: an agent operating in the IT zone, acting on natural-language or higher-level instructions, may emit OT commands that are individually well-formed but collectively outside the operator's intent. The FLAME bounded-deliberation stage of AUTHREX, applied at the bridge, is the natural countermeasure to this class of error: it forces a pause in the cross-boundary path long enough for an operator to confirm intent.

2.3 The Monterrey threat pattern (Dragos, May 2026)

The Dragos report on the Monterrey municipal water utility documents an intrusion in which an adversary used AI-assisted tooling to attempt lateral movement from a compromised corporate-IT host toward the

SCADA gateway feeding the utility's PLC fleet. The pivot attempt was blocked at the segmentation boundary; the adversary did not reach the PLC fleet. The report is informative in two ways. First, it confirms that the AI-assisted lateral-movement playbook is no longer theoretical. Second, it documents characteristic patterns of AI-generated lateral movement - bursty command issuance, atypical function-code mixes for the role of the originating host, and unusual session reuse patterns - sufficient to define a detection rule set for the ADARA stage of the bridge governance pipeline.

Three observations: (a) the Monterrey attempt is correctly characterized as an *AI-assisted IT-intrusion with attempted IT-to-OT pivot*, not as a successful OT attack; the pivot failed; (b) the strategic significance of the event lies precisely in the attempt, since the federal guidance had been published two quarters earlier and the threat model had until then been theoretical; and (c) the documented pattern is sufficient to specify a detection layer at the bridge without exposing operational details of the victim utility.

3 The AUTHREX framework

The AUTHREX framework [R17] organizes the authority lifecycle of an autonomous or semi-autonomous system into eight stages applied in sequence to every authority-bearing event: SATA (Sensor Authority Trust Allocation), ADARA (Adversarial Detection and Authority Risk Architecture), IFF (Identification, Friend or Foe), HMAA (Human-Machine Authority Architecture), MAIVA (Multi-Agent Independent Verification Authority), FLAME (Failure Latency Authority Management Envelope), ERAM (Escalation Risk and Adaptive Mitigation), and CARA (Causal Authority Recovery Architecture). The framework is domain-agnostic; it has been applied across defense, automotive, maritime, infrastructure, and orbital reference platforms.

Applied to the IT/OT bridge, the stages map onto the following questions, asked of every cross-boundary message:

SATA - what is the provenance score of the originating IT host? Has the host been previously authenticated, is its software state attested, is its session-history consistent with its declared role?

ADARA - does the message exhibit a signature consistent with AI-generated scripting or with documented adversarial patterns (per the Dragos taxonomy)? Is it a burst within an otherwise quiescent window? Does its function-code mix match the originating host's role?

IFF - is the OT-side target identified and authorized to receive this class of command from this class of host? Is the binding cryptographic or only network-topological?

HMAA - at what authority tier is the bridge currently operating? Does the operator (or operator role) on the originating host hold the necessary tier to issue this command?

MAIVA - do multiple inspection nodes in the redundant deployment agree on the verdict? If a regional deployment, do nodes across sectors (power / water / gas) see correlated activity?

FLAME - should the command be held in a bounded deliberation window for operator confirmation? FLAME is the deliberate counterweight to agentic-AI command emission.

ERAM - what is the aggregate risk of the action in the current posture? Is the bridge under elevated alert, and if so, does the risk model dictate a more conservative gate?

CARA - if the previous stages have detected compromise, what isolation action recovers the safe posture? CARA is the bridge's *circuit breaker*: an action that trips the inline path and forces the system to a known-good fallback.

4 Architecture of the BLADE-INFRA-OT appliance

4.1 Form and posture

The appliance is a 1U rack-mount (or DIN-rail) fanless device, industrial temperature range, conformal-coated. It is installed as a bump-in-the-wire element at the IT/OT segmentation boundary, with two inspection-path interfaces (IT-in and OT-out) and two mirror-tap interfaces (IT-mon and OT-mon). The default operational posture is *read-only inspection*: traffic that clears the AUTHREX pipeline is forwarded unchanged; traffic that triggers an isolation is dropped, with an out-of-band alarm. The conservative default reflects the operational principle that the bridge must never be the proximate cause of an outage; degraded operation is preferable to inadvertent disruption.

4.2 Plane separation

Internally, the appliance hosts two cryptographically separated planes: a **network plane** built on an x86 Atom-class fanless SBC, hosting the protocol-parser layer; and a **governance plane** built on a Xilinx Kria K26 SOM, hosting the AUTHREX pipeline. The planes communicate through a policy-enforced inter-plane channel that is monotonic, rate-limited, and audited. The two-plane structure is the architectural answer to the joint principles' segmentation requirement: even if the network plane is compromised through a parser flaw or a memory-safety bug, the governance plane retains its own root-of-trust and its own audit channel.

4.3 Protocol parser layer

Six OT protocols are supported in the initial design: Modbus TCP/RTU, DNP3, IEC 61850 (MMS, GOOSE, SV), OPC UA, EtherNet/IP CIP, and BACnet/IP. Each protocol receives a dedicated parser process that emits a uniform abstract-syntax-tree (AST) representation; the AUTHREX pipeline operates on the AST, not on the raw protocol envelope. This common-AST design is the architectural answer to the maintainability problem: governance logic is decoupled from protocol encoding details, which evolve faster than governance principles.

4.4 Audit ledger

Every decision is committed to an append-only, TPM-signed audit ledger before the message is propagated, deliberated, or dropped. The ledger schema captures the message-level metadata (source, destination, protocol, operation) and the AUTHREX verdict at each stage (SATA score, ADARA verdict, IFF status, HMAA tier granted, MAIVA consensus, FLAME state, ERAM risk, CARA action), plus the terminal outcome. Entries form a hash chain (BLAKE3) and are signed (Ed25519) by the TPM. The ledger is exportable in binary, JSON, and CIP-008-aligned XML formats; exports are themselves TPM-signed.

5 Pipeline application and latency

The pipeline is allocated between governance-plane software and governance-plane fabric. Stages with strict deterministic-latency requirements (HMAA tier lookup, FLAME timer, CARA cut-through) are implemented in fabric. Stages with richer state requirements (ADARA pattern matching against the Dragos taxonomy, MAIVA inter-node consensus) are implemented in software but on dedicated cores with bounded interrupt-latency.

Stage	Implementation	Latency target (typ.)
SATA	Kria - software	20 µs
ADARA	Kria - sw + LUT	40 µs
IFF	Kria - software	15 µs
HMAA	Kria - fabric	10 µs
MAIVA	Kria - sw (inter-node)	60 µs

FLAME	Kria - fabric + sw	5 µs (open/close)
ERAM	Kria - software	10 µs
CARA	Kria - fabric (cut-through)	5 µs (trigger)
Inline budget end-to-end	-	≤ 250 µs (Modbus)
Accelerated path	Kria fabric	≤ 100 µs (GOOSE)

Table 5.1 - AUTHREX pipeline allocation and latency targets. Values are design intent at fabricated-prototype maturity; not measured under this revision.

6 Simulation scenarios

The appliance is exercised in a browser-based simulation that scripts four scenarios. Each scenario produces a deterministic sequence of cross-boundary messages; the pipeline executes against each in turn; the audit ledger captures every verdict.

6.1 Scenario 01 - Nominal water-utility operation

A SCADA operator issues a Modbus write_holding_register to start a pump. SATA verifies operator credentials and originating-host provenance. ADARA confirms the message pattern is baseline. HMAA grants T3. The message propagates. The audit ledger entry is committed. The scenario demonstrates the appliance's transparency in normal operation: the inspection introduces no observable change in operator workflow.

6.2 Scenario 02 - Monterrey-pattern attack

The IT host corp-svc-04 is compromised. The adversary uses AI-assisted tooling to script Modbus writes against the OT side: a burst of write_holding_register calls, a sequence of write_coil operations against valve actuators, and an IEC 61850 MMS write against a feeder IED. ADARA recognizes the AI-generated pattern (bursty issuance, atypical function-code mix for the originating host's declared role). SATA scores degrade as the burst proceeds. HMAA denies. CARA isolates the compromised host. The tier collapses from T3 to T2 to T1 to T0 across the burst. Legitimate operator reads from scada-op-01 are now held pending manual confirmation per T0 policy. The scenario demonstrates the appliance's primary defensive value: the documented Monterrey threat pattern is recognized and contained.

6.3 Scenario 03 - Authorized vendor PLC firmware update

An authorized vendor remote-access for a PLC firmware update is scheduled. The vendor jump host vendor-jump-12 issues an OPC UA browse, then a CallMethod for firmware staging, then a write of the firmware image, then an activation call. HMAA verifies the vendor against the authorized-maintainer roster. FLAME opens a 120-second deliberation window. The operator confirms in-band. T2 supervised session is granted, time-bounded. The session terminates automatically at window expiry; the tier returns to T3. The scenario demonstrates that legitimate elevated activity is supported without removing governance.

6.4 Scenario 04 - Multi-utility coordinated probe

Probe activity appears simultaneously against power (IEC 61850 MMS), water (Modbus), and gas (DNP3) sectors. Each individual probe is below the threshold that any single inspection node would act on. MAIVA aggregates signals across the three sectors. ADARA correlates the pattern. CARA executes regional isolation. The scenario demonstrates the value of the multi-node consensus stage: no single utility sees enough signal to act unilaterally, but the consensus layer does.

7 Compliance mapping

Anchor	BLADE-INFRA-OT response
CISA Five Eyes OT (Dec 2025) - segmentation	Bump-in-the-wire inspection at the boundary with read-only default.
CISA Five Eyes OT (Dec 2025) - provenance	SATA scores originating-host provenance per message.
CISA Five Eyes OT (Dec 2025) - bounded autonomy	HMAA tiers + FLAME bounded deliberation window.
CISA Five Eyes OT (Dec 2025) - auditability	TPM-signed append-only ledger; three export formats.
CISA Five Eyes agentic AI (May 2026)	FLAME forces operator confirmation on agent-issued OT commands.
NERC CIP-005	Bridge instantiates and logs the electronic security perimeter.
NERC CIP-007	Configuration management for fail-bypass enable.
NERC CIP-008	Ledger XML export schema aligned for incident reporting.
NERC CIP-010	Signed policy bundles and dual-slot firmware support change management.
IEC 62443 zones and conduits	Conduit gateway between SL-1 IT and SL-3+ OT.
NIST SP 800-82 r3	Cross-walk documented in the ICD.
FIPS 140-2	Cryptographic boundary inherited from BLADE-INFRA HSM family.
DoDD 3000.09	HMAA tier model; human-on-the-loop posture maintained.

8 Related work and prior art

Three classes of prior systems are relevant. **Unidirectional gateways** (data diodes) provide the strongest physical IT-to-OT separation but at the cost of bidirectional flows that legitimate maintenance and supervisory-data exchange require. BLADE-INFRA-OT is intended to complement, not replace, data diodes; the appliance handles the bidirectional flows that must exist while applying governance. **Industrial IDS / passive monitoring** (Dragos, Claroty, Nozomi) supplies excellent situational awareness but does not enforce. The appliance's read-only default posture is intentionally a passive mode in spirit, but its enforcement-enabled mode applies authority in line. **OT firewalls with deep-packet inspection** (Fortinet, Palo Alto, industrial variants) enforce protocol rules but typically lack the operator-tier authority model and the agentic-AI counterweight that the AUTHREX pipeline supplies.

Academic work on runtime assurance (Schierman et al., the Simplex architecture, the work on shared autonomy in autonomous-vehicle platforms) provides the conceptual lineage for the FLAME and HMAA stages. Byzantine fault-tolerant consensus literature (PBFT, HotStuff) provides the conceptual lineage for MAIVA. The contribution of this paper is the application of these conceptual elements to the specific problem of the IT/OT bridge, with a concrete pipeline allocation, a concrete protocol parser inventory, and a concrete simulation against a documented threat pattern.

9 Limitations and threats to validity

Several limitations should be made explicit. **Hardware not fabricated.** The appliance described in this paper is a reference design at TRL 2–3. Latency budgets and reliability allocations are design intent; first-article test results are not available under this revision. **Simulation, not field trial.** The four scenarios are scripted against documented threat patterns; they are not the output of a red-team campaign against an operational utility. The simulation is informative for design verification, not for operational claims. **Detection coverage of AI-generated patterns is bounded by the published taxonomy.** The Dragos report supplies the initial taxonomy [R3]; future AI-assisted attack patterns will require extension. **Performance envelopes are protocol-specific.** The 250 µs end-to-end inspection budget assumes short-header control traffic; bulk OPC UA file transfers or large IEC 61850 datasets will exhibit different performance characteristics. **Reliance on TPM and HSM root-of-trust.** If the TPM or HSM supply chain is

itself compromised, the audit ledger guarantees do not hold; this is a known limitation of the broader hardware-root-of-trust paradigm and is not specific to BLADE-INFRA-OT.

10 Future work

Three directions are in scope for follow-on work. First, **prototype fabrication and first-article test**, contingent on funding outside the current fundamental-research posture. Second, **protocol parser extension** to cover S7Comm (Siemens proprietary, broadly deployed in European water and wastewater), MMS over IEC 62351 secured transport, and natively secured DNP3 per IEC 62351-5. Third, **extension of the ADARA pattern library** as additional AI-assisted attack patterns are documented by Dragos, Mandiant, and the CISA / FBI joint advisories; this is an open-ended commitment, since the threat landscape will continue to evolve faster than any one publication can capture.

11 Conclusion

The IT/OT segmentation boundary has been treated as a passive line: a place where two networks meet and where a firewall enforces a small number of static rules. The threat environment has moved past this. The joint Five Eyes guidance of December 2025 and the agentic-AI advisory of May 2026 articulated principles that the boundary must satisfy actively, not merely structurally. The Monterrey threat-intelligence report of May 2026 demonstrated that the threat model is no longer theoretical. BLADE-INFRA-OT is offered as a reference implementation of the joint principles: a bump-in-the-wire governance appliance, applying an eight-stage authority pipeline to every cross-boundary message, with a TPM-signed audit ledger designed for CIP-008 evidentiary use. The hardware is specified, the simulation is open, the ICD is open, and the working paper is open. The next maturity step - fabrication and first-article test - is outside the scope of this fundamental-research revision but is the natural follow-on under appropriate funding mechanisms.

12 References

[R1] Cybersecurity and Infrastructure Security Agency, Australian Signals Directorate Australian Cyber Security Centre, and National Security Agency. *Principles for the Secure Integration of AI in Operational Technology.* Joint publication, December 2025.

[R2] Cybersecurity and Infrastructure Security Agency and National Security Agency, with Five Eyes partners. *Careful Adoption of Agentic AI Services.* Joint advisory, 1 May 2026.

[R3] Dragos, Inc. *Monterrey water utility - AI-assisted IT-to-OT pivot attempt: threat intelligence report.* May 2026.

[R4] North American Electric Reliability Corporation. *CIP-005 / 007 / 008 / 010.* Critical Infrastructure Protection standards.

[R5] National Institute of Standards and Technology. *Guide to Operational Technology (OT) Security.* NIST SP 800-82 Rev. 3.

[R6] International Electrotechnical Commission. *IEC 62443 - Industrial communication networks - Network and system security.*

[R7] International Electrotechnical Commission. *IEC 61850 - Communication networks and systems for power utility automation.*

[R8] Institute of Electrical and Electronics Engineers. *IEEE 1815-2012 - DNP3.*

[R9] Modbus Organization. *Modbus Application Protocol Specification V1.1b3.*

[R10] OPC Foundation. *OPC Unified Architecture Specification, Parts 1–14.*

[R11] ODVA. *Common Industrial Protocol (CIP) and EtherNet/IP Specification.*

[R12] ASHRAE. *Standard 135 - BACnet.*

[R13] U.S. Department of Defense. *Directive 3000.09 - Autonomy in Weapon Systems.*

[R14] Federal Information Processing Standards. *FIPS PUB 140-2 - Security Requirements for Cryptographic Modules.*

[R15] Schierman, J. D., et al. *Runtime assurance for autonomous aerospace systems.* AIAA Journal of Aerospace Information Systems, 2020.

[R16] Castro, M., Liskov, B. *Practical Byzantine fault tolerance.* OSDI 1999.

[R17] Oktenli, B. *AUTHREX: A Unified Authority-Lifecycle Framework for Autonomous and Semi-Autonomous Systems.* Zenodo, 2026.

[R18] Oktenli, B. *BLADE-INFRA: Authority Governance for Critical Infrastructure.* Zenodo, 2026.

[R19] Mitre. *ATT&CK for Industrial Control Systems.*

[R20] Stouffer, K., Pillitteri, V., et al. *NIST Cybersecurity Framework 2.0.* NIST, 2024.

Colophon. Prepared by Burak Oktenli, independent researcher, Washington, DC, USA. ORCID 0009-0001-8573-1667. May 2026. Version 1.0. Released under CC BY 4.0. Companion artifacts: ICD-INFRA-OT-001 and the BLADE-INFRA-OT browser-based simulation, both deposited with this paper to Zenodo and SSRN.

