Broadcast-Gain: Two-Byte Neuromodulation for Long-Horizon Control

Stop-gradient overlay that trims latency under bursty loss

Dario Fumarola, Amogh Gaikwad

Amazon Web Services - Prototyping

Coordinating in Chaos

Cooperative MARL agents operating over bursty, lossy networks face a formidable challenge. Delayed rewards and high-variance gradients make learning stable policies difficult, often leading to “knife-edge” scenarios at intersections. Here, minor indecision can cascade into gridlock, creating a long tail of high-latency events that cripples system performance.

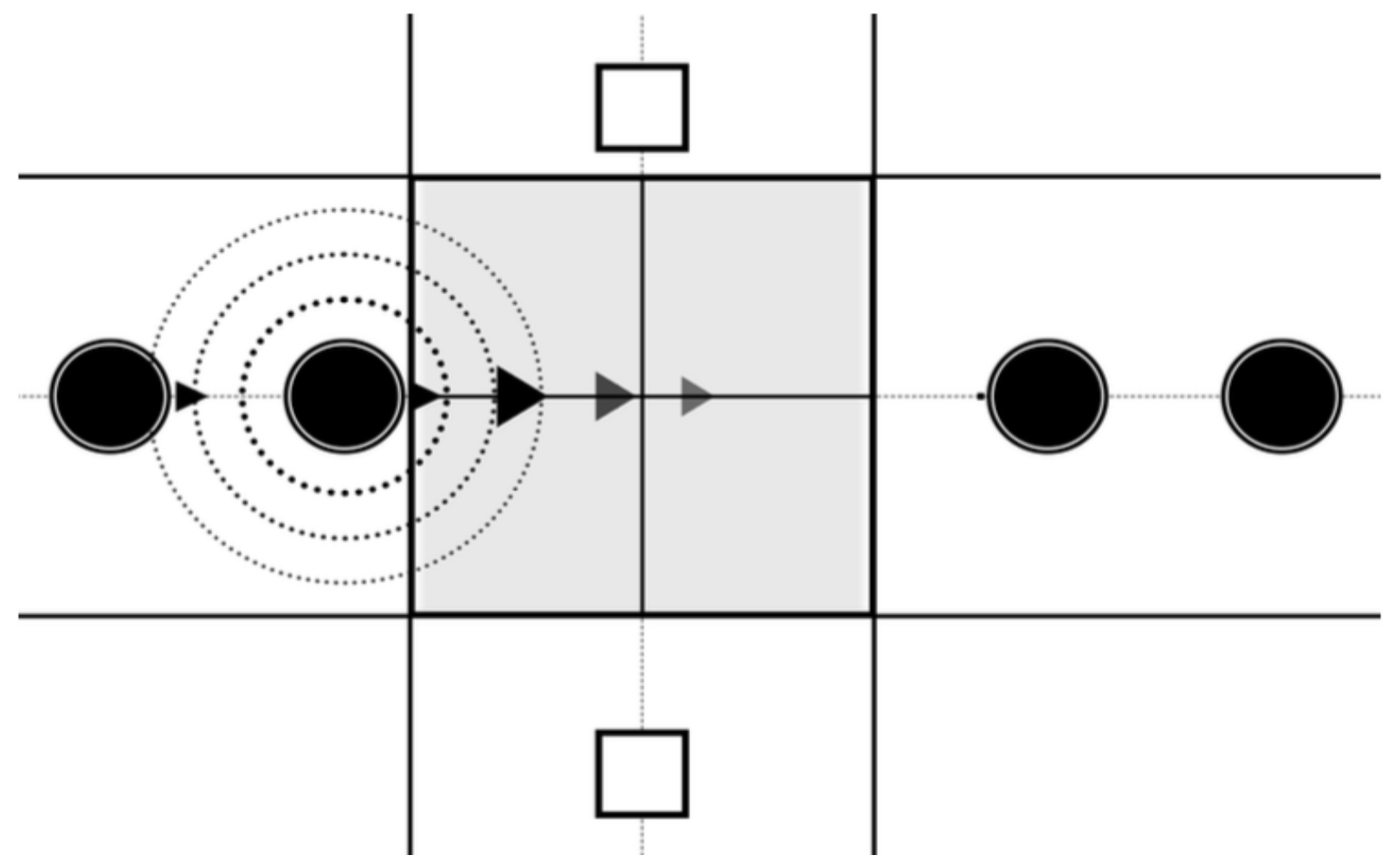

Figure 1. The ‘Knife-Edge’ Intersection

We introduce Broadcast-Gain (BG), a tiny, two-byte *neuromodulatory* signal that overlays a standard learned policy. It requires no changes to training, acting as a lightweight control plane to guide agents toward better, more robust coordination when it matters most.

The Challenge

Our testbed is a single, four-way intersection where agents must coordinate to cross from perpendicular directions. The rules are simple but create a difficult challenge:

- **One Green Light**: Only one direction (e.g., North-South) is given "green" at a time, while the other (East-West) is held at "red."
- **A Strict Safety Lock:** The intersection must remain completely empty for several steps before the light is allowed to change. This makes every decision critical and punishes hesitation severely.
- **The Goal:** Maximize throughput by minimizing wasted green time and preventing gridlock..

This setup is designed to stress agent decision-making under uncertainty, creating the long-tail latency events that BG targets.

The Key Insight: A Neuromodulatory Overlay

Our design is inspired by how the brain coordinates action under uncertainty. Rather than sending rich, point-to-point “messages,” neural systems use neuromodulators (low-bit, broadcast chemicals like noradrenaline or acetylcholine) to set network gain and decision thresholds. A short phasic pulse doesn’t say what to do; it asks the system to be more ready or more cautious for a moment, especially when many signals agree.

Broadcast-Gain copies that idea: a tiny, two-byte pulse creates a shared readiness state that rises when neighbors agree and quickly fades when evidence is weak or stale.

The Broadcast-Gain Mechanism

BG introduces a non-learning control plane that transforms internal RL signals into a robust, actionable consensus. It solves the coordination problem without altering the agent's reward function or policy gradients.

- **Broadcast: From Value Function to Coordination Signal.** Each agent broadcasts a 2-byte packet derived directly from its learning process:
  - **Byte 1: TD-Error Residual.** This byte encodes a quantized TD-error. A high TD-error signals a high-surprise state transition - it turns the agent's own learning signal into a broadcast for help.
  - **Byte 2: State Context.** An encoding of the agent’s axis and quantized distance to the intersection, providing spatial context.
- **Fuse: Confidence-Weighted Consensus.** Agents aggregate these stop-gradient signals from neighbors. The key is the fusion mechanism, which weights the emerging consensus by confidence. Confidence is a function of message freshness and coverage. Under heavy packet loss, confidence drops, causing the system to extend green times and act more conservatively.
- **Nudge: Targeted Logit Modulation.** The final consensus signal directly modulates the logit of the MOVE action in the agent's policy. This targeted *gain* provides a small, directional push. When the policy is indifferent, this nudge breaks the symmetry and resolve the conflict, turning a gridlock into a coordinated action.

Trimming the Long Latency Tail

Performance Under Extreme Burstiness

(Results from our harshest test: 120 agents with 70% packet drop)

Metric	Change with BG Overlay
Gridlock (p95 Wait Time)	-40.97 steps
Throughput (Crossings/1k steps)	+391.9

Table 1. Impact on Tail Latency and System Throughput

Mechanism Check: How BG Achieves its Gains

(Analysis of the system’s internal behavior)

Metric	Change with BG Overlay
Attention Shifted to Green	+16.2 pp
Realized Green Time	+19.9 pp

Table 2. From Agent Attention to Realized Green

These results validate our mechanism. The direct link between the shift in agent attention (+16.2 pp) and the increase in realized green time (+19.9 pp) shows precisely how BG works: it converts collective hesitation directly into coordinated action, providing a decisive advantage even under extreme communication loss.

Across all 108 test configurations, these gains were concentrated precisely where they are needed most: in scenarios with higher agent counts and longer communication cycles, highlighting its specific strength for long-horizon challenges

A Bounded Perturbation

BG is a bounded, stop-gradient hint. We nudge only the MOVE logit and clip that nudge at Δ, so the base policy stays in charge.

$$D_{TV}\Big(\pi_{BG}(\,\cdot\,|\,o),\,\pi(\,\cdot\,|\,o)\Big)\;\leq \tanh\Big(\frac{\Delta}{4}\Big)$$

Eq 1. Bounding the Maximum Behavioral Change

D_{TV} : Measures the maximum possible change in the agent's action probabilities.

Δ : The maximum strength of the "nudge" applied to the MOVE action's logit.

Softmax is like a dial that turns scores into probabilities. That dial has a limited steepness, so a small twist can only move probability a little - Local sensitivity is largest when the action’s probability is near mid-range (for small constant Δ).

Always-on guardrails

- **Hard stop near red:** no bias if a safety band would be crossed.
- **Minimum-green hold:** prevents flip-flopping under noise.
- **Confidence & freshness:** stale/weak signals (TTL) get down-weighted.
- **Stop-gradient:** We don’t backprop through BG; it influences learning only by changing on-policy trajectories..

What’s Next

Turn BG into an X-ray of coordination. Instrument the system to log when and where pulses fire, confidence rises, gates hold/flip, and near-gate ties break - then fold those into a simple “mood map” over time and space. Overlay it on the network (or road grid) so hotspots literally light up: we see the stretches that spawn tail risk.

Why it’s exciting: this map makes tuning obvious. If bursts create stale consensus, you’ll see it and shorten TTL; if ties persist at certain junctions, you’ll see them and adjust the near-gate nudge. Side-by-side with ε-max-pressure and no-BG, the map shows where BG prevents thrash and red encroachment. The same visualization travels to robot swarms, datacenter queues, and fleet routing: one small overlay, one universal way to reveal where coordination fails - and fix it.

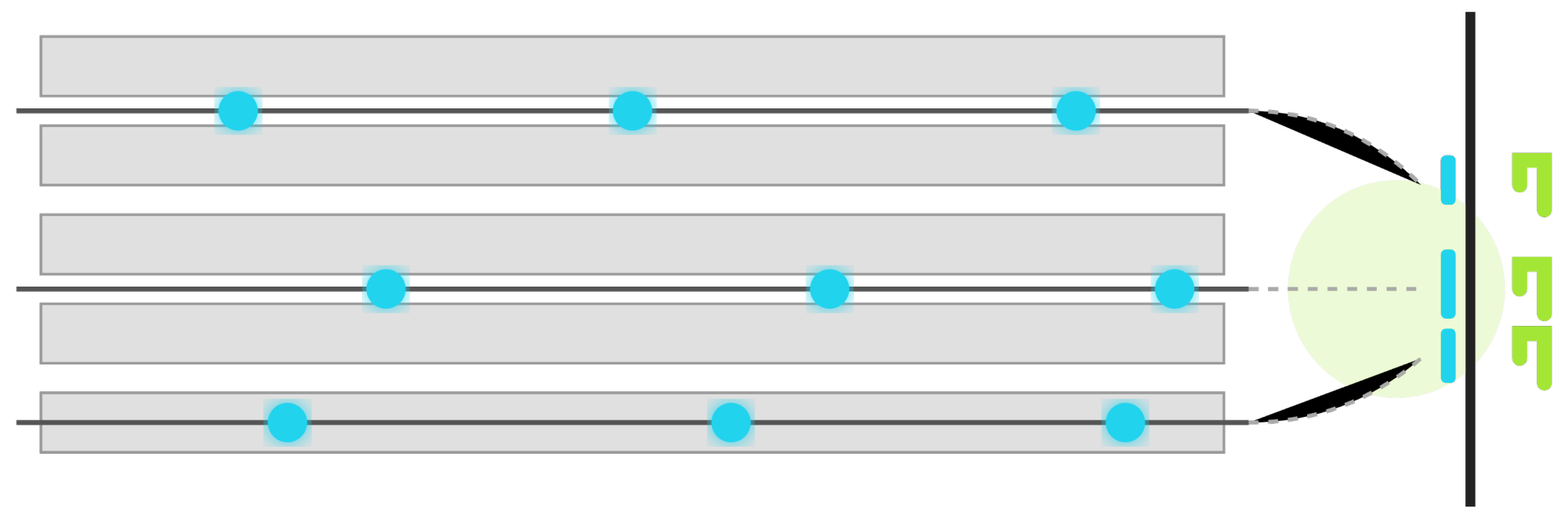

Figure 2. Coordination warehouse- robots (blue) reach a merge; a cyan pulse, brief pause, then go. Ties clear fast.

References

[1] Schulman, J., et al. "Proximal Policy Optimization Algorithms." arXiv:1707.06347, 2017.

[2] Schulman, J., et al. "High-Dimensional Continuous Control Using Generalized Advantage Estimation." arXiv:1506.02438, 2015.

[3] Foerster, J. N., et al. "Learning to Communicate with Deep Multi-Agent Reinforcement Learning." NeurIPS, 2016.