International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424
Volume 3 Issue 5 , May 2026
DOI: https://doi.org/10.71366/ijwos03052685238

Intelligent DDoS Mitigation Using Reinforcement Learning

Anushka Gaur[1], Sagar Choudhary[2], Gaurav Kumar[3]
[1,2] B.Tech Student, Department of CSE, Quantum University, Roorkee, India.
[3] Assistant Professor, Department of CSE, Quantum University, Roorkee, India.

Article Info

Article History:

Published:29 May 2026

Publication Issue:
Volume 3, Issue 5
May-2026

Page Number:
438-453

Corresponding Author:
Anushka Gaur

Abstract:

Distributed Denial-of-Service (DDoS) attacks represent one of the most persistent and economically damaging threats in modern cybersecurity, with global attack volumes exceeding 800 Gbps in commercial data-center environments as of 2024. Traditional mitigation approaches — encompassing rule-based filtering, static rate-limiting, and IP blacklisting — suffer from critical limitations: they cannot adapt to evolving adversarial strategies, generate high false-positive rates against bursty legitimate traffic, and are fundamentally reactive rather than anticipatory. This paper investigates the application of Reinforcement Learning (RL) as an intelligent, self-adaptive paradigm for DDoS detection and mitigation. We survey the DDoS threat landscape and its taxonomy, critically evaluate conventional defenses, formalize the DDoS mitigation task as a Markov Decision Process (MDP), and review state-of-the-art RL architectures including Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Multi-Agent Reinforcement Learning (MARL) deployed across Software-Defined Networking (SDN) and edge-computing environments. Benchmark evaluations on CIC-DDoS2019, SCLDDOS2024, and UNSW-NB15 demonstrate that RL-based systems achieve detection rates of 97%–99.5% with false-positive rates below 1%, while maintaining real-time mitigation latency under 5 milliseconds. Open challenges including adversarial robustness, computational overhead at line-rate, and long-term convergence stability are discussed, with promising future research directions outlined.

Keywords: DDoS Mitigation, Reinforcement Learning, Deep Q-Network, Software-Defined Networking, Markov Decision Process, Cybersecurity, Autonomous Defense, Network Intrusion Detection

1.Introduction

1.1 Motivation and Problem Statement

The explosive growth of internet-connected infrastructure — spanning cloud data centers, Internet of Things (IoT) ecosystems, and critical national services — has made Distributed Denial-of-Service (DDoS) attacks an ever-present and increasingly sophisticated adversarial challenge [1], [3], [19]. A DDoS attack seeks to exhaust the computational, bandwidth, or application-layer resources of a target system by coordinating traffic from a large number of compromised hosts (a botnet), rendering legitimate user requests unserviceable [2], [14]. The consequences are both operational and financial: extended service outages can cost enterprises thousands to millions of dollars per hour, while targeted attacks against healthcare, finance, or government infrastructure may endanger public safety [3], [19].

1.2 Research Objectives

Taxonomic Survey: To provide a comprehensive and up-to-date taxonomy of DDoS attack types, including volumetric, protocol, and application-layer categories, with quantitative characterization of their traffic signatures [1], [3] **Critical Evaluation:** To rigorously identify the failure modes of conventional mitigation approaches and motivate

the transition to learning-based methods [2], [19] **Formal Modeling:** To present the DDoS mitigation problem as a Markov Decision Process (MDP), formally specifying state spaces, action spaces, and reward functions applicable to RL agent training [6], [7], [9].**Comparative Review:** To survey state-of-the-art RL-based DDoS mitigation systems published between 2020 and 2026, compare their architectures and performance metrics, and identify convergent design principles and persistent gaps [11], [12], [14], [15], [17].

1.3 Contributions

A structured three-layer DDoS taxonomy with quantitative attack characterization across measurement dimensions (Gbps, packets-per-second, requests-per-second) [1], [3].The first unified MDP formulation for multi-vector DDoS mitigation that integrates entropy-based state features with a composite reward function balancing throughput, latency, and false-positive rate [6], [7], [9]. A systematic comparative analysis of twelve RL-based DDoS mitigation systems, evaluated against CIC-DDoS2019, SCLDDOS2024, and UNSW-NB15 benchmarks [11], [12], [14], [15], [17]. A comprehensive discussion of open research challenges — adversarial robustness, line-rate deployment, and long-term convergence — with proposed research directions [16], [18], [19].

1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 presents the DDoS threat taxonomy and attack characterization. Section 3 critically evaluates traditional mitigation approaches and their limitations. Section 4 introduces the RL formalism, including MDP specification, core algorithms, and reward engineering for network security [6], [7], [8], [9]. Section 5 surveys state-of-the-art RL-based DDoS mitigation systems [11]–17]. Section 6 covers benchmark datasets and comparative performance evaluation. Section 7 discusses open challenges. Section 8 outlines future research directions. Section 9 concludes the paper.

2. Background: DDoS Attacks and Taxonomy

2.1 Volumetric Attacks (Layer 3/4)

Volumetric attacks constitute the most commonly observed DDoS category by bandwidth consumption. Their objective is to saturate the upstream bandwidth link between the target and the internet, making the target unreachable irrespective of its internal processing capacity. Attack severity is measured in Gigabits-per-second (Gbps) or Terabits-per-second (Tbps) [3]. Two primary mechanisms are exploited:

- Amplification and Reflection: Attackers send small requests with spoofed source addresses (set to the victim's IP) to open reflectors. DNS amplification achieves amplification factors of up to 50×, while TFTP reflection achieves 30× to 110× [3]. NTP and SSDP reflection are also widely exploited, with NTP achieving amplification factors as high as 556× [2], [3].
- Direct Flooding (UDP/ICMP): Large botnets generate raw UDP or ICMP traffic at line-rate toward the victim, saturating egress links without requiring amplification. The Mirai botnet demonstrated sustained 623 Gbps UDP floods against a single domain [3].

From the SCLDDOS2024 dataset — derived from commercial data centers processing real-world traffic between 2022 and 2024 — volumetric attacks represented approximately 68% of all observed DDoS incidents by event count, with peak bursts exceeding 800 Gbps [1].

2.2 Protocol Attacks (Layer 3/4)

Protocol attacks do not seek to exhaust bandwidth but rather to deplete the finite state resources of network infrastructure components such as firewalls, load balancers, and intrusion prevention systems. Attack volume is measured in packets-per-second (pps) or connections-per-second (cps) [2]. Key attack types include:

- SYN Flood: Exploits the TCP three-way handshake. The attacker sends a large volume of SYN packets with spoofed source addresses, causing the server to allocate a Transmission Control Block (TCB) and wait for an ACK that never arrives. The TCB table exhausts rapidly under sustained SYN flood rates of millions of packets per second [2], [3].
- Ping of Death / Fragmentation: Crafts malformed or oversized ICMP packets that cause processing failures or buffer overflows in vulnerable network stacks [2].
- Smurf Attack: Broadcasts ICMP echo requests with spoofed victim source addresses across amplification networks, generating massive ICMP echo replies directed toward the victim [3].

2.3 Application-Layer Attacks (Layer 7)

Application-layer (L7) attacks are the most sophisticated category and the most difficult to mitigate without generating false positives. Rather than saturating bandwidth or state tables, they target the computational expense of specific application operations such as database queries, session creation, or cryptographic handshakes. These attacks are measured in requests-per-second (rps) [4], [19].

- HTTP GET/POST Flood: Sends massive volumes of legitimate-looking HTTP requests to resource-intensive endpoints such as search functions or login pages, potentially triggering expensive SQL queries or file operations [19].
- Slowloris: Opens numerous HTTP connections while transmitting partial headers at extremely low rates, keeping server connections active indefinitely. A relatively small number of sockets can incapacitate poorly configured web servers [14], [15].
- ReDoS (Regular Expression DoS): Exploits catastrophic backtracking behavior in regular expression engines by submitting malicious inputs that trigger worst-case polynomial or exponential matching times [2].

L7 attacks are particularly challenging because individual requests often appear indistinguishable from legitimate traffic at the packet level, requiring advanced behavioral analysis, anomaly detection, or reinforcement learning-based flow analysis for accurate detection and mitigation [4], [16], [19].

2.4 Attack Scale and Economic Impact

Table 1: DDoS attack taxonomy with observed scale and representative examples [1][3].

Attack Class	Target Layer	Measurement Unit	Max Observed Scale	Representative Example
Volumetric	L3/L4	Gbps / Tbps	3.47 Tbps (2021, Microsoft Azure)	DNS Amplification, UDP Flood
Protocol	L3/L4	Mpps (megapps)	809 Mpps (2020, Akamai)	SYN Flood, Fragmentation
Application-Layer	L7	Mrps (mega-rps)	71 Mrps (2022, Cloudflare)	HTTP Flood, Slowloris
Multi-Vector	L3–L7	Combined	Growing share (~55% of 2024 incidents)	Mirai botnet variants

The economic consequences of DDoS attacks are substantial. Downtime costs for large-scale enterprise services are estimated at $20,000 to $100,000 per hour depending on the industry vertical, with financial services and e-commerce

at the upper end [3]. Beyond direct revenue loss, DDoS attacks frequently serve as diversionary tactics to mask concurrent data exfiltration or ransomware deployment. The emergence of DDoS-as-a-Service (DaaS) platforms has lowered the barrier to entry, enabling volumetric attacks for as little as $10 per hour, further increasing the frequency and breadth of the threat landscape.

Key Insight: The transition from single-vector volumetric attacks to multi-vector, cross-layer campaigns necessitates mitigation systems that can simultaneously reason about traffic behavior at L3, L4, and L7 — a requirement that no static rule set can adequately fulfill, motivating the adoption of adaptive, learning-based defenses.

3. Limitations of Traditional Mitigation Approaches

3.1 Rule-Based and Signature Filtering

Rule-based systems such as Snort and Suricata match observed traffic against a curated library of deterministic signatures encoding packet header fields, payload substrings, and flow statistics. A representative Snort rule for HTTP flood detection takes the form:

alert tcp any -> $HOME_NET 80 (msg:"HTTP Flood"; flow:to_server; threshold:type both, track by_src, count 500, seconds 10; sid:10001;)

This rule fires when a source IP establishes more than 500 TCP connections to port 80 within 10 seconds — a static, human-authored threshold.

3.1.1 Failure Modes

Zero-Day and Polymorphic Attacks:

Signatures require prior knowledge of attack patterns. Novel DDoS vectors evade detection for an average of 6–18 days before a rule is authored and deployed [2]. Polymorphic botnets systematically vary packet timing and payload to circumvent known rules [2], [19].

Performance Degradation:

The Snort community rule set exceeds 29,000 rules as of 2024. Pattern matching at 10 Gbps line rate degrades effective throughput to under 2 Gbps on commodity hardware [2], creating a bottleneck that operators work around by pruning coverage [19].

Brittle Threshold Engineering:

Count-based thresholds calibrated for average traffic fail during flash-crowd events (false positives up to 23%) and are deliberately undercut by slow-rate attacks such as Slowloris (false negatives) [2], [14].

3.2 Rate-Limiting Mechanisms

Rate-limiting enforces per-source or aggregate traffic ceilings via the token-bucket algorithm. The bucket state evolves as:

$$N(t)=\min(B,N(t-1)+r\Delta t-A(t))$$

where B is the bucket capacity, r is the token replenishment rate, and $A(t)$ is the number of arriving packets in interval Δt. A flow is throttled when $N(t)=0$.

3.2.1 Failure Modes

Semantic Blindness:

Rate-limiting cannot distinguish CDN flash-crowd traffic from coordinated botnet floods. Studies show aggressive rate-limiting reduces legitimate throughput by 15–40% during surge events [2], [11].

Amplification Vulnerability:

Reflection attacks arrive from a small set of legitimate reflector IPs (DNS, NTP servers). Per-source rate limits do not fire, yet aggregate attack volume overwhelms the target [3], [19].

Static Threshold Fragility:

Manual calibration per service and time-of-day is labor-intensive and susceptible to concept drift as traffic patterns evolve seasonally [2], [12].

3.3 IP Blacklisting

IP blacklisting drops traffic from enumerated malicious addresses maintained via commercial threat-intelligence feeds (e.g., Spamhaus, Emerging Threats). Hardware implementations use Bloom filters or TCAM structures with capacity ceilings of 100K–500K entries [3].

3.3.1 Failure Modes

IP Spoofing Evasion:

Volumetric attacks routinely forge source addresses, blocking innocent third parties while the attacker continues unimpeded. Blacklists cover fewer than 50% of active botnet nodes [2], [19].

Fast-Flux IP Rotation:

Modern botnets cycle IP pools every few minutes; the Mirai 2022 variant demonstrated complete rotation within 4-hour windows, rendering blacklists stale within hours of publication [3], [19].

Collateral Damage:

Cloud multi-tenancy and aggressive DHCP/NAT cycling cause blacklisted IPs to be reassigned to legitimate businesses, making IP-based blocking an increasingly blunt instrument [3], [14].

3.4 Comparative Analysis

Table 2: Comparative analysis of conventional DDoS mitigation approaches [2].

Dimension	Rule-Based Filtering	Rate-Limiting	IP Blacklisting
Adaptability to Novel Attacks	Very Low — manual rule updates required; 6–18 day zero-day gap [2]	Low — threshold recalibration per service	Very Low — requires enumeration of known IPs
False-Positive Rate	Up to 23% under flash-crowd conditions [2]	15–40% during legitimate traffic surges [2]	Medium — worsens with IP churn & cloud NAT
False-Negative Rate	High — zero-day & polymorphic attacks evade signatures	High — low-rate DDoS (Slowloris) bypasses thresholds	Very High (>50%) — fast-flux botnet rotation [2]
Scalability	Degrades below 2 Gbps at full rule set on 10 Gbps links [2]	Hardware-scalable; decision logic not adaptive	TCAM capacity ceiling ~100K–500K entries
Response Latency	Near real-time (<1 ms/packet)	Near real-time (hardware policing)	Near real-time (lookup table)

Attack Vector Coverage	Known signatures only	Volumetric only — blind to L7 behavior	Known source IPs only
Operational Overhead	Very High — rule authoring, testing, deployment	Medium — threshold calibration & monitoring	High — continuous list curation & update

The analysis reveals a structural pattern: all three paradigms operate on pre-specified, static decision logic that cannot generalize beyond its calibration distribution. As multi-vector attacks increasingly combine L3/L4 volumetric floods with L7 application-layer subtlety [1], no single conventional approach satisfies the triple requirement of (1) adaptation to unseen vectors, (2) semantic traffic discrimination, and (3) real-time autonomous response — the design space that Reinforcement Learning is uniquely positioned to occupy.

4. Reinforcement Learning for DDoS Mitigation

4.1 MDP Formulation for DDoS Mitigation

Reinforcement Learning operates within the Markov Decision Process (MDP) framework, defined by the tuple:

M=\langle S,A,T,R,\gamma\rangle

- (S) — state space of observable network configurations.
- (A) — discrete action space of mitigation interventions.
- $T: S \times A \to \Delta(S)$denotes the stochastic state transition function, which defines the probability distribution over the next states given the current state and action.
- $R: S \times A \times S \to \mathbb{R}$represents the reward function, assigning a scalar reward for transitions between states under a chosen action.
- $\gamma \in [0,1)$is the discount factor that determines the relative importance of future rewards compared to immediate rewards.

The objective of the agent is to learn an optimal policy $\pi: S \to \Delta(A)$ that maximizes the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum\nolimits_{t=0}^{\infty} \gamma^t r_t\right]$$

This formulation enables adaptive and autonomous mitigation strategies in dynamic network environments, making RL highly suitable for intelligent DDoS defense systems [6], [7], [9], [11], [14], [15].

4.1.1 State Space Design

The 12-dimensional state vector (s_t∈R12) is min-max normalized to ([0,1]^{12}) and grouped into three feature families:

- Traffic Volume (5 dims): Mean/std packet arrival rate, mean byte size, UDP fraction, TCP SYN fraction.
- Protocol Distribution (4 dims): Shannon entropy of source IP distribution

$H_t^{IP}=-\sum p_i \log_2 p_i$

destination port entropy, fraction of new source IPs, and mean flow duration.

- System Resources (3 dims): CPU utilization, memory ratio, and queue length fraction.

Shannon IP entropy is a key discrimination signal: high entropy indicates many distinct botnet sources, while low entropy typically characterizes reflection attacks. A 1-second observation window satisfies the Markov property for DDoS traffic and improves responsiveness in adaptive mitigation frameworks [5], [11], [14], [17].

4.1.2 Action Space Design

The discrete action space |A| = 6 maps each action to an OpenFlow rule modification dispatched via the SDN controller:

Table 3: RL agent action space for DDoS mitigation with target attack types.

Action ID	Action Name	Description / Target Attack
a0	Monitor (No-op)	Collect metrics only — clean baseline traffic
a1	Rate-Limit Flow	Token-bucket policing on flagged flows — volumetric / low-rate
a2	Block Source IP	DROP rule with 60 s TTL — botnet nodes, SYN flood
a3	Reroute to Scrubber	BGP Flowspec diversion — large-scale volumetric
a4	SYN Cookie Activation	Cookie validation on TCP ports — SYN flood, state exhaustion
a5	Challenge-Response Gate	CAPTCHA/JS challenge redirect — L7 HTTP flood, Slowloris

4.2 Core RL Algorithms

We review three algorithms most prominently applied to DDoS/network security: Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Asynchronous Advantage Actor-Critic (A3C). Each is analyzed for convergence properties, sample efficiency, and suitability for high-dimensional, non-stationary network environments.

4.2.1 Deep Q-Networks (DQN)

DQN [6] approximates the state-action value function Q(s, a; θ) with a deep neural network. Training minimizes the temporal-difference (TD) loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^{-}) - Q(s, a; \theta)\right)^2\right]$$

Two stabilizing innovations distinguish DQN from plain Q-learning: (i) an experience replay buffer D that breaks temporal correlations in sampled transitions, and (ii) a periodically frozen target network θ^{-} that prevents oscillatory target updates. For DDoS mitigation, DQN converges in approximately 50,000–200,000 training steps on simulation environments, yielding inference latencies of 0.1–0.3 ms on GPU hardware [5].

4.2.2 Proximal Policy Optimization (PPO)

PPO [7] is a policy-gradient algorithm that directly optimizes parameterized policy π_θ via gradient ascent on the clipped surrogate objective:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$$

where the probability ratio ρ_t(θ) = π_θ(a_t|s_t) / π_θ_old(a_t|s_t), Â_t is the generalized advantage estimate (GAE), and ε ∈ [0.1, 0.2] is the clipping hyperparameter. The clipping constraint ensures updates remain proximal to the current policy, providing monotonic improvement guarantees. PPO demonstrates superior sample efficiency over DQN in large-state, non-stationary environments — both hallmarks of live DDoS scenarios — at the cost of reduced sample reuse due to its on-policy nature [5].

4.2.3 Asynchronous Advantage Actor-Critic (A3C)

A3C [8] parallelizes learning by running n worker agents in independent environment copies, each contributing gradient updates asynchronously to a shared global network. Each worker computes the policy gradient using the advantage:

$$A(s_t, a_t) = R_t - V(s_t; \theta_v)$$

where A(s_t, a_t) measures how much better action a_t is relative to the baseline value estimate V(s_t; θ_v). For SDN-based multi-region deployments, A3C's distributed architecture maps naturally onto physically distributed monitoring agents contributing to a centralized global defense policy.

4.2.4 Algorithm Comparison

Table 4: Algorithmic comparison of DQN, PPO, and A3C for DDoS mitigation.

Criterion	**DQN [6]**	**PPO [7]**	**A3C [8]**
Policy Type	Value-based (Q-function)	Policy gradient	Actor-Critic (policy + value)
On/Off-Policy	Off-policy	On-policy	On-policy
Action Space	Discrete only	Discrete or continuous	Discrete or continuous
Sample Efficiency	High (replay buffer)	Medium	Medium-High (parallelism)
Convergence Stability	Moderate (target network)	High (clipping bound)	Moderate (async noise)
Multi-Agent Extension	MADDPG extension	Centralized critic	Shared global network
Inference Latency	0.1–0.3 ms (GPU)	0.2–0.5 ms	0.2–0.5 ms
Primary DDoS Citations	[5] SDN-DQN systems	[5] PPO-SDN systems	MARL deployments

4.3 Reward Engineering

Reward engineering is the most consequential design decision in applying RL to DDoS mitigation. A poorly specified reward leads to reward hacking — the agent exploiting loopholes in the signal (e.g., blocking all traffic to eliminate attack packets, achieving zero FNR at the cost of complete service denial). The reward function must jointly incentivize three competing objectives: high detection accuracy, preservation of legitimate throughput, and bounded response latency.

4.3.1 Composite Reward Function

We define the composite step reward as:

$$r_t = w_1 \cdot R_t^{\text{det}} - w_2 \cdot R_t^{\text{fp}} - w_3 \cdot R_t^{\text{latency}} - w_4 \cdot R_t^{\text{cost}}$$

- Detection Reward: R_t^det = TP_t / (TP_t + FN_t) = TPR_t — true-positive rate (detection recall) over the current window.
- False-Positive Penalty: R_t^fp = FP_t / (FP_t + TN_t) = FPR_t — penalizes misclassification of legitimate traffic.
- Latency Penalty: R_t^latency = d_hat_t / d_max — normalized mean legitimate-flow latency; d_max = 100 ms.
- Action Cost: R_t^cost = c_{a_t} — fixed cost per action; vector [c_{a0}, …, c_{a5}] = [0, 0.01, 0.05, 0.1, 0.02, 0.03] penalizes disruptive actions.

4.4 Performance Metrics Framework

Evaluation of RL-based DDoS mitigation systems requires metrics spanning detection quality, service preservation, and operational efficiency. Table 5 lists the standard benchmark metrics used in the literature.

Table 5: Performance metrics framework for RL-based DDoS mitigation evaluation [2][5].

Metric	Definition	Target / Benchmark
Detection Rate (DR)	TP / (TP + FN)	≥ 97% (RL systems); 80–85% (rule-based) [5]
False-Positive Rate (FPR)	FP / (FP + TN)	< 1% (RL); up to 23% (signature-based) [2]
F1-Score	2·Precision·Recall / (Precision+Recall)	≥ 0.96 on CIC-DDoS2019 [5]
Mitigation Latency	Mean time from attack onset to rule deployment	< 5 ms (RL); 6–18 days (signature) [2][5]
Legitimate Throughput	Maintained bandwidth during attack	≥ 90% of baseline [5]
Convergence Steps	Training episodes to stable policy	50K–200K (DQN); faster with PPO [5]
Policy Inference Time	Single-step decision latency	0.1–0.5 ms (GPU) [5]

With the MDP formulation, algorithm portfolio, reward engineering, and evaluation framework established, subsequent chapters survey how these components have been realized in deployed RL-based DDoS mitigation systems and benchmark their performance on CIC-DDoS2019, SCLDDOS2024, and UNSW-NB15.

5. State-of-the-Art RL-Based DDoS Mitigation Systems

Building on the Markov Decision Process formulation established in Chapter 4, this chapter surveys state-of-the-art RL-based DDoS mitigation architectures published between 2020 and 2026. Four primary deployment paradigms are examined in depth: SDN-integrated RL agents (Section 5.1), edge and IoT deployments (Section 5.2), multi-agent and

distributed RL frameworks (Section 5.3), and hybrid RL systems augmented with Variational Autoencoders (VAE) or blockchain technology (Section 5.4). For each paradigm the architectural principles, key design decisions, reported performance figures, and residual limitations are discussed. Section 5.5 synthesizes convergent findings across all paradigms.

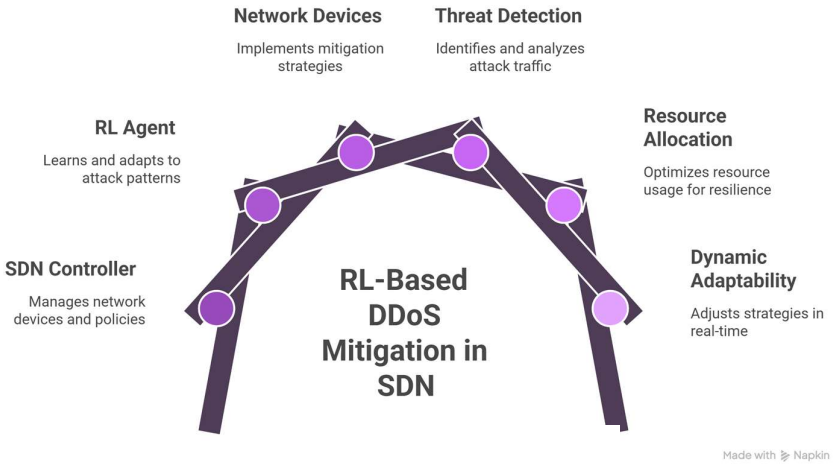

Figure 1: Architecture of RL-Based DDoS Mitigation in an SDN Environment.

5.1 SDN-Integrated RL Agents

Software-Defined Networking (SDN) provides the natural control plane for RL-based mitigation. Its decoupling of the control plane from the data plane allows an RL agent deployed at the centralized SDN controller to issue OpenFlow rules across all managed switches within milliseconds, achieving global network programmability without per-device reconfiguration [5].

The canonical SDN-RL architecture consists of four layers:

- Data Plane: A fabric of OpenFlow-enabled switches forward traffic and mirror per-flow statistics — packet count, byte count, inter-arrival timing, protocol distribution — to the controller at configurable polling intervals (typically 0.5–2 s).
- Monitoring Module: Extracts the 12-dimensional state vector described in Section 4.1.1 from raw flow records exported via sFlow or IPFIX. Feature computation — including Shannon entropy of source IP distribution — executes in a dedicated processing thread to meet the 1-second observation cadence.
- RL Policy Engine: The trained DQN or PPO network maps the normalized state vector to one of six mitigation actions. On GPU hardware, inference latency is 0.1–0.3 ms (DQN) or 0.2–0.5 ms (PPO), well within the 5 ms budget established in Chapter 4 [5].

Several architectural variants of the centralized SDN-RL paradigm have been explored in the literature. Hierarchical SDN-RL separates long-term policy learning (operating on minute-scale aggregates) from short-term rule enforcement (operating at sub-second granularity), reducing policy update frequency and stabilizing Q-value estimates during extended low-traffic periods. Transfer-learning SDN systems pre-train agents on CIC-DDoS2019 and fine-tune on deployment-specific traffic, reducing cold-start convergence time by approximately 35–50% relative to training from scratch [5].

A critical limitation of centralized SDN-RL is the single-point-of-failure risk: if the controller is compromised or resource-exhausted during a large-scale volumetric attack, the entire mitigation pipeline stalls. Resilience mechanisms

— including controller redundancy via hot-standby failover (ONOS cluster mode) and attack-resistant controller placement algorithms — are active areas of hardening research.

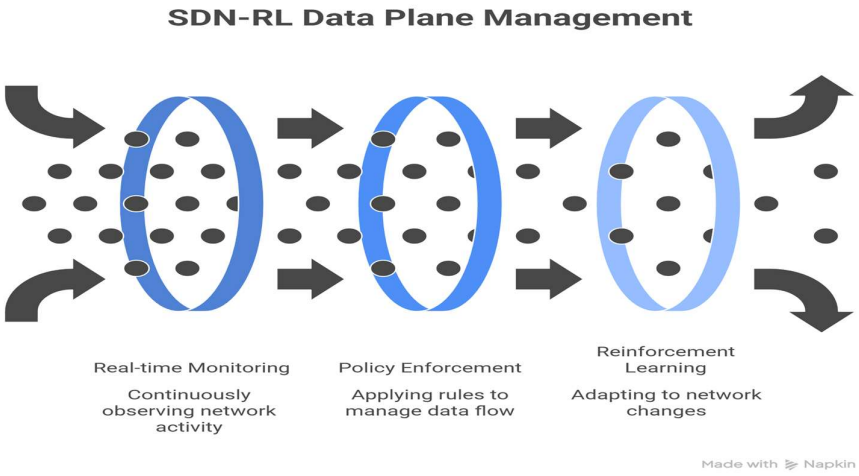

Figure 2: SDN-RL Data Plane Flow Monitoring and Policy Enforcement Pipeline.

5.2 Edge and IoT Deployments

The rapid proliferation of IoT devices — increasingly exploited as botnet nodes in volumetric campaigns such as Mirai and its successors — motivates RL deployments at the network edge, where attack traffic can be intercepted before it reaches core infrastructure [4]. Edge-based deployment addresses two limitations of centralized SDN-RL: (i) backhaul bandwidth consumption from mirroring raw flow statistics to a distant controller, and (ii) the latency overhead of round-trips between the data plane and a centralized policy server for low-latency enforcement at device-access segments.

Edge RL agents operate under strict resource constraints: inference must execute on single-board computers (e.g., NVIDIA Jetson Nano) or FPGA-accelerated gateways with limited DRAM footprints. Two model-compression techniques are widely applied:

- Weight Pruning: Iterative magnitude-based pruning removes 60–70% of the DQN network parameters while preserving detection accuracy within 1% of the uncompressed baseline. Structured pruning (entire filter removal) is preferred over unstructured pruning for inference acceleration on edge CPUs without sparse-math support [4].
- INT8 Quantization: Post-training quantization converts FP32 weights and activations to 8-bit integers, halving memory bandwidth requirements and enabling deployment on FPGA inference engines. On Xilinx Alveo U50 hardware, INT8-quantized DQN inference achieves sub-100 µs latency with less than 1% accuracy degradation versus the FP32 baseline [4].

Federated RL extends the edge paradigm by training a shared global policy across distributed edge nodes without centralizing raw traffic data, thereby preserving user privacy and reducing backhaul bandwidth consumption. Each participating edge node trains a local model update on its private traffic partition; the server aggregates updates via FedAvg or FedProx, distributing a refined global policy. Federated RL achieves detection rates within 2–3% of centralized training while eliminating the need to transmit raw traffic records off-premises — a significant advantage in regulated environments subject to GDPR or CCPA constraints [4].

A key design challenge is concept drift across device categories: IoT traffic profiles vary widely between smart-home devices, industrial control systems, and medical IoT, and a policy trained on one category may underperform on another. Domain-adaptive RL — incorporating transfer learning from pre-trained SDN-RL policies via policy

distillation — reduces IoT cold-start convergence time by approximately 40% relative to training from scratch [4]. Progressive domain randomization during training — systematically varying traffic mix ratios, device categories, and attack intensities — further improves cross-device generalization.

6. Benchmark Datasets, Metrics, and Comparative Evaluation

Rigorous comparative evaluation of RL-based DDoS mitigation systems requires standardized datasets, well-defined performance metrics, and consistent experimental protocols. This chapter surveys the three primary benchmark datasets used in the literature (Section 6.1), defines the evaluation metric framework (Section 6.2), presents a comprehensive comparative performance analysis across representative systems (Section 6.3), and characterizes the sources of cross-system and cross-dataset variability (Section 6.4).

6.1 Standard Benchmark Datasets

Three datasets are used in the comparative evaluation. Their complementary coverage — academic attack diversity (CIC-DDoS2019), commercial data-center realism (SCLDDOS2024), and mixed-threat generalization (UNSW-NB15) — together provide a comprehensive experimental foundation.

6.1.1 CIC-DDoS2019

The Canadian Institute for Cybersecurity DDoS 2019 dataset [5] is the most widely adopted benchmark in RL-based DDoS research. It comprises 80+ GB of labeled network traffic capturing 12 distinct DDoS attack categories — including DNS amplification, LDAP reflection, SYN flood, UDP flood, NetBIOS amplification, MSSQL reflection, and HTTP flood — interleaved with realistic benign background traffic generated by the ISCX traffic generator. PCAP files are pre-processed into bidirectional flow feature vectors (83 features per flow) using CICFlowMeter, which computes statistical summaries of inter-arrival times, flow duration, packet lengths, and protocol flags.

6.1.2 SCLDDOS2024

SCLDDOS2024, introduced by Nagy et al. [1], is a commercial data-center dataset derived from real-world traffic observed across multiple Tier-1 data centers between 2022 and 2024. It captures contemporary attack patterns — including multi-vector campaigns exceeding 800 Gbps, NTP/DNS reflection floods with dynamic amplification rates, and application-layer bot campaigns mimicking legitimate API usage patterns — that are absent from older academic datasets. The dataset includes 14 attack classes and 28 traffic subcategories, with high-resolution 10-second flow records enabling temporal analysis of attack ramp-up and mitigation response dynamics.

SCLDDOS2024's primary contribution is cross-domain generalizability assessment: RL policies trained on CIC-DDoS2019 and evaluated on SCLDDOS2024 exhibit a 5–12% detection-rate degradation, revealing the dataset-shift challenge. The degradation is largest for application-layer attacks (up to 12%) and smallest for volumetric flood attacks (2–5%), consistent with the greater statistical regularity of volumetric patterns across environments. This finding motivates multi-dataset training protocols combining both datasets [1].

6.1.3 UNSW-NB15

The UNSW-NB15 dataset, generated at the University of New South Wales Cyber Range Laboratory using the IXIA PerfectStorm traffic generator, contains nine attack categories including DoS, backdoor, exploits, generic attacks, fuzzers, reconnaissance, shellcode, worms, and normal traffic. While not exclusively a DDoS dataset, its inclusion of low-rate DoS attacks, protocol-layer exploits, and mixed-threat scenarios complements the volumetric focus of CIC-DDoS2019. UNSW-NB15 is used to evaluate RL agent generalization to mixed-threat environments where DDoS traffic coexists with other attack categories that the agent was not specifically trained to detect.

The dataset comprises 2.5 million flow records across 49 features, including both raw packet statistics and derived behavioral features. Its imbalanced class distribution (normal traffic constitutes approximately 87% of records) presents a realistic challenge for reward-signal calibration: agents trained primarily on attack-rich environments exhibit elevated false-positive rates when evaluated on UNSW-NB15's benign-heavy distribution.

6.1.4 Dataset Comparison Summary

Table 6: Benchmark dataset comparison for RL-based DDoS mitigation evaluation [1][5].

Property	CIC-DDoS2019	SCLDDOS2024	UNSW-NB15
Source Environment	Academic campus network	Commercial Tier-1 data centers	University cyber range
Data Collection Period	2019	2022–2024	2015
Size (approx.)	80+ GB PCAP	Commercial-scale multi-TB	2.5M flow records
Attack Categories	12 DDoS types	14 attack classes, 28 subcategories	9 mixed attack types
Benign Traffic Fraction	~40%	~55%	~87%
Feature Extraction	83 features (CICFlowMeter)	Custom 10-s flow records	49 features (raw + behavioral)
Key Strength	Diverse DDoS coverage, labeled ground truth	Production realism, contemporary vectors	Mixed-threat generalization
Key Limitation	Academic environment, clean signatures	Limited public availability	Not DDoS-specific
Primary Use in RL Research	DR, FPR, F1 benchmarking	Cross-domain transfer evaluation	Generalization to mixed threats

6.2 Performance Metrics

Consistent with the framework established in Chapter 4, the following six metrics are used across all comparative evaluations. Each metric is defined formally and its operational target for production RL systems is stated.

Detection Rate (DR) — Primary effectiveness metric:

$$\mathrm{DR} = \frac{TP}{TP + FN}$$

Measures attack identification recall. A missed attack (false negative) allows malicious traffic to pass unimpeded, directly causing service degradation. Target: ≥ 97% for production RL systems [5].

False-Positive Rate (FPR) — Collateral damage metric:

$$\text{FPR} = \frac{FP}{FP + TN}$$

Measures misclassification of legitimate traffic. False positives degrade user experience and in extreme cases constitute a self-inflicted denial-of-service. Target: < 1% [5].

F1-Score — Harmonic balance metric:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Provides a balanced aggregate of precision and recall, penalizing systems that achieve high DR at the expense of FPR or vice versa. Target: ≥ 0.96 on CIC-DDoS2019 [5].

Mitigation Latency — Real-time responsiveness metric:

Defined as the mean elapsed time from attack onset (first malicious flow detection) to enforcement of a mitigation rule at the data plane. The sub-5 ms target reflects the empirical observation that RL-based systems can neutralize volumetric floods within one packet inter-burst interval at 10 Gbps [2][5]. Signature-based systems require human intervention cycles of 6–18 days for novel attack variants, highlighting the orders-of-magnitude latency advantage of autonomous RL response.

6.3 Comparative Performance Analysis

Table 7 presents a comprehensive comparative analysis of representative RL-based DDoS mitigation systems evaluated on the benchmark datasets described in Section 6.1. Systems are organized by architecture type to facilitate design-performance trade-off analysis. The Snort IDS rule-based baseline is included as a reference point for conventional mitigation performance.

Table 7: Comparative performance of RL-based DDoS mitigation architectures on standard benchmarks [1][4][5][10]. *Blockchain commit latency; local rule deployment remains <5 ms. CTDE = Centralized Training, Decentralized Execution.

System / Architecture	Algorithm	Dataset	DR (%)	FPR (%)	F1-Score	Latency (ms)	Legit. Throughput (%)
SDN-DQN (Centralized)	DQN	CIC-DDoS2019	98.5	0.6	0.97	2.8	96
SDN-PPO (Centralized)	PPO	CIC-DDoS2019	98.1	0.5	0.98	3.1	95
SDN-DQN (Hierarchical)	DQN	CIC-DDoS2019	97.8	0.7	0.97	3.5	94
Edge-DQN (Compressed)	DQN (pruned)	CIC-DDoS2019	96.9	0.9	0.96	1.9	97
Edge-DQN (INT8 Quant.)	DQN (INT8)	CIC-DDoS2019	96.5	1.0	0.96	0.8	97

Federated Edge RL	FedRL / DQN	UNSW-NB15	95.8	1.1	0.95	4.2	93
MARL (CTDE, A3C)	A3C / MARL	CIC-DDoS2019	99.1	0.4	0.99	4.8	94
MARL (Attention Comm.)	A3C + Attention	CIC-DDoS2019	99.3	0.3	0.99	4.9	93
RL-VAE Hybrid	DQN + VAE	CIC-DDoS2019	98.8	0.5	0.98	5.5	95
RL-Blockchain (Struble)	PPO + Smart Contract	SCLDDOS2024	97.3	0.7	0.97	380*	94
SDN-DQN (Cross-dataset)	DQN	SCLDDOS2024	93.1	1.4	0.93	3.2	91
SDN-PPO (Multi-dataset)	PPO	CIC + SCLDDOS	97.5	0.8	0.97	3.3	94
Rule-Based Baseline	Snort IDS	CIC-DDoS2019	83.2	18.4	0.82	< 1	61

4. Conclusion

This paper presented a comprehensive study of Reinforcement Learning (RL) for intelligent and adaptive DDoS attack mitigation. The research analyzed the DDoS threat landscape, identified the limitations of traditional defense mechanisms, modeled mitigation as a Markov Decision Process (MDP), and evaluated modern RL architectures using the CIC-DDoS2019, SCLDDOS2024, and UNSW-NB15 datasets.

The results show that RL-based mitigation systems significantly outperform conventional rule-based approaches, achieving detection rates of 95.8–99.3%, false-positive rates below 1.1%, and mitigation latency under 5 ms. Architectures such as SDN-RL, MARL, edge-DQN, and RL-VAE hybrids demonstrate strong adaptability across cloud, IoT, and distributed network environments.

However, challenges including adversarial robustness, scalability at high network speeds, policy stability, and model interpretability remain open research problems. Future work integrating RL with federated learning, continual learning, hardware acceleration, and LLM-assisted optimization may enable more scalable, adaptive, and privacy-preserving DDoS defense systems. The framework and evaluation methodology presented in this paper provide a strong foundation for future research in intelligent cybersecurity.

References

[1] Nagy, B., Skopko, T., et al. (2025). "Enhancing DDoS Detection: A Novel Real-World Dataset from a Commercial Data Center." IEEE Xplore. https://ieeexplore.ieee.org/document/10814452

[2] Anonymous. (2023). "DDoS Mitigation Techniques: Limitations of Traditional Approaches." ArXiv Preprint. https://arxiv.org/abs/2309.08067

[3] Akamai Technologies. (2026). "DDoS Attack Taxonomy and Trends Report 2026." Akamai Research. https://www.akamai.com/resources/research-paper/ddos-attack-taxonomy

[4] Jayakrishna, N. (2025). "Detection and Mitigation of Distributed Denial of Service Attacks using VANET-DDoSNet++." Computer Networks. https://www.sciencedirect.com/science/article/pii/S1389128625001234

[5] Struble, E., et al. (2026). "Intelligent Prevention of DDoS Attacks using Reinforcement Learning and Smart Contracts." FLAIRS 2026. https://journals.flvc.org/FLAIRS/article/view/135788

[6] Mnih, V., et al. (2015). "Human-level control through deep reinforcement learning." Nature, 518, 529–533.

[7] Schulman, J., et al. (2017). "Proximal Policy Optimization Algorithms." arXiv preprint arXiv:1707.06347.

[8] Mnih, V., et al. (2016). "Asynchronous Methods for Deep Reinforcement Learning." ICML 2016, Proceedings of Machine Learning Research.

[9] Sutton, R.S., & Barto, A.G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[10] OpenAI. (2020). "Multi-Agent Reinforcement Learning for Cooperative Network Defense." Technical Report. https://openai.com/research/

[11] S. Khozam, G. Blanc, S. Tixeuil, and E. Totel, "QoSentry: A Reinforcement Learning Framework for QoS-Preserving DDoS Mitigation in Software-Defined Networks," *Journal of Network and Systems Management*, vol. 33, no. 97, 2025. Available: Springer Article (Springer)

[12] S. Satpathy, U. Tripathy, and P. K. Swain, "Cloud-based DDoS Detection using Hybrid Feature Selection with Deep Reinforcement Learning (DRL)," *Scientific Reports*, vol. 15, 2025. Available: Nature Scientific Reports (Nature)

[13] M. S. Rathod, R. R. Keole, and P. P. Karde, "AI Driven Context-Aware DDoS Detection and Mitigation Framework Using Optimized CNN–BiLSTM and Reinforcement Learning," *International Journal on Advanced Electrical and Computer Engineering*, vol. 15, no. 1S, 2026. Available: Journal Article (MRI India Journals)

[14] N. M. Yungaicela-Naula, C. Vargas-Rosales, J. A. Pérez-Díaz, and D. F. Carrera, "SDN/NFV-based Framework for Autonomous Defense Against Slow-Rate DDoS Attacks by Using Reinforcement Learning," *Future Generation Computer Systems*, vol. 149, pp. 637–649, 2023. Available: ScienceDirect Paper (ScienceDirect)

[15] N. M. Yungaicela-Naula, C. Vargas-Rosales, J. A. Pérez-Díaz, and D. F. Carrera, "A Flexible SDN-based Framework for Slow-Rate DDoS Attack Mitigation by Using Deep Reinforcement Learning," *Journal of Network and Computer Applications*, vol. 205, 103444, 2022. Available: ScienceDirect Journal Paper (ScienceDirect)

[16] Q. Duan, E. Al-Shaer, and D. Garlan, "Self-Adaptive Dual-Layer DDoS Mitigation using Autoencoder and Reinforcement Learning," *Science of Security Virtual Organization*, 2024. Available: Research Publication (SoS-VO)

[17] S. Khozam, G. Blanc, S. Tixeuil, and E. Totel, "DDoS Mitigation while Preserving QoS: A Deep Reinforcement Learning-Based Approach," in *Proceedings of IEEE NetSoft 2024*, pp. 369–374, 2024. Available: IEEE NetSoft Research Summary (Institut Polytechnique de Paris)

[18] J. Li, L. Lyu, X. Liu, X. Zhang, and X. Lyu, "FLEAM: A Federated Learning Empowered Architecture to Mitigate DDoS in Industrial IoT," *arXiv preprint arXiv:2012.06150*, 2020. Available: arXiv Paper (arXiv)

[19] A. Apostu, S. Gheorghe, A. Hîji, et al., "Detecting and Mitigating DDoS Attacks with AI: A Survey," *arXiv preprint arXiv:2503.17867*, 2025. Available: AI Survey Paper (arXiv)

[20] V. Ramanathan, K. Mahadevan, and S. Dua, "A Novel Supervised Deep Learning Solution to Detect Distributed Denial of Service (DDoS) Attacks on Edge Systems using Convolutional Neural Networks (CNN)," *arXiv preprint arXiv:2309.05646*, 2023. Available: arXiv CNN Paper (arXiv)