International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 2, 41-49, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I2P108

Original Article

Next-Generation Firewall (NGFW) Testing for Gen 6 and Gen 7 Devices

John Komarthi
San Jose, CA.

Received On: 02/03/2026 ***Revised On: 27/03/2026*** ***Accepted On: 04/04/2026*** ***Published On: 15/04/2026***

Abstract *- Next-generation firewalls (NGFWs) have to enforce the security policies and detect threats within predominantly encrypted and application traffic, while at the same time ensuring that the user performance at an enterprise scale. This paper presents a rigorous, scalable, and aligned with standards methodology for testing "Gen 6" and "Gen 7" NGFW devices. Building on the IETF's RFC 9411 benchmarking methodology for the next-generation network security devices, the test objectives are defined, including traffic profiles, Key Performance Indicators (KPIs), testbed architecture, data collection methods, and statistical treatments necessary to evaluate the factors. The factors include security effectiveness, connection scale, throughput, latency, and the cost of TLS/SSL inspection, along with the Deep Packet Inspection (DPI) and the accuracy in the application identification. This framework also assesses the robustness of the devices against the evasion techniques. This paper includes a test matrix, sample reporting templates, and automation guidance to support the results, which can be reviewed.*

Keywords *- Next-Generation Firewall, RFC 9411, QUIC, NGFW Benchmarking, Encrypted Clienthello, TLS Inspection, DPI, Application Identification, HTTP/3, Reproducibility, Evasion Resistance.*

1. Introduction

The baseline methodology is the IETF benchmarking standard for the next-generation network security devices, RFC 9411 [1], which explicitly positions the security effectiveness as a prerequisite to the performance benchmarking, and they define the realistic application traffic mixes, including HTTPS and HTTP/3, pass/fail validation criteria, and the required reporting artefacts. The terms "Gen 6" and "Gen 7" are primarily used as the vendor-specific generation labels, especially in SonicWall documentation for both appliances and NSv virtual firewalls. In this context, Gen6 and Gen7 correspond to the different operating system versions and the operations. For instance, the Gen 6 NSv runs on a SonicOS 6.5-based image, and they only support the classic mode, while the Gen 7 NSv utilizes the SonicOS 7 and, depending on the release, may support both classic mode and policy mode. Also, there is a published end-of-support date for the Gen6 NSv.

Gen6/Gen7 are addressed as an unspecified context for the outside vendors, and then the plausible generational attributes that need verification are outlined. These attributes include hardware acceleration, higher interface speeds, new cryptographic/TLS capabilities, and increased connection scalability. Testing plan is designed to support this verification, the threat landscape which drives the Next-Gen Firewall (NGFW) testing is characterized through several key factors, the near ubiquity of the web encryption, the prevalence of modern encrypted transport protocols, and the rapid exploitation of the high-severity vulnerabilities at the network perimeter particularly those related to the SSL VPNs, authentication, and management plane issues. These issues can lead directly to ransomware incidents and compromise the enterprises. The current telemetry, which is available publicly indicate that the HTTPS browsing on Chrome typically falls within the mid to high 90 percent range. Realistic NGFW tests have to include encrypted workloads and, when permissible, TLS interception. Protocols such as QUIC and HTTP/3 have been standardized (RFC 9000 [2], RFC 9001 [3], RFC 9114 [4]), and privacy enhancements like TLS-encrypted ClientHello (ECH, RFC 9849 [5]), which further reduce the passive visibility. This development has claims regarding the 'application identification' and 'domain visibility' increasingly reliant on the endpoint telemetry, policy-based handling, or active methods rather than solely depending on Deep Packet Inspection (DPI). The recent vulnerabilities exploited in the firewall platforms, including Gen6/Gen 7-labelled SonicOS devices, highlight the necessity of including management-plane hardening, policy correctness, and resilience testing under attack-like conditions in the test protocols.

2. EVOLUTION OF NGFWs

2.1. Evolutionary Baseline

A foundational difference in the firewall engineering between stateless packet filtering and stateful inspection. The industry benchmarking terminology formalised this shift through firewall-specific definitions and metrics in RFC 2647 [6], and firewall performance methodologies in RFC 3511 [7]. NIST guidance also frames the firewall policy [8], deployment considerations, default-deny posture, and the operational need to test firewalls as part of the lifecycle management, even if the document predates the modern NGFW feature sets.

2.2. NGFW in practice

Modern NGFWs extend beyond L3/L4 enforcement to include the inspection and classification at higher layers, which often combines the intrusion prevention, application identification, URL/content controls, TLS decryption, and centralised management. RFC 9411 [11] captures this broad network security device scope and explicitly targets the NGFW's and NGIP operating in-line with realistic application traffic mixes (HTTP/HTTPS/HTTP/3), which reflects the contemporary deployment expectations.

2.3. Standards-based benchmarking

A repeating problem in firewall benchmarking is non-comparability because of the inconsistent definitions, workloads, and reporting practices. The University of New Hampshire InterOperability Laboratory (UNH-IOL) [9] commentary on NGFW testing highlights how the inconsistent methods can produce incomparable results, and it motivates open standards-based approaches through the IETF benchmarking community. RFC 9411[11] is the current IETF document addressing this; it obsoletes RFC 3511 for next-generation devices and defines both testbed requirements and reporting requirements that are intended to improve the transparency and reproducibility.

3. Gen 6 & Gen 7 Device Characteristics

3.1. Generation terminology

"Gen6" and "Gen7" are unspecified as an industry taxonomy across all the vendors; instead, in accessible aprimary sources, Gen6/Gen7 clearly appear as vendor generation labels for specific product families and OS trains. Eg, SonicWall technical documentation [10] distinguishes Gen6 and Gen7 NSv models by the software image train (6.5 vs 7.x) and the operational mode support (classic vs policy mode), and publishes the lifecycle information such as the support end date for Gen6 NSv.

3.2. Vendor evidenced Gen6/Gen7 traits

In SonicWall's NSv FAQ, Gen6 NSv models are documented as running on a SonicOS 6.5.4-44v-based image, not on a SonicOS7 image, and support the classic mode only. Gen7 NSv is documented as supporting both the classic and policy modes from SonicOS 7.0.1 in the cited KB context, with partial configuration migration limitations while switching the modes. An upgrade KB notes that Gen7 NSv added the classic/global mode support, which was previously the only mode for Gen6 NSv, and it states the end of support for Gen6 NSv on 16 April 2026.

3.3. Performance and scale deltas

Gen 7 vs Gen 6 upgrade document for the TZ series shows the differences in the order of magnitude and associates the Gen 7 with SonicOS 7 features, including TLS 1.3 support and default BGP routing without additional license. Gen7 datasheets for higher-end series also explicitly advertise the TLS 1.3 support and 'millions of simultaneous TLS connections', along with high firewall inspection throughput targets.

3.4. Plausible but non-guaranteed Gen6 to Gen7 attributes

When the vendors label Gen-7, it plausibly correlates with the higher port speeds, e.g., multigig/10GbE on smaller form factors, more concurrent flow and TLS state capacity, improved cryptographic acceleration for the TLS inspection, and a redesigned management plane and the policy model. These plausibilities are consistent with the vendor claims of higher throughput, higher max connections, and explicit TLS 1.3 support on Gen7 product lines, but they are not universal properties across all the vendors and cannot be assumed without test evidence.

Table 1: Interpreting Gen 6 Vs Gen 7 Characteristics (Vendor-Specific Where Evidenced, Otherwise Unspecified)

Dimension	Gen6 (vendor label; example evidenced)	Gen7 (vendor label; example evidenced)	If unspecified: what to verify in the lab
OS / software train	Gen6 NSv documented as 6.5.4-44v-based image, classic mode only	Gen7 NSv documented as SonicOS 7.x; supports classic/global and policy modes (release-dependent)	Firmware train, feature parity, config migration paths, operational mode differences
Lifecycle / support	Gen6 NSv EOS date stated: 16 Apr 2026 (example)	Not stated in the same KB	Vendor lifecycle policy, patch cadence, CVE exposure window
Encrypted traffic support	Not specified globally; depends on product	Gen7 datasheets claim TLS 1.3 support and high TLS state capacity	TLS versions, cipher suites, TLS inspection throughput, handshake rate, failure modes
Scaling	Not specified globally	Example claims: higher max connections, VPN tunnels, VLAN interfaces	Max concurrent sessions, CPS, NAT table, policy scale, control-plane limits
Interfaces & throughput	Not specified globally	Example datasheet claims multigig/10GbE in small appliances; high throughput in enterprise	Throughput at realistic mixes, latency (TTFB/TTLB), PPS ceilings, jitter

The example evidence Gen6/Gen7 statements above are grounded in vendor KB datasheets and must not be generalised to non-vendor contexts without confirmation.

4. Threat Landscape and Testing Objectives

4.1. Encryption and default operating condition

Web encryption is universal at scale. Google's HTTPS transparency reporting [11] states that the Chrome browsing

time using HTTPS typically is in the mid-to-high 90 percent range in early 2026, reinforcing that the clear-text-only firewall benchmarks are no longer operationally representative. Vendor datasheets position the encrypted session processing as a dominant workload; e,g., a Gen7 datasheet stated that more than 70 percent of all the sessions are encrypted.

4.2. Modern encrypted transport protocols reduce middlebox visibility

QUIC (RFC 9000) standardises an encrypted-by-design transport over UDP, secured using TLS 1.3 mechanisms described for QUIC (RFC 9114)[2]. From a testing perspective, this translates to NGFW 'application identification' and 'policy enforcement' claims have to be evaluated under the HTTP/3 traffic profiles and under mixed HTTP/1.1-HTTP/3 realities as specified in the RFC 9411's guidance for QUIC and HTTP/3 emulation parameters.

4.3. ECH and shrinking set of passive identifiers

TLS Encrypted ClientHello (ECH) is now standardized in RFC 9849, which enables the encryption of ClientHello under a server public key and reduces the exposure of metadata historically used for domain-based policy. Operational analysis notes that the ECH relies on HTTPS DNS records to convey key material, highlighting that the network visibility and policy mechanisms are shifting and have to be tested.

4.4. Exploitation pressure on VPN and management planes

The recent high-severity vulnerabilities illustrate why the NGFW testing has to include "control plane + management plane" resilience and hardening, not just the data plane throughput. NVD descriptions for CVE-2024-40766 [12] explicitly scope impact to "Gen5 and Gen6 devices" and "Gen7 devices running SonicOS 7.0.1-5035 and older versions" (vendor-labelled generations). This demonstrates how the generation labels can intersect with the vulnerability exposure. Similarly, CVE-2024-53704 is described as an improper authentication issue in SSL, VPN, enabling the authentication bypass, and the CISA's KEV catalogue [13] includes the SonicWall SonicOS SSLVPN improper authentication, which indicates the exploitation relevance and the path urgency.

4.5. Macro threat drives shaping the workloads

ENISA's 2024 threat landscape identifies the prime threats as availability attacks and ransomware, the 2025 Verizon DBIR highlights the ransomware and vulnerability exploitation trends in the breach dataset, underscoring that the perimeter controls are evaluated not only on blocking but on sustaining the availability and limiting the blast radius under active exploitation.

4.6. Testing objectives

Consistent with the RFC 9411's structure [1], the security effectiveness validation, followed by the performance benchmarks, the tests for Gen6/Gen7 devices have to aim to:

- Validate the security effectiveness configuration, for example, how the vulnerability attack traffic is detected, blocked, and accurately reported, before interpreting the throughput benchmarks.
- Quantifying the performance cost of the security, inspected throughput under application mixes, connection establishment rate, transaction rate, and latency under HTTP/HTTPS/HTTP/3 with defined validation thresholds.
- The encrypted traffic handling limits have to be characterised, including TLS handshake rate, the concurrent TLS state, and the policy behavior under QUIC/HTTP/3 and under emerging metadata encryption shifts (ECH).
- Characterise encrypted-traffic handling limits: TLS handshake rate, concurrent TLS state, and policy behaviour under QUIC/HTTP/3 and under emerging metadata-encryption shifts (ECH).
- Demonstrating reproducibility, fixing the configurations, documenting traffic mixes, statistical treatment of variability, and isolated testbed constraints.

5. Testing Use Cases & Methodologies

Rigorous evaluation of the next-generation firewalls (NGFWs), across vendor-designated 'Gen6' and 'Gen7' devices, requires a methodology that moves beyond the synthetic throughput measurements and reflects the operational reality observed in modern cloud, enterprise, and hybrid environments.

5.1. Test design principles

The test cases are constructed around the foundational principles: (i) traffic realism, where the generated flows replicate the statistically accurate enterprise distribution, (ii) feature state isolation, ensuring that each security capability can be evaluated independently and in combination, and (iii) failure bound identification, where the objective is to determine the exact threshold at which the firewall deviates from the expected behavior. Each test is executed across multiple operational states of the device, the baseline forwarding mode, application-aware inspection mode, full security stack enabled (IPS+DPI+TLS inspection), and stress induced degradation conditions. This layered approach makes sure that the architectural efficiencies and security processing overheads are captured.

5.2. Encrypted enterprise traffic validation

The modern networks are full of encrypted traffic; this necessitates a primary test scenario that will evaluate the NGFW performance under HTTPS-heavy conditions. The firewall is subjected to a workload that consists of 70-90 percent of TLS-encrypted traffic, including both TLS 1.2 and TLS 1.3 sessions, with a smaller portion of unencrypted HTTP and API-driven traffic [14] [15]. The methodology involves the generation of concurrent client sessions that simulate realistic browsing and the application behavior. This includes the short-lived connections (REST calls), persistent sessions such as video streaming, and burst downloads. Traffic generation tools such as TRex are configured with stateful profiles to emulate the TCP handshake behavior, session reuse, and varying payload

sizes. Testing is performed incrementally, starting with TLS passthrough to establish a baseline, followed by full TLS inspection with the certificate interception enabled. During each phase, the throughput, transactions per second, TLS handshake rate, and CPU utilization are recorded. Attention is given to the degradation curve that is introduced by TLS inspection, as this directly reflects the cryptographic acceleration capabilities and the efficiency of the inspection pipeline. Additionally, the latency measurement, specifically the time to first byte (TTFB) and time to last byte (TTLB) are captured under a sustained load [1]. The objective is to quantify the trade-off between the inspection depth and user-perceived performance, which is a critical differentiator between generational firewall architectures.

5.3. Application identification and DPI accuracy testing

The application layer visibility is important, and its accuracy has to be evaluated under both normal and adversarial conditions [16]. The test case has to focus on validating the firewall's ability to correctly classify the traffic depending on the behavioral and payload signatures. A controlled dataset of the application traffic is constructed, including the web services, enterprise SaaS platforms, and custom-generated traffic patterns. Each flow is tagged with a ground truth label that enables the precise measurement of classification accuracy. The traffic is replayed using tcpreplay or generated dynamically using Scapy-based scripts that mimic the application-specific signatures. Firewall's classification logs are compared against the ground truth to compute the false positive and false negative rates. For example, the encrypted video streaming traffic has to be identified as its corresponding application, rather than being generically labeled as "SSL". In order to put more stress on the DPI engine, evasive traffic patterns will be introduced, such as protocol tunneling, payload obfuscation, and malformed headers [17]. The firewall's ability to maintain accuracy under these conditions provides insight into the robustness of its inspection engine and signature database.

5.4. Intrusion prevention and threat detection validation

Beyond mere classification, the next-generation firewalls (NGFWs) have to actively detect and block malicious traffic. The test scenario will evaluate the effectiveness of integrated intrusion prevention systems (IPS) through controlled attack simulations. The collection of known attack signatures is assembled, which includes the exploit payloads [18], command-and-control (C2) communication patterns, and web-based attack vectors. These payloads are delivered via both encrypted and unencrypted channels to assess the firewall's detection capabilities under different visibility conditions. The methodology involves replaying the attack traffic using the pre-captured PCAP files and generating the live exploit attempts where it is safe and controlled. The detection accuracy is measured by the true positive rate and the false negative rate [1]. The impact of IPS enforcement on throughput and latency is recorded, as signature matching introduces computational overhead. The crucial aspect of this test is maintaining a consistent inspection under load. The firewall is subjected to an increase in traffic volumes while the attack traffic is continuously injected. When the device starts to miss detections or allow malicious flows to pass, this is identified as the security degradation threshold, a critical metric that is often overlooked in vendor-reported benchmarks.

5.5. TLS handshake and cryptographic testing

The computational intensity of the modern encryption protocols, and the ability of NGFWs should be able to handle high volumes of TLS handshakes [15], becomes an important performance indicator. This particular test is directed towards the handshake rate and cryptographic processing efficiency. Clients are configured to initiate new TLS sessions at a high rate, avoiding session reuse to maximize the handshake load. RSA and elliptic curve cryptographic suites are tested to recreate the real-world pattern; the firewall's handshake rate, failure rate, and CPU utilization are recorded. The difference in performance of Gen6 and Gen7 is recorded, as the latest architectures incorporate hardware acceleration for cryptographic operations [15].

5.6. Failure mode and degradation analysis

The explicit identification of failure modes is an important feature [1]. Instead of reporting only the peak performance metrics, each of the tests seeks to find the specific conditions under which the firewall starts to fail. The failure is defined depending on the measurable criteria, that includes the pocket loss exceeding the acceptable thresholds, latency spikes beyond the predefined limits, security inspection bypasses, and session establishment failures. This enables a meaningful comparison between the devices, which highlights both their maximum capabilities and the operational stability under stress.

5.7. Reproducibility and validation

The test cases will be executed multiple times to ensure the statistical reliability, and the results will be reported as the mean values along with the confidence intervals. Detailed documentation of the test configurations, traffic profiles, and the tool versions are maintainted to enable reproducibility. The automation frameworks are employed to standardize the test execution and ensure consistency across the runs. Logs, packet captures, and performance metrics are archived for post-analysis validation [19].

6. Performance, Scalability, & Latency Benchmarking

The inspection of throughput, along with the application traffic mix, RFC 9411 test 7.1 defines the objective to determine the sustainable throughput after inspection, using a relevant application traffic mix [1]. The mix has to be documented, and if the TLS inspection is disabled, the report has to explain how it impacts the encrypted traffic in the mix. TLS/HTTPS benchmarking parameters, RFC 9411 specifies that the test clients have to use TLS 1.2 or higher, have to perform full handshakes with the session reuse disabled, and this provides the recommended cipher suite families that have to be documented to reflect the evolving use cases. It

also refers to the IANA-recommended TLS 1.3 cipher suites via RFC 8446 [15]. HTTP/3 (QUIC) considerations, RFC 9411 defines the QUIC stack emulation conformance to RFC 9000 and RFC 9001, which recommends specific QUIC parameter values for benchmarking, and includes validation criteria for the failed QUIC connections under HTTP/3. For the inspected throughput testing, RFC 9411 requires reporting of the inspected throughput and the application transactions per second as mandatory benchmarks and defines optimal TCP/TLS/QUIC KPIs, including the TCP connections per second, TLS handshake rate, and web transaction timing metrics, including time-to-first-byte (TTFB) & time-to-last-byte (TTLB).

7. Dpi & Application Identification Accuracy

NGFW policies are frequently hinged on the application identification (App-ID), URL categorisation, or IPS signatures. Misclassification causes false positives or false negatives. The research on endpoint aware inspection explicitly positions the false positives/false negatives [16] as the challenge when the security solution assesses without sufficient endpoint context, thus causing both passive fingerprinting and active methods.

7.1. Test approach for application identification and DPI accuracy

Establishing the ground truth, known application corpus (curated HTTP(S) workload, SaaS flows, enterprise apps), and known malicious corpus (CVE traffic set as per RFC 9411).

- Computation of confusion-matrix metrics for each application class (precision/recall/F1) and for each security category (malware/exploit/phishing classes) as a reporting layer, even though the RFC 9411's security effectiveness appendix is structured around blocked vs bypassed CVEs and reporting accuracy.
- Encrypted transport constraints have to be included; QUIC [2] obfuscates the handshake and SNI domain visibility when compared to TLS-over-TCP. The WUIC service classification accuracy can be materially impacted by the data drift and visibility limits. This requires testing App-ID accuracy across encrypted protocols rather than assuming the parity.

8. Evasion Techniques

8.1. Relevance

DPI and the intrusion prevention systems can be defeated or confused by the packet-level and protocol-level ambiguities. CCS 2025 study [17] on fingerprinting DPI devices enumerates probes that include overlapping of the IP fragments, TCP segments, and header mutations. This shows ambiguity-inducing transformations relevant to NGFW robustness tests. A 2025 arXiv study [17] analyses the overlapping IPv4/IPv6/TCP data and the prevalence of overlays during the reassembly. This supports building reassembly of test cases as a part of the robustness evaluation.

8.2. Techniques

- *Fragmentation:* Creation of overlapping IP fragments that will reassemble the malicious HTTP payload. For example, splitting a known exploit string across multiple IPv4 fragments so that the native reassembly might bypass the signatures. Use Scappy:

```
python

pkt1 = IP(src, dst, frag=1)/TCP()/b"malici"
pkt2 = IP(src, dst, frag=0)/TCP()/b"ous"
send(pkt1); send(pkt2)
```

- *TCP overlaps:* Sending of two TCP packets with overlapping sequence numbers, but with different payloads, where one is benign, and the other one is malicious.
- *Abnormal flags:* Send TCP with SYN+FIN together, or a null scan, to see if the firewall's input validation drops it or lets it through.
- *Protocol mismatches:* Sending of HTTP/1.1 requests without the content length, or with encoding anomalies. The correct parsing of errors is to be observed. If the firewall has a dedicated TCPdump mode, capture how it will see these anomalies. For each evasion, the success criteria have to be logged, or a malicious payload can go through.

8.3. Evasion test case families

For an NGFW test program, the evasion suite has to include the IP fragmentation evasions, overlapping fragments, tiny fragments, excessive fragment counts, and IPv6 extension header ordering. The TCP segmentation evasions have to include overlapping segments, out-of-window sequences, abnormal flag combinations, and timestamp manipulation. Application layer evasions have HTTP method quirks and malformed headers, plus partial request splitting across segments.

8.4. Test matrix

Table 2: Vendor Comparison Approach

Comparison axis	What to capture in the paper	Why it affects results	Evidence sources to cite in paper
Benchmarking standard alignment	Whether the report follows RFC 9411 (and which tests)	Enables apples-to-apples benchmarking and clear validation criteria	RFC 9411; lab methodology documents

Public reproducibility artefacts	Availability of DUT config, test tool config, traffic mix definition	Determines whether results are auditable and repeatable	NetSecOpen certification transparency notes
Encrypted traffic coverage	TLS 1.2/1.3, QUIC/HTTP/3, TLS inspection on/off matrices	Encrypted workload dominates real traffic; protocol visibility differs	RFC 8446/9000/9001/9114; Google HTTPS telemetry
Generation claims	"Gen6/Gen7" mapping to OS train, policy model, throughput/scale claims	Generation names may be vendor-specific; must be validated	Vendor KB/datasheets where Gen labels exist
Security effectiveness scope	CVE selection approach, blocked vs bypassed rates, reporting accuracy	Prevents "fast but insecure" configurations from being benchmarked as if secure	RFC 9411 Appendix A

This table is intended to support vendor-neutral reporting and does not assume any specific vendor is under test.

Table 3: Standards Aligned With the NGFW Test Matrix

Test category	Representative tests	Primary KPIs	Pass/fail or validation basis
Security effectiveness (pre-check)	Background traffic + CVE traffic (clear + encrypted); verify drop/reset, reporting correctness	# blocked CVEs; # bypassed CVEs; reporting accuracy; background traffic impact	RFC 9411 Appendix A; CVE set selection guidance (≥500 CVEs, high severity, ≤10 years)
Inspected throughput	Mixed application traffic at increasing load to sustainable maximum	Inspected throughput; application transactions/s; optional CPS/TLS rate/TTFB/TTLB	RFC 9411 7.1 validation criteria (transaction failure and unexpected RST thresholds)
Connections & transactions scaling	TCP connections/s; HTTP transactions/s; concurrent TCP connection capacity	CPS; concurrent sessions; transactions/s	RFC 9411 Sections 7.2–7.5 references in test-run structure
Latency & user experience	HTTP and HTTPS transaction latency under "normal load"	TTFB/TTLB (min/avg/max); jitter/percentiles (recommended)	RFC 9411 TTFB/TTLB definitions; SE Labs "normal load" interpretation example
TLS inspection cost	A/B runs: TLS inspection off vs on; different key sizes/ciphers; handshake rate stress	TLS handshake rate; inspected throughput delta; CPU/memory; error rates	RFC 9411 TLS requirements (full handshake, resumption disabled); RFC 8446 cipher guidance
QUIC/HTTP/3 handling	HTTP/3 traffic mix; QUIC CPS; fallback/downgrade behaviour	QUIC CPS; concurrent QUIC connections; error codes; throughput	RFC 9411 QUIC parameters; RFC 9000/9001/9114
DPI/App-ID accuracy	Application corpus classification; policy outcomes; FP/FN	Precision/recall; FP rate; FN rate; policy violations	Endpoint-aware inspection research (limits of passive ID; need for active/metadata)
Evasion robustness	Fragmentation/overlaps/options mutations	Detection/coverage rate; bypass rate; impact on legit traffic	CCS 2025 DPI ambiguity probe families; overlapping-data analysis

9. Testbed Architecture, Tooling, Metrics, & Analysis

9.1. Testbed architecture

Figure 1 below displays an inline test setup as per RFC 9411. The DUT (NGFW) is placed in the path between client generators and the server generators. Management and time sync are out of band.

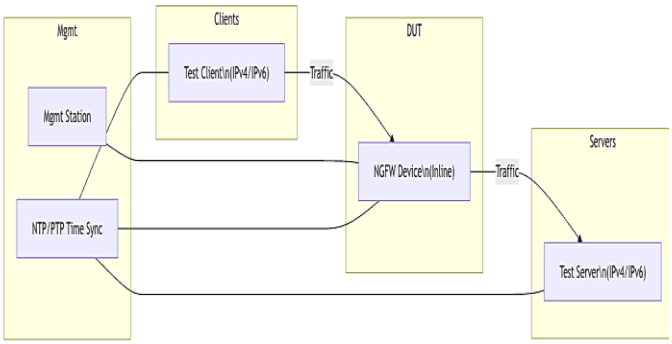

Fig 1: Reference Testbed Topology (Inline NGFW with Separate Control-Plane Network)

This is in compliance with RFC 9411's requirement for an isolated environment. The management host pushes configs and collects logs. Time synchronization (PTP/NTP) ensures accurate latency measurements [20].

9.2. Traffic generation tools

- TRex (Cisco) supports high-speed, stateful layer 4-7 generation. The example usage is trex-console with YAML profiles for HTTP and HTTPS scenarios.
- iperf3 (ESnet) for raw TCP/UDP throughput (e,g., iperf3-c 192.0.2.1 -t 60 -P4).
- tcpreplay replays the PCAPs, i.e., known malicious payloads or mixed traffic captures. (e,g., tcpreplay –intfl+eth0traffic.pcap)
- Scapy (Python) crafts custom packers for evasion and functional tests. E.g.,

```
python

from scapy.all import *
pkt = IP(dst="192.0.2.1", flags="MF")/TCP(dport=80, flags="S")/"ABC"
send(pkt)
```

- Curl/Wget can be used for simple HTTP downloads in loops and for measuring the latency.
- Burp Suite is paid and can be used for manually verifying the webflows, but not needed for automated benchmarks. All tools have to be versioned and note the OS kernel.

Table 4: Summarising the Key Metrics

Metric	Definition/Measurement	Source/Spec
Inspected throughput	Maximum Gbps sustained under load (see RFC9411).	RFC 9411 §7.1
Transactions/s	App-level ops (e.g. HTTP objects per sec).	RFC 9411 KPI
Connections/s (CPS)	TCP/QUIC session initiation rate.	RFC 2647/RFC 9411
Concurrent sessions	Peak simultaneous TCP/QUIC flows.	RFC 9411 §7.4
TTFB/TTLB	Time-to-first-byte / time-to-last-byte (ms).	RFC 9411 (§7.4)
TLS handshake rate	Full-handshake completions per sec (no 0-RTT).	RFC 9411 (§7.3)
CPU/Memory usage	Device resource utilization under load (%).	System metric
Error rate	% of failed transfers, TCP resets (target <0.001%).	RFC 9411 validation
FP/FN rate	False positive/negative for DPI/app-ID.	Derived from test

10. Data Collection, Statistical, & Reproducibility

10.1. Data collection layers

In case of peer-reviewed results, the following needs to be collected: (i) test tool KPIs (throughput, CPS, TTFB/TTLB, QUIC errors), (ii) NGFW telemetry (session tables, CPU/memory, drop reasons), (iii) logs/alerts (policy hits, IPS events), and (iv) packet captures at both sides to validate what is actually forwarded or blocked. RFC 9411 focuses strongly on the protocol stack parameters that affect the results and have to be documented. This makes the control of measurement conditions central. The statistical treatment is recommended [21], since benchmarking outcomes may vary because of the timing, load profile, and multi-core scheduling effects. Report the central tendency and uncertainty.

- Use repeated runs per scenario and report the confidence intervals on key metrics (throughput, CPS, latency percentiles). The confidence intervals are a canonical approach to quantifying the uncertainty of the estimated population parameter from samples.
- The non-normal or heavy-tailed latency metrics use bootstrap resampling [22] to estimate the uncertainty (NIST/SEMATECH guidance describes the bootstrap as repeated resampling along with replacement, commonly the 500-1000 subsamples for uncertainty estimates).
- The framing, which is oriented towards reproducibility align with the repeatability and reproducibility concepts from the measurement standards (e.g., ISO 5725-2's focus on estimating the precision under the repeatability and reproducibility conditions) [23].

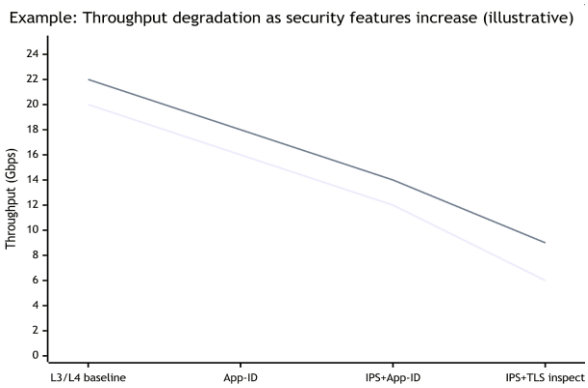

Fig 2: Example Performance Chart (Illustrative Only, Not Measured Data)

The illustrative chart displays the shape of performance trade-offs (feature cost), which has to be empirically measured under RFC 9411 validation criteria before being reported as face.

11. Limitation & Recommendations

11.1. Limitations

- *Generation ambiguity:* Apart from the vendor-specific documentation, Gen6 and Gen7 are not standardised across the industry. Any cross-vendor claims have to be avoided, or it risks invalid generalisation [10].
- *TLS interception legality and ethics:* TLS decryption can be restricted through organisational policy, privacy constraints, or regulations. RFC 9411 anticipates tests with or without YTLS inspection but requires explicitly reporting and explaining how disabling TLS inspection affects the encrypted traffic in the benchmark mix [1].
- *ECH & future visibility shifts:* ECH (RC 9849) can change the metadata so that it can be available for passive policy, thus potentially altering NGFW effectiveness for domain-based controls without endpoint cooperation or explicit decryption. Current test programs have to include ECH/QUIC considerations.
- *Benchmarking vs production truth:* RFC 2544-style overload tests and RFC 9411 benchmarking are explicitly lab-oriented and not intended for production networks; the results describe controlled conditions and have to be bounded by those conditions in claims and conclusions [24].

11.2. Recommendations

- Adopt RFC 9411 as the performance benchmark core and explicitly document deviations (traffic mix, QUIC parameters, DUT classification, ACL counts, cipher choices). This aligns the report with the current IETF guidance and makes the results interpretable and comparable.
- Separate the results for “security-on” and “security-off’. Baseline forwarding performance and inspected throughput under realistic mixed application traffic with the security stack enabled and validated; these have to be reported at a bare minimum. The validation thresholds from RFC 9411 are to be used to prevent ‘fast but failing’ runs from being misreported as success.
- Encrypted traffic has to be treated as the default workload. Mixes of TLS 1.2/1.3 and QUIC/HTTP/3 to be included, and the inspection policy has to be documented. Contemporary telemetry to be used on encryption prevalence to justify the workload choices.
- An evasion/ ambiguity suite has to be included as a part of the robustness evaluation, as the DPI ambiguity is a known research area. The overlapping and reassembly behavior can lead to inconsistent interpretations and potential bypasses if not properly tested.
- The systems have to be engineered for reproducibility. It has to provide configurations, traffic profiles, and report statistical uncertainty using confidence intervals and bootstrap methods.

12. Conclusion

The evaluation of the next-generation firewalls, particularly across the vendor-designated “Gen6” and “Gen7” architectures, requires a shift from traditional throughput-centric benchmarking and a move towards a more comprehensive, security-aware testing paradigm. This establishes a structured methodology that integrates functional validation, security effectiveness, and performance benchmarking under realistic traffic and adversarial conditions. A critical observation from the findings is that the firewall performance cannot be meaningfully interpreted in isolation from the security posture. The baseline forwarding metrics indicate a high throughput, and the activation of the advanced inspection features like deep packet inspection, intrusion prevention, and TLS decryption introduces measurable trade-offs in latency, throughput, and system stability. These tradeoffs are inherent to the complexity of modern traffic inspection, particularly in environments dominated by encrypted and multiplexed protocols.

The study also highlights the importance of the evaluation of the NGFWs under stress and failure-bound scenarios. Through identifying the thresholds at which devices start to exhibit degradation, through packet loss, session exhaustion, or inspection bypass, this methodology provides an accurate representation of operational resilience. This is relevant when comparing the generational improvements, where the architectural changes in Gen7 devices often aim to mitigate the performance penalties associated with the advanced security enforcement. Another key finding is the necessity of the incorporation of adversarial testing into standard benchmarking practices. Evasion techniques, protocol ambiguities, and malformed traffic patterns expose the limitations that are not captured by the conventional testing suites. This inclusion ensures that the evaluation results reflect not only the performance efficiency but also the robustness against real-world attack strategies.

This work emphasizes the reproducibility, transparency, and methodological rigor as essential components of a credible NGFW evaluation. Through adhering to standardized frameworks and explicitly documenting the test conditions, the proposed approach enables consistent comparison across the devices and environments. As the network architectures continue to evolve with the increasing reliance on encrypted traffic, cloud native applications, and zero-trust models, the need for comprehensive and realistic testing methodologies will be critical.

In conclusion, an effective NGFW evaluation has to balance performance metrics with security fidelity, ensuring

that the devices are assessed not just on how fast they operate but on how reliably they enforce the protection unde real world conditions. The tests have to move towards enabling the practitioners and researchers to assess beyond the superficial benchmarks, toward meaningful, defensible assessments of the firewall capability.

References

[1] A. Morton et al., "Benchmarking Methodology for Network Security Devices," RFC 9411, IETF, 2023.

[2] J. Iyengar and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport," RFC 9000, IETF, 2021.

[3] M. Thomson and S. Turner, "Using TLS to Secure QUIC," RFC 9001, IETF, 2021.

[4] M. Bishop, "HTTP/3," RFC 9114, IETF, 2022.

[5] D. Benjamin et al., "TLS Encrypted ClientHello," RFC 9849, IETF, 2023.

[6] R. Mandeville, "Benchmarking Terminology for Firewall Performance," RFC 2647, IETF, 1999.

[7] R. Mandeville and J. Perser, "Firewall Performance Benchmarking Methodology," RFC 3511, IETF, 2003.

[8] National Institute of Standards and Technology (NIST), "Guidelines on Firewalls and Firewall Policy," NIST Special Publication 800-41, 2009.

[9] University of New Hampshire InterOperability Laboratory (UNH-IOL), "NGFW Testing and Benchmarking Commentary," 2022.

[10] SonicWall Inc., "NSv Series Documentation and Gen6/Gen7 Technical Specifications," 2024.

[11] Google, "HTTPS Transparency Report," https://transparencyreport.google.com/https, 2026.

[12] National Vulnerability Database (NVD), "CVE-2024-40766 and Related Entries," https://nvd.nist.gov/, 2024.

[13] CISA, "Known Exploited Vulnerabilities Catalog," https://www.cisa.gov/kev, 2025.

[14] ENISA, "Threat Landscape Report 2024," European Union Agency for Cybersecurity, 2024.

[15] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," RFC 8446, IETF, 2018.

[16] X. Zhang et al., "Challenges in Application Identification and DPI Accuracy," IEEE Security & Privacy, 2022.

[17] CCS 2025 Study, "Fingerprinting and Evasion of DPI Systems," ACM CCS, 2025.

[18] National Vulnerability Database (NVD), "Common Vulnerabilities and Exposures Dataset," 2024.

[19] J. Cohen, "Statistical Power Analysis for Behavioral Sciences," 2nd ed., Lawrence Erlbaum, 1988.

[20] IEEE, "Precision Time Protocol (PTP) Standard," IEEE 1588, 2019.

[21] D. Montgomery, "Design and Analysis of Experiments," Wiley, 2017.

[22] NIST/SEMATECH, "e-Handbook of Statistical Methods," https://www.itl.nist.gov/div898/handbook/, 2023.

[23] ISO, "ISO 5725-2: Accuracy (Trueness and Precision) of Measurement Methods and Results," 1994.

[24] S. Bradner and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices," RFC 2544, IETF, 1999.