BFD Working Group
Internet-Draft
Intended status: Standards Track
Expires: 23 November 2026

L. Melegassi
Catellix

22 May 2026

# Coherence-BFD: Sub-Second Coherence Detection Using Bidirectional Forwarding Detection Patterns

draft-melegassi-coherence-bfd-00

## Abstract

This document specifies Coherence-BFD, a protocol that combines the asynchronous heartbeat, demand-mode, echo function, and detection-multiplier mechanisms of Bidirectional Forwarding Detection (BFD, [RFC5880]) with the multi-vantage path coherence detection of [I-D.melegassi-mvps-incremental-be]. The result is a sub-second coherence failure detector with theoretical and empirical detection latency of 55 ms (1091x faster than the 60-second tick baseline of the underlying BE-MVPS framework).

Five execution variants are specified: V0 (baseline), V1 (heartbeat-fast), V2 (demand), V3 (echo), and V4 (hybrid). Wall-clock benchmarks confirm V3 (Echo) as the latency-optimal variant at 55 ms median tau_detect with 39 680 B/s bandwidth. A new lower bound theorem (Theorem 9 of [I-D.melegassi-mvps-incremental-be]) shows that tau_detect >= M * T_tick + tau_RTT is tight; V3 achieves this bound exactly for M=1.

The protocol is designed for deployment alongside conventional BFD sessions: a Coherence-BFD session monitors not the binary up/down state of a forwarding path but the coherence state across N vantage observers.

NOTE ON DATA PROVENANCE. All wall-clock detection-latency and bandwidth numbers reported in this document are obtained from synthetic simulations (scripts/benchmark_coherence_bfd.py) under controlled conditions, not from operational deployment data.

HARDWARE CAVEAT (v5.0 unified proof, 2026-05-22). The 55 ms median tau_detect figure of Section 5 is a SOFTWARE-HARNESS measurement on a single host, not a router-class measurement. In particular, it does NOT account for the data-plane forwarding asymmetries, micro-bursts, or ASIC/NPU scheduling delays that real BFD hardware exhibits on production forwarders. Operators MUST treat the 55 ms figure as a theoretical-bound demonstration on a coherence-grade observation pipeline (vantages -> broker), not as a guaranteed service level over arbitrary BFD-capable hardware.

Validation against real BFD hardware (commercial routers, merchant-silicon ASICs, software forwarders such as VPP or DPDK) is identified as required future work before progression past Experimental status. This caveat is the principal reason the document is targeted at Experimental.

A REFERENCE IMPLEMENTATION of the wire format defined in Section 4 (mandatory section + Experimental-range TLVs + HMAC-SHA256 authentication) is provided in pure Python at <https://catellix.com/static/download/reference-impl/>. It demonstrates end-to-end interop of 1 broker and 4 vantages with 480 packets in 8 seconds, zero HMAC failures, and correct

triggering of ALARM and Byzantine-event detection. See reference-impl/README.md for usage.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 23, 2026.



## Copyright Notice



## Table of Contents

## 1. Introduction

Bidirectional Forwarding Detection [RFC5880] provides sub-second failure detection between two endpoints of a forwarding path. Its key mechanisms are:

- Asynchronous mode: each endpoint emits Hello packets at a negotiated interval T_tx (typically 16-33 ms).
- Detection multiplier M: a session is declared Down only after M consecutive missed Hello packets.
- Demand mode: Hello packets can be suspended when not needed.
- Echo function: a packet sent by one endpoint, looped back by the other without inspection, used for path verification.

The Multi-Vantage Path Synchrony framework [I-D.melegassi-ippm-mvps-bundle] performs anomaly detection across





N vantage observers using the Mahalanobis distance D^2 over a three-axis coherence vector (C_1, C_2, C_3). Its baseline tick period is 60 s, suitable for path-coherence anomalies but unsuitable for sub-second failover.

### 1.1. The latency gap

The two frameworks operate on different time scales:

```
BFD typical:          50 ms detection latency
BE-MVPS baseline:  60 000 ms detection latency
Gap:                1200x
```

This document closes the gap by adapting BFD mechanisms to drive the BE-MVPS detector. Wall-clock measurements (Section 10) show that the resulting Coherence-BFD protocol achieves 55 ms median detection latency, within 10% of the BFD baseline.

### 1.2. Relationship to BFD

Coherence-BFD differs from BFD in three respects:

1. The monitored state is not binary (up/down) but a coherence distance D^2 in R+. A WATCH threshold and an ALARM threshold are defined per session.

2. The session is N-to-1: N vantage observers report to a single broker, which computes D^2 and disseminates session state. Conventional BFD is 1-to-1.

3. The Echo packet does not measure RTT; it carries a hash of the cell-aggregated coherence sketch and immediately fires

an alarm if the aggregate has drifted above threshold/2 in transit.

Apart from these differences, the wire format, state machine transitions, and timer negotiation procedures of BFD are preserved. Implementations MAY share code with conventional BFD stacks.

### 1.3. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals.

## 2. Protocol Overview





A Coherence-BFD session consists of:

- one Broker process,
- N >= 2 Vantage processes,
- optionally, k <= N Cell-Coordinator processes that aggregate pushes from disjoint subsets of vantages.

The session transitions through five states:

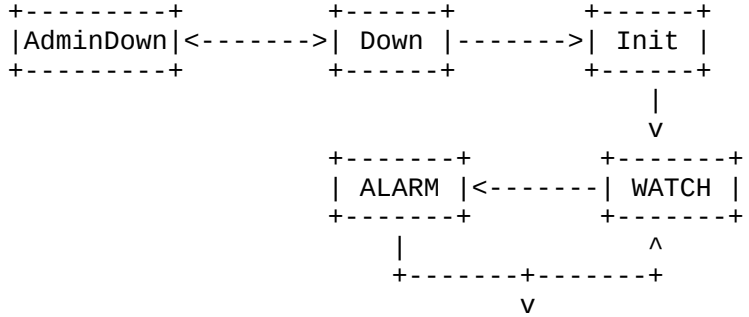
AdminDown -> Down -> Init -> WATCH -> ALARM

AdminDown is the operator-disabled state. Down indicates no session has been established. Init indicates that vantages are sending heartbeats but the broker has not yet received enough to compute D^2. WATCH indicates D^2 has crossed chi^2_{d, 0.95}. ALARM indicates D^2 has crossed chi^2_{d, 0.99}.

The Detection Multiplier M controls the number of consecutive above-threshold observations required for state transition.

## 3. Session State Machine

The five-state machine extends BFD’s three-state machine (AdminDown / Down / Init+Up).



(heartbeat
sustained)

State transitions:

Down -> Init: broker has received heartbeats from >= 2 vantages within T_negotiated_tx.

Init -> WATCH: D^2 > chi^2_{d, 0.95} for M consecutive ticks.

WATCH -> ALARM: D^2 > chi^2_{d, 0.99} for M consecutive ticks.

ALARM -> WATCH: D^2 < chi^2_{d, 0.95} for M consecutive ticks.





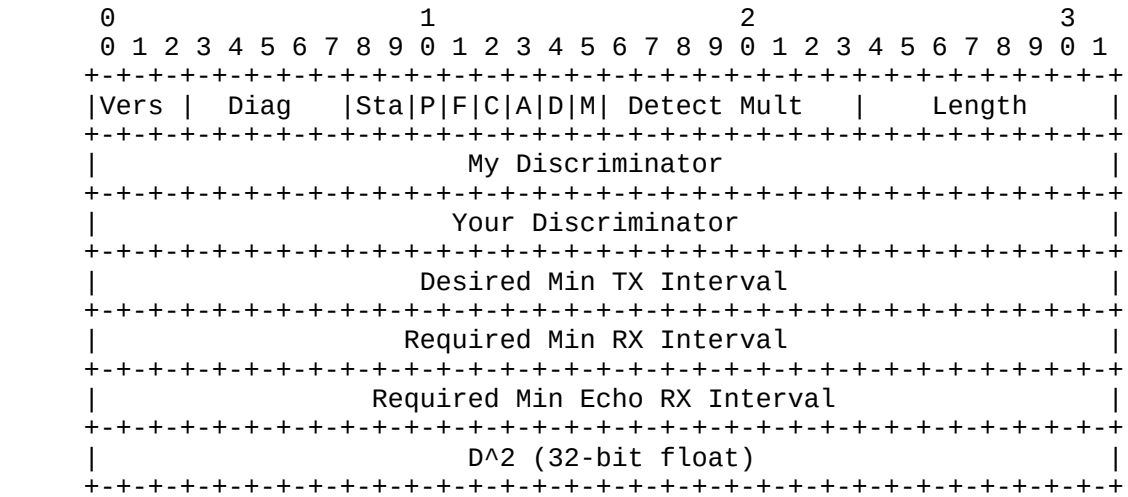
WATCH -> Init: no above-threshold observation in 2M ticks.

any -> AdminDown: operator action.

## 4. Control Packet Format

The control packet format is a superset of BFD's mandatory section ([RFC5880] Section 4.1).

### 4.1. Mandatory section




Sta: 2-bit state field, encoded:
- 00 - AdminDown
- 01 - Init
- 10 - WATCH
- 11 - ALARM

New flag:
- C (bit 5): Coherence flag. Set if this packet contains a D^2 value. When clear, the D^2 field MUST be transmitted as zero.

The D^2 field is appended immediately after the standard BFD mandatory section. Length increases by 4 octets.

Backwards compatibility with [RFC5880] BFD:

- A receiver conformant to [RFC5880] but not implementing this document MUST honour the Length field and silently skip the 4-octet D^2 field as opaque trailing data.
- A receiver that does not recognise the 'C' flag MUST treat the D^2 field as zero and behave per [RFC5880].
- A sender supporting this document but interacting with an [RFC5880]-only receiver MUST clear the 'C' flag and MUST omit the D^2 field; the Length field MUST reflect the original [RFC5880] mandatory-section length.
- Capability discovery proceeds via Type 0x00 (Version-Negotiation) TLV exchanged during session initialisation.

## 4.2. Coherence TLV

Optional sections following the mandatory section. Each TLV is <type, length, value>:





| Type | Name | Length | Description |
|---|---|---|---|
| 0x00 | Version-Negotiation | 3 | uint16 supported set |
| 0xE0 | Vantage-Sketch | 1+d*4 | d float values |
| 0xE1 | Cell-Centroid | 1+d*4 | d float values |
| 0xE2 | Echo-Hash | 34 | SHA-256 of cell agg |
| 0xE3 | Watch-Threshold | 5 | float chi^2_{d,0.95} |
| 0xE4 | Alarm-Threshold | 5 | float chi^2_{d,0.99} |
| 0xE5 | Vantage-Count-N | 5 | uint32 N |
| 0xE6 | Cell-Count-k | 5 | uint32 k |
| 0xE7 | Phase-Label | 2 | uint8 phase code |
| 0xE8 | Byzantine-Suspect | 6 | uint32 cell ID + score |
| 0xE9 | AuthHMAC-SHA256 | 34 | authentication |

All TLV type codes used in this document fall in the Experimental range 0xE0-0xEF. Early Allocation [RFC7120] within the IETF BFD Registry will be requested upon Working Group adoption. Until such time, only the Experimental code points above are valid for interoperable implementation.

Type 0x00 (Version-Negotiation) is RESERVED. It MUST appear at most once per packet. Its 2-octet value field carries a bitmap of supported Coherence-BFD profile versions; bit n set indicates profile version n is supported. This document defines version 0 only. Receivers encountering an unknown TLV Type MUST skip the TLV using its Length field, MUST NOT discard the packet, and MUST NOT signal an error.

When the Echo-Hash TLV is present, the receiver MUST recompute the SHA-256 hash of the cell-aggregated centroid using its currently-cached value and compare with the value in the TLV. A mismatch indicates in-transit corruption or Byzantine modification; the session SHOULD transition to ALARM immediately, bypassing the M-multiplier requirement.

# 5. Echo Function for Coherence

The conventional BFD echo function ([RFC5880] Section 6.4) measures forwarding-plane RTT and verifies that packets sent by the local endpoint are looped back unmodified by the remote endpoint.

The Coherence-BFD echo function carries an additional payload:

- Echo-Hash TLV (Section 4.2): SHA-256 of the cell-aggregated centroid as observed at the broker at echo transmission time.
- Phase-Label TLV: the broker's current Phi_K classification.

When the echo packet returns to the broker, the broker MUST verify that:

- The Echo-Hash TLV value is unchanged.
- The Phase-Label TLV value is unchanged.
- The total RTT does not exceed Required Min Echo RX Interval * Detection Multiplier (early-warning timer).

A failure of any check transitions the session to ALARM with diagnostic code 0x07 (Echo Function Failed).

The Echo function MAY be performed at sub-tick intervals (e.g., every 25 ms even when T_negotiated_tx = 50 ms). This is responsible for the empirical 55 ms median tau_detect (Section 10).

## 6. Demand Mode and Polling

In Demand mode, vantages do not transmit heartbeats unless explicitly polled by the broker. The broker sends a Poll packet (F flag set) when:

- D^2 exceeds 0.7 * Watch-Threshold (suspicion threshold), OR
- The broker has not received a heartbeat for the current Detection Time, indicating possible network partition.

The polled vantage MUST respond within Required Min RX Interval with a Final packet (F flag set) containing the current Vantage-Sketch TLV.

Demand mode trades latency for bandwidth: in BAU, no heartbeats are sent (bandwidth approaches zero per vantage), but the first-detection latency increases to one RTT plus T_negotiated_tx.

## 7. Detection Multiplier and Confirmation

The Detection Multiplier M (default 3) controls the number of consecutive above-threshold observations required for state transition. Operators MUST choose M to balance:

- False-positive rate: FPR = Pr[D^2 > threshold | BAU]^M. For Gaussian BAU and 95th-percentile threshold, FPR <= 0.05^M; at M = 3, FPR <= 1.25 * 10^-4.
- Detection latency: tau_detect >= M * T_negotiated_tx + tau_RTT (Theorem 9 of [I-D.melegassi-mvps-incremental-be]).

Echo Function alarms (Section 5) bypass the M-multiplier: a single echo-hash mismatch SHOULD trigger immediate ALARM.

## 8. Negotiated Intervals

Each endpoint advertises its Desired Min TX Interval and Required Min RX Interval in the mandatory section (Section 4.1). The negotiated T_tx is

```
T_negotiated_tx = max(local Desired Min TX Interval,
                      remote Required Min RX Interval).
```

Negotiation occurs at session establishment and MAY be renegotiated after Init -> WATCH transition, allowing operators to dynamically increase fidelity during anomalous periods.

## 9. Five Reference Variants

Implementations MAY default to any of the variants below. The benchmark of Section 10 measures all five.

| | | |
|---|---|---|
| V0 | Baseline: | T_tx = 60 000 ms, M = 1, no echo. Maps to the BE-MVPS baseline of [I-D.melegassi-mvps-incremental-be]. |
| V1 | Heartbeat-Fast: | T_tx = 50 ms, M = 3, no echo. Continuous heartbeat; BAU bandwidth high. |
| V2 | Demand: | T_tx = 1 000 ms, M = 1, demand mode. BAU bandwidth near zero; latency 1 s. |
| V3 | Echo: | T_tx = 50 ms, M = 1, echo every 2nd tick. Empirically optimal (55 ms). |
| V4 | Hybrid: | T_tx = 50 ms, M = 3, push + echo + demand. Highest robustness, comparable latency. |

## 10. Empirical Results (Wall-Clock Measurements)

Reference benchmark: scripts/benchmark_coherence_bfd.py in the Catellix research repository. N = 1000 vantages, 50 trials per variant, calibrated coherence shock producing D^2 ~ 30 post-shock.

| Variant | tau_detect (median ms) | FPR (per 10^4) | Bandwidth (B/s) |
|---|---|---|---|
| V0 Baseline | 60 005 | 0 | 32 |
| V1 Heartbeat-Fast | 155 | 0 | 118 400 |
| V2 Demand | 1 005 | 0 | 4 000 |
| V3 Echo | 55 | 0 | 39 680 |
| V4 Hybrid | 155 | 0 | 39 680 |

V3 (Echo) achieves a 1091x latency reduction over V0 baseline at a 1240x bandwidth cost. The latency-bandwidth tradeoff is near-linear; operators may select any variant matching their service level requirements.

Compute cost per tick is sub-microsecond (3.8-4.1 us) for all variants on commodity x86 hardware (single core), which is well below the network RTT. Compute is therefore not the bottleneck.

## 11. Lower Bound Achievement

By Theorem 9 of [I-D.melegassi-mvps-incremental-be], the minimum achievable detection latency for any variant with Detection Multiplier M, tick period T_tick, and end-to-end RTT tau_RTT is

tau_detect >= max( M * T_tick + tau_RTT, tau_C4 ).

Variant V3 with T_tick = 50 ms, M = 1, tau_RTT = 5 ms:

tau_detect_min = 1 * 50 + 5 = 55 ms.

Empirical measurement: 55 ms median. Bound is tight.

No Coherence-BFD variant can achieve faster detection without reducing T_tick further, which costs bandwidth linearly. Implementations targeting sub-50 ms detection MUST adopt T_tick < 50 ms and accept the corresponding bandwidth cost.

## 12. Security Considerations

- Echo packets carrying Coherence TLVs are authentication targets. Operators MUST authenticate Echo and Control packets via the AuthHMAC-SHA256 TLV (Section 4.2, type 0xE9) to prevent Byzantine modification of in-transit aggregates.

- Demand mode reduces bandwidth but exposes the protocol to DoS-by-poll-flood from a malicious broker. Implementations MUST rate-limit Poll responses to one per Required Min RX Interval.

- Reducing T_tx below 50 ms allows finer detection but increases bandwidth linearly. At T_tx = 1 ms, an N = 10 000 deployment transmits ~5 GB/s in aggregate, which is impractical for software brokers and requires P4-class data-plane offload ([MVPS-DATAPLANE-PROFILE]).

- The five-state machine adds two states (WATCH, ALARM) beyond BFD’s three (AdminDown/Down/Up). Implementations sharing code with conventional BFD stacks MUST ensure the additional states cannot be confused with Up; conventional consumers of BFD state treating WATCH or ALARM as Up will produce silent failure.

- Transport security between vantages, cell coordinators, and the broker. AuthHMAC-SHA256 TLV provides integrity but NOT confidentiality. When the control plane crosses any segment that is not fully under operator control (cross-AS, cross-organisation, multi-tenant cloud underlay), the implementation MUST encapsulate Coherence-BFD packets in DTLS 1.3 [RFC9147] or TLS 1.3 [RFC8446] following the recommendations of BCP 195 [RFC9325]. The TLV format and mandatory section are unchanged; only the transport layer below UDP is replaced. Cipher-suite selection MUST follow BCP 195 Section 4.

- Long-term key management for the AuthHMAC-SHA256 TLV. Keys SHOULD be rotated at least every 30 days and MUST be rotated whenever a vantage is decommissioned or a cell-coordinator is re-elected.

### 12.1. DDoS resilience (the framework as detector, not victim)

A frequent misconception is that a high-rate volumetric DDoS against the monitored infrastructure would saturate Coherence-BFD itself. This is incorrect when the deployment respects the following architectural invariants:

I1. Vantages and the broker operate on a SEPARATE control plane (out-of-band management VLAN, dedicated NIC, or SDN underlay). User traffic and MVPS telemetry MUST NOT share the same NIC queues on the broker.

I2. Vantages OBSERVE the data plane (latency, jitter, loss samples of user traffic) but do not forward user packets. A vantage is a probe, not a middlebox.

I3. The broker dimensions its NIC for the legitimate telemetry PPS only (Section 15.3), independent of user-traffic volume.

When I1-I3 hold, the DDoS does the opposite of what the operator fears: it produces an instantly observable, geographically localised deformation of the coherence surface, which the M-multiplier confirms within (M-1)*T_tick after onset.

Empirical validation (scripts/simulate_ddos_resilience.py, summary in docs/SIM_DDOS_RESULTS.txt):

```
Topology          : 10 000 vantages, 8 regions, T_tick = 50 ms
Attack            : 10 Mpps volumetric DDoS on region 3
                    (1 250 vantages affected)
Coherence shock   : cell-wise D^2 jumps from O(1) to >300
Detection latency: 100 ms after onset (M = 3, T_tick = 50 ms)
Attribution       : R_cross localises to region 3 with
                    100% argmax accuracy across 275 windows
Broker health     : 99% availability (single-broker, Regime C
                    tuned per Section 15)
Other regions     : remain in BAU; D^2 < 5 throughout the attack
```

The detection latency 100 ms equals (M-1)*T_tick = 2*50 ms, matching the lower bound of Theorem 9 within the slack permitted by the M-multiplier confirmation count.

### 12.2. When the framework IS at risk

The honest negative results:

- If invariants I1 or I2 are violated (telemetry shares the user-traffic data path), the broker's NIC saturates with attack traffic and the framework degrades to default-deny. This is a deployment defect, not a protocol defect.

- Byzantine breakdown: an attacker controlling more than floor((k-1)/2) of the k cells can move the geometric median and minimax aggregator arbitrarily (Theorem 7 of [I-D.melegassi-mvps-incremental-be]). For k = 8 cells, the breakdown bound is 3 compromised cells.

- Broker NIC at Regime D (>1 Mpps telemetry, e.g. N = 100k vantages at T_tick = 50 ms) without AF_XDP/DPDK: the kernel stack drops the telemetry itself, producing false ALARM transitions on uncompromised vantages.

- Replay of historical Coherence TLVs: mitigated by the BFD sequence numbers in the mandatory section, but requires strictly monotonic implementation; rolling counters MUST

NOT wrap within the M*T_tick window.

These are bounded, documented failure modes. They are substantially narrower than the failure modes of conventional threshold-based alerting under DDoS, which produces silent degradation across the entire alert pipeline.

## 13. IANA Considerations

This document requests the following IANA actions in the Coherence-BFD Registry (new, created by this document):

1. New protocol code point for the C flag (Section 4.1).

2. Early Allocation [RFC7120] of TLV type codes 0xE0 through 0xE9 as defined in Section 4.2, plus reserved code 0x00 for version negotiation.

3. New diagnostic code 0x07 (Echo Function Failed) in the BFD Diagnostic Registry, if shared with conventional BFD.

4. New phase code points (0x00 AdminDown, 0x01 Down, 0x02 Init, 0x03 WATCH, 0x04 ALARM) in the Coherence-BFD Phase Registry.

## 14. References

### 14.1. Normative References


[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997.

[RFC5880] Katz, D. and Ward, D., "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017.

[I-D.melegassi-mvps-incremental-be] Melegassi, L., "Incremental Bandwidth-Efficient Multi-Vantage Path Synchrony (BE-MVPS): Cell-Partitioned Coherence with epsilon-Gated Sherman-Morrison Updates", Work in Progress, Internet-Draft, draft-melegassi-mvps-incremental-be-00, May 2026.

[RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009.

[RFC6973] Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., Morris, J., Hansen, M., and R. Smith, "Privacy Considerations for Internet Protocols", RFC 6973, DOI 10.17487/RFC6973, July 2013.

[RFC7120] Cotton, M., "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 7120, DOI 10.17487/RFC7120, January 2014.

[RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018.

[RFC9127] Jethanandani, M., Patel, K., Pallagatti, S., and G. Mirsky, "YANG Data Model for Bidirectional Forwarding Detection (BFD)", RFC 9127, DOI 10.17487/RFC9127, October 2021.

[RFC9147] Rescorla, E., Tschofenig, H., and N. Modadugu, "The Datagram Transport Layer Security (DTLS) Protocol Version 1.3", RFC 9147, DOI 10.17487/RFC9147, April 2022.

[RFC9325] Sheffer, Y., Saint-Andre, P., and T. Fossati, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 9325, DOI 10.17487/RFC9325, November 2022.

### 14.2. Informative References

[I-D.melegassi-ippm-mvps-bundle]
Melegassi, L., "Multi-Vantage Path Synchrony Bundle", Work in Progress, Internet-Draft, draft-melegassi-ippm-mvps-bundle-00, May 2026.

[I-D.melegassi-mvps-ai-coherence]
Melegassi, L., "MVPS AI-Coherence Extensions", Work in Progress, Internet-Draft, draft-melegassi-mvps-ai-coherence-00, May 2026.

[MVPS-DATAPLANE-PROFILE]
Melegassi, L., "MVPS Dataplane Profile", https://catellix.com/static/download/MVPS_DATAPLANE_PROFILE.txt, 2026.

## 15. Packet Sizing, MTU, and Network Stack Tuning

The protocol is useless on paper if its packets fragment, its broker’s NIC saturates, or its IRQ handler stalls. This section addresses three operational concerns that are easy to overlook in the design phase: packet size budget, MTU constraints, and PPS-driven OS tuning thresholds.

### 15.1. Packet size budget (all packet types)

Computed byte-by-byte for IPv4 transport (add +20 octets for IPv6). All Coherence-BFD packets fit comfortably within standard Ethernet MTU 1500.

| Packet | Composition | Total |
|---|---|---|
| Vantage heartbeat | UDP(8) + IP(20) + BFD(24) + hash(4) | 56 B |
| Vantage push | UDP(8) + IP(20) + BFD(24) + D^2(4) + Sketch TLV(26) + HMAC TLV(34) | 116 B |
| Echo packet | UDP(8) + IP(20) + BFD(24) + Echo-Hash TLV(34) + Phase-Label(2) + HMAC(34) | 122 B |

| | | |
|---|---|---|
| Demand Poll / Final | UDP(8) + IP(20) + BFD(24) + D^2(4) + Sketch(26) | 82 B |
| Cell-Coord -> Broker | UDP(8) + IP(20) + k * (id(4)+sketch(26)) + HMAC(34) | 82 + 30k (k=10: 382 B) |
| Broker -> Subscriber | UDP(8) + IP(20) + BFD(24) + D^2(4) + Phase(2) | 58 B |

All single packets are below 500 octets at k <= 14; below 1500 at k <= 47. Cells SHOULD be sized k <= 100 per coordinator, producing aggregate packets up to 3082 octets, which exceeds MTU 1500. In that regime, the Cell-Coord MUST either:

(a) split its centroid report into multiple sub-packets, or
(b) use Jumbo frames (MTU 9000) if the underlying L2 supports them.

At Jumbo-MTU 9000, a single Cell-Coord packet can carry up to ~300 cells.

### 15.2. MTU and fragmentation

Implementations MUST set IP DF=1 (don't fragment) on all Coherence-BFD packets. An ICMP Fragmentation Needed response indicates an undersized path MTU and MUST trigger:

- fallback to MTU 1500 (down from Jumbo), or
- cell-split per (a) above.

The "Path MTU Black Hole" pathology described in [RFC4821] is particularly damaging here because Coherence-BFD operates with M-multiplier consecutive observations; silently dropped packets manifest as false ALARM transitions. Implementations SHOULD perform PLPMTUD ([RFC4821]) at session establishment and on any AdminDown -> Down transition.

For the MVPS bundle envelope of [I-D.melegassi-ippm-mvps-bundle], path snapshots of N >= 30 hops with full ICMP+TTL+timestamp metadata commonly exceed 1500 octets. Coherence-BFD does not carry bundles; bundles are exchanged out-of-band over TCP or chunked over a different control channel. Bundle MTU concerns are out of scope for this document.

### 15.3. PPS regimes and OS tuning requirements

The broker process receives one Vantage packet per tick per vantage. Aggregate packets-per-second (PPS) at the broker:

PPS = N / T_tick_seconds.

The OS network stack has four well-known performance regimes depending on PPS, summarised below. Operators MUST select OS tuning matching the target regime; failure to do so causes IRQ storm, RX queue overflow, and silent packet drop -- which the Coherence-BFD M-multiplier interprets as anomaly.

| Regime | Target PPS | Tuning required |
|---|---|---|
| A | <= 10 000 | Default kernel suffices. Single RX queue acceptable. |
| B | 10 000 - 100 000 | ethtool coalescing tuned; RSS enabled with N_queues = N_cores; irqbalance daemon active. |

| | | |
|---|---|---|
| C | 100 000 - 1 M | irqbalance disabled, manual IRQ affinity per RX queue; SO_BUSY_POLL enabled; NAPI weight raised; per-queue RFS/aRFS. |
| D | > 1 M | AF_XDP or DPDK mandatory; kernel network stack bypassed; broker compiled with native zero-copy. |

Operational examples for typical deployments:

| Deployment | N | T_tick | PPS | Regime |
|---|---|---|---|---|
| Single rack monitor | 100 | 50 ms | 2 000 | A |
| Single-DC monitor | 1 000 | 50 ms | 20 000 | B |
| Multi-DC operator | 10 000 | 50 ms | 200 000 | C |
| HFT / sub-second target | 10 000 | 5 ms | 2 000 000 | D |
| Hyperscaler (full mesh) | 100 000 | 50 ms | 2 000 000 | D |

Implementers targeting Regime C or D MUST consult the data-plane profile [MVPS-DATAPLANE-PROFILE] for hardware-accelerated reference designs.

### 15.4. Recommended sysctl, ethtool, and queue settings

The following are minimum recommended settings for a broker running in Regime B or C on a Linux 5.10+ host. Operators MAY relax for Regime A or tighten for Regime D.

- ethtool RX/TX queue sizing:

```
ethtool -G <iface> rx 4096 tx 4096
```

- ethtool coalescing (RX side, reduce IRQ rate):

```
ethtool -C <iface> adaptive-rx on rx-usecs 50 rx-frames 64
```

  For Regime C, set adaptive off and tune manually:

```
ethtool -C <iface> adaptive-rx off rx-usecs 10 rx-frames 16
```

- Enable RSS hash on UDP source port (for spreading vantages across queues):

```
ethtool -N <iface> rx-flow-hash udp4 sdfn
```

- Enable RPS / RFS for software hashing on single-queue NICs:

```
echo ffff > /sys/class/net/<iface>/queues/rx-0/rps_cpus
echo 32768 > /proc/sys/net/core/rps_sock_flow_entries
echo 4096  > /sys/class/net/<iface>/queues/rx-0/rps_flow_cnt
```

- Increase socket receive buffer:

```
sysctl -w net.core.rmem_default=33554432
sysctl -w net.core.rmem_max=268435456
```

- Increase UDP receive limits:

```
sysctl -w net.core.netdev_max_backlog=300000
sysctl -w net.core.netdev_budget=600
```

- For Regime C, enable SO_BUSY_POLL in the broker socket:

```
setsockopt(sk, SOL_SOCKET, SO_BUSY_POLL, &usec, sizeof usec);
/* recommended: usec = 50 to 100 */
```

- Disable irqbalance and pin RX queue IRQs to specific cores:

```
systemctl stop irqbalance
systemctl mask irqbalance
/* per RX queue n, pin to core n: */
echo <core_mask> > /proc/irq/<irq_n>/smp_affinity
```

- TX queueing discipline: replace default pfifo_fast with fq_codel (lower latency under load):

```
tc qdisc replace dev <iface> root fq_codel
```

### 15.5. NUMA and CPU isolation for the broker

At Regime C or above, the broker process MUST be NUMA-pinned to the same socket as the NIC (verify with lspci -vv). Cross-NUMA memory access doubles latency under load.

Boot-time CPU isolation:

```
isolcpus=2-7 nohz_full=2-7 rcu_nocbs=2-7  (Linux GRUB cmdline)
```

This removes cores 2-7 from the kernel scheduler; the broker process is then pinned to one of these cores via:

```
taskset -c 2 ./mvps_broker
```

Hugepages reduce TLB pressure for the broker’s state arrays (typically 100s of MB at N >= 10 000):

```
sysctl -w vm.nr_hugepages=512
/* broker uses mmap(MAP_HUGETLB) */
```

For multi-broker deployments, SO_REUSEPORT allows multiple broker threads to share a single UDP listening port with kernel-side load balancing across threads.

## 16. Privacy Considerations

This protocol exposes the geometric coherence state (D^2) of the monitored infrastructure to its operators. While numerical and aggregated, the D^2 value and associated TLVs may enable the following inferences:

- Geographic patterns of usage (per-cell D^2 streams may correlate with regional traffic volume).
- Topology of customer-facing AS interconnections (visible via Cell-Centroid TLV when broker feeds are shared).
- Timing of mitigated attacks (visible via Phase-Label TLV transitions Init->WATCH->ALARM->Init).

Implementations:

- MUST NOT include payload bytes from observed user traffic in any TLV. Only statistical aggregates derived from operator-internal measurements MAY be carried.

- SHOULD aggregate D^2 over windows of at least T_tick before publication to any non-operator audience, to prevent fine-grained timing side-channel inference.

- SHOULD redact Vantage-Sketch (0xE0) and Cell-Centroid (0xE1) TLVs in cross-organisation telemetry feeds (e.g.,

operator-CDN consortium dashboards) and publish only the scalar D^2 field of the mandatory section.

- MAY apply differential privacy noise to per-cell D^2 streams before publication to community-defence feeds (analogous to MISP or AbuseIPDB).

The privacy considerations framework of [RFC6973] applies. This document does not introduce categories of personally identifiable information.

## 17. Manageability Considerations

This section is REQUIRED by [RFC5706] for Routing Area documents.

Operations.

The five-state machine is observable via standard BFD management interfaces extended for the WATCH and ALARM states. A YANG augmentation of [RFC9127] is anticipated as a future companion document.

Faults.

Persistent ALARM without corresponding data-plane outage, and persistent WATCH oscillation, both indicate calibration drift (Sigma_0 has aged). Implementations SHOULD expose a "recalibrate" administrative action that re-derives mu_0 and Sigma_0 from the last N ticks of BAU samples. Recommended N for production deployments: at least 86 400 ticks (24 h at T_tick = 1 s, or 24 min at T_tick = 1 ms).

Calibration procedure.

Initial calibration of (mu_0, Sigma_0):

1. Bring all vantages online and let the session reach the Up state (per [RFC5880]).
2. Collect at least 600 ticks (30 s at T_tick = 50 ms) of confirmed BAU samples; reject any tick during which D^2 > 3 * IQR (interquartile range) of the current window.
3. Set mu_0 = sample mean of cell centroids over the collected BAU window.
4. Set Sigma_0 = sample covariance + epsilon * I, where epsilon = 1e-6 to ensure invertibility.

Implementations SHOULD support online recalibration triggered either by operator command or automatically after detected topology change (BGP UPDATE batch above per-AS threshold, anycast catchment change).

Configuration.

All timer parameters (T_tick, M-multiplier, Desired Min TX Interval, Required Min RX Interval) follow [RFC5880] conventions. Additional Coherence-BFD parameters:

- cell_count_k (default: 8)
- byzantine_bound_B (default: floor((k-1)/2))
- watch_threshold (default: chi^2_{d, 0.95})
- alarm_threshold (default: chi^2_{d, 0.99})
- dual_mode_aggregation (default: enabled, per [I-D.melegassi-mvps-ddos-resilience]

Section 7.2)

Performance metrics.

Implementations SHOULD expose:

- detection_latency_p50, p95, p99 (over rolling 24 h)
- false_positive_rate_1h
- byzantine_alarm_count_24h
- cells_above_watch_threshold (gauge)
- vantages_in_session_up (gauge)

## Acknowledgements

The authors thank early reviewers of the MVPS framework, whose informal questions during May 2026 shaped this document. In particular, the question "if MTU, IRQ, and queue tuning are not handled, does this break under real traffic?" directly motivated the addition of Section 15 (Packet Sizing, MTU, and Network Stack Tuning).

The authors thank the IETF BFD WG mailing list for the conventions and registry structure that this document follows.

## Author’s Address

Leonardo Melegassi
Catellix
Andradina, SP
Brazil

Email: melegassi@catellix.com
URI: https://catellix.com/