LINX127 · 19 NOVEMBER 2026

Maybe it is time to consider layer 3 internet exchanges

James A. T. Rice · Jump Networks Ltd (AS8943)

The expectation - traffic is delivered to, and only to, the network it was addressed to

Disclosure and disruption are both safety risks: misforwarding exposes people's browsing activity and can drop VoIP calls to emergency services, including UK 999 calls.

Not all traffic is encrypted, and encrypted traffic is still subject to harvest-now-decrypt-later.

The reality - shared layer 2 exchanges give no guarantee of correct delivery

One large exchange has stated in writing that it gives no guarantee that traffic is delivered, or delivered only to the intended port, and that it knows of no exchange that does.

Every neighbour entry fault breaks availability, and incorrect entries break confidentiality and integrity when unknown unicast is flooded

FAULT	CAUSES	EFFECT	C	I	A	NEEDS FLOODING
Persistent missing	Filtering: an exchange ND filter expecting modified EUI-64 link-local addresses the participant does not use, or a participant on an outdated netmask ignoring ARP from outside it	Traffic to prefixes learnt from the route servers silently discarded; bilateral sessions fail to establish			●	No
Transient missing	ARP, ND, broadcast or multicast policers in the fabric, a remote peering carrier or a participant; aggregate policing cannot tell necessary ARP and ND from the rest	Entries disappear briefly and repeatedly, with a silent discard each time			●	No
Transient incorrect	A participant changes its MAC address, and its peers keep the old one until they re-resolve	Silent discard until peers re-resolve; with flooding, sent to every port until then	○	○	●	No
Persistent incorrect	Misbehaving devices forward flooded frames to the destination, which counts as forward progress (RFC 4861 7.3.1); a misbehaving ND proxy answers NS for a departed MAC with NAs that leave the entry unchanged (7.2.5)	Silent discard indefinitely; with flooding, sent to every port and delivered through the misbehaving devices	○	○	●	Forward progress: yes ND proxy: no

● broken ○ broken only while unknown unicast flooding is enabled

Why the wrong MAC address stays REACHABLE - misbehaving devices supply both kinds of confirmation that RFC 4861 accepts

Forward progress

RFC 4861 section 7.3.1

With unknown unicast flooding enabled, a misbehaving device processes flooded frames it should have discarded at the layer 2 ‘my station’ check, and forwards them to the destination. The destination’s replies count as forward progress, so no solicitation is sent and the entry is never re-resolved.

A solicited NA with no target link-layer address

RFC 4861 section 7.2.5

A misbehaving ND proxy at the fabric edge answers NS sent to a MAC address that has gone, with an NA that carries no target link-layer address option, which confirms reachability without correcting the MAC address.

Either way the participant’s router is following RFC 4861 as specified, and the entry is kept wrong by a device it has no control over.

Faults have hit traffic on one large exchange’s layer 2 peering LANs

2010 Loop during a port upgrade, 25 minutes unstable
2010 Port security cut members off over stray frames
2010 Ring master forwarded on its blocked port, a loop
2010 Leaked MAC addresses learned on the wrong port
2010 Switch reload made a brief loop in the ring
2010 Ring flap in fibre work, two minutes of disruption
2010 Ring unstable twice in two days, members shut ports
2011 Ring instability flooded traffic to idle ports
2011 Faulty port destabilised two rings, broadcast storm
2011 Five ring flaps in ten weeks, sessions dropped each time
2011 Loop made in maintenance, heavy packet loss
2012 MAC limit bug black-holed members' traffic
2012 Maintenance left one-way paths, ARP storm followed
2012 Other members' traffic flooded to one switch's ports
2012 Loop in an attached network stormed both LANs
2012 Members' MACs no longer learned on their own ports
2012 Traffic between two members delivered to a third
2012 Ring config error, one site unreachable for 83 minutes
2012 MAC learning fault, then traffic sent to the wrong LAG
2012 One member's LACP frames flooded, others' LAGs reset
2012 A MAC not learned, traffic filled another member's port
2012 Failover left LAGs stuck, member traffic flooded
2012 Card crash, then traffic flooded to idle ports
2013 Ring switchover flushed MAC tables, traffic flooded
2013 Misprogrammed MAC entries flooded a whole switch
2013 MAC entries aged out every 5 minutes, traffic flooded
2013 MAC programming fault, peers unreachable
2013 Card crash, flood, then a member's MAC not learned
2014 Flooded traffic across the LAN, cause not given
2014 Sponge answering ARP for live members
2014 Loop in the quarantine VLAN, 24 minutes of loss
2014 Flapping link upset the rings, flooding and loss
2014 One router's ND storm dropped other members' sessions
2015 MAC learning failed on a switch, 90 minutes down
2015 Inconsistent MAC state, ARP failed for two days
2015 Ring left unprotected in maintenance, traffic flooded
2015 Ports live on two routers at once, packet loss
2016 Flooding in the quarantine VLAN, sessions flapped
2016 New port skipped quarantine, an hour of packet loss
2017 MAC learning fault split a LAN for an hour
2018 One packet looped round the ring for 90 minutes
2019 One packet type swamped a switch CPU, traffic dropped
2019 MAC filter blocked a member's new router
2020 MAC filter blocked a member's upgraded router
2020 Packets punted to switch CPUs, 40 to 60% loss
2020 Routers stopped answering ND after a switch upgrade
2020 Switch bug broke ARP and ND during MAC changes
2021 Static entry lost in a port move, five hours down
2021 Router hit its neighbour cache limit, IPv6 turned off
2021 Removing the old overlay cut traffic between switches
2021 Stale entry after a MAC change, traffic flooded to all
2021 MAC learning fault after onsite work, an hour of flaps
2022 Stale switch MAC entries, peers unreachable for months
2022 Duplicate address detected on reboot, IPv6 down
2022 Peers held the old MAC after a port move
2023 MAC entries pointed at the wrong switch
2023 LAN ARP volume tripped a router's ARP policer
2023 Peers held the old MAC after a router change
2023 Flooded frames re-forwarded by other routers
2023 Multicast policer dropped neighbour solicitations
2023 Inconsistent MAC tables, some peers unreachable
2023 Switches advertised invalid MAC-to-IP mappings
2023 Management LAN flood restarted peering line cards
2023 Patch left MAC tables out of sync for six hours
2024 Peers held a member's old MAC
2024 Routers still re-forwarding flooded frames
2024 Other members' traffic arriving at a new port
2025 Proxy ARP and ND entries not learnt, switch cut off
2025 Peers held the old MAC after a MAC change
2025 Old netmask on some routers, new addresses unreachable
2026 MAC tables out of sync, traffic dropped
2026 Switches answered ND for a MAC that had gone
2026 Route servers didn't enforce-first-as
2026 ND packets crashed a process on the switches
2026 Missing ND entries black-holed route-server routes
2026 Static MAC left on two switches in a port move
2026 Link-local filter dropped real neighbour discovery

Internet exchanges have presented as shared layer 2 networks for thirty years, because in the 1990s that was the only real choice

THE 1990S	NOW
Routers with high-capacity ports were expensive and forwarded in CPU, and layer 2 switches were cheap	Switches forward at layer 2 and layer 3 at line rate in hardware
A few dozen engaged, technical participants found and fixed faults between themselves	The largest exchanges have close to a thousand participants, most of whom buy a port and expect it to just work

The rest of the industry has spent years replacing large layer 2 domains in its own networks with routed ones, for the isolation between ports and the easier debugging.

The proposal - forward at layer 3, so each exchange port is a routed point-to-point link like a PNI or transit port

Your router
any router, any vendor

/31 and /127, or link-local with RFC 8950 · BGP and BFD

The exchange's router
the exchange's AS, one published policy, RPKI, IRR, ASPA and OTC filtering on every port

the same port and policy for every participant

Every other participant
reached through the exchange's routing

What you configure

- One point-to-point link per port, with one neighbour entry, for the exchange's router
- One BGP session per address family with the exchange, with BFD, in place of sessions with every peer or with route servers and collectors
- Selective announcement by community, as with route servers today, with no new sessions when a peer joins or replaces a router

What you receive

- Every peer's routes on one session, each identified by its AS path and by a community naming the peer
- Traffic delivered to the connected network and from there by its routing, the assurance a PNI or transit gives

THE PROPOSAL

Transit-standard features on every port, and new services for the exchange

RTBH in the exchange

Discarded in the exchange's forwarding plane, so no peer has to accept blackhole routes

Graceful shutdown

RFC 8326 honoured for every peer at once: make before break in place of session culling

MEDs and more-specifics

Honoured by one published policy, for traffic steering between a participant's ports

Nearest exit and ECMP

Closest-exit routing across the exchange's sites, and ECMP across a participant's ports with no LACP

FlowSpec

RFC 8955 rules from a participant, validated against its own prefixes and applied in the exchange

DOWNGRADE community

draft-ietf-grow-downgrade-bgp-community: tagged prefixes queued at lower priority, dropped first under congestion in place of a blackhole

Congestion withdrawal

Opt-in withdrawal of a participant's prefixes when its port is congested, ordered by large communities

Routing profiles

Chosen per participant, such as a peering profile of customer routes only, settlement-free or paid

Later, BGP-LU with a label per egress peer for egress peer selection; FlowSpec rules and large numbers of FIB entries can be charged for, as transit providers charge for FlowSpec today.

Many existing exchange switches already forward at layer 3 at line rate, and route scale is what needs planning

- Exchanges with an EVPN or VPLS overlay already run a routed underlay, and a flat layer 2 fabric with spanning tree would need routing between its switches first
- Routing platforms hold the full table; switching platforms with smaller FIBs may have to be selective, taking large participants' routes in part and sending the rest via transit
- Blackhole routes, more-specifics for steering and FlowSpec rules also use FIB space, and large numbers of them can be charged for

Migration - routed sub-interfaces next to the peering LAN, for as long as both are wanted

- Each participant gets a point-to-point link on an 802.1Q sub-interface next to its existing port, and removes its peering LAN interface once its layer 3 sessions are up
- The peering LAN stays for the participants who want it, so the exchange runs both, as the largest exchanges expect to for years
- Migrating every participant is comparable in effort to a peering LAN renumbering

Is there an appetite among networks and exchanges for layer 3 internet exchanges?

- Given the choice, would your network prefer a layer 3 exchange port to a layer 2 one?
- Which of the transit-standard features would you use on an exchange port?
- Would you take part in a measured trial on sub-interfaces next to the peering LAN?

Questions and discussion

Appendix - the RFCs describing ARP, ND and /64 weaknesses on shared links since 2004

RFC 3756 (2004)	Where nodes do not trust each other, every node “must be considered able to send falsified ND and RD messages”; spoofed NS and NA redirect traffic for routers and hosts
RFC 5227 (2008)	A malicious host can answer every ARP request with its own hardware address, “claiming ownership of every address on the network”
RFC 5963 (2010)	Gives /64 as the maximum prefix length for a peering LAN, noting /127 in use; its security section covers loops and denial of service from outside the LAN, and does not mention ND spoofing or cache exhaustion
RFC 6164 (2011)	Recommends /127 for inter-router links, a year after RFC 5963; peering LANs stayed /64, 2^64 addresses for about a thousand routers
RFC 6583 (2012)	“Four simultaneous nmap sessions from a low-end computer” made a router’s ND process unusable, on the largest routers available; recommends sizing subnets to the addresses in use
RFC 6980 (2013)	Switch ACLs, RA-Guard and ND monitoring: “current implementations of [...] mitigations for NDP attacks can be trivially evaded” by fragmenting the packets

LINX127 · 19 NOVEMBER 2026

Thank you

Maybe it is time to consider layer 3 internet exchanges

James A. T. Rice · Jump Networks Ltd (AS8943)