TECHNICAL WHITE PAPER

Seeing Past the Link Light

Why 25G and 100G Ethernet Fail Quietly and What It Takes to Diagnose Them

SECTION 01

Executive Summary

Networks built on 10G Ethernet forgave a great deal. A slightly contaminated connector, a marginal patch cord, a fiber plant certified years earlier for a lower rate — conditions like these stayed invisible because 10G carried enough signal margin to absorb them. At 25G per lane, that margin is gone.

The consequence is not that high-speed links fail more often. It is that they fail differently. Instead of dropping, they degrade. Instead of logging errors, they present as application slowness, as intermittency, as a link that should come up and does not. Forward error correction compounds the problem by design: it masks a lane running at the edge of its correction budget so completely that every counter the switch exposes reads clean.

This paper follows six tickets through a single shift at a mid-size regional hospital running 25G to the server and 100G between racks. None of the six arrived described as what it actually was. Four presented as degradation rather than failure. Critically, not one of them was solvable from the switch CLI — every resolution turned on a measurement the switch either does not expose or reports as healthy.

From those six cases we derive a three-question diagnostic framework and a minimum measurement set for operating a 25G/100G plant. The argument is straightforward: at these rates, parts swapping stops working as a troubleshooting strategy, and physical-layer visibility stops being optional.

WHAT'S INSIDE

6
Tickets in one shift, none described as what it actually was

4
Presented as degradation rather than outright failure

0
Solvable from the switch CLI alone

SECTION 02

Why 25G Changes the Diagnostic Problem

LANES, NOT SPEEDS

Above 10G, Ethernet scales by bundling lanes rather than by raising the rate of a single path. A lane is one electrical or optical path running at a fixed signaling rate, and the common variants assemble as follows:

10G	one 10G lane
25G	one 25G lane (IEEE 802.3by)
40G	four 10G lanes (IEEE 802.3ba)
100G	four 25G lanes (IEEE 802.3bm and later 802.3cd variants)

The number four appears in both 40G and 100G short-reach optics, but it does not describe the same physical thing across variants — and that distinction determines where you look when a link misbehaves.

KEY CONCEPT

Four fibers or four colors decides the instrument: end-face inspection and per-strand power, or per-wavelength channel power.

FOUR FIBERS OR FOUR COLORS

40GBASE-SR4 and 100GBASE-SR4, the short-reach variants most data centers deploy, use parallel fiber: four separate transmit and receive pairs inside a single MPO/MTP connector, each carrying one lane over multimode. No wavelength multiplexing is involved. 40GBASE-LR4 and 100GBASE-LR4/ER4 do the opposite — four CWDM wavelengths multiplexed inside the transceiver and carried over one duplex single-mode pair.

SR4 and LR4 both say “four,” but one means four fibers and the other means four colors. That single fact decides whether a degraded lane is a fiber-pair problem or a wavelength problem — and therefore whether the right instrument is an end-face inspection and per-strand power reading or a per-wavelength channel power measurement. Two of the six tickets that follow hinge on exactly this distinction.

WHERE THE MARGIN WENT

A 25G lane is substantially less tolerant of loss and reflection than a 10G lane over the same physical plant. OM4 multimode that has carried 10G cleanly for a decade is not automatically qualified for 25G or 100GBASE-SR4. A connector with light contamination or minor end-face damage will barely register at 10G and will cause consistent loss at 25G.

Because 100GBASE-SR4 is four independent 25G lanes sharing one link light, any marginal condition a single lane tolerates is multiplied across the link — and any one of eight strands can degrade without bringing the interface down. A plant that was good enough for 40G, which runs four more-forgiving 10G lanes with NRZ signaling, can fail outright at 100G.

Two further properties of these links matter operationally. First, on optical modules using FEC, it is a manual setting on both ends — there is no negotiation, and a mismatch produces a link that should work and does not. Second, there is no auto-MDIX equivalent on fiber; TX/RX polarity must be managed correctly across cassette, trunk, and patch cords, and getting it wrong throws no error at all.

> **Nothing about the plant changed. The margin that existed at 10G per lane never existed at 25G per lane.**

SECTION 03

Field Evidence: Six Failures in One Shift

The following cases are drawn from a composite mid-size regional hospital environment, assembled from patterns common to 25G/100G enterprise deployments. The tickets are ordinary. The diagnostic paths are where lane architecture stops being background reading.

THE ENVIRONMENT

Meridian Regional: 480 beds, a main campus data center, and an imaging annex 800 meters away across a parking structure. Four spine-leaf racks, each leaf offering 48 ports of 25G to servers and six 100G uplinks to the spine. Inter-rack uplinks are 100GBASE-SR4 over an OM4 multimode plant originally installed for 10G in 2014 and re-terminated into MPO cassettes during a 2023 refresh. The run to the imaging annex is 40GBASE-LR4 over single-mode. Optics are mixed — the core vendor's own modules on the spine, third-party on roughly half the leaf uplinks. Riding on top: the EHR, a PACS archive under constant load, real-time video from four operating rooms, and a nightly backup window that has never finished early.

TICKET 1 · 7:12 AM

“Imaging pulls are slow this morning”

SYMPTOM

Retrieving a prior study takes roughly forty seconds instead of four — but only about one pull in three. The same study retrieves instantly from a workstation on a different rack, which moves the problem off the PACS server and onto the path between two racks.

WHAT THE SWITCH SHOWED

The leaf-to-spine 100GBASE-SR4 uplink is up. No CRC errors, no discards, no log entries. By every indicator the platform exposes, the link is healthy.

WHAT MEASUREMENT SHOWED

Pre-FEC bit error rate per lane told a different story: three lanes near 1e-12, lane 2 at 4e-5. Seven orders of magnitude worse, and still fully corrected by RS-FEC — which is precisely why the counters were clean. Per-lane received optical power on that lane was 3.2 dB below its neighbors, in one direction only.

ROOT CAUSE

End-face inspection at the patch panel found oil contamination across the core of the corresponding strand. The MPO connector had been reseated during a 02:00 UPS maintenance window. Clean, re-inspect, re-seat: lane 2 returned to 1e-12 and the intermittent slowness stopped.

TAKEAWAY

A 100GBASE-SR4 link is four independent 25G links sharing one link light. One degraded strand out of eight will not take the link down — it will make the applications riding on it feel unreliable, which is a far harder ticket to close. Per-lane pre-FEC BER and per-strand optical power, available from a handheld tester such as NetAlly LinkRunner 100G, are what make the failing quarter of the link visible at all.

TICKET 2 · 9:05 AM

“New OR workstation has no network”

SYMPTOM

A contractor added a 25G drop to Operating Room 4 over the weekend. Module present, interface enabled, configuration identical to the working port in OR 3. No link, no errors, no log entries.

WHY THE COPPER REFLEX FAILS

Swapping the patch cord accomplishes nothing, because no fiber is broken. Transmit and receive are simply landing on the wrong strands: the closet cassette is wired Type B and the trunk the contractor installed is Type B as well. On copper, auto-MDIX would correct this silently and nobody would ever know. On fiber there is no such mechanism — the link does not come up, and it does not say why.

RESOLUTION

Fiber mapping and polarity verification confirmed the mismatch in roughly ninety seconds. Replacing the Type B with a Type A at one end brought the link up immediately.

TAKEAWAY

On fiber, “no link and no error” is not a mystery — it is a polarity signature. Verifying it takes a minute and a half with the right instrument, or an afternoon without one.

SIGNATURE

```
link: down · errors: 0 · optics: nominal
→ polarity or FEC, not hardware
```

TICKET 3 · 11:20 AM

“Vendor says the optics are bad”

SYMPTOM

A new pathology-lab switch from a different vendor than the core. QSFP28 at both ends, correct SR4 type, correct multimode fiber, polarity verified. Both modules report present and read nominal transmit and receive power. The link will not establish. The core vendor’s support line has already suggested the third-party module on the lab end is at fault.

WHAT MEASUREMENT SHOWED

Reading the running FEC state on both devices — not the datasheet — showed RS-FEC enabled on the core and FEC disabled on the lab switch, which shipped that way by default. On optical links using clause 82/91 encoding, FEC is a manual setting at both ends. Nothing negotiates it, and two devices that each believe they are configured correctly will sit there indefinitely.

RESOLUTION

Enable FEC on the lab switch. The link comes up in under a second.

TAKEAWAY

When the fiber tests clean, the optics read nominal, and polarity is confirmed, a link that refuses to establish is a configuration mismatch. Check FEC at both ends before anyone swaps hardware — and note that this ticket is where vendor-neutral test data earns its keep, because it ends the third-party-optics argument with evidence rather than assertion.

TICKET 4 · 13:40 PM

“Third-floor closet went dark after the core upgrade”

SYMPTOM

The core received a new 100G QSFP28 line card over the weekend. Three distribution closets still running 40G QSFP+ were moved onto it. Two came back up; the third did not — module present, no link, configuration identical to the other two.

ROOT CAUSE

This is the gap between mechanical compatibility and operational compatibility. A QSFP+ module fits a QSFP28 cage because they share a 38-pin connector and physical envelope. Whether the switch ASIC will run that port at 40G is a separate question, and on this platform ports are grouped in fours with every port in a group required to operate at the same rate. The third closet landed in a group whose neighbor was already provisioned at 100G.

RESOLUTION

Relocate to a port in an all-40G group. The optics were never the issue, which is why examining them was a dead end — the answer was in the port-group rate map.

TAKEAWAY

Confirm mixed-rate support at the port-group level with your vendor before cutover. “It physically fits” and “it will link” are different claims, and the rate map should be planned before the maintenance window rather than discovered during it.

TICKET 5 · 15:10 PM

“Backups are no faster after the 100G upgrade”

SYMPTOM

The nightly backup window still runs six hours. Finance would like to know what the upgrade purchased.

WHAT MEASUREMENT SHOWED

A line-rate traffic generation test between the two racks returned 98 Gbps sustained, clean, with pre-FEC nominal on all four lanes. The network is delivering what it was bought to deliver.

ROOT CAUSE

The backup host has a single 100G NIC in a PCIe Gen3 x8 slot, which ceilings out near 63 Gbps before protocol overhead, and the target array’s capacity tier is spinning disk that tops out well below that. The link was never the constraint — and was not the constraint at 40G either.

98 Gbps
sustained and clean over the link under test

63 Gbps
host ceiling from a PCIe Gen3 x8 slot

TAKEAWAY

Prove the link at line rate first. Twenty minutes of generated traffic converts an argument about the network into a conversation about the host, and it is the fastest available route to innocence when the network is being blamed for an application problem.

TICKET 6 · 16:30 PM

“Imaging center link degrading all afternoon”

SYMPTOM

The 40GBASE-LR4 span to the imaging annex has accumulated errors since roughly 14:00, worsening through the afternoon and clearing overnight.

WHY THE MORNING’S METHOD DOES NOT APPLY

The temptation is to treat this like Ticket 1 — a fiber problem. But LR4 is not SR4. There are no four fiber pairs here; there is one duplex single-mode pair carrying four CWDM wavelengths multiplexed inside the transceiver. Four lanes means four colors.

WHAT MEASUREMENT SHOWED

An OTDR trace on the span was clean and total insertion loss was comfortably within budget, ruling out the plant. Per-wavelength channel power exposed the asymmetry: three channels nominal, the fourth drifting progressively lower through the afternoon and recovering overnight. The span crosses a rooftop conduit that heats through the day, and a laser with failing thermal control will hold at ambient and wander as it warms.

ROOT CAUSE

The failing channel was inside the core vendor’s own module, not the third-party optic on the far end — which ended the conversation about third-party optics before it started. Replacing the transceiver stopped the errors.

TAKEAWAY

On LR4 and ER4, diagnosing lanes means per-wavelength optical power, not fiber testing. Knowing which of the two architectures is in front of you determines whether you reach for an OTDR or a channel power measurement — and reaching for the wrong one costs a day.

SECTION 04

The Pattern, and a Framework

Five of the six tickets arrived described as something other than what they were, and four presented as degradation rather than outright failure. That is not coincidence. At 10G per lane there was enough margin that marginal conditions stayed invisible until they became outages.

At 25G per lane there is not, so the plant reports its problems gradually — through application performance — well before it reports them through a down interface. Three questions resolve most of a day like this one, and they resolve it in order:

01

Is it one lane or all four?

One lane points at a strand, a wavelength, or a connector — a physical condition local to part of the link. All four points at configuration, the module, or the port. This single question separates Ticket 1 and Ticket 6 from Tickets 2, 3, and 4, and it is answerable only with per-lane data.

MEASUREMENT

Per-lane pre-FEC BER and per-lane optical power

02

Is the plant dirty, or is the configuration mismatched?

Clean fiber plus nominal optics plus no link means somebody's FEC setting, polarity, or port-group rate map is wrong. Establishing the plant is clean is what makes the configuration answer credible — and what prevents the hardware-swap spiral.

MEASUREMENT

End-face inspection, polarity verification, FEC state at both ends

03

Is the link the bottleneck, or is the host?

Answering this first prevents the majority of misdirected escalations. A line-rate test that comes back clean redirects the investigation to where the constraint actually lives, as it did in Ticket 5.

MEASUREMENT

Line-rate traffic generation

Applied in sequence, these three questions triage a 25G/100G plant faster than any amount of configuration review — because each one is answered by a measurement rather than an inference.

What the Switch Cannot Tell You

Worth noting what none of the six tickets was solvable from: the switch CLI alone. Every one turned on a measurement the switch either does not expose or reports as healthy. The counters stayed clean all morning while a quarter of a 100G link ran at the edge of correction.

The minimum working measurement set for a plant of this kind:

MEASUREMENT	WHAT IT ESTABLISHES	WHY THE SWITCH CAN'T PROVIDE IT
Per-lane pre-FEC bit error rate	Remaining margin rather than present state — the earliest warning available.	Counters report post-FEC state; a corrected lane reads as clean.
Per-lane and per-wavelength optical power	Distinguishes a strand problem (SR4) from a channel problem (LR4/ER4).	Aggregate module readings mask a single degraded lane or wavelength.
Fiber mapping and polarity verification	The ninety-second answer to “no link, no error”.	No auto-MDIX equivalent exists on fiber; no error is generated.
End-face inspection	Whether a 10G-era plant is actually qualified for 25G per lane.	Contamination is invisible to every electrical indicator until it isn't.
OTDR and insertion loss	Rules the plant in or out before a vendor ticket is opened.	Not a switch function.
Line-rate traffic generation	Separates the network from the host.	A switch cannot prove its own throughput to an application owner.
FEC state read from both ends	Whether the two devices agree — not what the datasheet says.	Each end reports only its own setting; nothing negotiates or reconciles them.

CONSOLIDATING THE TOOLKIT

Most of this measurement set can be carried in one hand. The NetAlly LinkRunner 100G addresses the electrical and optical layer directly — per-lane pre-FEC bit error rate, per-lane optical power, FEC state, and line-rate traffic generation up to 100G — which covers the evidence that closed Tickets 1, 3, and 5. Paired with fiber inspection and fiber mapping, it also answers Tickets 2 and 6. The point is not instrument count; it is that all of these measurements are vendor-neutral by construction. When a support line points at a third-party module, independent physical-layer data is what ends the conversation instead of extending it.

SECTION 06

Pre-Deployment Practice

01 **Qualify the plant for the rate you are deploying, not the rate it was certified for.**

Clean, inspect, and test every connector before cutover. A plant that has been quiet at 10G has not been tested at 25G.

02 **Baseline pre-FEC BER per lane at turn-up and re-read it periodically.**

Post-FEC clean tells you the link works today; pre-FEC tells you how much margin is left, which is the operative question at 25G per lane.

03 **Verify FEC state and polarity at both ends before declaring a link faulty.**

Both are manual, neither is negotiated, and neither generates a useful error.

04 **Obtain the port-group rate map from your vendor before the maintenance window.**

Mechanical compatibility between QSFP+ and QSFP28 does not imply operational compatibility.

SECTION 07

Conclusion

High-speed Ethernet did not introduce new categories of failure. It removed the margin that used to hide the existing ones. A contaminated end face, a reversed polarity, an FEC mismatch, a mis-provisioned port group — none of these are new problems. What is new is that they now surface as degradation rather than as outages, that forward error correction conceals them until the moment it cannot, and that the switch reports every one of them as healthy right up to the point where users start complaining.

The strategy that fails under these conditions is the parts swap. Replacing modules until symptoms move is slow, expensive, and at 25G per lane frequently converges on the wrong component — as Tickets 3, 4, and 6 each demonstrate. What replaces it is measurement: per-lane, physical-layer, vendor-neutral, and taken before the argument starts rather than after.

> **None of this is exotic. It is the difference between a diagnosis and a parts swap and at 25G per lane, the parts swap stops working as a strategy.**

ASIC (switch ASIC)
The dedicated chip inside a switch that actually moves traffic. Its design and the module you plug in decide which combinations of port speeds a switch can run at once.

BER (Bit Error Rate)
Bit Error Rate is what fraction of bits arrive wrong, written in scientific shorthand: 1e-12 is one bad bit per trillion and healthy, 4e-5 is one per 25,000 and a lane about to become a ticket.

Cassette, trunk, patch cord
The three pieces of a fiber channel. A trunk is the permanent multi-fiber cable running between rooms; a cassette is the panel-mounted box that breaks the trunk out into connectors; a patch cord is the short jumper from panel to equipment. Polarity has to be right across all three, and swapping one can break a channel that worked.

Clause 82 / Clause 91
Two sections of the IEEE 802.3 standard: how a 100G stream is encoded and split across lanes, and how its errors are corrected. You don't need the internals — the operationally relevant fact is that on links using this encoding, FEC is a setting a human turns on at each end, and nothing in the protocol reconciles the two ends if they disagree.

Correction budget
Margin as FEC sees it: how much error the correction can repair before it runs out. A lane at the edge of its budget reads perfectly clean right up until it doesn't.

CRC errors and discards
Two things the switch does count. A CRC error is a frame that failed its checksum; a discard is a frame dropped deliberately, usually for congestion. FEC repairs damage before the checksum is ever checked, so neither counter moves on a degraded lane.

CWDM (Coarse Wavelength Division Multiplexing)
Putting several signals on one fiber by giving each its own color of light. “Coarse” means the colors are spaced far enough apart to keep the optics cheap.

EHR and PACS
The clinical record system, and the medical imaging archive that stores and retrieves X-ray, CT, and MRI studies. PACS retrievals are large and sustained, so they feel a degraded uplink before anything else does.

FCS errors (CRC) and discards
An FCS error is what the switch reports when bit errors exceed the ability of FEC to correct them in their entirety. When FEC cannot correct all the errors, the frame is bad, and the FCS error is the indication. FCS and CRC refer to the same thing — CRC is the general term, and the FCS is a CRC — so most switch Command Line Interfaces print this counter as “CRC.”

FEC and RS-FEC
Forward Error Correction is extra data sent alongside the real data so the receiver can repair damaged bits itself, with no retransmission. RS-FEC is the specific Reed-Solomon scheme used at 25G and 100G, and it's strong enough that a badly degraded lane gets repaired completely — which is why the switch's counters can read zero on a link in trouble.

Insertion loss
How much signal strength is lost end to end across a channel, in decibels. Every connector and splice costs a little; the question is whether the total stays inside the budget for the speed you're running.

Margin
How much abuse a link can absorb and still work: the gap between the signal quality you have and the minimum you need. It's the paper's central idea. A 10G lane had enough margin to swallow a dirty connector without anyone noticing; a 25G lane doesn't, so the same connector now causes errors. Margin is invisible on a healthy-looking link, which is why it has to be measured rather than inferred.

MPO / MTP connector
One connector housing many strands of fiber at once, typically 12, of which SR4 uses 8. It's what makes parallel-fiber links practical, and it means one reseating mistake can affect several lanes.

Multimode and single-mode fiber
Multimode has a wide glass core and is used for short runs inside a building; single-mode has a narrow core and carries a signal much further. OM4 is a multimode grade rated for 100G to about 100 meters — a rating on the glass, not on the current condition of the connectors.

NIC (Network Interface Card)
The network adapter inside a server. A 100G NIC only delivers 100G if everything behind it in the server keeps up.

NRZ (Non-Return-to-Zero)
The simplest way to send bits over light: on for a one, off for a zero, one bit at a time. Both 10G and 25G lanes use it, but the 10G lanes inside a 40G link run it with far more headroom than the 25G lanes inside a 100G link.

OTDR (Optical Time-Domain Reflectometer)
An instrument that fires a pulse of light down a fiber and times what comes back, mapping every connector, splice, bend, and break by distance. It rules the cable plant in or out — necessary but not sufficient evidence that a link will carry 25G per lane.

PCIe Gen3 x8
The server slot a network card sits in, described by generation and lane count. That combination tops out near 63 Gbps of real throughput, so a 100G card in one cannot reach 100G however healthy the network is.

Polarity, Type A / Type B
Polarity is making sure each transmitting strand lands on a receiving strand at the far end. Type A and Type B are standardized wiring methods that differ in how fiber positions are arranged; neither is correct on its own, but the cassettes, trunks, and patch cords in a channel have to combine into a working path.

Spine-leaf
A design where every access switch connects to every core switch and none to each other. The leaf-to-spine uplinks are shared paths, so one degraded lane affects many applications at once.

Strand
One individual fiber: a single thread of glass carrying light one direction (except when using a bi-directional transceiver, like the Cisco 40/100 SRBD (short range bi-directional) which transmit light in both directions simultaneously). An SR4 link uses eight strands, and any one can degrade without the link going down.