Stop using 100GBASE-LR4

A Guide To Modern 100G Optical Technologies

Richard A Steenbergen <ras@petabitscale.com>

Last Updated: July 2026

INTRODUCTION

Why This Talk?

- The interconnection world runs on **100GBASE-LR4**. IXP port? LR4. Peering cross-connect? “LR4 please.”
  - For many networks in this room, it's the only 100G optic anyone has ever asked them for — and the only one they know to ask for.

- Meanwhile, the rest of the industry quietly abandoned it **more than a decade ago**.
  - Hyperscalers dropped LR4 within one product cycle of an alternative existing (CWDM4/PSM4, 2014).
  - Every claim on its datasheet has since been beaten by cheaper optics — including reach.

- Goal: walk the whole story — baud rates, NRZ vs PAM4, chromatic dispersion, FEC, DSPs, and how the standards actually get written — so this deck can be **the reference you point people at**.

- **TL;DR: LR4 is the worst possible way to buy 100G in 2026.**

INTRODUCTION

An Overview of 100G Optical Technologies for Client Interfaces

GEN 1 (2010)

100GBASE-LR4

- 4 × 25G NRZ lanes
- LAN-WDM grid (±1 nm)
- 4 cooled lasers / 4 TECs
- No host RS-FEC — raw BER or nothing

Designed like 2009; priced like it too.

GEN 2 (2014)

CWDM4 / 4WDM

- 4 × 25G NRZ lanes
- CWDM grid (±6.5 nm)
- 4 uncooled lasers
- Host RS-FEC: 5–6 dB margin

Same result, but half the price.

GEN 3 (2018+)

100G Single Lambda

- 1 × 53 GBd PAM4 lane
- One 1310 nm wavelength
- One laser, zero filters
- DSP + mandatory KP4 FEC

One laser plus math.

Note: I'm specifically not going to talk about more exotic DWDM technologies, like 100G 2L PAM4 , or 100ZR Coherent.

Everyone's Favorite Part

But First, Some Physics

Baud, PAM4, chromatic dispersion, and FEC

GEN 1 · LR4 · 2010

GEN 2 · CWDM4 · 2014

GEN 3 · SINGLE λ · 2018+

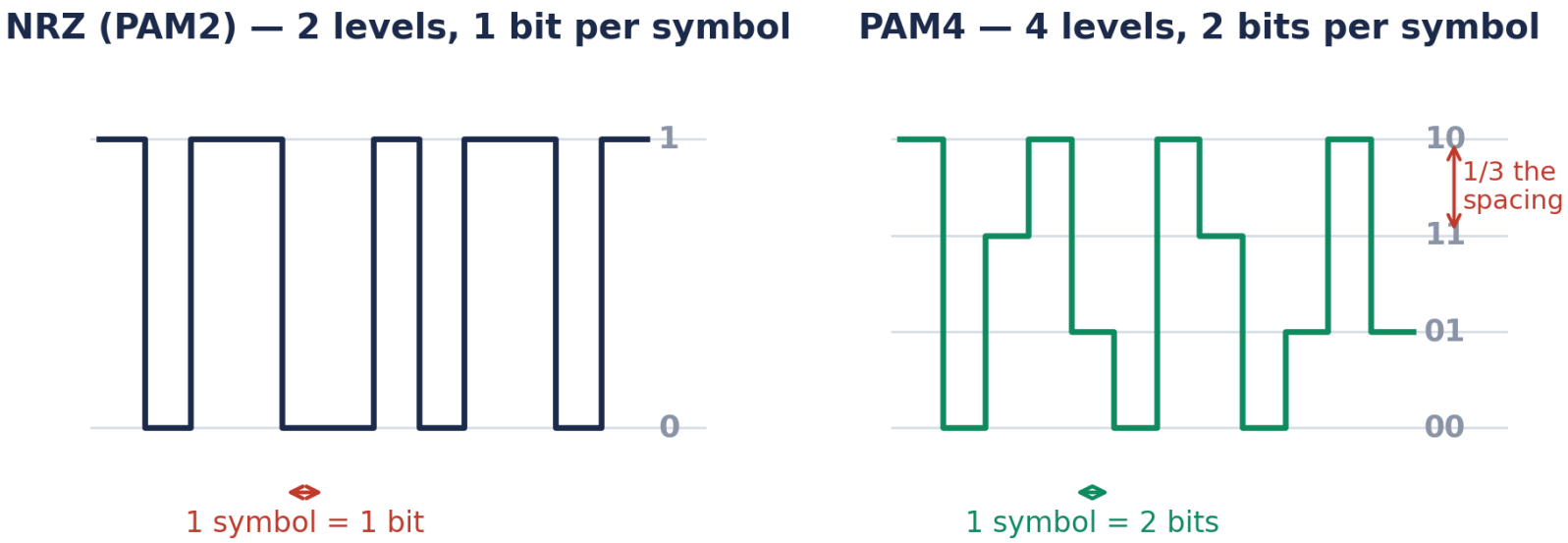
BACKGROUND PHYSICS

Baud Rate and Modulation

Baud is symbols/sec, how many times per second the transmitter changes the signal. Each symbol carries one or more bits.

NRZ (aka PAM2): two levels, 1 bit per symbol → 25.78 GBd = 25.78 Gb/s. Simple enough that a bare photodiode can decode it.

PAM4: four levels, 2 bits per symbol → double the bits at the same baud rate. But the levels are 3× closer together — each "eye" needs ~9.5 dB more SNR.



Same baud rate, twice the bits — but the levels get 3x closer together

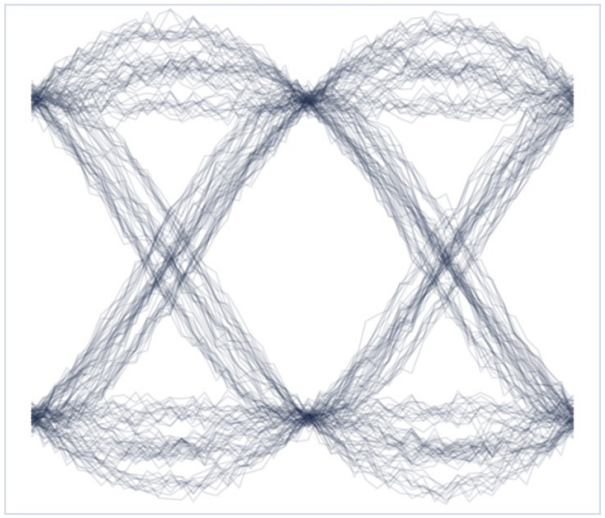
BACKGROUND PHYSICS

What the Receiver Actually Sees

- An **eye diagram** overlays every symbol transition. An open “eye” means the receiver can reliably tell the levels apart; a closed eye means errors.
- 25 GBd NRZ: one big eye — decodable with basic receiver technology. 53 GBd PAM4: three squeezed eyes at twice the speed — **the raw signal is borderline garbage.**

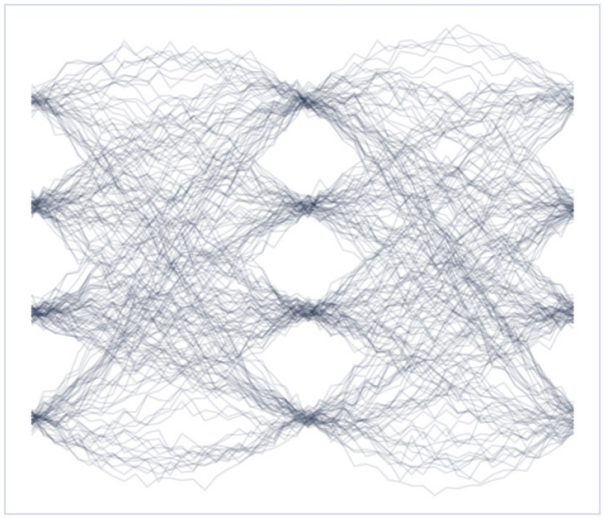
This picture is why Gen 3 took eight extra years: it needed a DSP revolution, not better lasers.

25.8 GBd NRZ



one eye, wide open — decodable by a photodiode

53.1 GBd PAM4

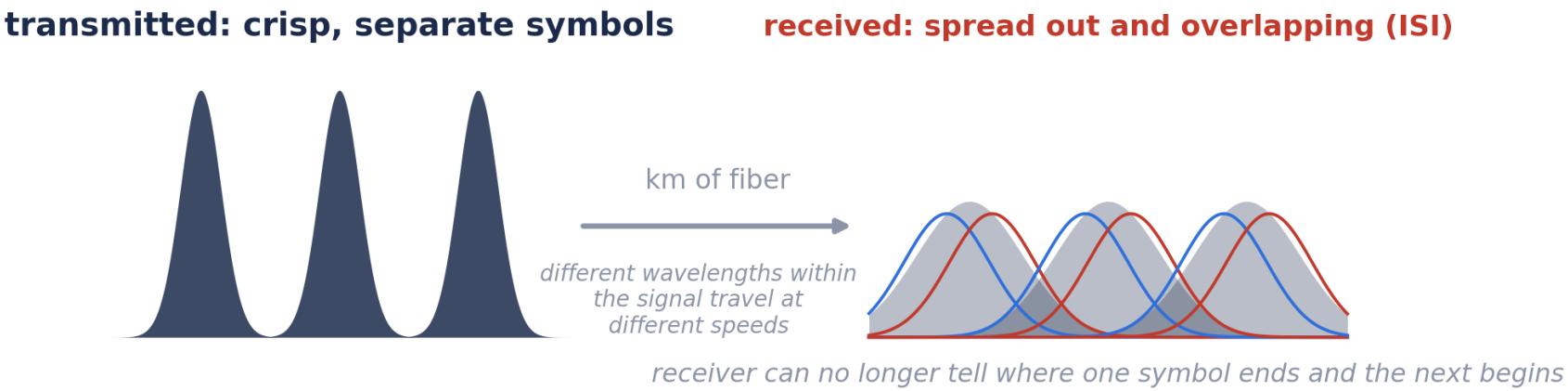


three squeezed eyes at 2× the speed — needs a DSP

BACKGROUND PHYSICS

Chromatic Dispersion (CD)

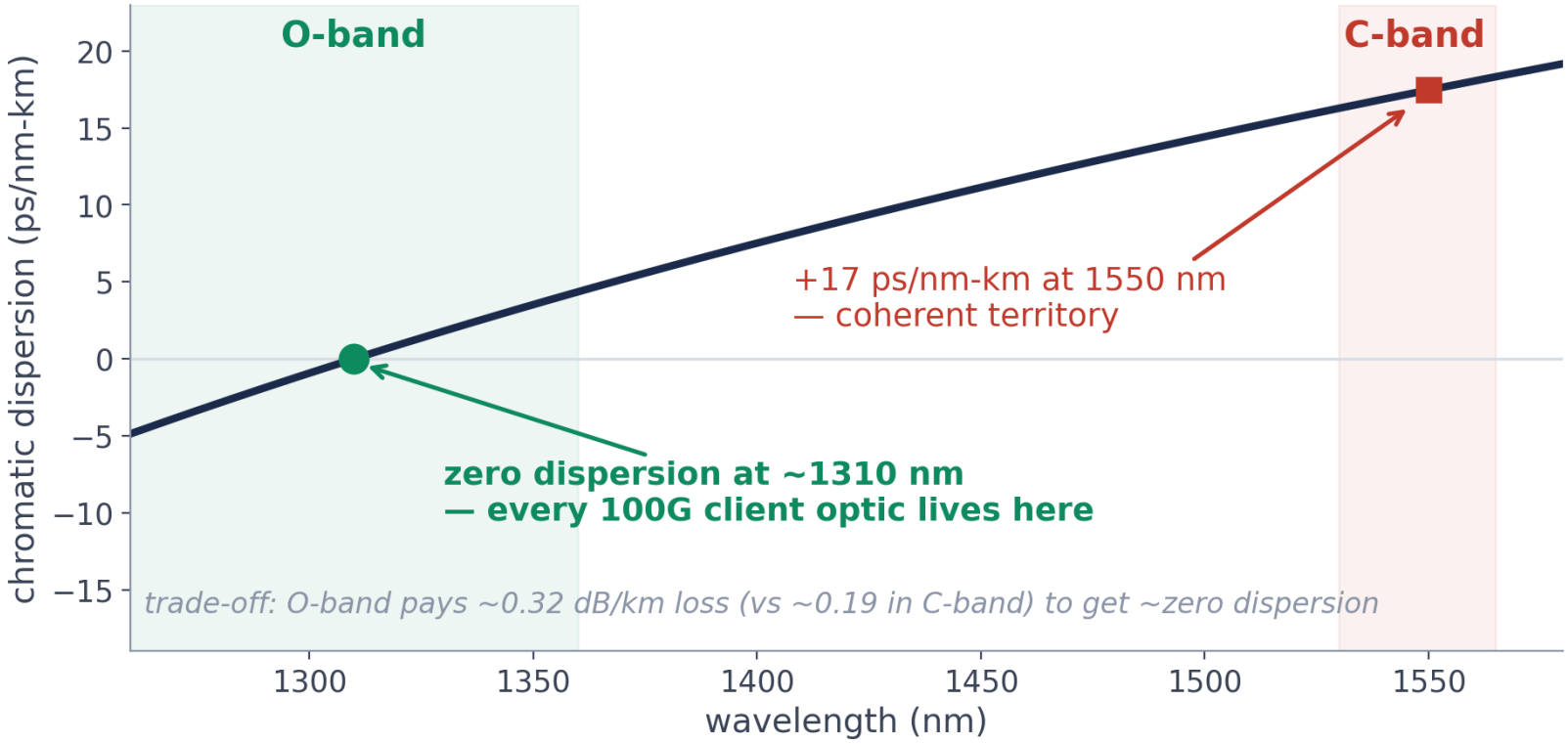
- No laser is perfectly monochromatic — and the act of modulating it spreads its spectrum further. Every signal occupies a small **range** of wavelengths.
- Different wavelengths travel through glass at slightly different speeds (same physics as a prism). Over distance, the fast parts of a symbol outrun the slow parts.
- Symbols smear into their neighbors → **inter-symbol interference (ISI)** → the receiver can't find the symbol boundaries anymore.



BACKGROUND PHYSICS

CD Impact Depends on Where You Sit in the Spectrum

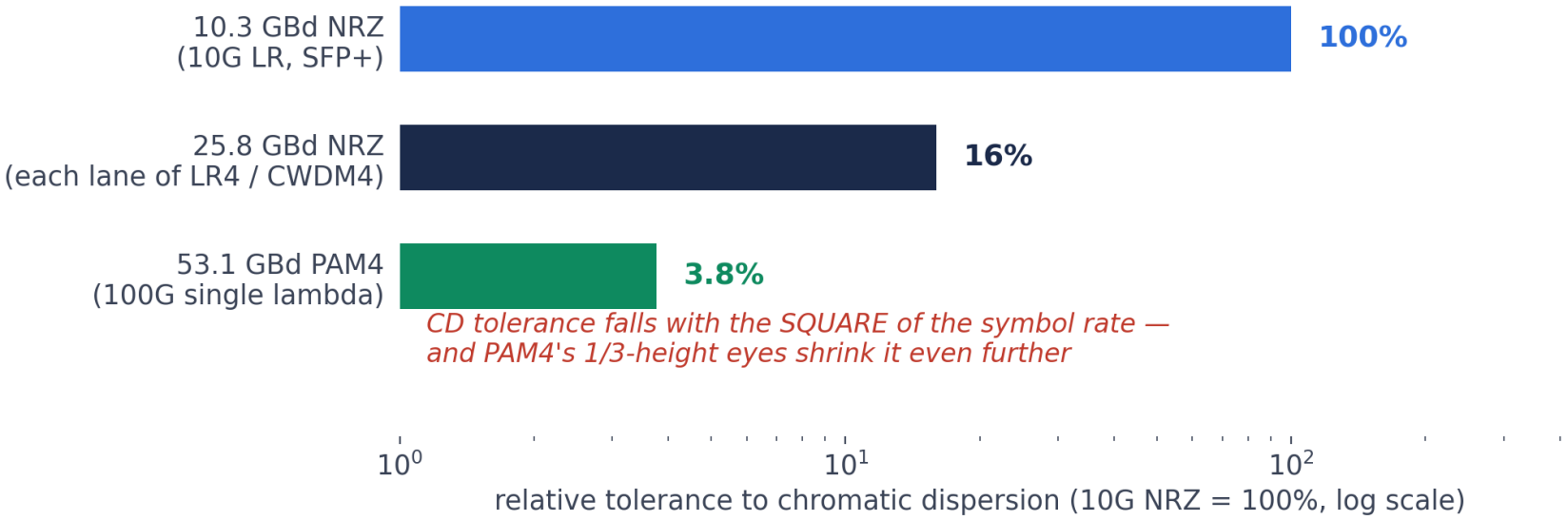
Standard G.652 fiber has a **zero-dispersion point at ~1310 nm**, which is why every 100G client optic lives in O-band.



BACKGROUND PHYSICS

CD Tolerance Collapses as Baud Rate Rises

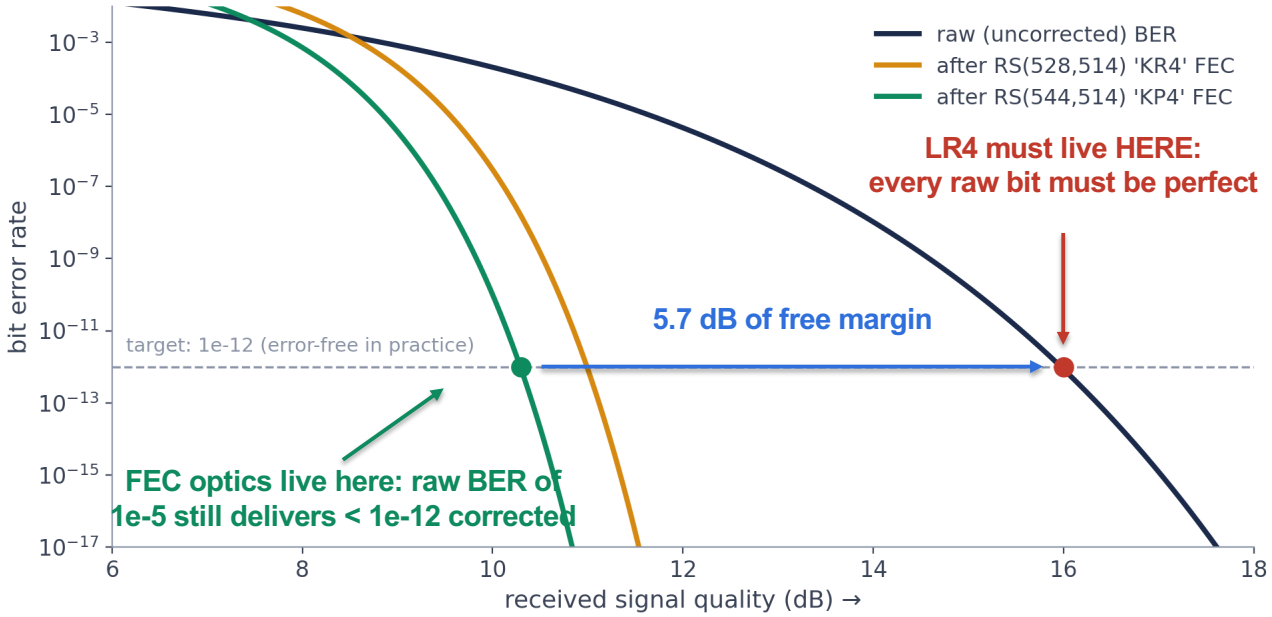
- Double the symbol rate and two things happen at once: symbols get half as long, **and** the spectrum gets twice as wide. The same fiber smearing now eats 4× more of each symbol.
- **CD penalty scales with the square of the baud rate.** This single fact drives the design of every generation that follows.



BACKGROUND PHYSICS

Forward Error Correction (FEC) – Margin From Math

- FEC appends parity symbols so the receiver can **mathematically reconstruct** corrupted data (Reed-Solomon codes — same family as CDs and QR codes).
- It converts “every bit must arrive perfect” into “a surprisingly bad day is fine”: raw BER of 10^{-5} still delivers a corrected link better than 10^{-12}.



PART 1 · PHYSICS

FEC Changed the Economics of Optics

	RS(528,514) “KR4”	RS(544,514) “KP4”
Standardization	IEEE 802.3bj (2014)	IEEE 802.3bs/cd (2017–18)
Corrects up to	7 symbol errors / codeword	15 symbol errors / codeword
Pre-FEC BER budget	~5×10^{-5}	~2.4×10^{-4}
Effective gain	~5.4 dB	~6.4 dB
Used by	CWDM4, 4WDM, SR4	Everything PAM4 (Gen 3, 400G, 800G)

- The gain is **free**: RS-FEC has been baked into every 25G+ SerDes switch/router ASIC shipped since ~2015. It's already in your box, idle, waiting.
- Bonus: **pre-FEC BER is free telemetry.** You watch a link degrade for weeks before it takes a single real error — the best early-warning gauge in optical ops.
- 100GBASE-LR4 was standardized **four years before Ethernet FEC existed**. It is one of the last optics in service designed on the assumption that FEC would never come.

Gen 1: 100GBASE-LR4 (2010)

The last optic designed for a world without FEC.

GEN 1 · LR4 · 2010 | GEN 2 · CWDM4 · 2014 | GEN 3 · SINGLE λ · 2018+

GEN 1 · 100GBASE-LR4 (2010)

Where LR4 Came From

- **IEEE 802.3ba (June 2010)**: the first 100G Ethernet standard. 100GBASE-LR4 = the 10 km duplex-SMF PMD: 4 × 25.78 Gb/s NRZ, wave-division-muxed onto one fiber pair.

- Born in the **CFP** form factor: 20+ W, hand-assembled gold-box optics, four-digit prices.

- Later shrunk through CFP2/CFP4 into **QSFP28** (SFF-8665, 2014; volume ~2015).
  - The QSFP28 made 100G a widespread commodity, LR4 was the incumbent optic that rode in with it.

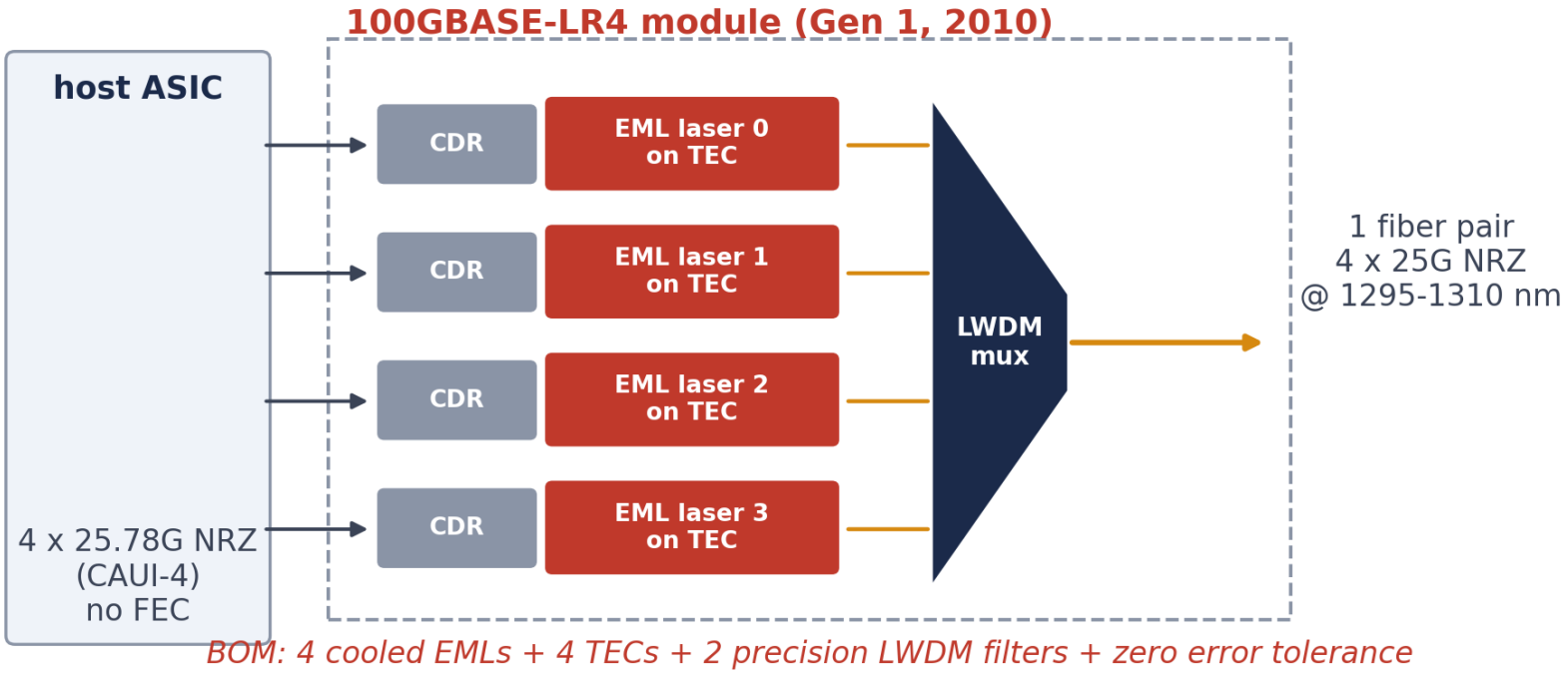
- Crucially: designed **before Ethernet had FEC** (802.3bj arrived 2014). The optical link must run error-free raw, at BER 10^{-12}, with zero help.

- Every LR4 design decision flows from that one constraint.
- It wasn't a bad design - it was a pre-FEC design.

GEN 1 · 100GBASE-LR4 (2010)

LR4 Under the Hood

- Four independent 25G NRZ optical lanes, each on its own precision wavelength, muxed onto one fiber pair through LAN-WDM filters:

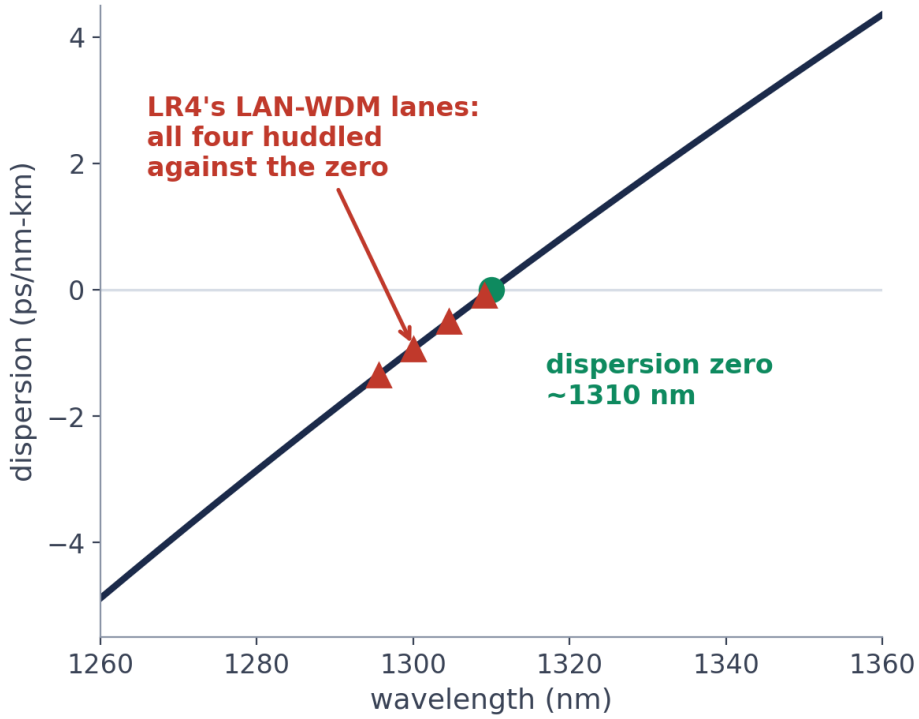


Receive path mirrors this: precision demux → 4 photodiodes. Every block above is a cost the next generations delete.

GEN 1 · 100GBASE-LR4 (2010)

Why the LAN-WDM Grid?

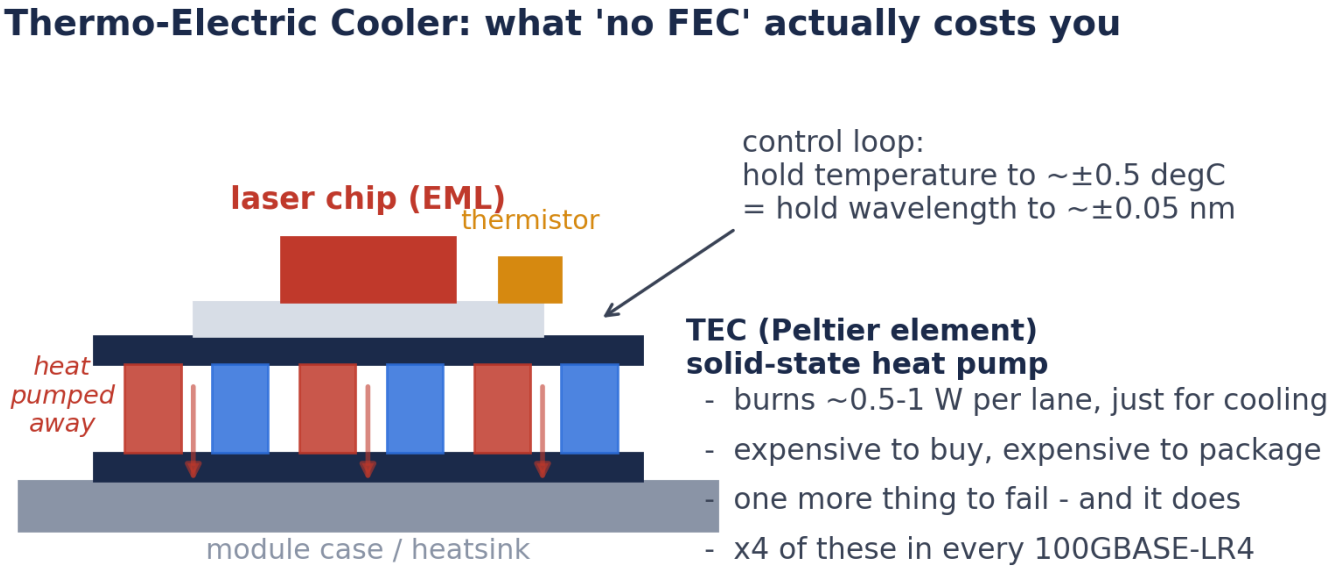
- No FEC → zero tolerance for ISI → **all four lanes must sit at minimum dispersion** to survive 10 km raw.
- LAN-WDM: 800 GHz (~4.5 nm) spacing at 1295.56 / 1300.05 / 1304.58 / 1309.14 nm — huddled tight against the G.652 zero-dispersion point.
- Each lane gets a window of roughly **±1 nm**. An uncooled laser drifts ~7 nm across its temperature range.
- Every laser is welded to a **TEC** and held at constant temperature for the life of the module.
- CD solved by brute force: **buy precision, cool it, and pay for it forever.**



The only way to survive 10 km with no FEC behind you.

GEN 1 · 100GBASE-LR4 (2010)

The Thermo-Electric Cooler (TEC) Tax



- A TEC is a solid-state heat pump squeezed between the laser and the case, running a control loop to hold the laser's temperature — and therefore its wavelength — constant.

- It is one of the most expensive, power-hungry, failure-prone components in any transceiver.

Gen 2: CWDM4/4WDM (2014)

The first FEC-native 100G — same speed, same fiber, half the price.

GEN 1 · LR4 · 2010 | **GEN 2 · CWDM4 · 2014** | GEN 3 · SINGLE λ · 2018+

GEN 2 · CWDM4/4WDM (2014)

Spend FEC Instead of Hardware

- 2014: QSFP28 hosts arrive with **KR4 RS-FEC already in the switch ASIC**. The CWDM4 MSA (Avago, Finisar, JDSU, Oclaro, Sumitomo) redesigns LR4 around that fact.

- CWDM4/4WDM changes exactly **two** things versus LR4:
  - **Turn on host FEC**: Tolerate pre-FEC BER of 5×10^{-5} rather than 10^{-12}, worth ~5-6 dB of margin.
  - **Relax the grid**: Swap LAN-WDM (±1 nm) for CWDM (1271/1291/1311/1331, ±6.5 nm)
    Let's you use uncooled lasers, cheap wide filters, no TEC.

- Everything else is identical: 4 × 25.78G NRZ lanes, duplex SMF, LC connectors.

- The FEC margin also quietly pays the CD penalty of the outer CWDM lanes (1271/1331 sit off the dispersion zero) — **math absorbing what hardware used to do.**

GEN 2 · CWDM4/4WDM (2014)

So Why Didn't Gen 2 (CWDM4/4WDM) Kill LR4?

- It did — **everywhere that real money was at stake.**

  Hyperscale 100G went CWDM4/PSM4 essentially overnight;

  LR4 shrank to a rounding error of their 100G volume within a couple of years.

- But the interconnection world never moved:
  - It's relatively easy to change what you deploy, when you control both sides of the link.
  - It's much harder to make a change when you have to coordinate with a remote party.
  - Customers don't know to ask their providers, providers don't know to offer it to their customers, and peers don't want to slow down an upgrade to negotiate yet another item.

- The Interconnection ecosystem spent a decade paying the TEC tax, essentially **out of habit**.

PART 4 OF 5

Gen 3: 100G Single Lambda (2018+)

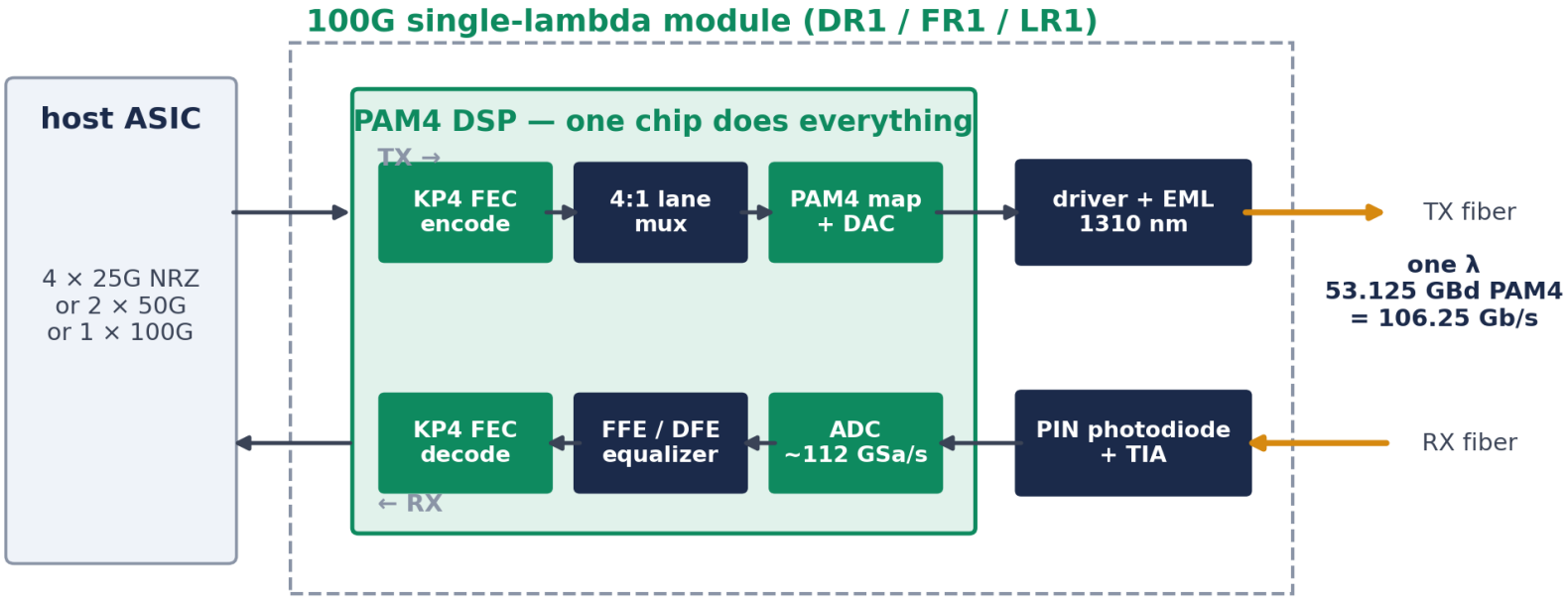
One laser, one wavelength, and a DSP putting it all together.

GEN 1 · LR4 · 2010 | GEN 2 · CWDM4 · 2014 | **GEN 3 · SINGLE λ · 2018+**

GEN 3 · 100G Single Lambda (2018+)

The Concept: All 100G on One Wavelength

- One optical lane at **53.125 GBd PAM4** = 106.25 Gb/s line rate (KP4 overhead included). No mux, no wavelength plan — **one laser instead of four**, and one chip doing all the hard parts:



the ENTIRE optical BOM: one uncooled laser + one photodiode. No mux, no TECs, no filters.

GEN 3 · 100G Single Lambda (2018+)

Why This Is Brutally Hard

	25.8 GBd NRZ (Gen 1/2 lane)	53.1 GBd PAM4 (Gen 3 lane)
Symbol duration	38.8 ps	18.8 ps
Levels to distinguish	2	4
Eye height (relative)	100%	~33% × 3 eyes
SNR required	baseline	~+9.5 dB
CD tolerance (vs 10G)	~16%	~4% — before the eye penalty
Raw decodability	photodiode + slicer	hopeless without DSP

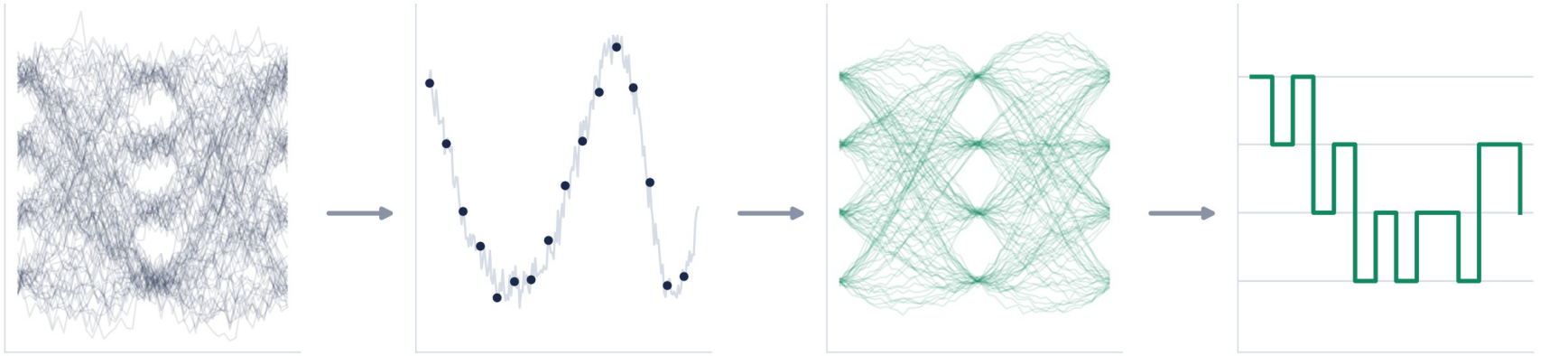
- Twice the baud **and** one-third the eye, simultaneously. After even a few km of fiber, the symbols are no longer separable by any analog receiver.

- The signal that arrives is not “degraded” — it is **mush, by design**, on the assumption that silicon will un-mush it.

GEN 3 · 100G Single Lambda (2018+)

The Enabler: The PAM4 DSP

The receiver stops trying to “see” the signal optically. It **digitizes everything and reconstructs the data with math** — then FEC mops up whatever survives:



what arrives

53 GBd PAM4 after the fiber: ISI everywhere, eyes closed. Useless to any analog receiver.

1 · ADC

digitize the entire waveform, ~112 GSa/s — the hardest analog circuit in the module

2 · FFE / DFE equalizer

weighted taps mathematically cancel ISI — tap count sets the CD window (= the reach limit)

3 · decision + KP4 FEC

re-slice the recovered eyes; RS(544,514) repairs stragglers → BER < 1e-12 out

The photonics stayed CWDM4-grade: an uncooled laser and a photodiode. The breakthrough was CMOS signal processing, not optics.

GEN 3 · 100G Single Lambda (2018+)

The PAM4 DSP Is Uniquely “Frozen" Silicon

- The hard block is the **ADC**: ~112 GSa/s with >26 GHz of analog bandwidth. Analog front-end performance does not ride Moore's Law — it's a physics-and-heroics circuit.
- The equalizer's reach is fixed at architecture time: **tap count = how much smear it can undo = the CD window.** More taps cost power, latency, and die area the commodity market won't pay for.
- Process shrinks (16 nm → 7 nm → 5 nm → 3 nm) deliver lower power, smaller die, lower cost — **never more dB and never a wider CD window.** A 2018 DR1 and a 2026 DR1 reach the same distance.
- Module power: the DSP burns ~1.5–2.5 W of a Gen 3 module's ~3.5–4.5 W total.
- Consequence: There is only one type of “PAM4 DSP” in existence, and every vendor’s DR1/FR1/LR1 are identical internally.

GEN 3 · 2018+

DR1 / FR1 / LR1: Three Labels, One Optic

	100G DR1	100G FR1	100G LR1
Standard	802.3cd (2018)	802.3cu (2021)	802.3cu (2021)
Label reach	500 m	2 km	10 km
Loss budget	3.0 dB	4.0 dB	6.3 dB
Modulation	53.1 GBd PAM4	53.1 GBd PAM4	53.1 GBd PAM4
Wavelength	1310 nm	1310 nm	1310 nm
DSP + FEC	same DSP, KP4	same DSP, KP4	same DSP, KP4
Actual difference	—	Tx power / Rx binning	more launch power

- Under the hood these are the **same design** — same DSP, same modulation, same laser technology — binned and tested to three different acceptance reports.
- You are largely buying a test report and a handle color. (The MSA even extended the family to ER1-30/ER1-40 — same DSP again, plus an SOA and APD.)

THE STANDARDS GAME

“But the Label Says 2 km”

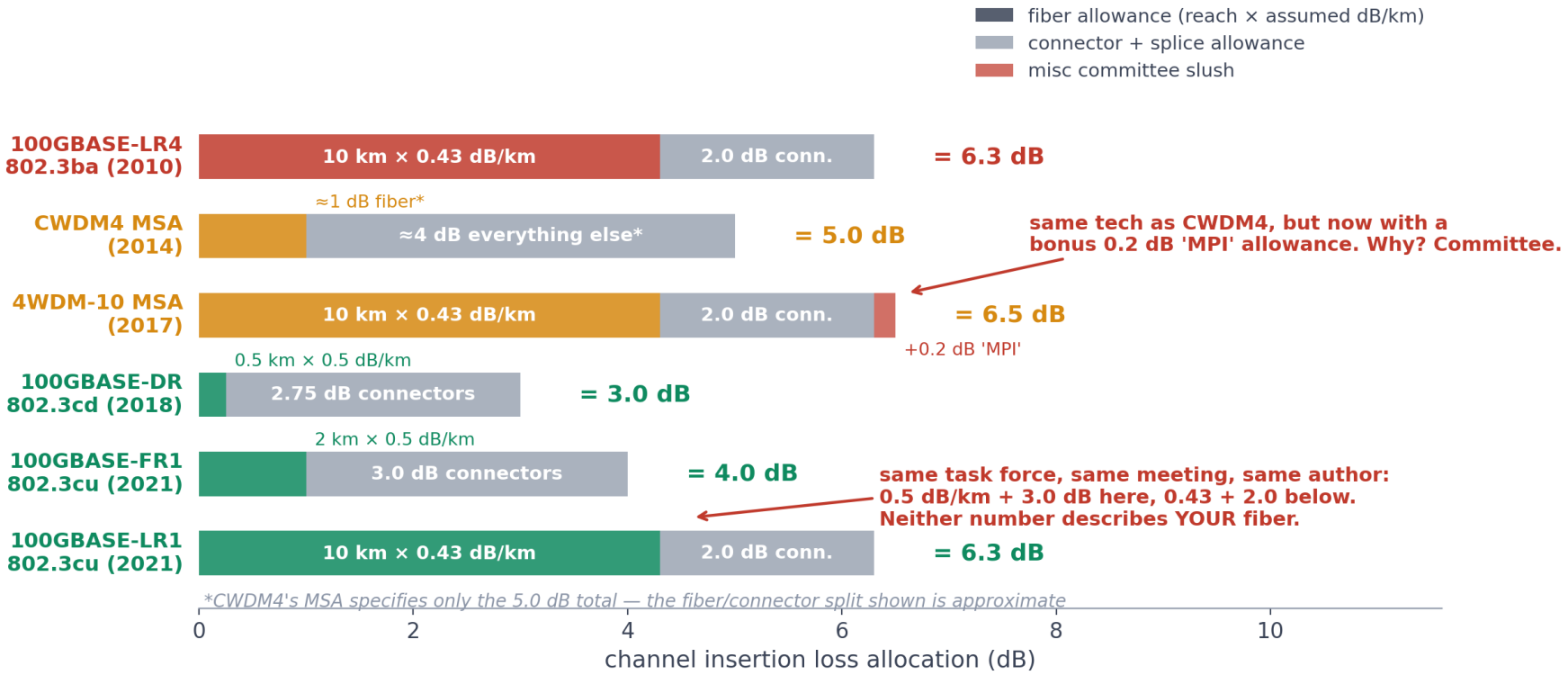
	100GBASE-LR4	CWDM4	4WDM-10
Label reach	10 km	2 km	10 km
Loss budget	6.3 dB	5.0 dB	6.3 dB
FEC behind it	none	KR4 (~5–6 dB)	KR4 (~5–6 dB)
Effective headroom	0 dB spare	~5+ dB spare	~6+ dB spare

- CWDM4's optical budget is nearly LR4's — **plus** an entire FEC engine standing behind it.
- **4WDM-10 MSA (March 2017)**: the SAME CWDM technology, formally specified to 10 km and interoperable with CWDM4. The “LR4 killer” — published, shipping, ignored.
- Field reality: a FEC'd “2 km” optic routinely outruns an LR4 on the same span — and degrades gracefully (correctable errors) instead of falling off a cliff.

THE STANDARDS GAME

How an Optical "Reach" Number Actually Gets Made

- The standards are never really designed around “kilometers”, they’re designed around **channel insertion loss**
- Then converted with some assumptions: Distance × (Assumed fiber dB/km) + (Assumed connector/splice allowance):



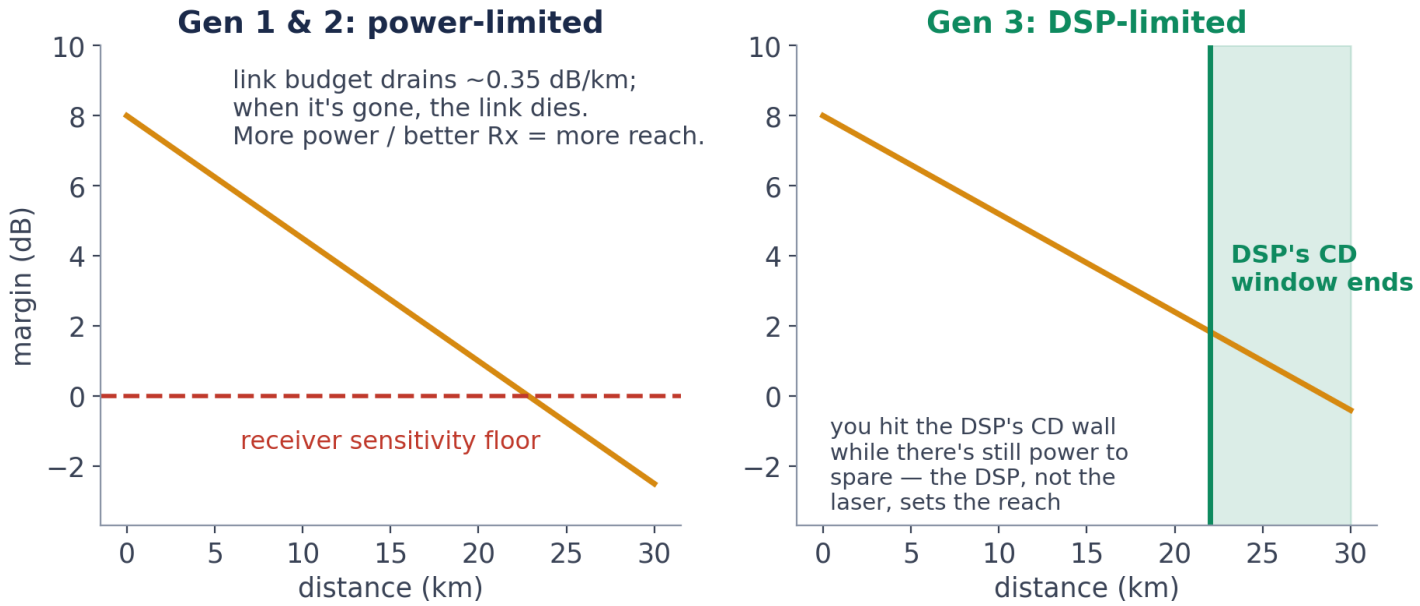
Your span is compliant at ANY distance where measured loss fits the budget (and CD stays in window). Real metro G.652 measures ~0.32–0.35 dB/km — every label above assumes worse fiber than yours.

GEN 3 · 2018+

With PAM4, Optical Power No Longer Sets Max Reach

- Gen 1/2 reach was primarily based on optical power budget: add launch power or a better receiver, go further.
- **Gen 3 reach is primarily constrained by the DSP's chromatic-dispersion equalization capabilities.**

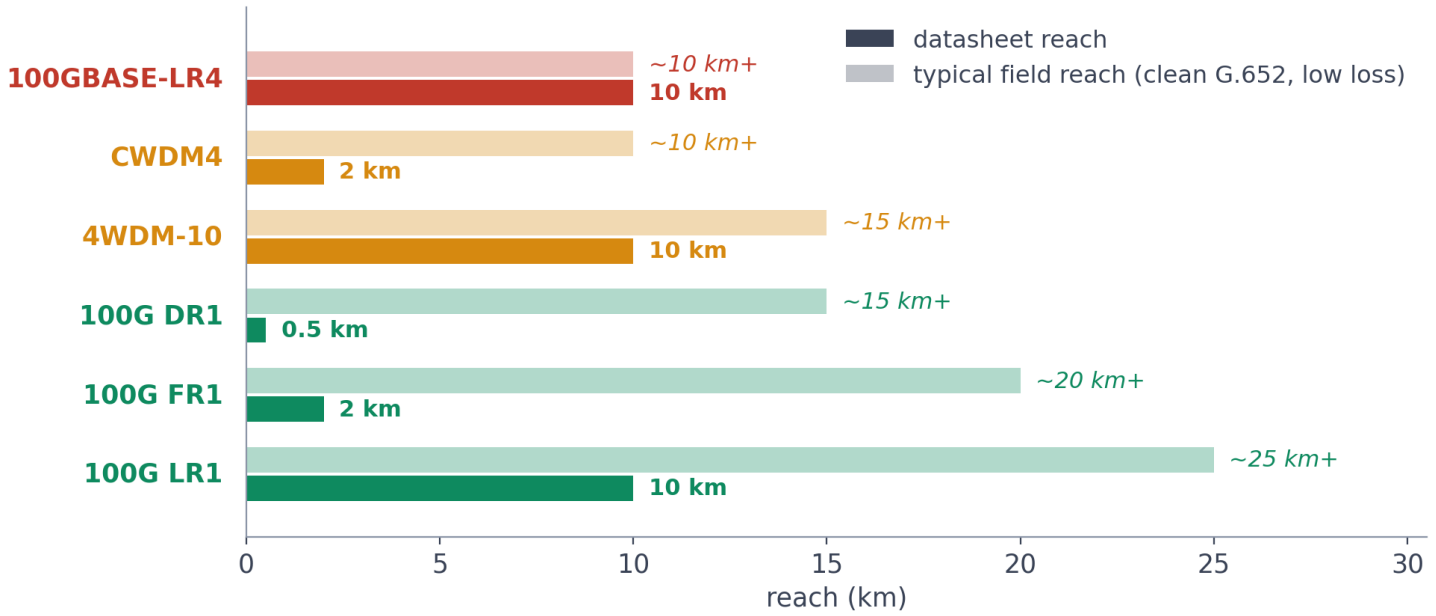
- The first client optics whose reach limit is **computational, not optical.**



GEN 3 · 2018+

What the Labels Mean in the Field

- At 1310 nm on G.652, dispersion accumulates slowly — so a “500 m” DR1 carries enough power and enough DSP window for **15+ km on any sane metro span.**



Typical results on clean G.652 — not a guarantee. The asymmetry is the point: FEC/DSP optics beat their labels; LR4's label is a ceiling.

THE STANDARDS GAME

Standards Are Written Years Before the Products

- An IEEE task force starts **3–6 years before volume product exists**: 802.3ba chartered 2007 → QSFP28 volume 2015. 802.3bs/cd chartered 2014–16 → PAM4 volume 2020. 802.3dj chartered 2022 → completing ~2026.
- At charter time nobody knows what the lasers, receivers, or DSPs will actually achieve. The committee **negotiates a best guess**: loss budgets, penalty allocations, reach tiers — with margin stacked on margin, filtered through vendor politics.
- That's why the DR / FR / LR ladder exists at all: the tiers were carved out when they plausibly required **different designs**. By the time the silicon converged on one design, the tiers were frozen into standards and purchasing habits.
- It's also why the assumptions drift randomly between projects (0.43 vs 0.5 dB/km; 2.0 vs 3.0 dB of connectors): **different meeting, different compromise.** Nobody went back to reconcile.
- **Datasheet labels encode committee history, not physics. You should engineer with dB budgets and CD windows, not with a distance estimate in the name of the standard.**

GEN 3 · 2018+

“So Are the Vendors Scamming Me?”

- Fair question: why do three products that are the same thing under the hood exist at three prices? **Answer: no — you're just not their target customer.**
- The tiers are frozen in the standards (see previous slide); the silicon converged underneath them. What's left is three test reports of one design.
- The overwhelming majority of optical transceiver business is going to hyperscalers interconnecting within very large buildings (500m). And the hyperscalers already know what works and what to buy.
- Interconnection demand is **a rounding error of a rounding error** of that volume. The market is simply not designed around this room — and never will be.
- The upside: buy the commodity SKU and you inherit hyperscale economics for free.

The Road Ahead

The story so far — and how it ends for LR4.

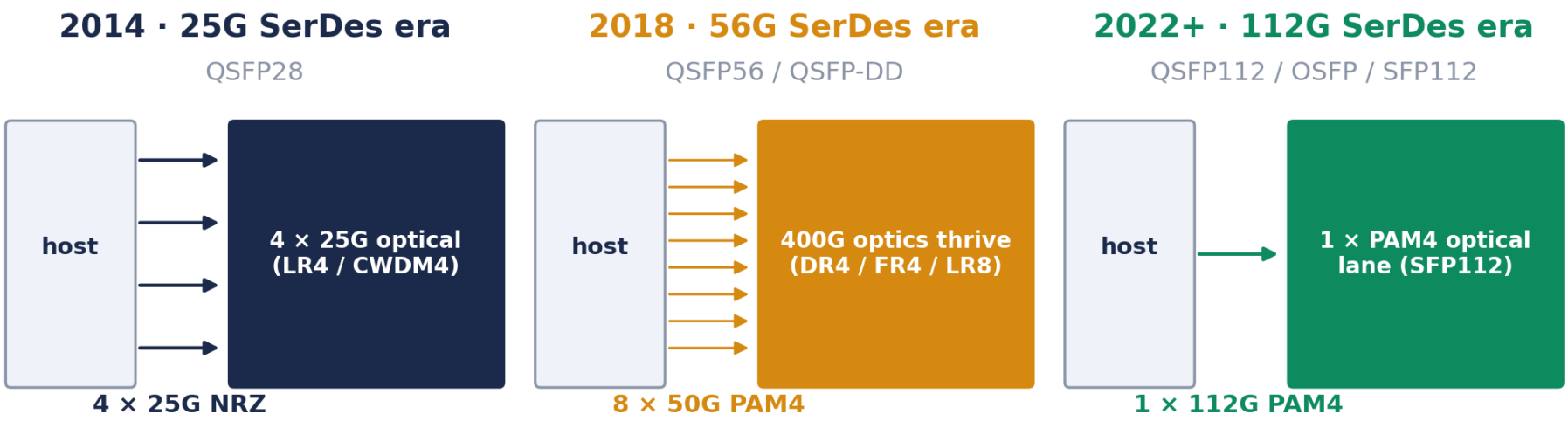
GEN 1 · LR4 · 2010

GEN 2 · CWDM4 · 2014

GEN 3 · SINGLE λ · 2018+

THE ROAD AHEAD

But The Ultimate End-of-Life Move Will Come from SerDes Speeds



electrical = optical, 1:1
100G is a native port.
No DSP anywhere.

100G = 2 × 50G: built exactly once (a bespoke SFP56-DD). No traction — 100G stays on legacy 4 × 25G ports.

1 : 1 again — no gearbox, minimum BOM, cheapest 100G ever. Next up 224G: 100G is half a lane.

The 56G era is the awkward middle child: it built the entire 400G ecosystem and exactly ONE new 100G optic — a bespoke SFP56-DD 2×50G DR1/FR1 (Eoptolink, for Juniper). Horrible pricing, no traction: the exception that proves the rule.

THE ROAD AHEAD

SFP112: The Endgame for Pure 100G

- **SFP112** = the classic SFP cage with a single 112G electrical lane feeding a single 53 GBd PAM4 optical lane. No gearbox, no lane conversion, the minimum possible BOM.
  - Shipping now from multiple vendors: SR1 / FR1 / LR1 / ER1 flavors covering 100 m to 30–40 km.
  - This will be the **cheapest 100G optic ever made** — it's a 25G SFP28's parts count running 4× the bits.

- Today, every modern switch built to support AI is based around the 112G SerDes (e.g. 64x 800G = 51.2T on a single chip / box), primarily consumed as 8-lane OSFP112 or QSFP112-DD today.
- Ever modern NIC built for AI is going to QSFP112 as a 4-lane optic that’s 112G natively end-to-end.

- This is the hardware that will eventually make its way down to Internet networks.
  - Implementing 4x100G xR1 QSFP112 or 8x100G xR1 OSFP will cost 90% less per 100G this way.
  - And on those platforms, an LR4 optic can only run by taking a 400G or 800G port and running it in 100G mode.

- **When this technology finally lands, LR4 optics become a stranded asset with zero long-term value.**

THE ROAD AHEAD

What To Actually Do (Peering / IX Edition)

- **Stop writing “LR4” into port specs, LOAs, and cross-connect docs!**
- **DR1** covers essentially every metro cross-connect already. You usually only need **FR** if you’re talking about **400G FR4**, and want the 4-lanes muxed onto a single LC port. All three are the same silicon. And every transceiver vendor who sells to Google already has an extra 3dB of margin above spec baked in to support their photonic switching layer.
- **Interop is per-span**: 100G Single Lambda does not talk to LR4/CWDM4 on the wire — different signals entirely. But all 100G Single Lambda optics talk to all other 100G Single Lambda optics.
- You should probably be **graphing pre-FEC BER**: you’ll get an early warning on degraded links before you ever drop a single packet.
- Price-check reality: FR1 street pricing reached parity with LR4 years ago; SFP112 will land below that. **You are no longer saving money by being conservative — you're just paying the TEC tax.**

THE ROAD AHEAD

The Scorecard

	Gen 1: LR4	Gen 2: CWDM4/4WDM	Gen 3: DR1/FR1/LR1
Era	2010	2014–2017	2018 → now
Signal	4 × 25G NRZ	4 × 25G NRZ	1 × 53 GBd PAM4
Grid (Passband)	LAN-WDM ±1 nm	CWDM ±6.5 nm	single λ, 1310 nm
Lasers / TECs	4 / 4	4 / 0	1 / 0
WDM filters	2 × precision	2 × relaxed	none
FEC	none	KR4 (host)	KP4 (mandatory)
Module power	~3.5 W (TEC, but no DSP)	~2.5–3.5 W (No TEC or DSP)	~3.5–4.5 W (DSP ~2 W)
Reach limit	optical power, no margin	optical power + FEC	DSP CD window
Label vs field	10 km = ceiling	2–10 km → ~10–15 km	0.5–10 km → ~15–25 km
Trajectory	stranded by 112G+ SerDes	fading with 25G lanes	SFP112 → cheapest 100G ever

Summary: Let It Go Already!

- 100GBASE-LR4 (2010–2026): the last FEC-less optic still shipping in volume.
- Obsolete on **cost** since 2014, on **performance** since 2018, stranded by **112G+ SerDes** going forward.
- For everything LR4 can do, there is something better and cheaper available - for over a decade.
- The next person who asks you for 100GBASE-LR4: **send them this deck!**