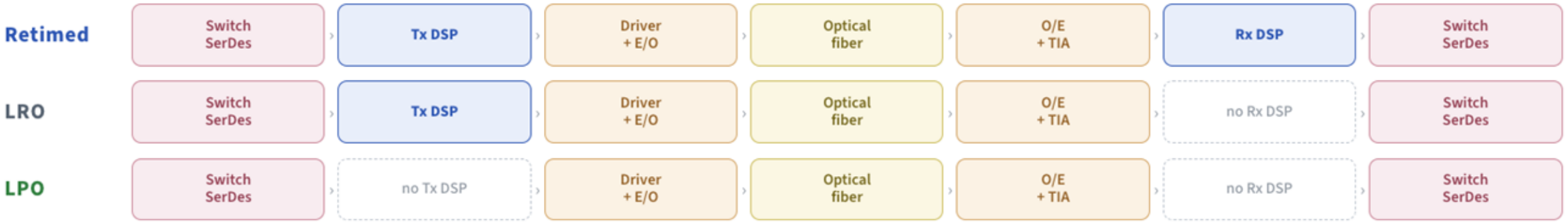
Operating Linear Pluggable Optics at Hyperscale: From Link Management to Production Reliability

Amin Yekani *[Optical Engineer, Meta Platforms]*

Qing Wang *[Optical Engineer, Meta Platforms]*

OCP PLATINUM Member

LPO Benefits and Challenges



The problem it solves

- At AI scale, the per-module retimer DSP is a repeated tax in power, heat, and cost across millions of links.

The challenge

- No DSP safety net in the module. The host must own equalization, lost triage capabilities.

The question we answer

- Can linear optics beat retimed reliability at fleet scale?

We shaped it, we productized to it, now we put it to the test.

OFC/OCP 2024
Definition
Studied feasibility and defined Meta's LPO architecture and spec for the module, informed by the industry MSA and broader ecosystem activity.

OCP 2025
Productization
How we productize: qualification, host and module calibration, diagnostics, and the path to production.

OCP 2026
Reliability at scale
Checked LPO reliability at datacenter scale: fleet evidence from live production and a controlled lab.

This evidence is from 100G/lane LPO in large-scale production, but the lessons carry forward to next-gen linear: 200G/lane LRO/LPO, NPO, and beyond.

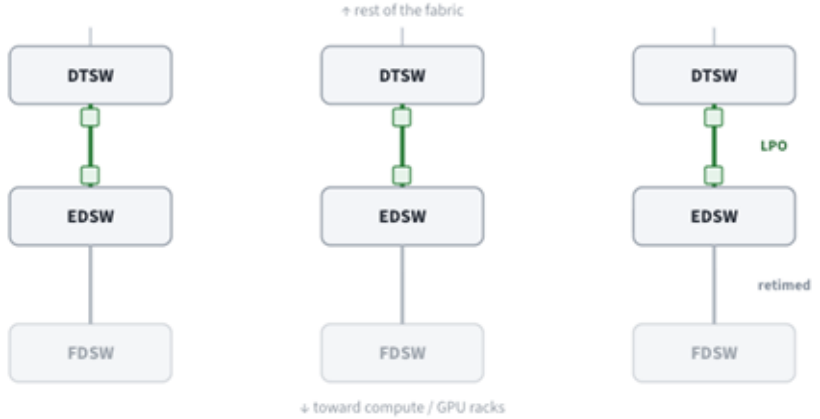
What changed across the three years.

Item	OFC/OCP 2024 Definition	OCP 2025 Commitment	OCP 2026 Data (this talk)
Host data rate	2x400G, 8x100G, 1x800G	Same	Same
Media data rate	2x400G, 8x100G, 1x800G	Same	Same
Interop between optics	LPO-LPO; retimed-LPO	Required: LPO-LPO. In progress: retimed-LPO	Limited deployment LPO-LPO
Host SerDes setting	Same settings per switch vendor/MPN, independent of LPO vendor/MPN (per-port settings allowed)	Tx FIR calibration depends on both switch and modules; work in progress on decoupling the two contributions	Decoupling now proven, but not used in this deployment (production runs the combined switch + module calibration)
Pre-FEC BER target	<3E-6	Same	Same
Fiber link budget	Reach = 500 m, IL = 2.5 dB	Same	Same
TP1/a compliance	oif2021.405.08 + minor update	oif2021.405.11	Same
TP2 compliance	Retimed lite spec (internal)	LPO-specific spec: tightened optical and electrical parameters vs. the retimed lite spec (details internal)	Same
TP3 compliance	Retimed lite spec (internal)	Same	Same
TP4/a compliance	oif2021.405.08 + minor update	oif2021.405.11	Same
200G/lane	Retimed Tx required to be TP2 compliant	Same	Same
Diagnostics	New features needed at Driver and TIA	Same	Same

One question, checked two ways: live production and a controlled lab.

Setup 1 Live DC production

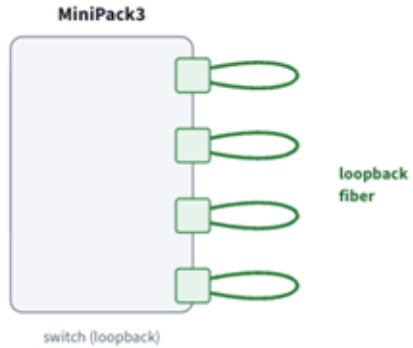
LPO on one hop of a DSF network fabric, LPO on both ends.
LPO runs the DTSW-EDSW hop, All the rest are retimed.



Link 2x400G FR4, 1x800G mode	**Placement** DTSW ↔ EDSW edge
Scale 4k modules - 4 months in production	**Software** FBOSS (production network OS)

Setup 2 Controlled lab isolation

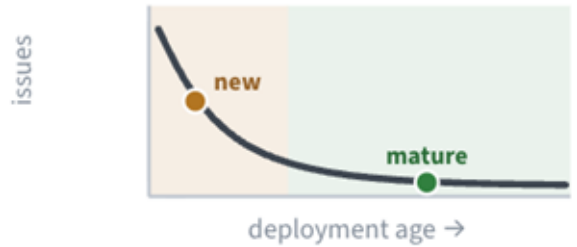
~1k modules in loopback on our MiniPack3 (51T) platform.
Strips away DC-specific exposures to isolate module behavior.



Modules ~1k in loopback	**Placement** MiniPack3 switches
Mode 2x400G	**Software** Diag OS stack

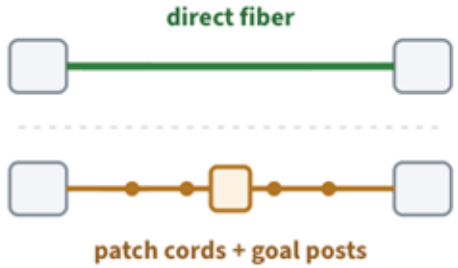
Comparing Linear Optics to retimed is not a hookup-and-count.

Age of the setup



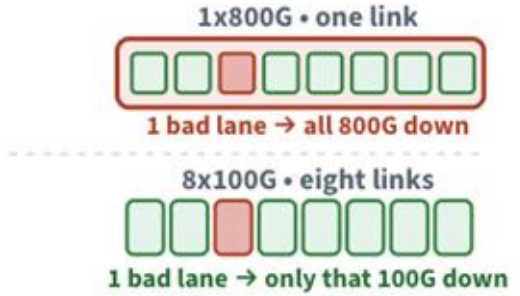
New fleets settle as they mature. Compared at equal fleet maturity on the verdict slide.

Network layer



Each layer takes a different physical path.

Link mode



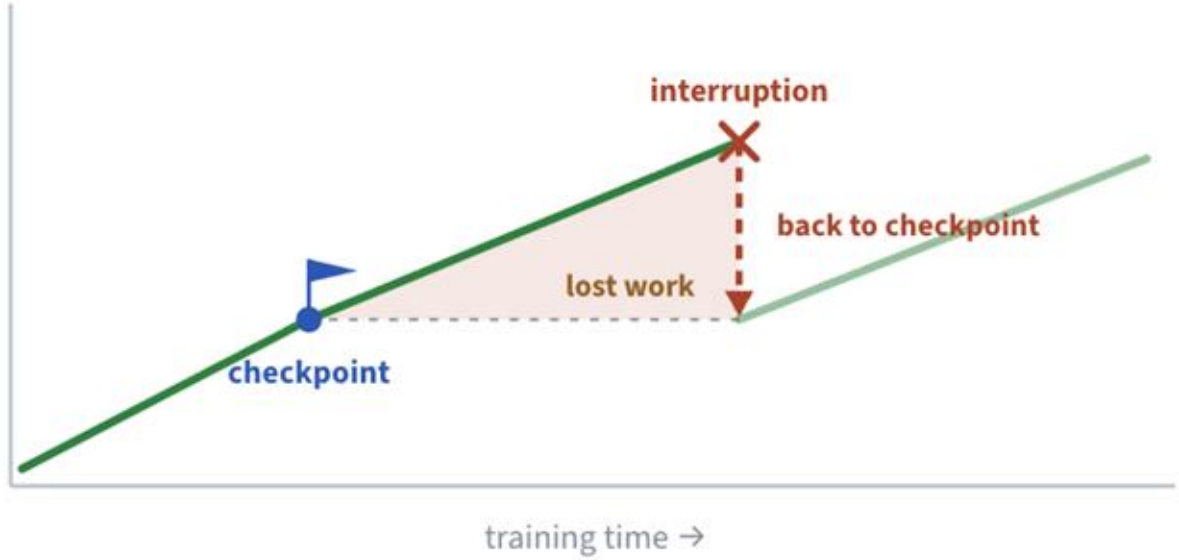
How lanes are grouped into links sets both FEC sharing and failure blast radius. Any LPO vs. retimed comparison must therefore use the same link mode.

Control for age, layer, and mode and the comparison reflects the optics, not the deployment.

At AI scale, any interruption can potentially send training back to the last checkpoint.

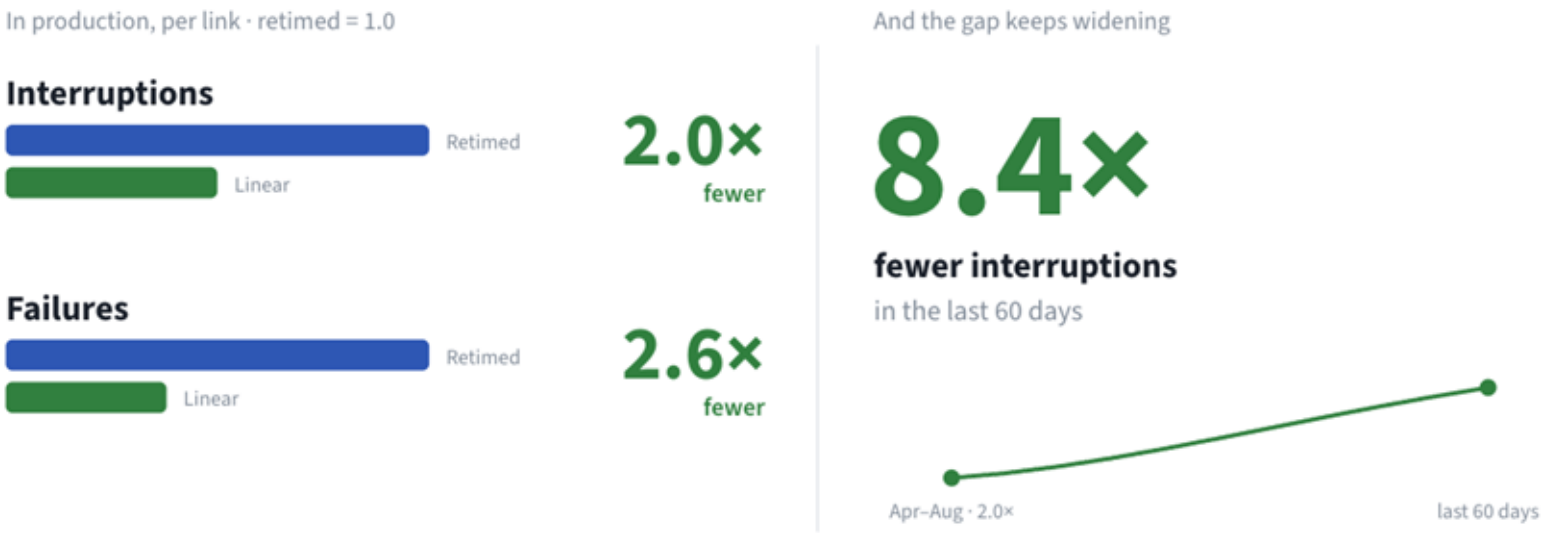
Failure · the old bar
Crosses the data center incident threshold. Triaged and service-affecting.

Interruption · the new bar
Any brief link event, even one below today's triaged threshold. As training tightens what it can tolerate, that threshold has to tighten with it, and interruption rate, not failure rate, is the metric that keeps up.



And the bar only rises. As optics move on-package (NPO, co-packaged), a failed link can no longer be pulled and swapped like a pluggable. Giving up that serviceability only pays off if reliability climbs to match.

Failure Rate vs. Interruption Rate (Live Production)



Same link mode, live production: linear runs ~2× fewer interruptions and ~2.6× fewer service-affecting failures than retimed.

- At the module level the two optics are on par, so the gain is architectural: no DSP means one fewer class of failure.
- The advantage keeps widening as the fleet matures: ~2× over the full record, more than 8× in the last 60 days.

Controlled lab environment

Setup

- continuous loopback soak at our lab, monitored end to end.
- Fan speed is held fixed for the entire run
- Stability: 8 lanes, 104 days: optical power holds flat (≤0.25 dB peak-to-peak, ≤0.04 dB per-lane drift)

Clean module

- Same typical FEC load as retimed, tighter tail, wide margin to codeword loss
- Median pre-FEC BER ~1×10−9
- 0 module-attributable failures over 2.96M link-hours

More efficient to run (fixed fan speed)

- −46% power per module
- −17.5°C cooler at that same fan speed





Call to Action

- Problem to solve
  - Deploy LPO at scale without giving up reliability
- Build LPO module ecosystem
  - Try our recipe. Qualify your module/switch and share performance
- Where to find us
  - Amin: ayekani@meta.com
  - Qing: qingwang@meta.com

Thank You!

