The emergence of 112G PAM4 for cloud data centers

Steve Taranovich - April 20, 2018

David Maliniak of Teledyne LeCroy wrote an excellent article in 2016 that explained the fundamentals of PAM4. We want more data and we want it faster, Maliniak says. NRZ encoding is not fast enough for us in the coming era of 5G and NRZ-type encoding will not meet these needs. My colleague, Martin Rowe, commented that NRZ is dead, but not everywhere according to a DesignCon 2018 panel entitled ‘The Case of the Closed Eye.’ Well, at 56 GHz and short-reach links over copper, NRZ (known as PAM2) is still adequate and useful. But everywhere else, PAM4’s four-level pulse amplitude modulation rules! Especially at 56Gbps and 112Gbps.

Broadcom had run experiments on PAM4, PAM8, and PAM16 and found that PAM8 and PAM16 have eyes that are too small; this is where PAM4 stepped in at 56Gbps. Rowe made another observation following DesignCon 2018: 112 is the new 56.

Other panel members had comments such as PAM4 being made possible with forward-error correction (FEC) and that DSP processing will enable 112 Gbps. I’ll talk about that later in this article.

The importance of high-speed cable/connectors

Sometimes we forget that the cable and connectors might be some of the weakest links in the high-speed channel. Samtec and Credo presented a demo at OFC 2018 this year with Samtec’s Flyover System. They set up two ports of 112 Gbps PAM4 data sourced from Credo clock and data recovery (CDR) chips. The signals were routed through RF jumpers to an SI characterization card for Samtec’s FOSFP-DD connector. The signal then traveled through 12 inches of Samtec ultra-low Skew Twinax cable and then on to a Samtec accelerate high speed cable assembly SI characterization card.

Then the signal path finally passed through a second set of RF jumpers, to a second Credo CDR and we finally see the channel output on the GUI which shows that the 112 Gbps PAM4 data was run at a pseudo-random bit sequencing of 31 at 2e⁻⁷. Voila--we have proof of concept for signals at high speed over a fairly long-reach cable.

Check out the video from Samtec’s site:
Challenges and opportunities

There is an ever-increasing demand for bandwidth (BW) at a blinding pace; the ICs, system, and optics industries are trying to converge on the next generation signaling rate to cater to this surge in BW demand. IEEE 802.3 is working on 100G signaling for 100GBASE-DR1 and the 400GBASE-DR4 & OIF PLL group has started to work on CEI-112G-PAM4-VSR. MACOM has envisioned this transition early and has had the opportunity to “play” with this technology in the lab for over a year.

Data centers

Optical packet switching can provide an energy-efficient way to communicate within the data center, especially at such rates of 112 Gbit/s. An analysis has been done using PAM4 and PAM8 with pre-distortion and a look at three kinds of optical receivers. A single-stage system allows the connection of 48 servers using PAM8 along with a semiconductor-optical-amplifier (SOA)-PIN with a 62.5 GHz grid. By using a two-stage configuration, the number of servers that could be connected jumps to 1488. The number of connected servers is limited by two parameters: (1) the optical power budget, which depends upon the type of optical interface and (2) the number of wavelength channels that can be addressed with respect to the tuning band of the lasers used, the resolution of the tuning mechanism, and channel spectral occupancy.

The internal DC interconnection networks account for 23% of the total data center power and most networks use electrical packet switches (EPS), which are connected optically up to 10 Gbit/s. However, higher bit rates are now possible at 40 to 100+ Gbit/s. Performance/cost ratio is important here and the present systems make use of parallel links like 4×25Gbit/s or 10×10 Gbit/s, or even multi-level formats based upon intensity modulation and direct detection (IM-DD).

I was happy to see that the use of low-power analog (circuit or packet level) optical switching reduced power consumption here instead of digital (bit-level) optical switching--take that, digital dweebs. In addition, optical switching should help lower latency introduced by intra-DC connections.
The authors of Reference 2 chose to use optical-packet switching (OPS) to suppress EPS in intra-data center connections. Passive-optical-pod interconnect (POPI) has a simple infrastructure that is passive using an optical star coupler. POPI can be used to connect racks and servers depending on the desired transfer capacity (Figure 1).

![Diagram of POPI architecture.](image)

**Figure 1** In the POPI architecture, the servers in Rack 1 and 2 share a whole wavelength (red/green). Servers from rack r use different wavelengths. (Image courtesy of Reference 2)

Techniques that rely on fast-wavelength-tunable lasers like time wavelength interleaved network (TWIN), especially in metropolitan area network (MAN) applications, are being studied and are getting a great deal of attention by design engineers, especially since they can reduce power consumption and have reduced latency, as compared to other solutions, which is critical to server systems. Increasing the interconnection bit rate up to 112 Gbit/s facilitates fast server migration and will allow some servers to be turned off according to the available electrical power and workload.

**Long reach**

Inphi developed the first Gigabit Ethernet PAM4 IC chips for cloud interconnects in 2015. This 100G, optical PAM4 modulation scheme reduces the optics count by doubling the bits/symbol at the same baud rate as four fibers/wavelengths at 25Gbps each that 100G data centers have used to date. The complex functionality is transferred into CMOS with PAM4 encoding, real-time DSP and FEC technology. This improves bandwidth at a lower cost as compared to the NRZ solutions presently used.

The new PAM4 chipset was used in the Reference 4 paper and link performance with real-time DSP leading to a small silicon size was studied for standard OM4 and wideband multimode fiber (WBMMF) to achieve upgrades of 40/50 Gbps and 100/200 Gbps. The PAM4 transmission was able to reach record-proportions at 550m with the use of a Ge/Si avalanche photoDiode (APD) and a
record aggregated rate of 212.5 Gbps over WBMMF with real-time DSP processing (Figure 2).

Figure 2 The PAM4 test architecture is shown here with (a) 1λ 40/50 Gbps vertical-cavity surface-emitting laser (VCSEL), (b) 2λ (or greater) 100/200 Gbps VCSEL, with 850 nm TX optical eye diagram, and RX DSP recovered histograms shown in the inset; (c) shows the EMB vs. wavelength for conventional OM4 and WBMMF. (Image courtesy of Reference 4)

The test showed error-free transmission with an excellent margin under KP4 FEC threshold at 1310nm wavelengths for up to 40km. This solution enables small form factor modules like CFP4 and QSFP28 enabling a much smaller footprint with improved performance over existing designs.

**Short reach**

I was Nortel Account Manager/Application Engineer for Burr-Brown in the late '90s and saw Nortel's first coherent transponders in the industry run at 40G speeds. The telecom industry was not quite ready then for 40G because the telecom providers said that there was so much "dark" fiber in the ground and they wanted to fill that data pipe now before investing in a faster system.

In the last 10 years or so, developers ran various different baud rates with quadrature amplitude modulation (QAM). Key research efforts recently have focused upon coherent QAM systems with DSP processing to reach longer transmission distances at higher bit rates.

Problems that are seen in the kind of systems mentioned above are chromatic dispersion (CD) and polarization mode dispersion (PMD), which are corrected by fixed and adaptive linearequalizers, respectively (Figure 3).
Adaptive filters are typically used in noise and echo cancellation, sinusoidal rejection, equalization, and so many more applications.

In Figure 3, the symbols \( s \) get sent over a channel with a transfer function \( H \). The equalization filter can have different structures with coefficient vector \( w \) and input sample vector \( x \), where \( M \) is a filter span that indicates intersymbol interference (ISI) spreading (\( M \) sampling periods). In this case for Reference 5, we assume \( M \) is an odd number = \( 2K + 1 \). \( w_{dc} \) is the dc component. See Reference 5 for more details.

**Volterra and Wiener equalizer filters**

The Wiener filter has the Volterra filter as one of its subsets. In this paper, the authors used a stochastic gradient adaptive algorithm that is based on the discrete nonlinear Wiener model.

Volterra equalizer filters, a subset of the Weiner filter, are notoriously good for modeling the distortion in semiconductor laser diodes, the transfer function of single-mode filter, the non-linear propagation within multimode interference couplers, and more. The least mean square (LMS) algorithm, used in the Volterra system, is a stochastic steepest-descent algorithm in which the true gradient vector is approximated by an estimate obtained directly from the input and output signals, and is very simple. However, when the autocorrelation matrix eigenvalues have a large spread, slow convergence is unavoidable. Using the discrete fourier transform (DFT) or Gram-Schmidt (lattice) procedure can achieve better orthogonality. In Reference 5, the restricted Volterra filter was determined best for bandwidth-limited systems. Band-limited systems need a duobinary PAM4 equalizer filter; orthogonalization is not possible here.

In the final analysis, a third-order Volterra filter was determined sufficient in all test cases in Reference 5 for low-cost 112 Gb/s PAM4 receivers for metro and data centers’ network segments.

It looks like 112G is maturing nicely and not a moment too soon as cloud data centers are multiplying across the global landscape. At OFC 2018, Credo showed off their 112G PAM4 offerings.

High speed and low power consumption rule in this area and I expect to see more innovative techniques in both of those areas in the coming years.

**References**
1. 100G transition: Electrical & Optical, Challenges & Opportunities, Francesco Caggioni, MACOM, 2017, IEEE

Also see:
- Samtec: DesignCon 2018 Parting Thoughts---How real is 112 Gbps PAM4?
- Xilinx: 112Gbps Serial Transmission over Copper—PAM4 vs PAM8 Signaling
- Journal of Lightwave Technology: 112Gb/s for future 5G Ethernet-based Fronthaul Architecture

Related articles:
- The fundamentals of PAM4
- High-speed digital: 112 is the new 56
- 112G, PCIe Gen4 Highlight DesignCon 2018
- QSFP-DD pluggable modules boost data density
- NRZ is dead, but not everywhere
- What good is state-of-the-art high-speed electronics without good connectivity?
- Construct a 56 Gbaud PAM4 signal source
- Cloud data center server power and optical transceivers: a dynamic duo
- Building a cloud-native 5G core will be essential
- How 5G is changing data centers