Towards Neural Network Equalizer Implementations for IM/DD Transceivers

Christian Bluemm
Optical & Quantum Communications Laboratory
Huawei Research Center
Munich, Germany
christian.bluemm@huawei.com

Bo Liu
Chair for Electronic Design Automation, Technical University
Munich, Germany

Talha Rahman
Optical & Quantum Communications Laboratory
Huawei Research Center
Munich, Germany

Bing Li
Chair for Electronic Design Automation, Technical University
Munich, Germany

Ulf Schlichtmann
Chair for Electronic Design Automation, Technical University
Munich, Germany

Stefano Calabrò
Optical & Quantum Communications Laboratory
Huawei Research Center
Munich, Germany

***Abstract*— Next generation DCI/DCN reaches 800Gb/s. Distances up to 10km make chromatic dispersion the bottleneck for O-band PAM4 CWDM4. Neural network equalizers provide effective compensation. We compare against state-of-the-art and discuss implementation trade-offs for real-world transceivers.**

***Keywords*— Digital Signal Processing, Intensity Modulation Direct Detection, Chromatic Dispersion Mitigation, Neural Network Equalizer**

I. Introduction

Within the past 20 years, the peak compute rates (in hardware flops) scaled up about 3000 times more than interconnect bandwidths, as shown in [1]. Reducing this imbalance of progress in computation hardware vs. interconnects is especially crucial for data centers and cloud services, where the transport of data is equally important to its processing. In order to remain the key enabler and backbone for our information age, the capacity of data center interconnects (DCI) and data center networks (DCN) must therefore grow much more rapidly, while staying affordable and power-efficient.

Most of today's DCI/DCN technologies are based on intensity modulated/direct detect (IM/DD) transceivers, which support IEEE 802.3 Ethernet options. The most urgent next generation IM/DD use case in this context will be 200Gb/s per lane in a four times multi-lane coarse wavelength division multiplexing (CWDM) configuration [2].

The main limitations for such high speed IM/DD systems are two nonlinear phenomena, namely four-wave mixing (FWM), which is the crosstalk between multiple CWDM channels, and more crucial, chromatic dispersion (CD), which is the frequency dependency of phase velocities of the lightwave pulses [3]. Photo-detection turns the linear CD problem into a nonlinear one and makes it the main bottleneck for higher reaches and rates. Its mitigation is essential and primarily done with digital signal processing (DSP). State of the art techniques, however, make transceivers quickly too complex.

In this paper, we discuss upon the example of 10km IM/DD O-band CWDM4 transmission with 200Gb/s per lane options of low-complexity DSP at the receiver (Rx) to mitigate CD with neural network (NN) nonlinear equalization (NLE).

In order to yield 200Gb/s per lane, we choose the line rate of 112GBd PAM4, accounting for typical forward error correction (FEC) overheads. Which type of FEC to choose for this use case is yet to be decided [4] and will not be further discussed here.

We show upon measurement results, how NN-NLE can meet Volterra NLE performance with 30% less hardware multiplier complexity. When also applying magnitude weight pruning, an additional 43% reduction is possible without performance loss across all CWDM4 lanes. If needed, an added MLSE stage can further push performance in both cases. In any of these configurations, a key enabler against strong CD penalties is duobinary training, which is applicable to all feed forward equalization architectures.

The remainder of this article is structured as follows. Section II introduces NLE designs against CD perturbations in IM/DD systems and how NLEs can benefit from duobinary (DB) training targets. Section III outlines our measurement setup, including results. A concept for area-efficient adjustability is introduced in section IV. Section V discusses hardware parallelization and section V draws conclusions.

Parts of the work presented here have been published before in [5] and [6]. This paper, however, extends previously published results by discussing in greater detail implementation aspects, including hardware parallelization for real world IM/DD transceivers or adjustable designs.

II. Equalizers for Chromatic Dispersion

CD is caused by varying propagation speeds (different group velocities) of different frequency components of light wave signals. Its effect is pulse widening, which may span over 100's or 1000's of symbols. It is a nearly static linear effect of the fiber with signal distortions in phase, not in amplitude.

By means of photo-detection, IM/DD receivers generate the square of the received lighwave signals. Within this step, the phase information gets lost, which is the reason why linear CD distortions become a nonlinear challenge after detection [7]. The main impairment here are spectral nulls in the transfer function of the detected signal. Due to its nonlinear nature, linear equalizers cannot recover it.

A. Volterra Nonlinear Equalization

Classical nonlinear equalizers are usually based on Volterra

series. With $x(n)$ and $y(n)$ representing system input and output, respectively, the Pth order discrete time Volterra series with M_p memory taps for order p is [8]:

$$y(n) = \sum_{p=1}^{P} \sum_{m_1=0}^{M_1} \cdots \sum_{m_p=0}^{M_p} h_p(m_1, \cdots, m_p) \prod_{k=1}^{p} x(n - m_k) \quad (1)$$

An NLE based on this Volterra series maps current input samples $x(n)$ and historic samples $x(n-m)$ as a linear combination of nonlinear functions (kernels). Implemented as such, each pth-order Volterra kernel $h_p(m_1,\ldots,m_p)$ describes all possible combinations of a product of p time shifts of the input signal up to the memory m_p. Pruned versions with reduced kernel sets are not considered here.

The V-NLEs in this article identify optimal kernels in terms of the least squares (LS) error criterion upon training data, which are received sequences with their known transmitted counterparts [9].

According to [8], the number of multipliers for V-NLE equals the sum of unique kernels across all orders, with C_{M+1}^1 unique kernels for first, $C_{M+1}^2 + C_{M+1}^1$ for second and $C_{M+1}^3 + 2C_{M+1}^2 + C_{M+1}^1$ for third order,

$$with \qquad C_m^p = \frac{m!}{(m-p)!p!} \quad (2)$$

B. Neural Network Nonlinear Equalization

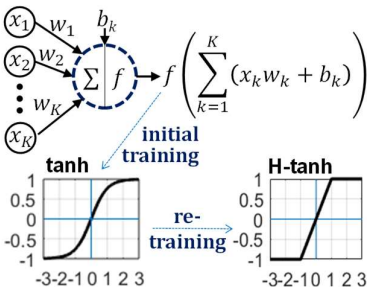

Fig. 1: Artificial neural network with *tanh* and *H-tanh* activation

The top of Fig. 1 shows the basic processing unit of any NN, the artificial neuron. Weights $w_1,\ldots,w_K$ scale the input signals $x_1,\ldots,x_K$ before feeding their sum to a (non-)linear activation function. The bias b_k offsets the output.

For our NN-NLEs, *tanh* serves as activation for initial training and is then replaced by its low-cost variant *hard tanh (H-tanh)* [10] for further re-training, as illustrated in the bottom of Fig. 1. Re-training is motivated by the simplification of *tanh* to *H-tanh*.

For hardware implementations, *H-tanh* requires only one hardware multiplier for its linear part, defined by m, and two comparators for clipping. Some equalization tasks, unlike the one presented, might even allow to set $m=1$ to end up with no need for any multipliers at all. In either case, *tanh* would be significantly more complex to implement.

Fig. 2 shows a corresponding equalizer configuration, which maps memory at the input with a tapped delay line and then nonlinearities with hidden layers (HL) and an output layer, which is chosen as a single linear neuron for optimal performance in our case. In sec. III, we consider both, fully connected HLs, as well as pruned subsets. Pruning has proven powerful compression capabilities for NN-NLEs for IM/DD systems [11][12] and can even speed up the convergence of the training process [13].

For identification of our adjustable NN-NLEs parameters (weights, biases) against CD distortions, we apply minibatch backpropagation with the ADAM optimizer [14] within

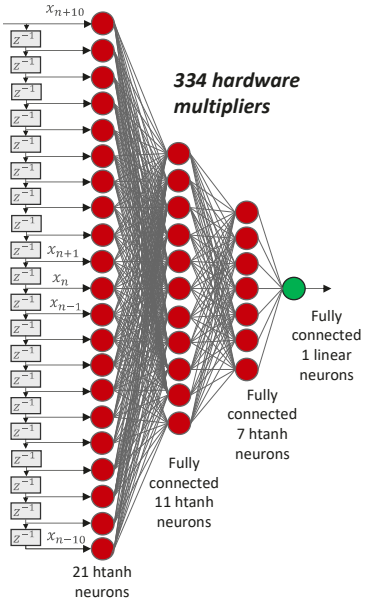

Fig. 2: A preferred feed forward NN nonlinear equalizer configuration

numerous iterations (epochs) about 50k to 70k until convergence for initial training with *tanh* activation and another 5k to 10k to retrain with *H-tanh* activation.

With d defining the number of layers including input and output layer and $s = s_1|s_2|\ldots|s_d$ describing the NN design with s_i neurons in the i-th layer [10]. The total required number of multipliers for a NN-NLE with *H-tanh* activation function is defined as:

$$mul_{NN-NLE} = \sum_{i=1}^{d-1} s_i s_{i+1} + \sum_{i=2}^{d-1} s_i \quad (3)$$

The first sum relates to the number of weights and the second sum to the number of *H-tanh* activations.

C. Duobinary Training Target

In order to recover transmitted data sequences, classical equalizers are trained to model the inverse linear/nonlinear channel transfer function towards a flat frequency response of the channel and equalizer combined.

For better performance, we suggest instead to train the equalizer towards a spectral output, which is similar to the frequency response of the optical channel, i.e. with attenuated high frequency components. This is achieved, by training the equalizers not towards estimating the transmitted data sequences directly, but its DB version, which has an overall narrower signal BW [15].

The advantage within the equalization process is less noise enhancement of the high frequency components, which suffer the most from channel-induced BW limitations and CD, i.e. those frequency components with lowest signal to noise ratio.

Changing the equalizer training target from recovering the original transmitted sequences b_n to their DB counterpart c_n means in practice to apply a simple $1+D$ DB partial response filter, as described in [16], on the known training sequences used for data-aided training:

$$c_n = b_n + b_{n-1} \quad (4)$$

Being trained like this, forces the equalizer later during operation to output an estimate of $\hat{c}_n$ instead of an estimate of $\hat{b}_n$. This is advantageous for linear equalizers, V-NLE and NN-NLE equalizer.

In case of PAM4 transmission, four modulation levels {0, 1, 2, 3} become seven levels {0, 1, 2, 3, 4, 5, 6} for the training target c_n and the equalizer output $\hat{c}_n$. Without loss of generality, an offset or scaling on the modulation levels is not

TABLE I: EXAMPLE SEQUENCES OF PAM4 TRANSMISSION WITH DUOBINARY-TARGET EQUALIZATION (EQ). MODULO 4 RECOVERY REQUIRES A TX PRECODING COUNTERPART

Original sequence	a_n		3	2	0	1	0	2	1	1	3
Tx Precoding	$b_n = (a_n - b_{n-1}) \bmod 4$	*0*	3	3	1	0	0	2	3	2	1
- Optical Transmission -											
DB-target EQ hard output	$\hat{c}_n = \hat{b}_n + \hat{b}_{n-1}$		3	6	4	1	0	2	5	5	3
Modulo 4 recovery	$\tilde{a}_n = (\hat{c}_n) \bmod 4$		3	2	0	1	0	2	1	1	3

considered here.

The simplest way to recover PAM4 symbols from the estimated seven-level equalizer output is modulo 4 reduction (Mod4) [17]. Mathematically, it makes no difference if hard symbol decision is taken before or after this Mod4 operation. For hardware implementations, hard symbol decision makes more sense before Mod4 recovery, i.e. on the seven-level equalizer output. This way, when using the two's-complement representation of the words in the digital circuit, Mod4 can be implemented without additional hardware costs, by truncating all bit levels but the two least significant bits. Table I gives an example of PAM4 transmission sequences with duobinary-target equalization and Mod4 recovery.

A more complex, but more effective alternative to Mod4 is classical MLSE, which recovers the enforced ISI with theoretically best possible performance.

In case of Mod4, a differential DB precoder is required on the transmitter side to generate a transmittable sequence b_n from the original PAM4 modulated sequence a_n as follows:

$$b_n = (a_n - b_{n-1}) \; mod4 \qquad (5)$$

For MLSE recovery, this DB precoder is optional. For the examples presented here, precoding was applied for both types of recovery. More about DB training is outlined in [5].

III. 800GB/S MEASUREMENTS

This section introduces details of our measurement setup and measurement results of V-NLEs and NN-NLEs.

A. IM/DD Measurement Setup

Fig. 3 depicts the IM/DD measurement setup and off-line DSP of our 10 km PAM4 transmission with 112 GBd per lane. 3 dB BWs are mentioned below each electro/optical component.

At the Tx, pseudorandom binary sequences (PRBS) are Gray-mapped to PAM4 symbols. In case of choosing Mod4 recovery over MLSE recovery at the Rx side, precoding is the next Tx step, before pulse-shaping the sequences with a raised cosine filter with a roll off factor of 0.14. While this roll-off factor is an optimized compromise between timing jitter robustness and spectral efficiency, we do not observe major end-to-end performance fluctuations with decent variations. After resampling, the sequences match the 120 GS/s arbitrary waveform generator (AWG), which converts to analogue signals. A 60 GHz driver amplifier (DA) amplifies towards O-band Mach Zehnder modulation (MZM). While the focus is on standard O-band CWDM4 wavelengths 1270nm, 1290m, 1310nm and 1330nm [18], illustrated with their carriers in Fig. 3 (a), further captures at in-between wavelengths allow for better insights of the performance/wavelength relationship.

After 10 km transmission over SSMF, a variable optical attenuator (VOA) controls the received optical power (ROP) at the input of a Praseodymium-doped fiber amplifier (PDFA). An optical filter suppresses the broadband noise of the PDFA. Fig. 3 (b) shows good noise suppression without cutting the optical signal spectrum.

The filtered optical signal is fed to a photo diode (PD) and its electrical output is digitized at 256 GS/s by a real time digital oscilloscope. The ROPs at the PDFA input were tuned to yield approximately 7 dBm optical power at the PD for optimal performance.

The offline Rx DSP starts with resampling and timing recovery. Equalization is done under one sample per symbol signalling. Fig. 3 (c) shows histograms of the received PRBS data before and after DB-targeted equalization. MLSE with Euclidian distance metric and complexity-optimized memory length of only 1 or Mod4 DB recovery is applied to the equalizer outputs before BERs are counted.

For the results presented next, V-NLE is compared against NN-NLE and Mod4 recovery is compared against MLSE recovery upon identical equalizer output data. This means, the same data-aided training (see sec. III A and III B) has been applied to the V-NLE or NN-NLE equalizers, independent of the choice of subsequent symbol recovery. In order to allow not only for MLSE recovery, but also for Mod4 recovery, we applied in both cases precoding on the Tx side. Although precoding is only a requirement for Mod4 recovery, also MLSE performance can benefit from Tx precoding.

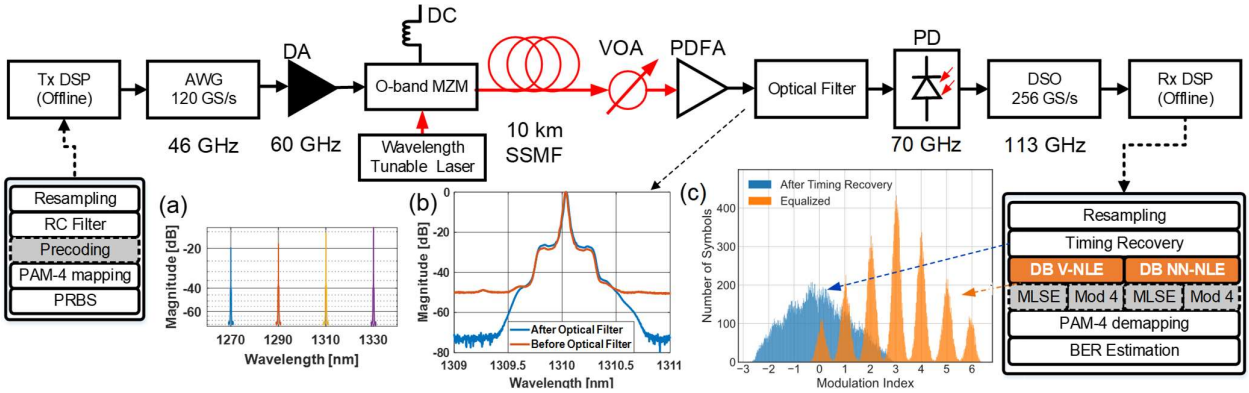

Fig. 3: Experimental 10km setup with DSP. Inset (a) shows classical unmodulated CWDM4 O-band carrier wavelengths, (b) optical spectra before and after optical filtering and (c) DB-targeted PAM4 equalization input and output with its seven modulation levels

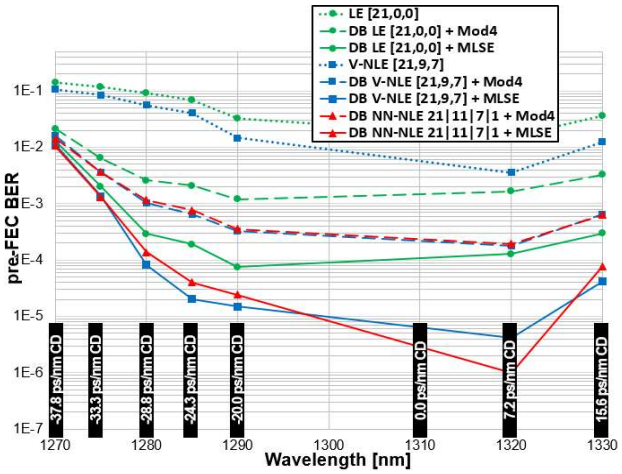

Fig. 4: Linear equalization (LE) vs. V-NLE vs. NN-NLE with and without DB target, with and without MLSE

B. Neural Network vs. Volterra Equalization Results

Fig. 4 presents pre-FEC BERs vs. wavelengths for three DSP configurations, linear equalization (LE) in green colour, V-NLE with [1st, 2nd, 3rd] order memory taps in blue and NN-NLE in red. Accumulated CD values for 10 km, as defined in [18], are added to the x-axis.

Dotted lines represent classical training for LE and V-NLE, dashed lines DB training with Mod4 and solid lines with added MLSE. Massive DB gains are obvious.

The NN-NLE architecture 21|11|7|1, as depicted in Fig. 2 matches the pre-FEC BER performance of V-NLE [21/9/7], but requires only 334 hardware multipliers instead of 486 for V-NLE, thus >30% less.

For clarity, Fig. 4 shows only a small subset of vast sweep studies regarding memory taps for LE and V-NLE and number of hidden layers and neurons per layer for NN-NLEs. As with other NN disciplines, NNs as NLEs have the additional challenge of a vast parameter space, which is hard to manage. Nonetheless, all plots of Fig. 4 are results of careful optimizations of performance vs. complexity, meaning more complex variants improve results only insignificantly.

C. Neural Network Weight Pruning

For further complexity reduction of NN-NLEs, an alternative to cutting entire neurons or even layers is pruning of individual weights $w_1,\ldots,w_K$ (see Fig.1), by gradually zeroing them out during the training process. This achieves NN sparsity and reduces multiplier counts, while keeping the NN architecture the same.

We adopt magnitude-based weight pruning, which is a simple but effective way to identify in NNs those weights which contribute least to the predictions. Other selection and regularization techniques exist [19], but could not proof higher effectiveness in our use case.

Compared to fully connected NN-NLEs (see Fig. 2), a slightly higher number of input memory taps prove beneficial as an architectural basis for weight pruning. Fig. 5 shows pruning results on the basis of DB NN-NLE configuration 23|11|7|1. Results were obtained based on initial training with *tanh* and then *H-tanh* as before, plus additional 300 to 500 epochs of gradually increasing magnitude threshold to prune weights up to a pre-defined percentage. In our NN

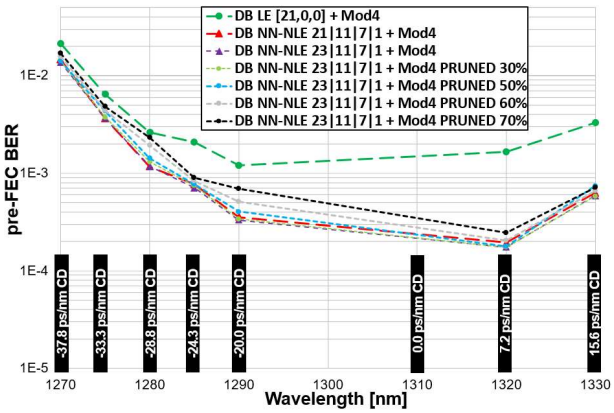

Fig. 5: Magnitude-based weight pruning on DB NN-NLE + Mod4

architecture, there are only 7 connections towards the output layer. Pruning any of them has shown significant performance drops. This is why pruning was limited to connections from input to first hidden layer and from first to second hidden layer. Thus, the percentage numbers 30%, 50%, 60% and 70% translate into NN-NLEs with 257, 191, 158 and 127 hardware multipliers, respectively. Fig. 5 shows that the configuration 23|11|7|1 with 191 hardware multipliers (equivalent to 50% pruned in the figure) can still match the performance of the fully connected baseline configuration 21|11|7|1, which would require 334 hardware multipliers. This equals a reduction by 43%, or, compared to the previously analysed V-NLE configuration with 486 hardware multipliers, a reduction by >60%. The comparison against V-NLE, however, should be taken with care, as V-NLEs can also allow for kernel pruning for complexity reduction. This has been shown several times in the past [20][21] and will be out of scope for this article.

IV. Area Efficient Adjustability

Often, equalizers require adjustability to different scenarios, which may vary in their source of (non-)linearity. For our DCI/DCN use case, CD is a quasi-static distortion. Thus, NLEs can be pre-trained before deployment in an ASIC. Further, also the typical source for channel dynamics, namely temperature variations, can be neglected for DCI/DCNs with their well-controlled stable environmental temperatures. Yet, channel characteristics might still change, for example in case of path rerouting in switched networks. Here, channel dynamics are limited to few configurations, which are well-known in advance. This makes pre-training on multiple configurations an adequate solution for later switching of NN parameters in the field. Conventionally, this means to train parameter sets of weights *plus* biases and multiplex between them. Instead, our suggestion for lowest hardware complexity is to switch only between sets of biases and share a common set of fixed weights. Both, the classical and the suggested parameter switching concept are illustrated in Fig. 6.

To define such a commonly shared set of weights, a coefficient training process, called multi-task learning (MTL) [22], can be applied to train a single NN equalizer on datasets representing multiple scenarios jointly [5].

Keeping fixed weights for all scenario options means the related multipliers are also fixed, which then allows for

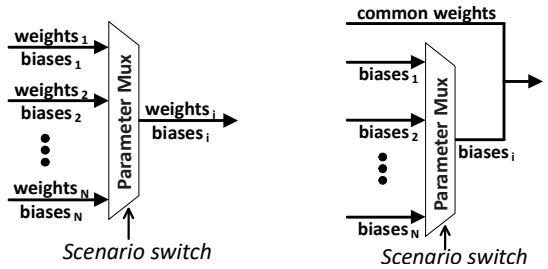

Fig. 6: Conventional NN parameter switching (left) and bias-only switching for lower complexity (right)

hardware simplifications without losing NN adjustability. Fig. 7 illustrates the ASIC area ratio reduction of simplified vs. full multipliers after logic synthesis for all possible two's complement fixed point representations for an 8-bit example in the range [-128,127]. For these results, we apply Synopsis' Design Compiler with a representative 45nm process technology. For today's IM/DD transceivers, other technologies with smaller nodes down to 7nm or 5nm would be preferred, but relative area savings are comparable. Simplified multipliers can yield an average reduction of about 76.2% across all 8-bit number representations.

The reason for area reduction is that multiplier factors of 2^n allow in two's complement fixed point number representations to shift one place to the left for each n^{th} power of *2* multiplication. This means, fixed multiplier factors of 2^n do not increase hardware complexity. Further, this can also simplify fixed multipliers of numbers other than 2^n, by applying distributive multiplier properties.

For performance comparison and in order to mimic adjustability requirements to different channel scenarios, we choose to train with MTL on all wavelengths introduced in sec. III: 1270nm, 1275nm, 1280nm, 1285nm, 1290nm, 1320nm and 1330nm. As shown in Fig. 8, performance matches between separately trained DB NN-NLE equalizers (red line) and MTL DB NN-NLEs with shared weights for all wavelengths (orange line). Further details of MTL training for low complex adjustability are outlined in [5].

V. Hardware Parallelization

The potential of NN-NLEs to outperform classical equalizers in terms of complexity and/or performance has been proven many times. However, with the efforts to make NN-NLEs a general tool for future high-speed transceiver implementations, the importance and necessity of hardware parallelization is often overseen or sometimes only considered as a further tool for hardware complexity reduction. In fact, for high-speed ASIC designs for coherent or IM/DD transceivers, hardware parallelization is unavoidable. A typical operating clock could be around 1GHz or less for today's transceivers. In order to support the much higher baud rate/ data rate, parallel processing is required. For example,

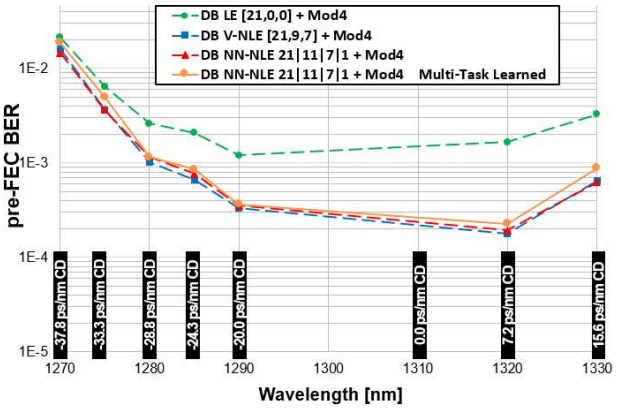

Fig. 8: DB NN-NLEs trained separately (red line) and with MTL for shared weights (orange line); reference: DB V-NLE, DB LE

112GBd with 1 sample per symbol processing would require 112 parallel hardware pipelines with a 1GHz clock.

With classical linear or nonlinear equalizers, the traditional solution is unconnected parallelization, meaning 112 equalizers would be required in our example.

[23] describes instead a mix of multi-symbol output (two outputs) and parallelization (4x20 times) for a 50Gb/s real time FPGA implementation with a 325 MHz clock.

A key question is, if NN-NLEs, which process and output *k* consecutive symbols at a time can be less complex than the equivalent of *k* parallel NN-NLEs. With the example *k=112*, we compare as our best-performing low complexity fully parallel reference the 50% pruned configuration 23|11|7|1 of Fig 5. vs. a configuration with 112 multi-symbol outputs. To define such a multi-symbol architecture, we choose a very large configuration 151|1232|784|112 and prune it as described in section III C with 1500 to 3000 epochs down to 99% sparsity. The remaining 1% translates into 12418 required hardware multipliers for the multi-output design. This corresponds to complexity saving of 42% hardware multipliers versus the fully parallel solution with its 112 x 191 = 21392 hardware multipliers. Fig. 9 plots mean BER performances, averaged over all 112 multi-symbols (purple line) vs. the slightly worse performing fully parallel reference (blue line). This shows that multi-symbol feed forward architectures have massive complexity advantages compared to parallelized feed forward architectures.

While recurrent neural network (RNN) NLE architectures often outperform feed forward NN-NLE architectures in optical transmission use cases [24-26], realizing numerous multi-symbol outputs seems to be an unsolved problem [27]. In [28], RNN equalizers based on gated recurrent units (GRU-RNN) for PAM4 IM/DD can support up to 8 multi symbol outputs. In [29], the authors compare different RNN variants for 212Gb/s PAM4 IM/DD transmission over 1km with up to 6 multi symbol outputs. Finally, in [30], the authors propose

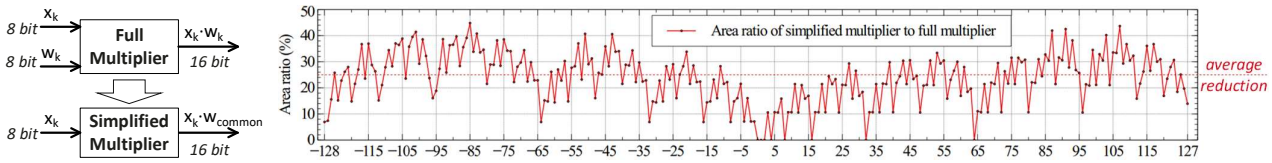

Fig. 7: *Area ratio reduction simplified vs. full multipliers after logic synthesis with the tool Synopsys Design Compiler for all possible 8-bit two's complement fixed point numbers; the average simplification is down to 23.8% compared to a full multiplier*

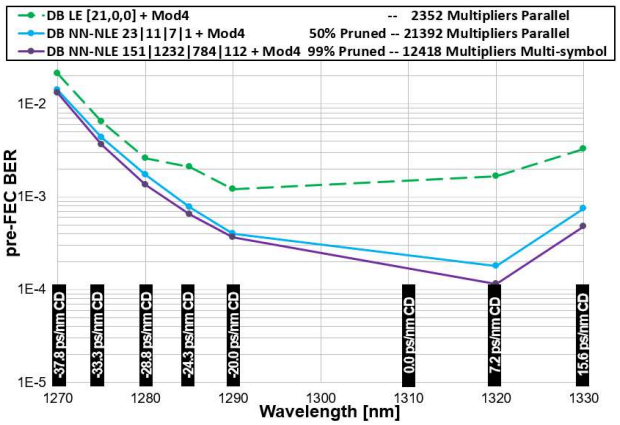

Fig. 9: 112 symbol outputs realized by 50% pruned fully parallel architecture 23|11|7|1 and 99% pruned multi-symbol output architecture 151|1232|784|112; reference: DB LE

to solve the RNN parallelization problem by avoiding recurrent structures entirely and using feed forward architectures instead. They merely apply RNN references for training assistance. Feed-forward NN-NLE architectures provide better scalability, also shown in [12] with up to 9 multi-symbol outputs.

VI. Conclusions

This paper introduces neural network nonlinear equalizer (NN-NLE) options for 10km IM/DD CWDM4 transceiver implementations with 200Gb/s per lane, where CD is the main mitigation target. We compare against Volterra nonlinear equalizers (V-NLE) in terms of performance and implementation complexity, with and without additional MLSE processing.

The concept of duobinary (DB) training target against noise enhancements during the equalization process improves the performance of both V-NLE and NN-NLE significantly.

DB NN-NLEs match the pre-FEC BER performance of 3rd order DB V-NLEs with more than 30% complexity reduction in hardware multipliers. Magnitude weight pruning allows for an additional 43% complexity reduction without performance loss across all CWDM4 lanes.

A NN-NLE solution for area-efficient adjustability to dynamic channel scenarios has been presented in form of adapting only biases and not adapting biases plus weights to account for changes. Post-synthesis VLSI comparisons of a reference implementation show massive area reductions of 76.2% for simplified hardware multipliers.

Another implementation specific necessity, hardware parallelization, has been discussed. In an example for 112 parallel outputs, one multi-output feed forward architecture shows 42% less hardware multiplier complexity compared to 112 parallel single-output feed forward architectures.

References

[1] A. Gholami, et al., “AI and Memory Wall” RiseLab Medium Blog Post, University of Califonia Berkeley, 2021

[2] T. Ninomiya, et al., "Advancement in optical interconnect technology for high speed data transmission," SPIE 12007, Optical Interconnects XXII, 120070R, 2022

[3] Liu, H., Zhou, X., Lam, C. “Coherent or non-coherent: what’s the story?”, *OFC,* 2022

[4] X. He, “FEC Architecture of B400GbE to Support BER Objective” IEEE 802.3 Beyond 400G Study Group, 2021

[5] C. Bluemm et al., "800Gb/s PAM4 Transmission over 10km SSMF Enabled by Low-complex Duobinary Neural Network Equalization," *ECOC*, 2022

[6] B. Liu et al., “Area-Efficient Neural Network CD Equalizer for 4×200Gb/s PAM4 CWDM4 Systems,” *OFC*, 2023

[7] M. Chagnon, "Direct-Detection Technologies for Intra- and Inter-Data Center Optical Links," *OFC*, 2019

[8] L. Guan, “FPGA-Based Digital Convolution for Wireless Applications”. Berlin, Germany: Springer, 2017

[9] Kim, J. and Konstantinou, K. “Digital predistortion of wideband signals based on power amplifier model with memory”, Electronics Letters, 37, (23), 2001

[10] M. Schädler et al. "Soft-Demapping for Short Reach Optical Communication: A Comparison of Deep Neural Networks and Volterra Series," *Journal of Lightwave Technology*, vol. 39, no. 10, 2021

[11] L. Ge, et al., "Compressed Neural Network Equalization Based on Iterative Pruning Algorithm for 112-Gbps VCSEL-Enabled Optical Interconnects," *Journal of Lightwave Technology*, vol. 38, no. 6, 2020

[12] Xu, Z., et al., "Computationally-Efficient Sparsely-Connected Multi-Output Neural Networks for IM/DD System Equalization", 2022

[13] J. Xiao,et al., "Fastly converged transfer learning using neuron-pruning nonlinear equalizer for intra data center networking," OECC, 2022

[14] D. P. Kingma and L. J. Ba, ”Adam: A Method for Stochastic Optimization,” *ICLR*, 2015

[15] Gutiérrez-Castrejón, et.al. “Systematic Performance Comparison of (Duobinary)-PAM-2,4 Signaling under Light and Strong Opto-Electronic Bandwidth Conditions”, *MDPI Photonics*, 8, 81, 2021

[16] T. Wettlin, et al., "Beyond 200 Gb/s PAM4 transmission using Tomlinson-Harashima precoding," *ECOC*, 2019

[17] M. S. -B. Hossain, et al., "Transmission Beyond 200 Gbit/s With IM/DD System for Campus and Intra-Datacenter Network Applications," *IEEE Photonics Technology Letters*, vol. 33, no. 5, 2021

[18] ITU-T Rec. G.652 “Characteristics of a single-mode optical fibre and cable”, Series G: Transmission Systems and Media, Digital Systems and Networks – Optical fibre cables, 2016

[19] S. Han, H. Mao and W. J. Dally, ”Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” *ICLR*, 2016

[20] W. J. Huang *et al.*, ”93 % complexity reduction of Volterra nonlinear equalizer by L 1-regularization for 112-Gbps PAM-4 850-nm VCSEL optical interconnect,” *OFC*, 2018

[21] C. Chuang *et al.*, ”Sparse Volterra Nonlinear Equalizer by Employing Pruning Algorithm for High-Speed PAM-4 850-nm VCSEL Optical Interconnect,” *OFC*, 2019

[22] S. Vandenhende, et al., “Multi-task Learning for Dense Prediction Tasks: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, 2022

[23] N. Kaneda, et al., "Fixed-Point Analysis and FPGA Implementation of Deep Neural Network Based Equalizers for High-Speed PON" *Journal of Lightwave Technology*, vol. 40, no. 7, 2022

[24] Huang, L., et al., "Performance and Complexity Analysis of Conventional and Deep Learning Equalizers for the High-Speed IMDD PON," Journal of Lightwave Technology, vol. 40, no. 14, 2022

[25] Xu, Z, et al. "Feedforward and Recurrent Neural Network-Based Transfer Learning for Nonlinear Equalization in Short-Reach Optical Links," Journal of Lightwave Technology, vol. 39, no. 2, 2021

[26] M. Schädler, et al., "Recurrent Neural Network Soft-Demapping for Nonlinear ISI in 800Gbit/s DWDM Coherent Optical Transmissions," *Journal of Lightwave Technology*, 39, 2021

[27] P. J. Freire, et al., "Reducing Computational Complexity of Neural Networks in Optical Channel Equalization: From Concepts to Implementation," *Journal of Lightwave Technology*, 2023

[28] S. L. Murphy, et al., "High Dynamic Range 100G PON Enabled by SOA Preamplifier and Recurrent Neural Networks," Journal of Lightwave Technology, 2023

[29] B. Sang, et al, "Low Complexity Neural Network Equalization Based on Multi-Symbol Output Technique for 200+ Gbps IM/DD Short Reach Optical System," *Journal of Lightwave Technology*, vol. 40, 2022

[30] S. Srivallapanondh, et al. "Knowledge Distillation Applied to Optical Channel Equalization: Solving the Parallelization Problem of Recurrent Connection." *OFC*, 2023