Energy-Efficient Implementation of Probabilistic Shaping

Tsuyoshi Yoshida(1), Magnus Karlsson(2), and Erik Agrell(2)

(1) Mitsubishi Electric Corporation, Kamakura, Japan, Yoshida.Tsuyoshi@ah.MisubishiElectric.co.jp
(2) Chalmers University of Technology, Gothenburg, Sweden

Abstract *We propose a modified low-complexity architecture for hierarchical distribution matching (DM), achieving 2^{22}-ary quadrature amplitude modulation or 4.32-Tb/s data rate on a field programmable gate array with a limited rate loss. Such a performance–complexity balanced DM helps minimizing the power consumption with probabilistic shaping. ©2023 The Author(s)*

Introduction

Energy-efficient communications are required for achieving green sustainable societies. In the optical fiber communication field, the typical solution is increased throughput per optical wavelength channel, i.e., higher and highly granular baudrate and information rate (IR). Here, we usually employ polarization-division multiplexed in-phase and quadrature modulation with coherent detection and digital signal processing (DSP) [1]. The latest standard utilizes soft-decision (SD) forward error correction (FEC) and flexible quadrature amplitude modulation (QAM) [2,3]. More flexibility in IR and a better performance are realized by probabilistic constellation shaping (PCS) [4,5].

For practical reasons, the PCS coding is usually placed outside the FEC coding. PCS coding is called distribution matching (DM), because it explicitly or implicitly matches the probability distribution of channel-input symbols to a desired distribution. Several DM techniques have been proposed, such as constant-composition DM [6], enumerative sphere shaping [7], hierarchical DM (HiDM) [8], and prefix-free code DM [9]. Among these, only HiDM and prefix-free code DM have been implemented on a field-programmable gate array (FPGA) [10–15].

To enhance the energy efficiency, we require not only low complexity but also high performance, because the throughput (i.e., net data rate) becomes limited at a given optical wavelength when the performance of the DSP and analog devices is poor. The redundancy at a given required signal-to-noise ratio (SNR) with low-performance PCS/FEC is larger than that with a high-performance one. This higher redundancy results in higher baudrate and higher power consumption.

The original contribution of this work consists of clarifying the dependence of the energy efficiency on the performance–complexity tradeoff of PCS, proposing a low-complexity hierarchical DM, and implementing the proposed DM at record-high data rates up to 4.32 Tb/s.

Energy efficiency in systems with PCS

This section explains the energy efficiency dependence on the PCS implementation. First, we quantify the baudrate increase in systems with PCS. We assume fixed net data rate and required optical SNR (ROSNR). The target channel-input entropy is 3.8 bits per two-dimensional (2d) symbol without the PCS implementation penalty (i.e., at zero rate-loss). There are various realizations of the performance and complexity combinations. The performance is given by the required SNR (RSNR) penalty in the PCS implementation, α. The DM complexity is quantified by additional power consumption of optical transceiver functions, β, compared with a system without PCS at a given baudrate. For example, a high-performance DM has the combination (α, β) = (0.1 dB, 10%), while a low-complexity DM could have (0.5 dB, 1%). For a large α, to satisfy the net data rate and ROSNR, the PCS redundancy must be large, which requires a higher baudrate. Note that the performance improvement by the PCS should cope with the noise bandwidth increase caused by the baudrate increase (in other words, PCS must have a net coding gain larger than 0 dB).

Fig. 1 shows an example of the baudrate increase r as a function of RSNR penalty in the PCS implementation. There are three cases: a) fixed RSNR, b) fixed ROSNR without transceiver (TRx) noise, and c) fixed ROSNR with TRx noise (–20 dB at α = 0 dB and proportional to the baudrate). The result for case b) is practically more relevant than a), because the system performance is usually characterized by OSNR rather than SNR. While case c) is more specific and practical than b), the baudrate difference between b) and c) is small.

Fig. 2 shows the additional power consumption with PCS, $(1+\beta)(1+r)-1$, at a given ROSNR without TRx noise as a function of α and β. Interestingly, the additional power consumption with DM having parameters of (α, β) = (0.2 dB, 5%) is lower than that of (0.6 dB, 0.1%). Thus, an energy-efficient PCS system

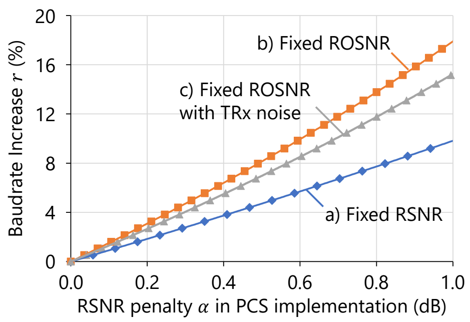

Fig. 1: Baudrate increase $\boldsymbol{r}$ as a function of RSNR penalty $\boldsymbol{\alpha}$ in PCS implementation under requirements a), b), and c).

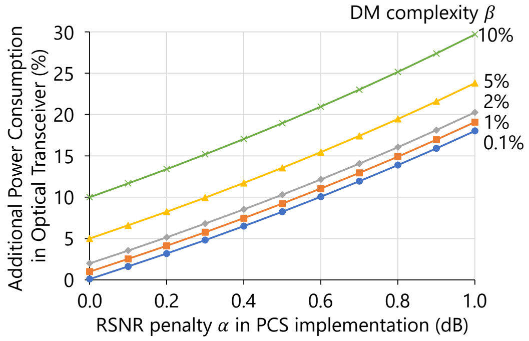

Fig. 2: Additional power consumption with PCS at a given ROSNR for various performance–complexity realizations.

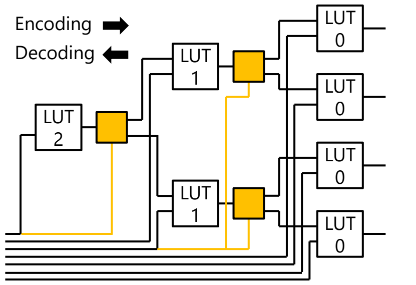

Fig. 3: HiDM modification: orange denotes the new function.

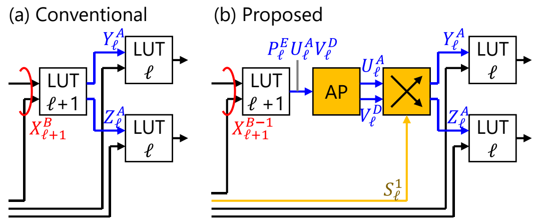

Fig. 4: Inter-layer connection in HiDM encoding.

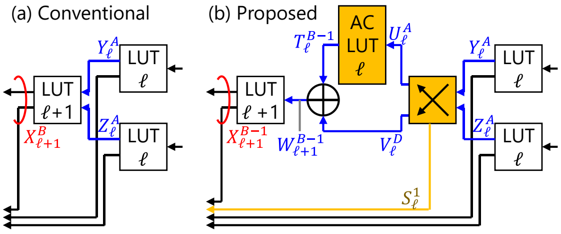

Fig. 5: Inter-layer connection in HiDM decoding.

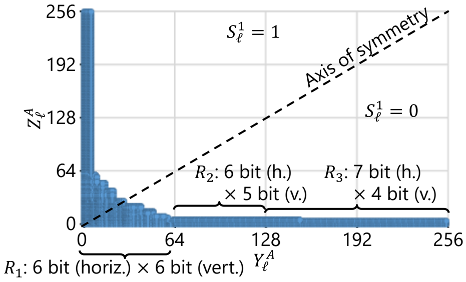

Fig. 6: Exemplified combinations of (Y_ℓ^A, Z_ℓ^A).

requires a high-performance DM with a moderate complexity.

Principle of lower-complexity HiDM

The PCS performance significantly influences the energy efficiency in entire systems as described above. HiDM, having tree-structured and layered look-up tables (LUTs), has been implemented on an FPGA several times [10,13–15] because of the good balance between performance and complexity. This section proposes a modified HiDM for reducing the complexity further. Fig. 3 shows the HiDM schematic, where the encoding process goes from left to right and the decoding in the opposite direction. In the modified HiDM, small functions indicated in orange are added between the layers to reduce the number of bits at each LUT interface.

Figs. 4 and 5 show (a) the conventional and (b) the proposed inter-layer connections in encoding and decoding, respectively. In the conventional encoding in Fig. 4(a), an LUT at layer $\ell+1$ converts the input symbol $X_{\ell+1}^B$ into output symbols Y_ℓ^A and Z_ℓ^A, which are fed to two LUTs at layer ℓ. Here the superscript A denotes the required number of bits for labelling the symbol. While there are totally 2^{2A} possible combinations of (Y_ℓ^A, Z_ℓ^A), the actual combinations are limited to 2^B, where B denotes the number of bits fed to the LUT at layer $\ell+1$ and $B < 2A$ to perform PCS. Fig. 6 shows an example of a (Y_ℓ^A, Z_ℓ^A) combination for $(A, B) = (8,12)$, where there are 65536 candidates (entire region) and 4096 used cases (blue markers).

Fig. 4(b) shows the proposed more efficient schematic with additional functions exploiting the features in Fig. 6. When either Y_ℓ^A or Z_ℓ^A is large, the other is small (see for example [8, Tab. III]). To accomplish this we introduce an adaptive partitioning (AP) of the (Y_ℓ^A, Z_ℓ^A) space into several regions, here denoted by R_1 to R_3. LUTs store the prefix P_ℓ^E for the selected region and $U_\ell^A V_\ell^D$ for the location inside the region. The total number of bits $C = E + A + D$ can be designed to be smaller than $2A$. To ensure the desired symmetry between Y_ℓ^A and Z_ℓ^A, indicated by the dashed line in Fig. 6, LUT $\ell+1$ encodes (U_ℓ^A, V_ℓ^A) as follows: (i) If $U_\ell^A \neq V_\ell^D$, then $Y_\ell^A = U_\ell^A$ and $Z_\ell^A = V_\ell^D$ for $S_\ell^1 = 0$, and $Y_\ell^A = V_\ell^D$ and $Z_\ell^A = U_\ell^A$ for $S_\ell^1 = 1$. (ii) If $U_\ell^A = V_\ell^D$, then $Y_\ell^A = Z_\ell^A = 2U_\ell^A + S_\ell^1$. Case (i) ensures that the top-left and

Tab. 1: Resource utilization of modified HiDM with compressed shaping on a single FPGA chip.

(a) hyper-scale QAM, 200 MHz on VCU128

Item	Used	Available
LUT as Logic	640377	1303680
Register	1004969	2607360
Block RAM	770	2016

(b) 4.32 Tb/s 16-QAM, 264 MHz on VCU118

Item	Used	Available
LUT as Logic	594328	1182240
Register	813476	2364480
Block RAM	2032	2160

bottom-right parts of Fig. 6 are encoded similarly, whereas case (ii) takes care of the diagonal, to which no symmetry applies. In the latter case, S_ℓ^1 serves as the least significant data bit.

Fig. 5(b) shows the modified decoding schematic. As the inter-layer operation, cases (i) and (ii) above are reversed to recover (U_ℓ^A, V_ℓ^D). The address conversion (AC) LUT and the addition follows to give a unique index $W_{\ell+1}^{B-1} = T_\ell^{B-1} + V_\ell^D \in \{0,1,\dots,2^{B-1}-1\}$, where the AC-LUT derives $T_\ell^{B-1} = O[U_\ell^A]$ with the offset function $O[k] = O[k-1] + count\,if(U_\ell^A = k-1)$ for integers $k \geq 0$ and $O[-1] = 0$.

The complexity of HiDM is mainly characterized by the memory size $\mathbb{S} = n_{\mathrm{out}} 2^{n_{\mathrm{in}}}$ in the LUTs, where n_{in} and n_{out} denote the number of input and output bits, respectively. At layer $\ell + 1$, $(n_{\mathrm{in}}, n_{\mathrm{out}})$ are $(B, 2A)$ and $(2A, B)$ for the conventional and $(B-1, C)$ and $(B-1, B-1) + (A, B-1)$ for the modified encoder and decoder, respectively. For example, the sum of $\mathbb{S}$ at layer $\ell + 1$ is reduced by 16 times (from 851968 to 51968) by the modified HiDM in case of $(A, B, C) = (8,12,13)$.

Example of implementation and experiment

We implemented the modified HiDM encoding and decoding on a single FPGA chip. The first example (a) supports various QAM formats including hyper-scale QAM up to 2^{22} constellation points and data rates up to 1.6 Tb/s on the Xilinx® Virtex® Ultrascale+™ VCU128. Example (a) consists of 8 sub-DMs, which each treats up to 768 shaped bits (for being combined with 256 unshaped sign-bits) at 200 MHz. The second example (b) is for the minimal constellation of binary amplitude-shift keying, achieving 4.32 Tb/s 16-QAM on the VCU118. Example (b) consists of 16 sub-DMs, which each treats up to 512 shaped bits (for being combined with 512 unshaped sign-bits) at 264 MHz. The resource utilization including logic elements ("LUT as Logic"), registers, and random access memory (RAM) and the utilized chip areas are summarized in Tab. 1 and Fig. 7, respectively.

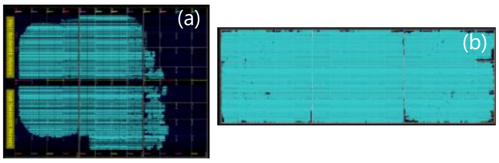

Fig. 7: Utilized areas of a single FPGA chip for (a) hyper-scale QAM and (b) 4.32 Tb/s 16-QAM.

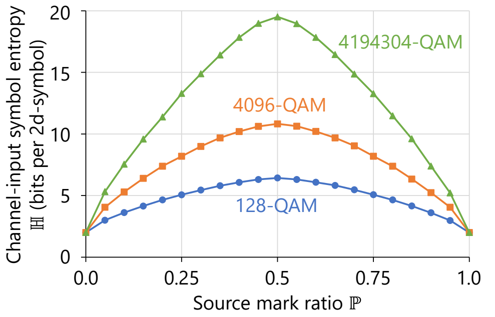

Fig. 8: FPGA evaluation results for the example (a).

We also implemented the additional feature of compressed shaping [16], which reduces the channel-input symbol entropy $\mathbb{H}$ by exploiting the source sparseness with a compression stopper to avoid excess transmission penalty [17].

Fig. 8 shows experimental results with the FPGA for example (a). The examined constellations are 128-, 4096-, and 2^{22}-QAM, having an IR $\mathbb{R}$ of 6.375, 10.75, and 19.5 bits per 2d symbol, respectively. We swept the probability of a '0'-bits fed to the encoder (i.e., the source mark ratio) $\mathbb{P}$ and observed $\mathbb{H}$. As in previous work for small-order QAM [16], $\mathbb{H}$ was reduced when $\mathbb{P}$ was different from 0.5, which showed the proper operation of the implemented FPGA. At $\mathbb{P}$=0.5 (without compressed shaping), the rate loss $\mathbb{H} - \mathbb{R}$ was 0.055, 0.074, and 0.075 bits per 2d symbol for 128-, 4096-, and 2^{22}-QAM, respectively, where the block length was 128, 64, and 32 2d symbols and the SNR penalty was around 0.2 dB. Regarding example (b), the status is that the FPGA implementation has been completed and the experimental evaluation has not yet, which is the next step. Based on the design, 16-QAM shows comparable rate loss with example (a).

Conclusions

HiDM was modified for low complexity by reducing the memory size to, e.g., 1/16. With the modified architecture a record-high QAM-order of 2^{22} or 4.32 Tb/s data rate was implemented on a single FPGA with an SNR penalty around 0.2 dB.

Acknowledgements

This work was in part supported by the commissioned research of National Institute of Information and Communications Technology (NICT), Japan, grant number 01401.

References

[1] K. Roberts, M. O'Sullivan, K.-T. Wu, H. Sun, A. Awadalla, D. J. Krause, and C. Laperle, "Performance of dual-polarization QPSK for optical transport systems," Journal of Lightwave Technology, vol. 27, no. 16, pp. 3546–3559, Aug. 2009, doi: 10.1109/JLT.2009.2022484.

[2] Open ROADM MSA, [Online]. Available: www.openroadm.org/home.html

[3] 400ZR, [Online]. Available: https://www.oiforum.com/wp-content/uploads/OIF-400ZR-02.0.pdf

[4] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," IEEE Transaction of Communications, vol. 63, no. 12, pp. 4651–4665, Dec. 2015, doi: 10.1109/TCOMM.2015.2494016.

[5] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: an experimental demonstration," Journal of Lightwave Technology, vol. 34, no. 7, pp. 1599–1609, Apr. 2016, doi: 10.1109/JLT.2015.2510034.

[6] P. Schulte and G. Böcherer, "Constant composition distribution matching," IEEE Transactions on Information Theory, vol. 62, no. 1, pp. 430–434, Jan. 2016, doi: 10.1109/TIT.2015.2499181.

[7] Y. C. Gültekin, F. M. J. Willems, W. J. van Houtum, S. Şerbetli, "Approximate enumerative sphere shaping," in Proc. IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, Jun. 2018, pp. 676–680, doi: 10.1109/ISIT.2018.8437623.

[8] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," Journal of Lightwave Technology, vol. 37, no. 6, pp. 1579–1589, Mar. 2019, doi: 10.1109/JLT.2019.2895065.

[9] J. Cho, “Prefix-free code distribution matching for probabilistic constellation shaping,” IEEE Transactions on Communications, vol. 68, no. 2, pp. 670–682, Feb. 2020, doi: 10.1109/TCOMM.2019.2924896.

[10] T. Yoshida et al., "FPGA implementation of distribution matching and dematching," in Proc. European Conference on Optical Communication (ECOC), Dublin, Ireland, Sept. 2019, Paper M.2.D.2.

[11] Q. Yu, S. Corteselli, and J. Cho, "FPGA implementation of prefix-free code distribution matching for probabilistic constellation shaping," in Proc. Optical Fiber Communication Conference and Exhibition (OFC), Mar. 2020, Paper Th1G.7, doi: 10.1364/OFC.2020.Th1G.7.

[12] Q. Yu, S. Corteselli, and J. Cho, "FPGA implementation of rate-adaptable prefix-free code distribution matching for probabilistic constellation shaping," Journal of Lightwave Technology, Feb. 2021, doi: 10.1109/JLT.2020.3035039.

[13] T. Yoshida, K. Igarashi, Y. Konishi, M. Karlsson, and E. Agrell, "FPGA implementation of hierarchical subcarrier rate and distribution matching for up to 1.032 Tb/s or 262144-QAM," in Proc. Optical Fiber Communication Conference and Exhibition (OFC), Paper Tu6D.4, doi: 10.1364/OFC.2021.Tu6D.4.

[14] L. Zhang, W. Wang, W. Qian, K. Tao and Y. Cai, "Real time FPGA investigation of probabilistic shaping 16QAM with HiDM and OFEC," Optical Fiber Communications Conference and Exhibition (OFC), San Francisco, CA, USA, Mar. 2021, Paper Tu6D.5, doi: 10.1364/OFC.2021.Tu6D.5.

[15] L. Zhang, K. Tao, W. Qian, W. Wang, J. Liang, Y. Cai, and Z. Feng, "Real-time FPGA investigation of interplay between probabilistic shaping and forward error correction," Journal of Lightwave Technology, vol. 40, no. 5, pp. 1339–1345, March 2022, doi: 10.1109/JLT.2021.3128490.

[16] T. Yoshida, K. Igarashi, M. Karlsson and E. Agrell, "Compressed shaping: Concept and FPGA demonstration," Journal of Lightwave Technology, vol. 39, no. 17, pp. 5412–5422, Sept. 2021, doi: 10.1109/JLT.2021.3085974.

[17] T. Yoshida, T. Inoue, K. Igarashi, M. Binkai, Y. Konishi, N. Suzuki, M. Karlsson, and E. Agrell, "Nonlinear fiber transmission of compressed shaping signals," 2022 European Conference on Optical Communication (ECOC), Basel, Switzerland, 2022, Paper Mo3D.6.