Tu2c2: Machine Learning II

Zero-Multiplier Sparse DNN Equalization for Fiber-Optic QAM Systems with Probabilistic Amplitude Shaping

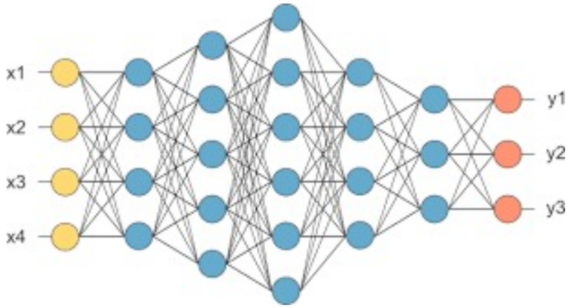
Toshiaki Koike-Akino, Ye Wang, Keisuke Kojima, Kieran Parsons, Tsuyoshi Yoshida

September 14, 2021

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

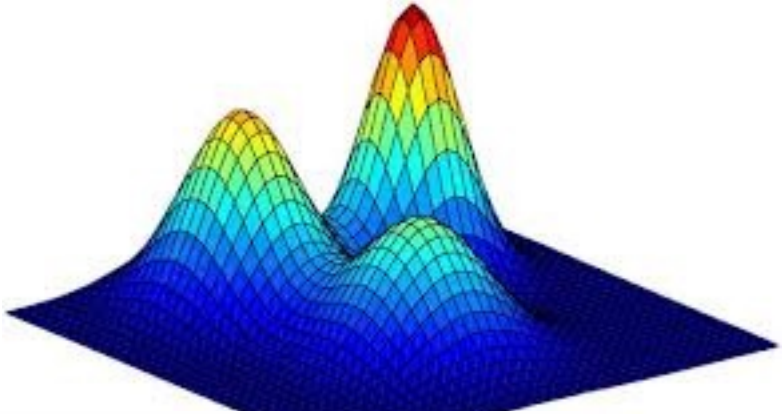
Outline

- Machine learning for optical communications
  - Research trend
  - Optics applications
  - Nonlinearity compensation
- Deep neural network (DNN) for shaped DP-QAM
  - From maximum-likelihood to machine learning
  - Multi-label binary cross-entropy loss
  - Architecture comparison
- Multiplier-less DNN
  - Additive powers-of-two quantization achieving floating-point performance
  - Multiply-accumulate to shift-accumulate
- Sparse DNN
  - Lottery-ticket hypothesis (LTH) pruning
  - **99% reduction** of arithmetic operations
- Summary

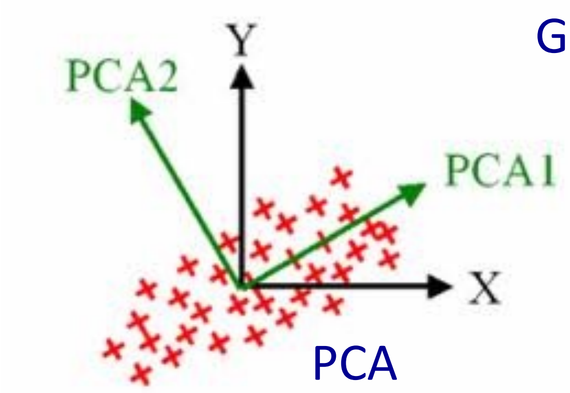


Machine Learning (ML)

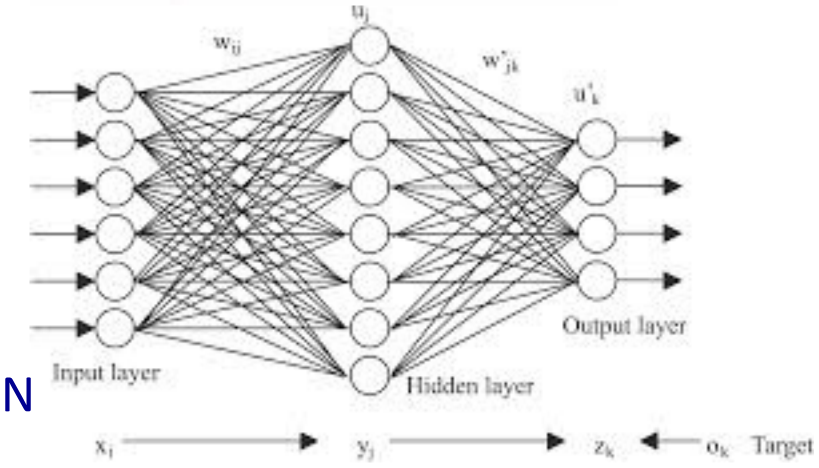
- K-means
- Gaussian mixture model (GMM)
- Principal component analysis (PCA)
- Independent component analysis (ICA)
- Support vector machine (SVM)
- Self-organizing map (SOM)
- Hidden Markov model (HMM)
- Artificial neural networks (ANN)
- **Deep learning (DL)**
- …



GMM

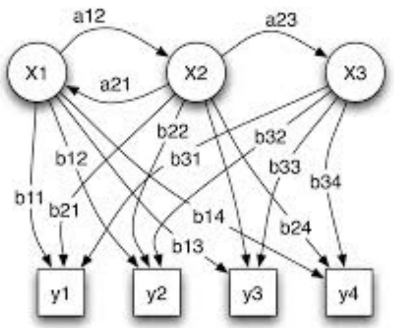


PCA



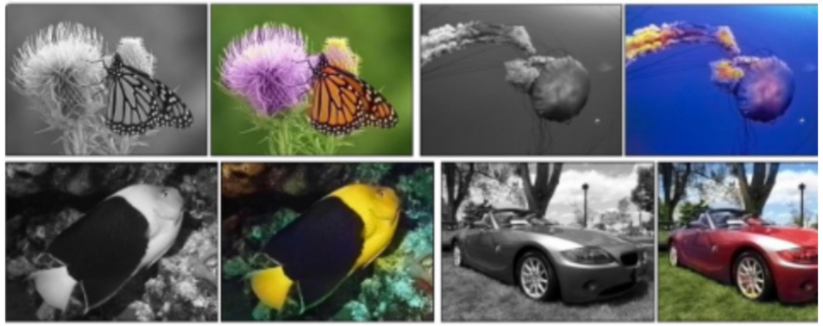
ANN

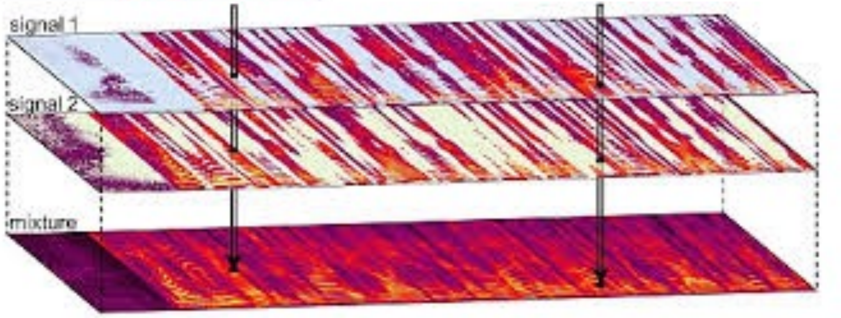
HMM

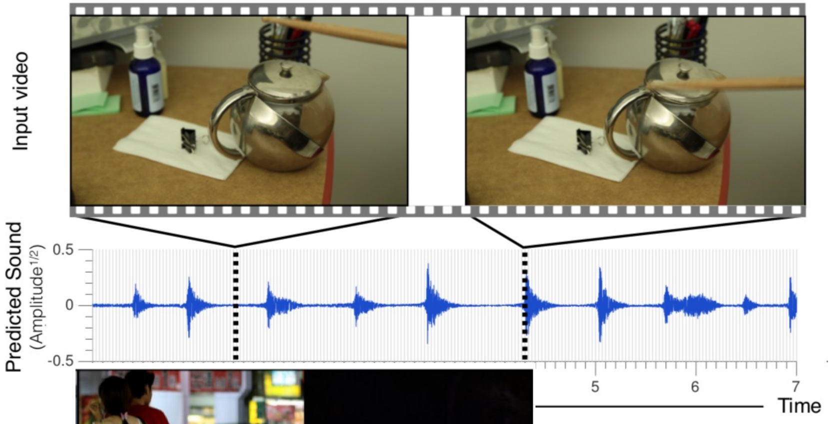


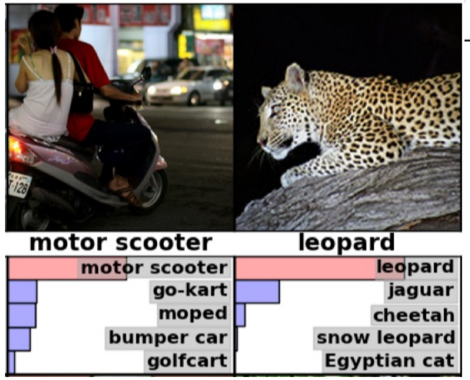
ML Success in Audio & Visual Signal Processing

- Denoising, segmentation, classification, translation, dialog, recognition, decomposition, generation, super-resolution, …





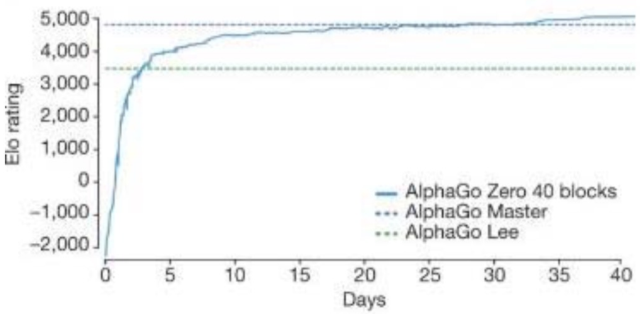


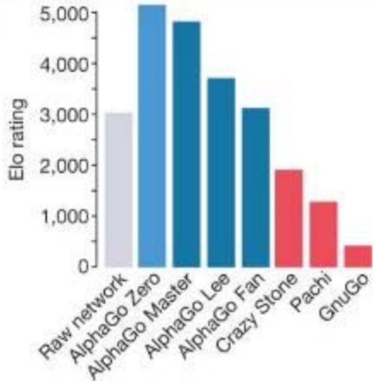


"man in black shirt is playing guitar."

ML Surpassed Human-Level Performance

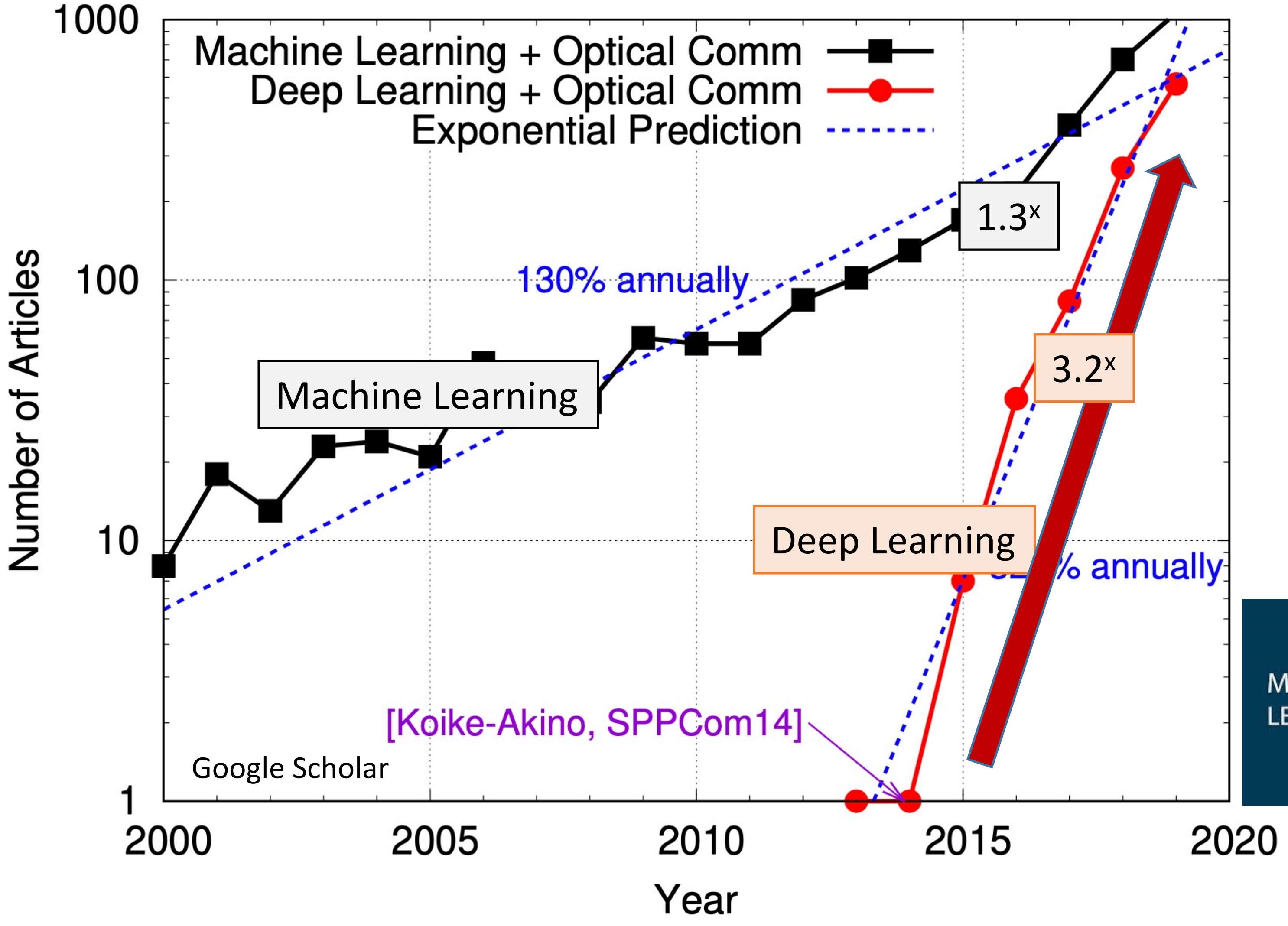
- For some applications, …

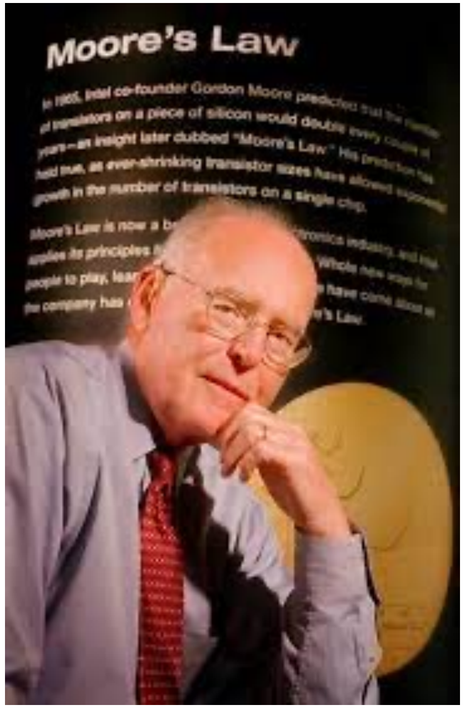


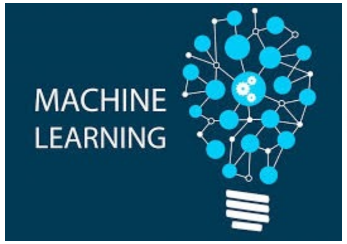


ML Meets Optical Communications

- New **Moore's Law** rediscovered here:
  Number of articles grows exponentially, nearly **tripling** every year



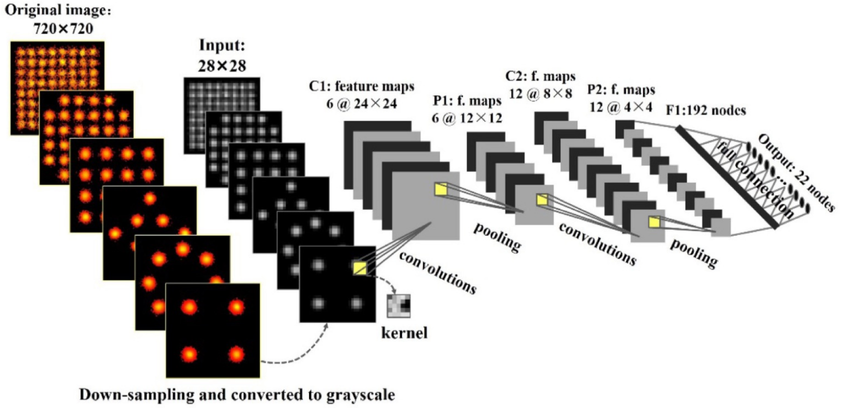


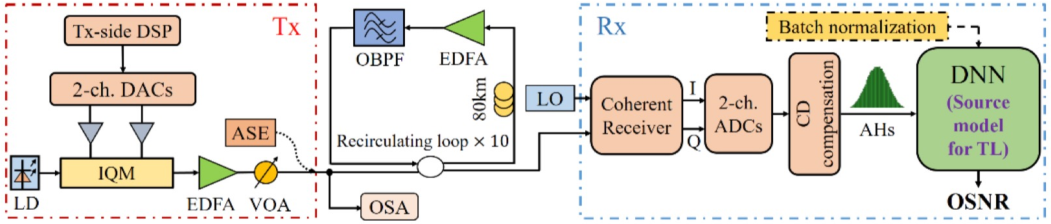


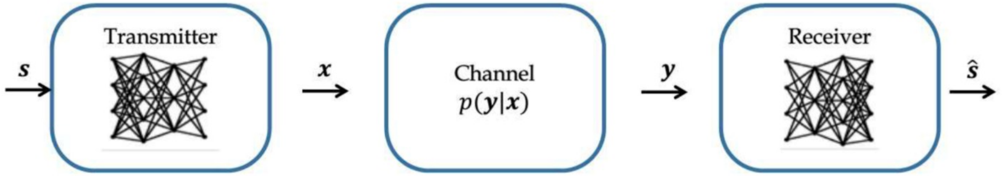
Deep Learning Applications for Optics

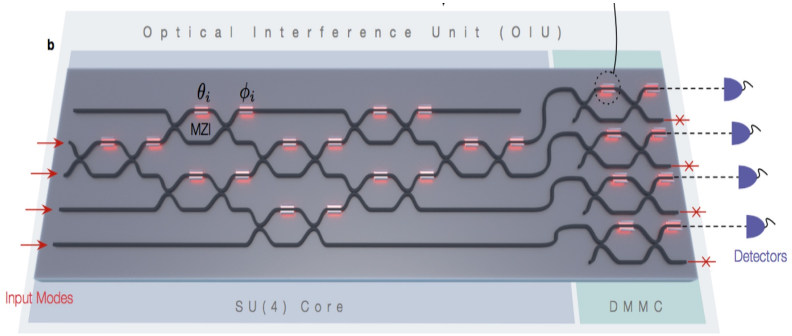
Already approx. 1000 related articles annually:

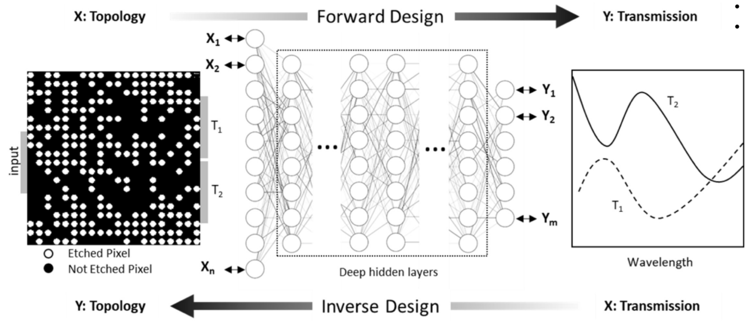
- Modulation classification
- Link quality monitoring
- Resource allocation
- Signal detection
- End-to-end design
- **Nonlinear compensation**
- Photonic circuit design
- Optical neural processor

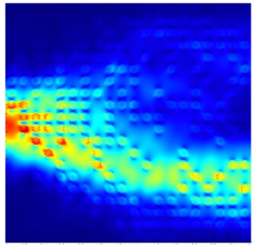












Why ML for Nonlinearity Compensation?

- Fiber channels are governed by **nonlinear physics in nature**
  - Self-phase modulation, cross-phase modulation, four-wave mixing, etc.
- Spectral efficiency can be improved by nonlinearity compensation
  - **Complicated model-based approaches** are required to capture real physics
- **Terabit-class massive data within a second** can be obtained
  - Deep learning: New **data-driven approach**. Suited for **massive parallel computing**

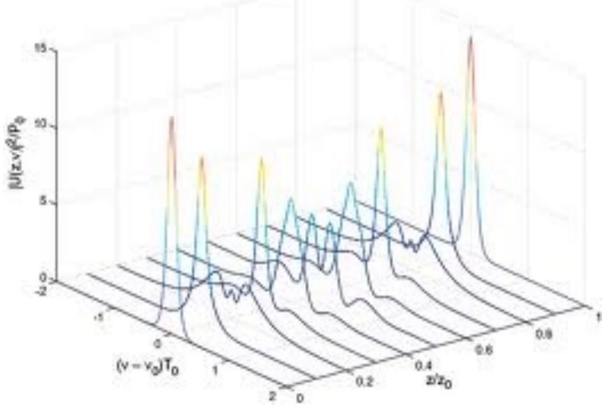
Nonlinear Schrodinger Equation:

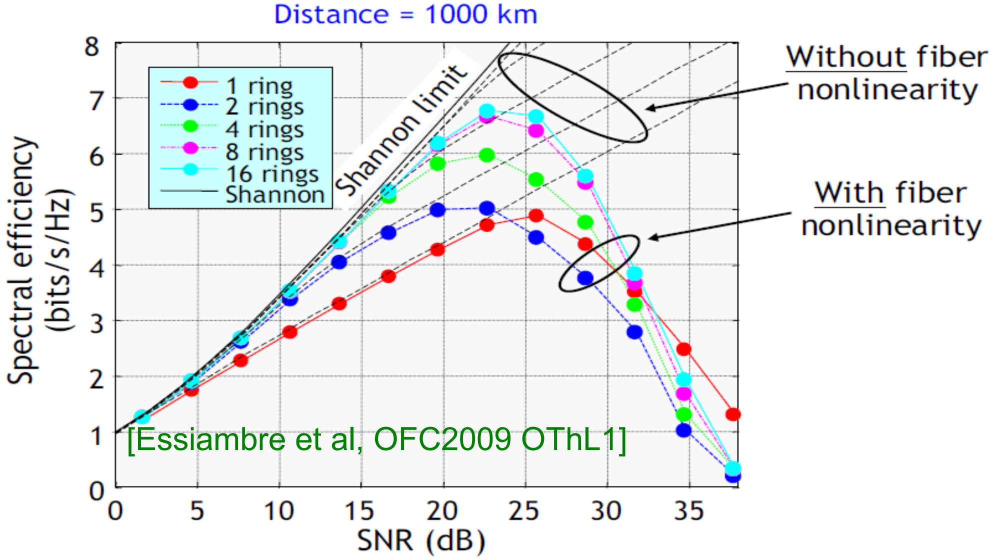
PMD CD (GVD) SPM/XPM

$$\frac{\partial \boldsymbol{E}}{\partial z} = \underbrace{\left(-\frac{1}{2}\boldsymbol{\alpha} - \boldsymbol{\beta}_1\frac{\partial}{\partial t} - \jmath\boldsymbol{\beta}_2\frac{1}{2!}\frac{\partial^2}{\partial t^2} + \frac{1}{3!}\boldsymbol{\beta}_3\frac{\partial^3}{\partial t^3}\right)}_{\hat{\boldsymbol{D}}}\boldsymbol{E} + \underbrace{\jmath\gamma\left(\|\boldsymbol{E}\|^2\boldsymbol{I} - \frac{1}{3}(\boldsymbol{E}^\dagger\boldsymbol{\sigma}_3\boldsymbol{E})\boldsymbol{\sigma}_3\right)}_{\hat{\boldsymbol{N}}}\boldsymbol{E}$$

CD slope XPolM

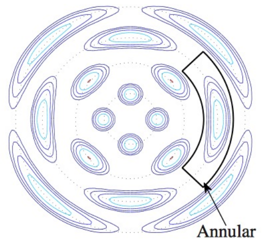
Nonlinear propagation

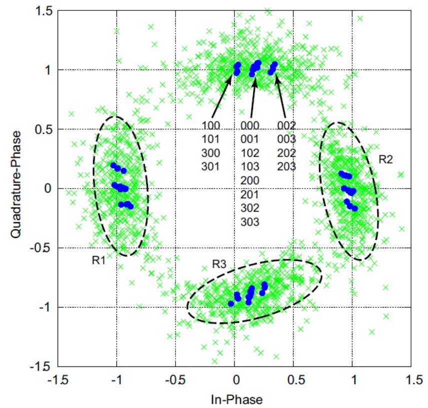


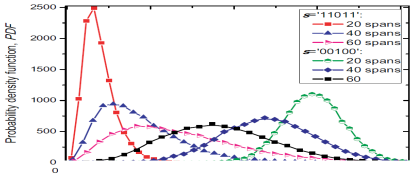


Nonlinear Equalization

- Nonlinear impairments may be compensated by *nonlinear equalization:*
  - Decision feedback equalizer (DFE)
  - Maximum-likelihood sequence equalizer (MLSE)
  - Volterra equalizer
  - Digital back-propagation (DBP)
  - Turbo equalizer (TEQ)
  - **Deep neural networks (DNN)**



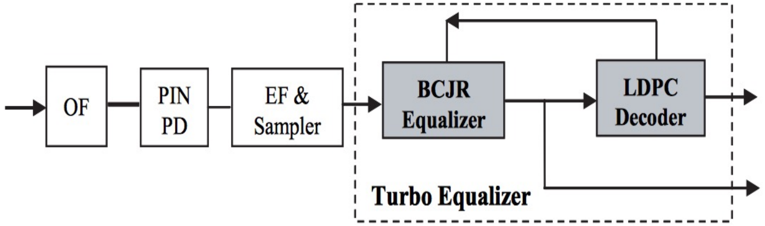




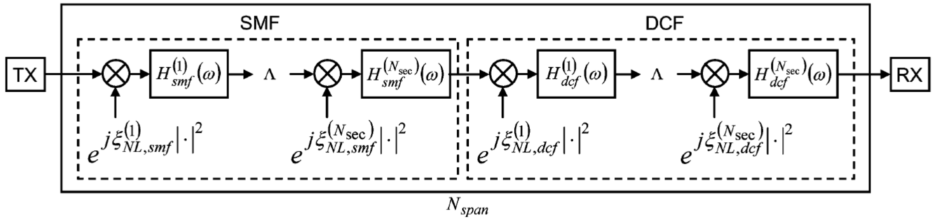
$$y(n) = \sum_{p=0}^{P} \sum_{l_1,\ldots,l_p=0}^{L_p} h(l_1,\ldots,l_p) \boxed{x(n-l_1)x^*(n-l_2)\cdots x(n-l_p)} + z(n)$$

Volterra series expansion

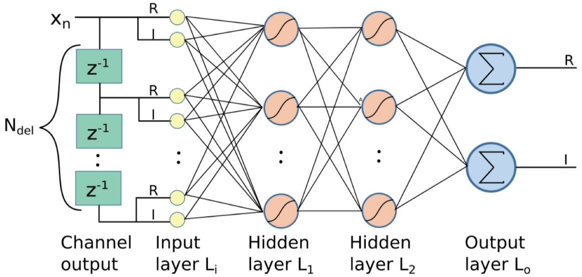
Volterra [Peddanarappagari '97]



TEQ [Haunstein '04, Djordjevic '07]



Digital back-propagation [Li et al '08, Ip-Kahn '08]



DNN [Sidelnikov '18, Koike-Akino '18, Kamalov '18]

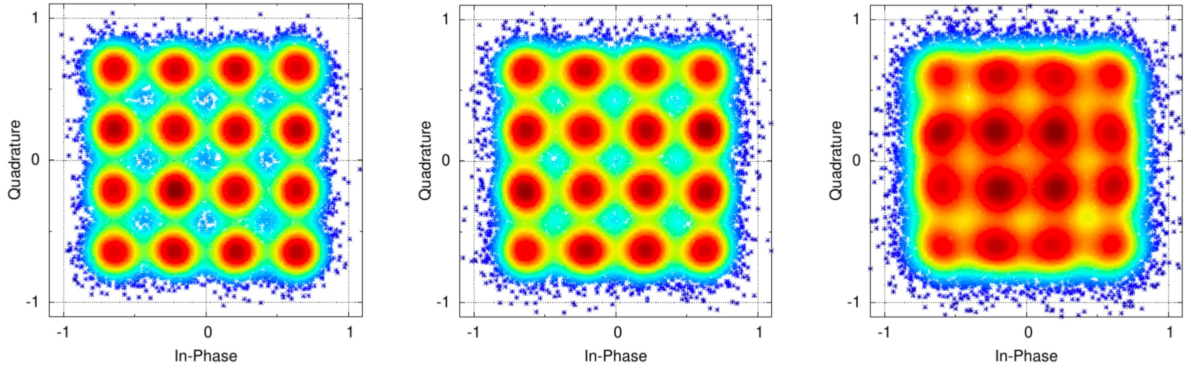
ML2ML: Maximum-Likelihood to Machine Learning Paradigm Shift

- Nonlinear equalization based on **maximum-likelihood (ML)**
  - Log-likelihood maximization, depending on nonlinear channel statistics

Maximum-Likelihood (ML)

$$\max_i \log \Pr(x_i|y)$$

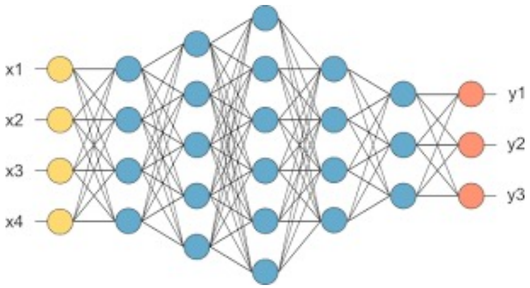
How to determine?
Model based?
Model mismatch?



(a) -5 dBm Launch (b) -3 dBm Launch (c) -1 dBm Launch

Post-Linear Equalization Distortion (16 spans)

- Cross-entropy minimization based on **machine learning (ML)**
  - Learning nonlinear channel statistics given massive data
  - Lower bound maximization of **GMI** (generalized mutual information)
  - Analogy to SSFM: sequence of **linear** transform and **nonlinear** operation

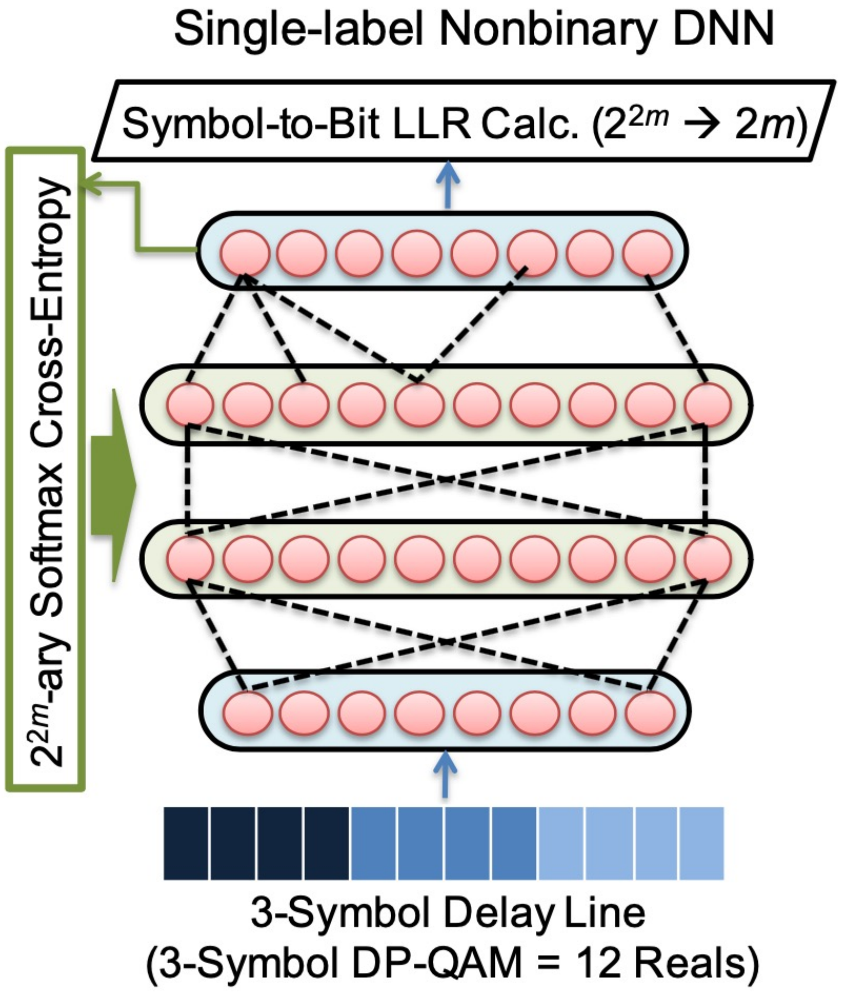


Binary cross entropy (BCE) corresponds to GMI

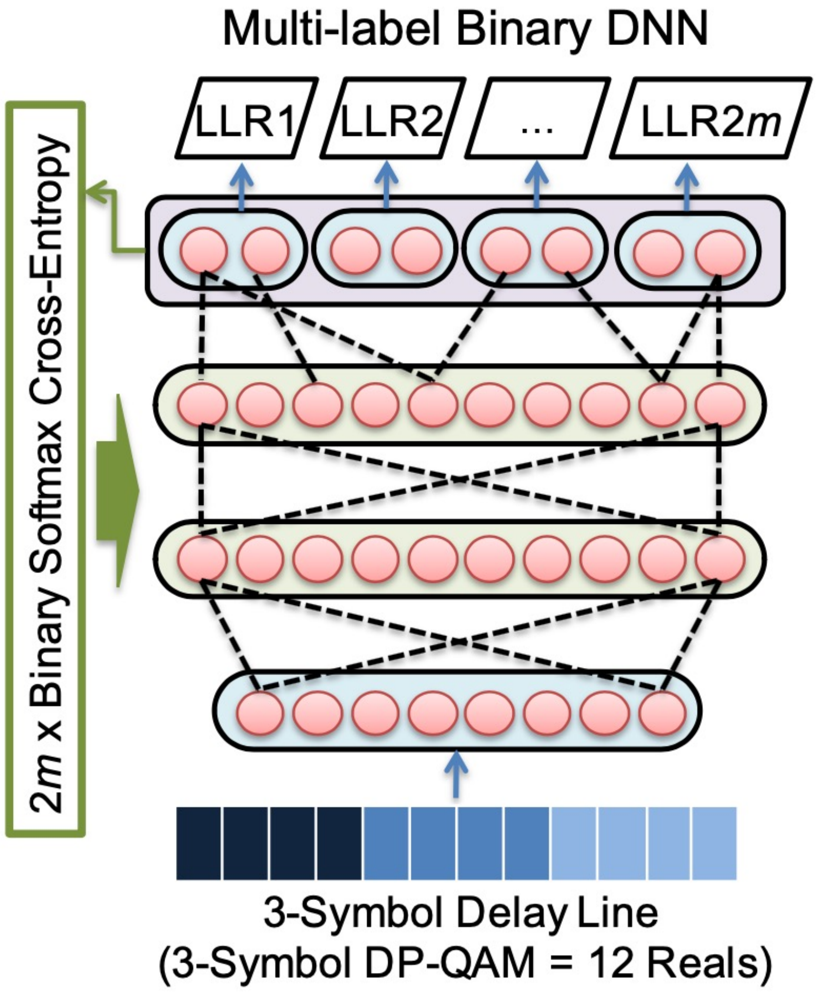
$$\mathbb{E}[\sum_i -\log \Pr(x_i|y)] \rightarrow 1 - \mathrm{GMI}$$

Non-Binary vs. Binary Cross-Entropy

- DNN nonlinear equalizer with NBCE/BCE



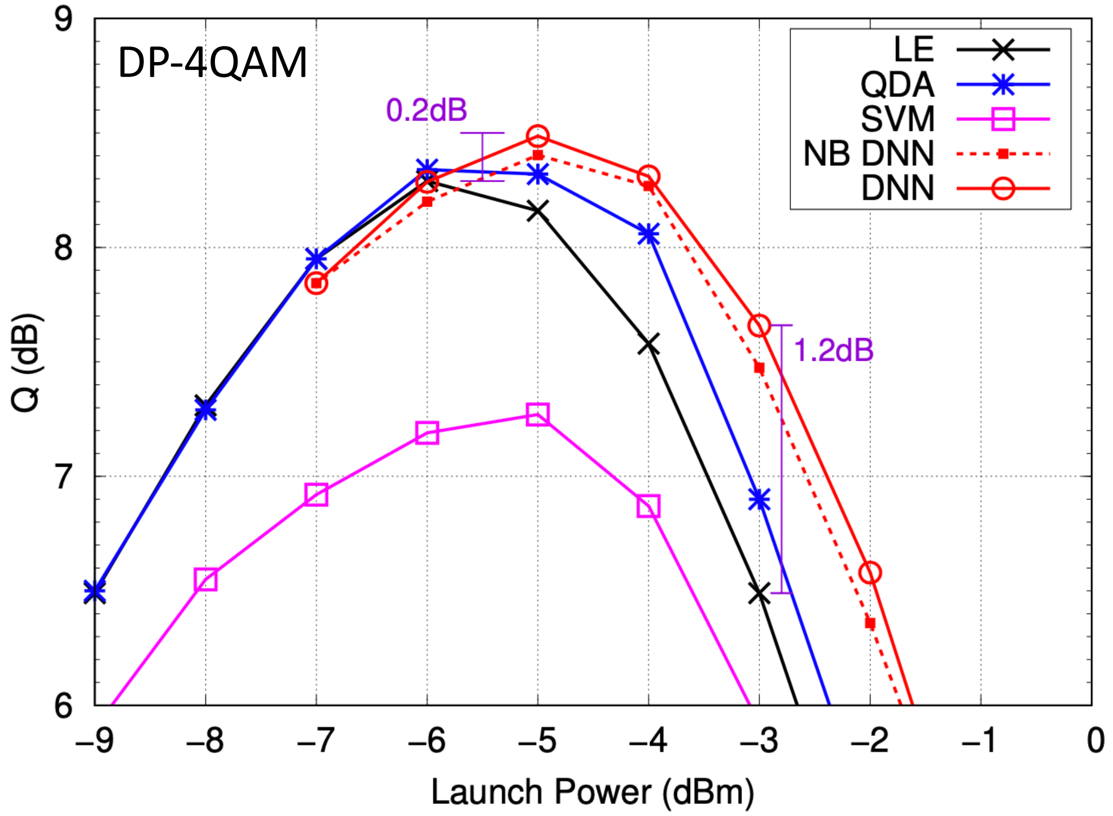
(a) One 2^{2m}-ary softmax



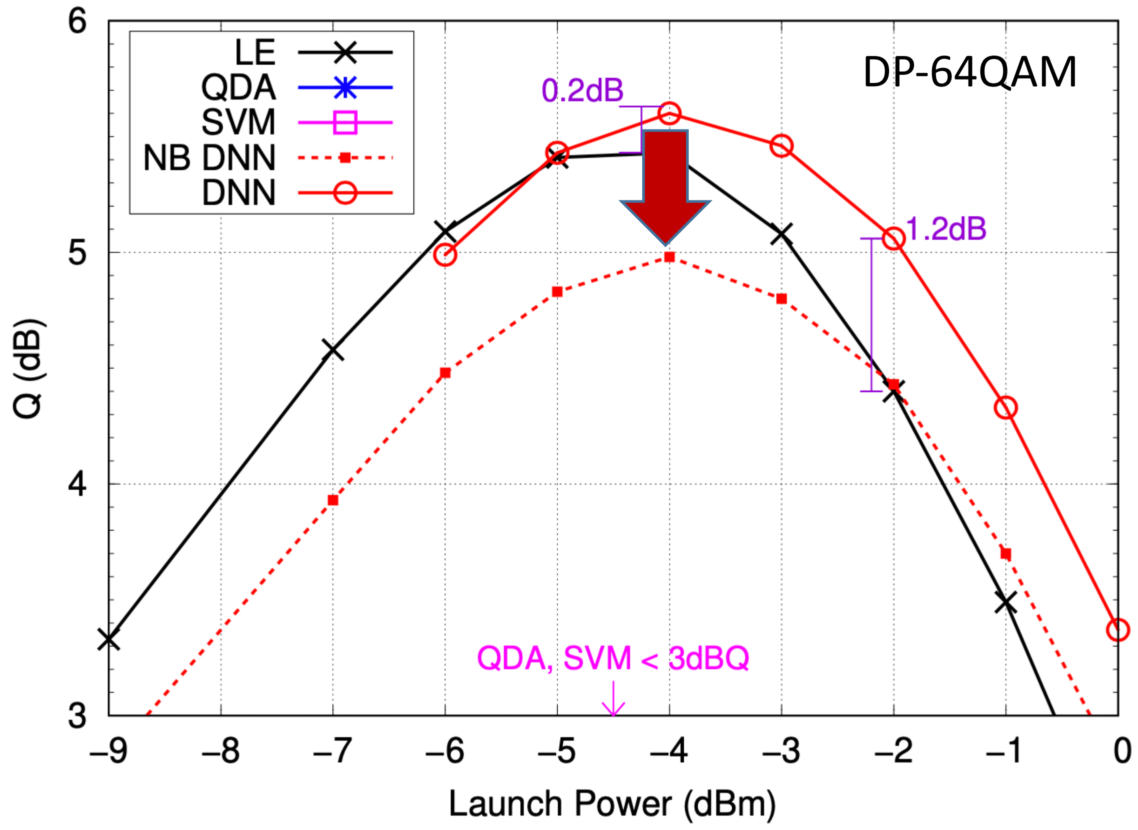
(b) $2m\times$ binary softmax

DNN Equalizer Performance

- Nonbinary cross-entropy does not work for high-order QAM



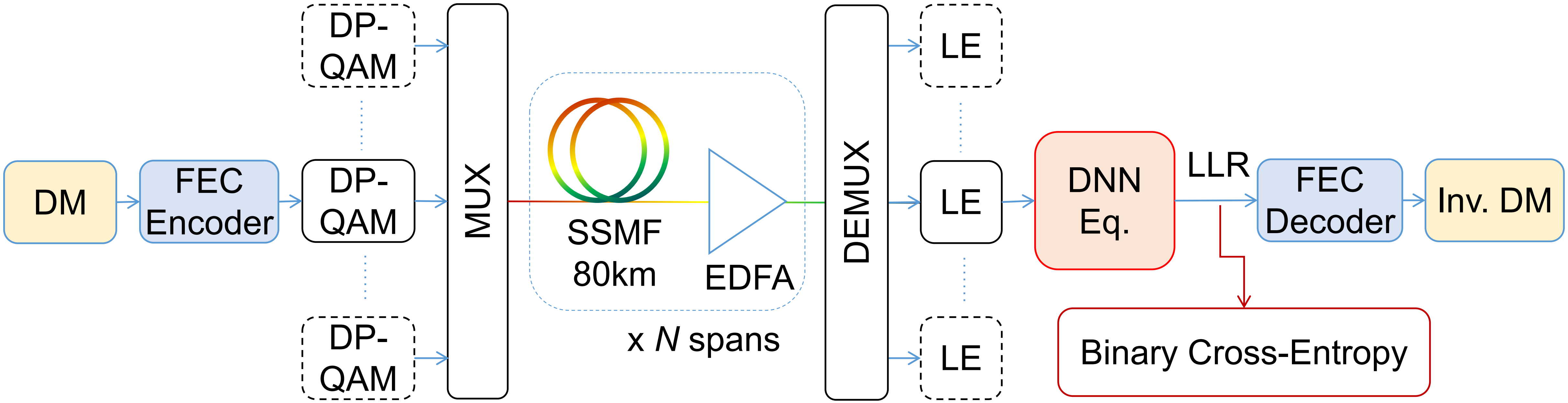
Q factor comparisons for DP-4QAM 50-span NZDSF.



Q factor comparisons for DP-64QAM 8-span NZDSF.

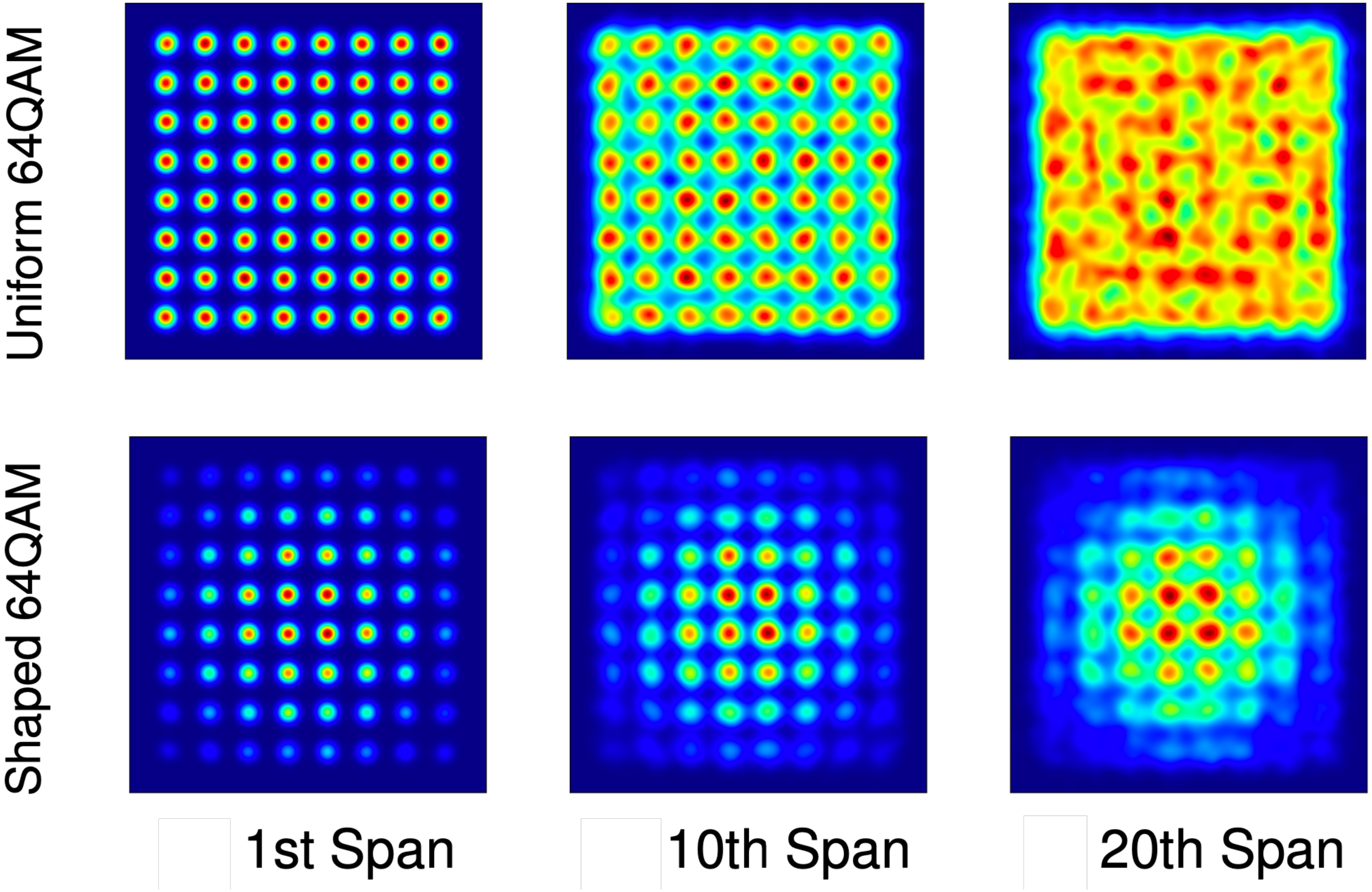
System Model

- We learn nonlinear statistics over 500,000 symbols on system model:
  - Dispersion unmanaged standard single-mode fiber (SSMF) 80km x *N* spans
    - 17ps/nm/km, 1.2/W/km, 0.2dB/km
  - Erbium-doped fiber amplifier (EDFA) 5dB noise figure
  - 11-channel DP-QAM at 34GBd, root-raised cosine role-off 2%
  - 61-tap least-squares linear equalizer (LE) prior to DNN nonlinear compensation
  - **Probabilistic amplitude shaping (PAS)** with Maxwell-Boltzmann distribution



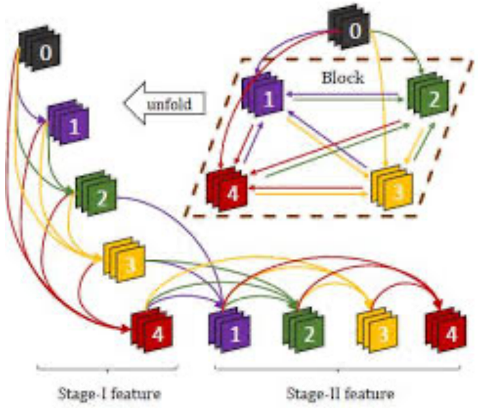
Nonlinear Distortion

- Unshaped/shaped DP-QAM



DNN Architectures

- Residual Multi-Layer Perceptron (ResMLP)
- Residual Convolutional Neural Network (ResCNN)
- Bidirectional Long Short-Term Memory (BiLSTM)
- Transformer, U-net, …



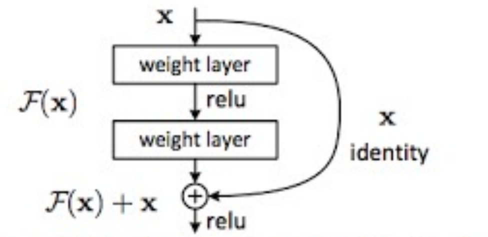
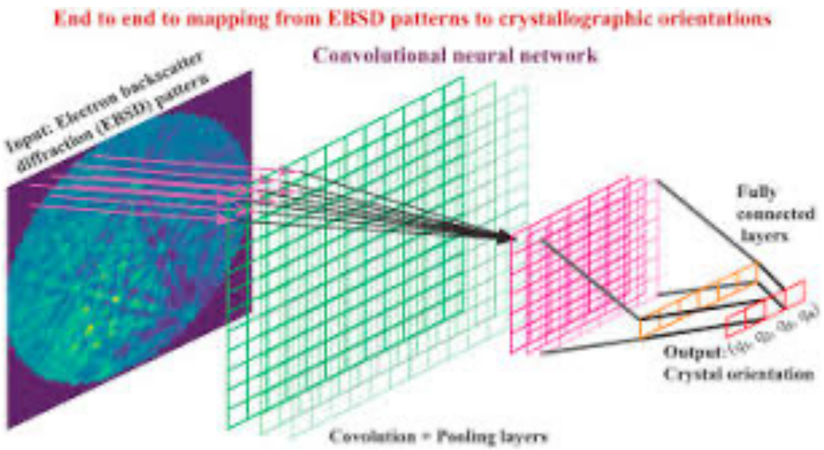
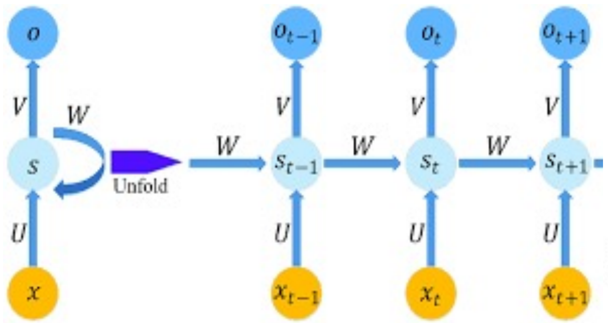
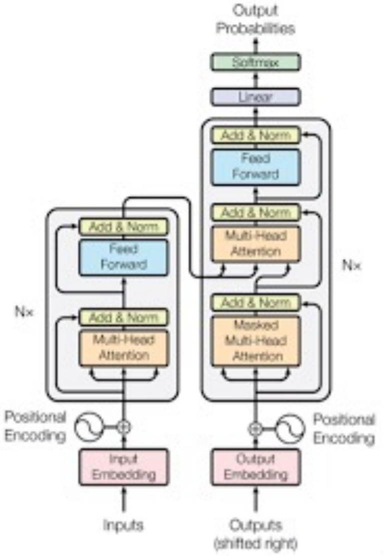
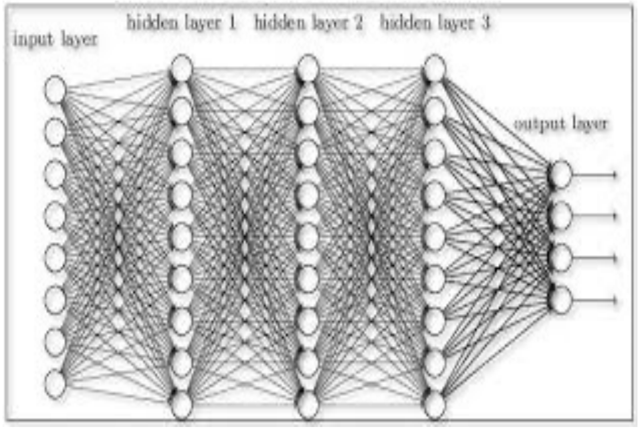


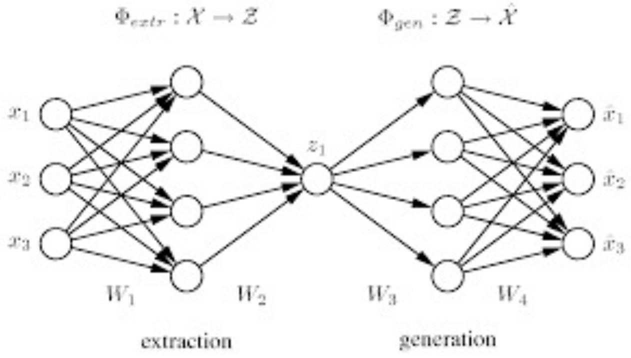
Figure 2. Residual learning: a building block.

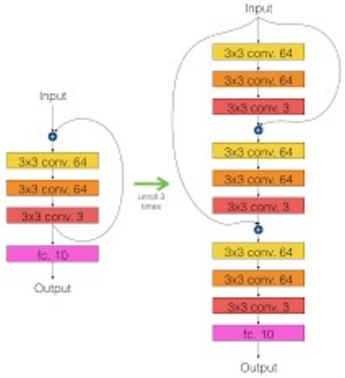


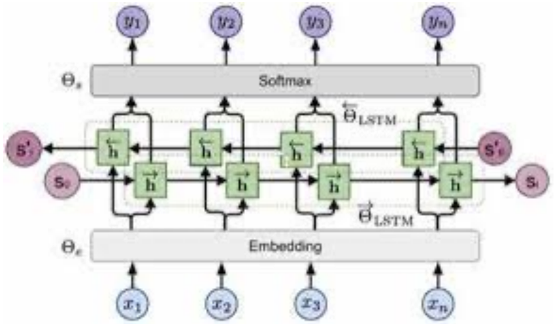


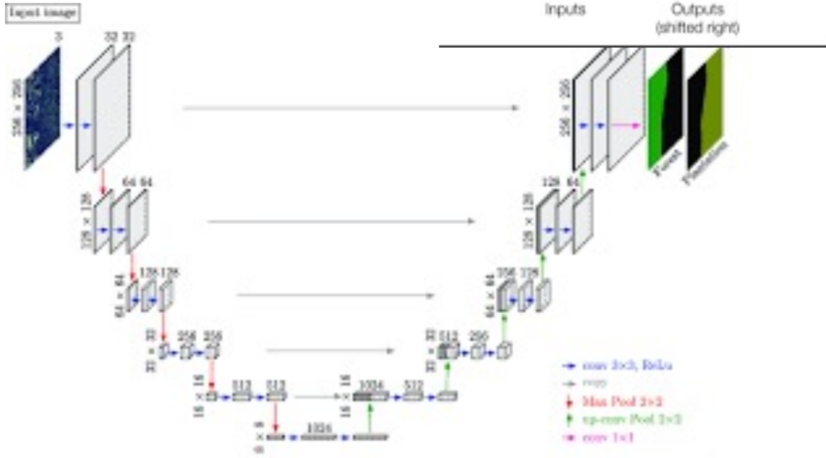




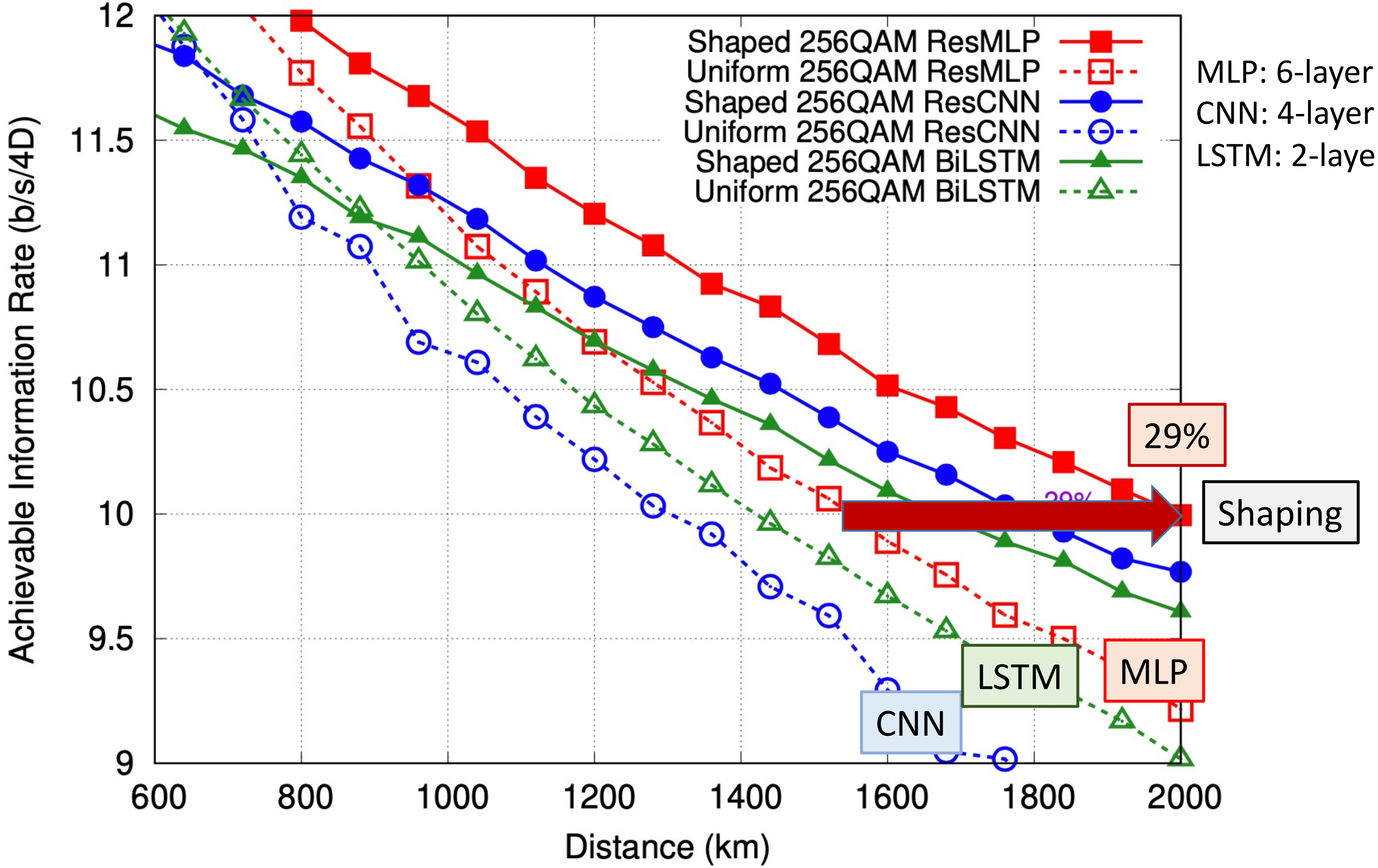








DNN Architecture Comparison (DP-256QAM)

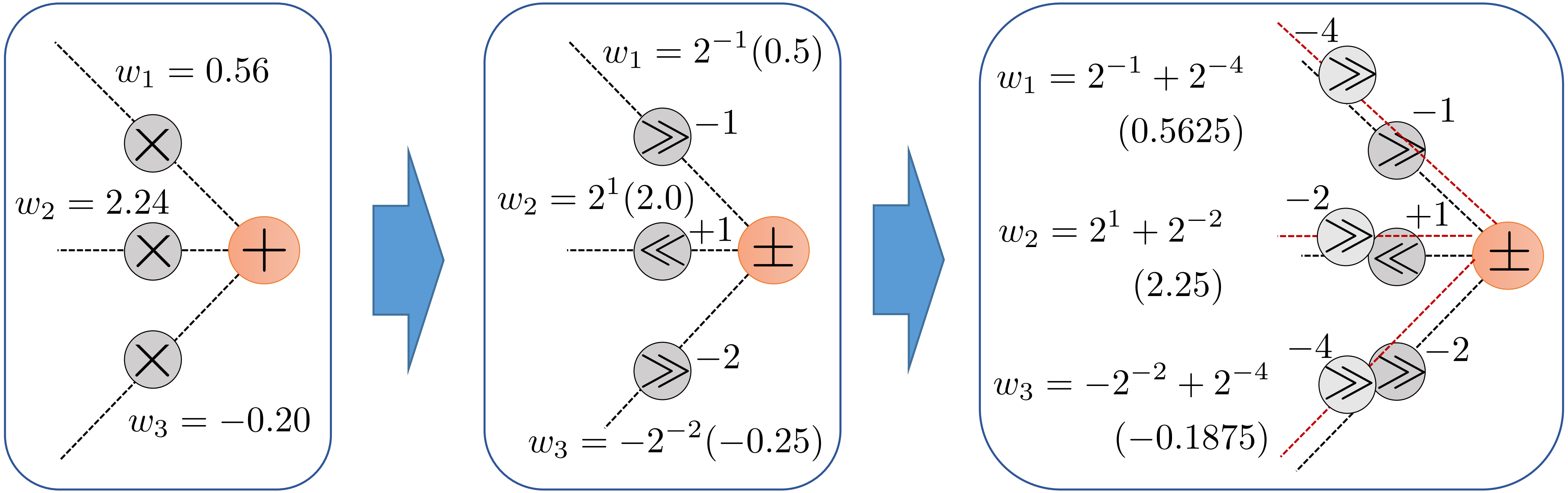


MLP: 6-layer 100-node
CNN: 4-layer kernel-3
LSTM: 2-layer 100-mem

Towards Multiplier-Less DNN

- DNN employs affine transforms requiring multiply-accumulate operations: $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$
- DeepShift [Elhoushi 2019]: Multiplier-less affine transforms with signed power-of-two (PoT) weights, realizing shift-accumulate $w = \pm 2^u, \quad u \in \mathbb{Z}$
- We improve it with additive PoT (APoT) for reducing the quantization error

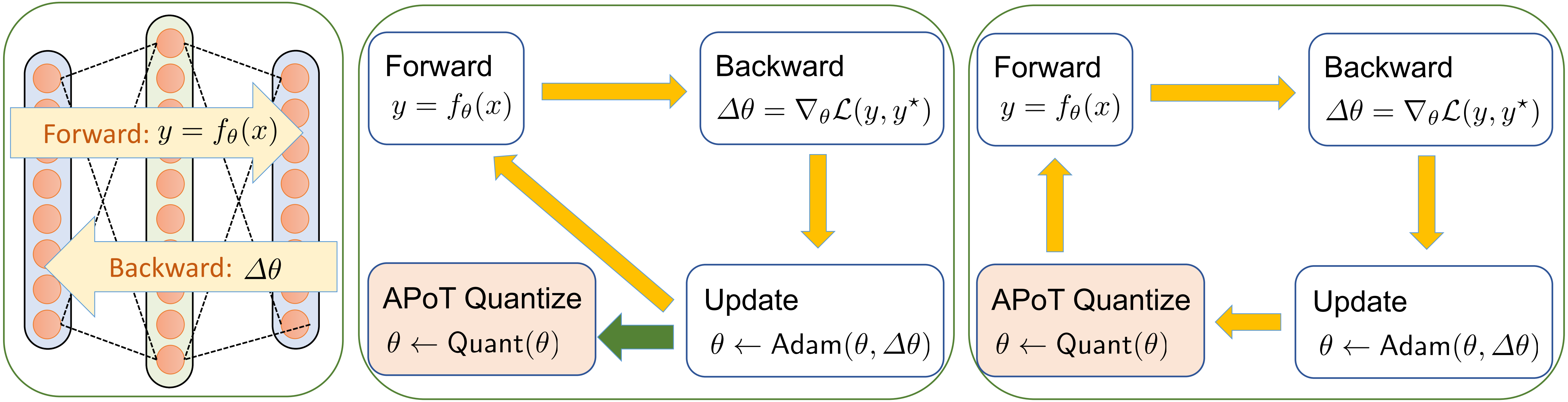
$$w = \pm 2^u \pm 2^v, \quad u, v \in \mathbb{Z}$$



(a) FP: Multiply & Add (b) PoT: Shift & Add (c) APoT: Dual Shift & Add

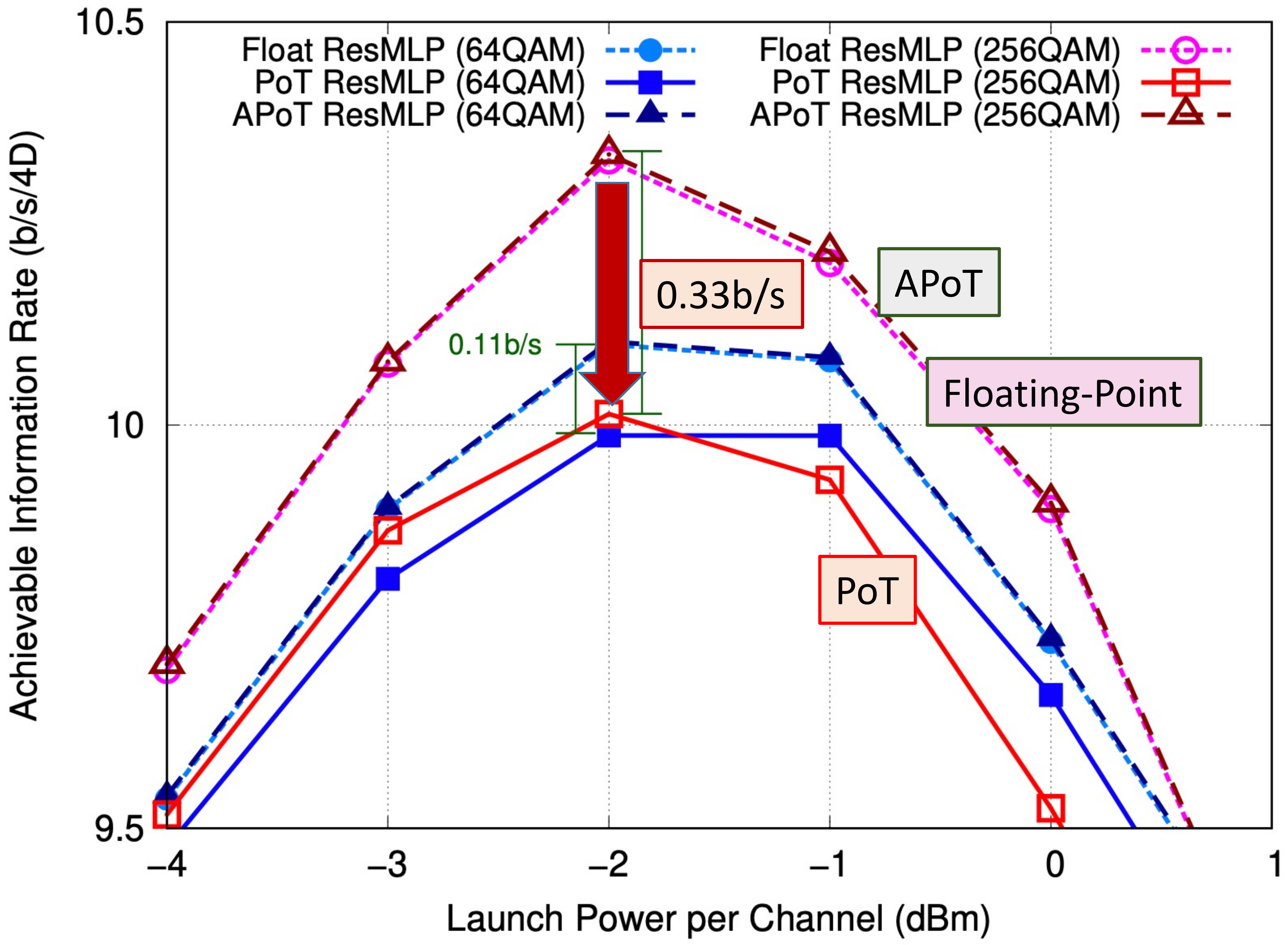
Quantization-Aware Training (QAT)

- Update with quantization: straight-through rounding in the loop of stochastic gradient
  - Finding best signs and integer shifts for affine transforms in training loop
- QAT overcomes quantization errors due to static/dynamic quantization



(a) Quantized DNN (b) Static Quantization (c) Quantization-Aware Training

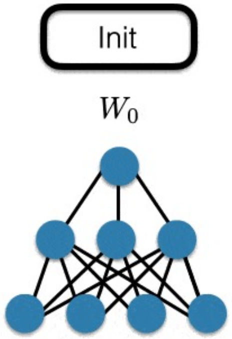
Multiplier-Free DNN Performance



Sparse DNN with Lottery-Ticket Hypothesis (LTH)

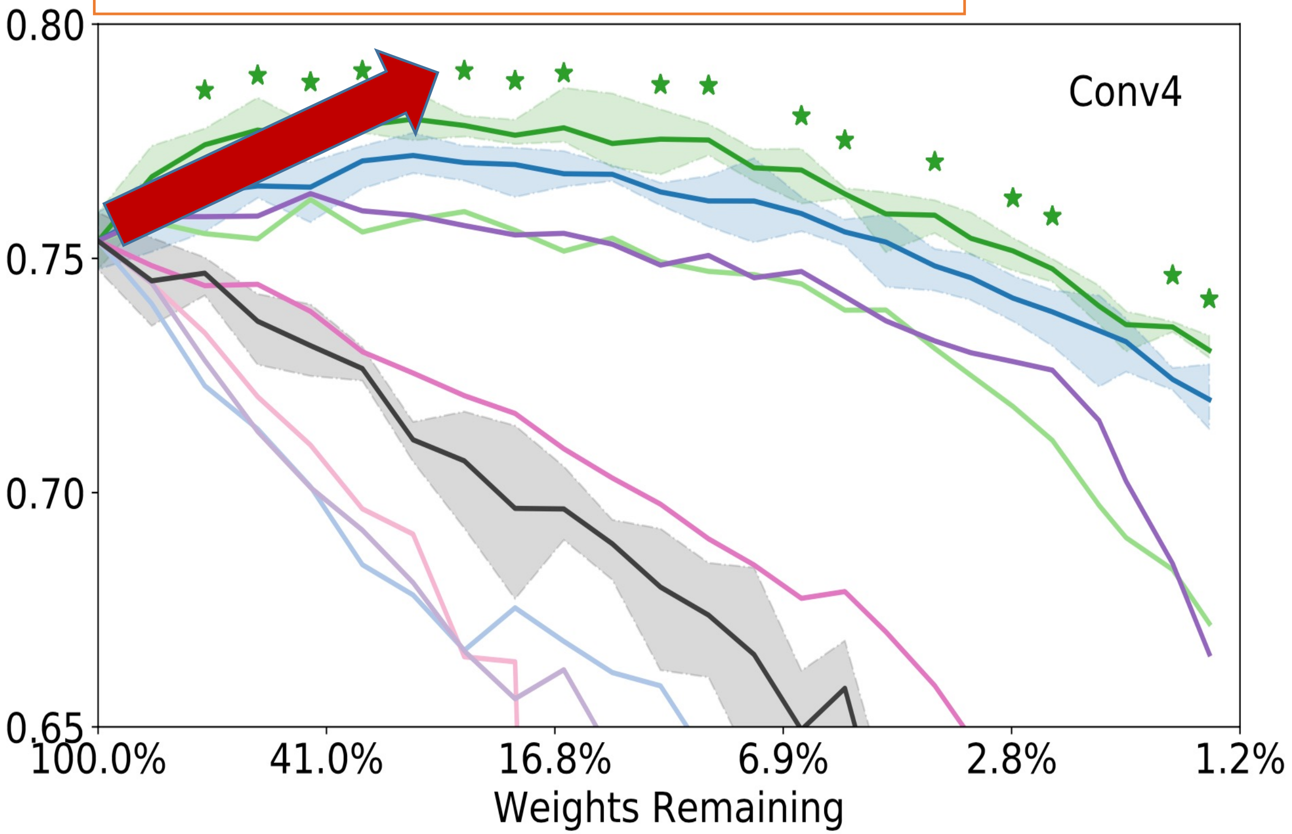
- LTH pruning [Frankle 2018]: Sparse DNN can outperform dense DNN with trained mask and rewinding weights

Iterative Magnitude Pruning with Rewinding

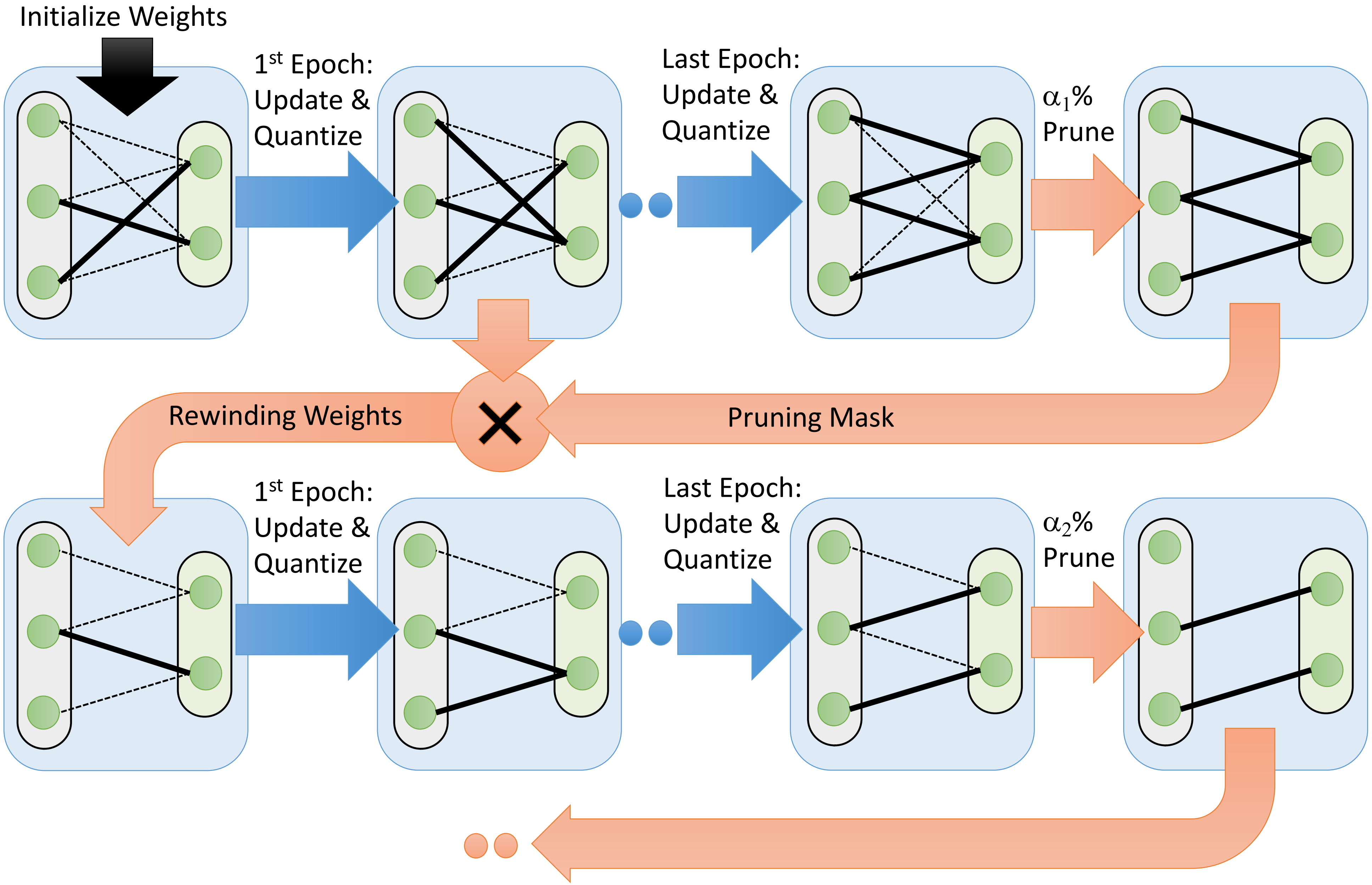


Frankle et al., 2019
Viz: @RobertTLange

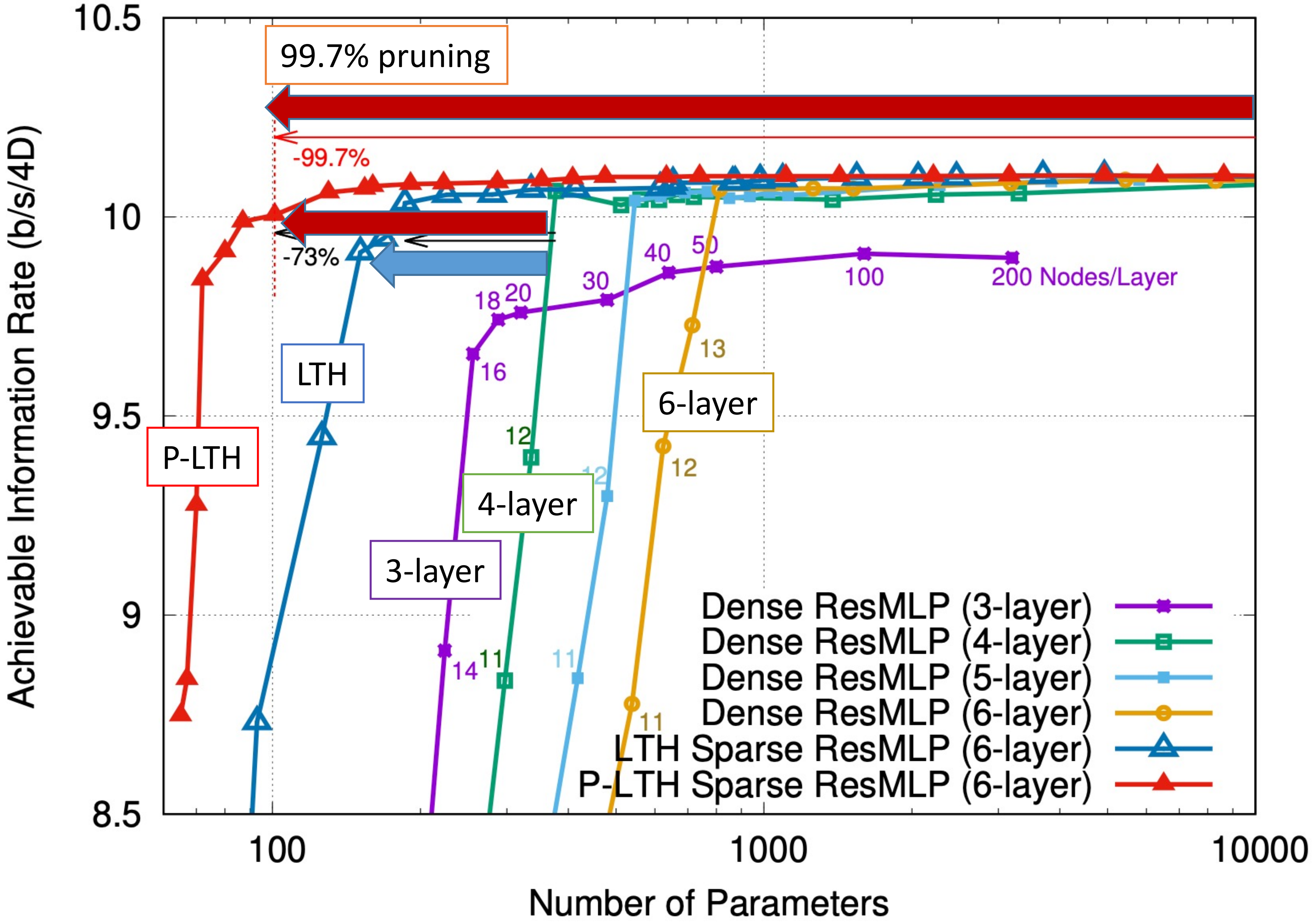
Lower complexity and better performance



Progressive LTH Distillation: Incremental Sparsity

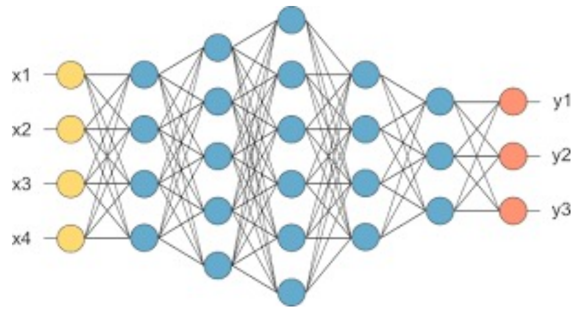


Sparse DNN Performance (Shaped DP-64QAM, 22 Spans)



Summary

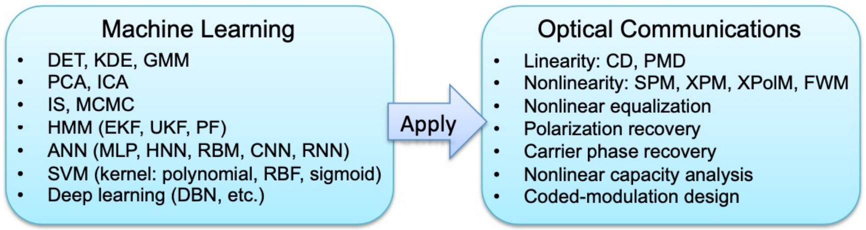
- We showed some perspectives of deep learning techniques for nonlinear optical fiber communications
  - Nonlinear fiber distortion may call for **nonlinear** signal processing
  - **Data-driven approach** can be a viable alternative to model-based approaches as massive data are available in high-speed optical transmission
- We proposed **multiplier-less sparse DNN equalizer** for low-power real-time operations
  - Compared different DNN architectures for PAS systems
  - Zero-multiplier APoT QAT achieves slight improvement over floating-point weights
  - 99% weights can be eliminated by progressive LTH pruning
- There are a great amount of open research fields to apply deep learning techniques to optical communications because of the nature of nonlinear physics

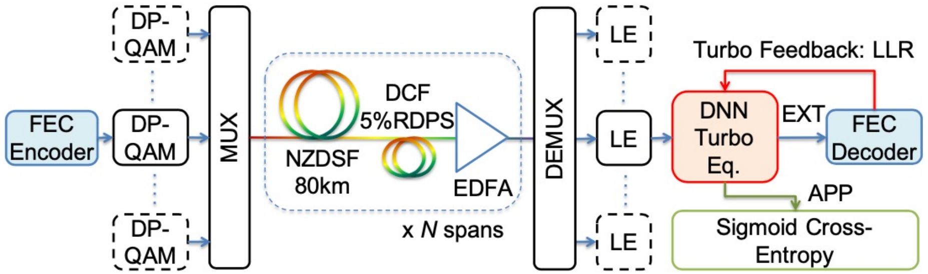


Backups

DL/ML Works for Optical Communications

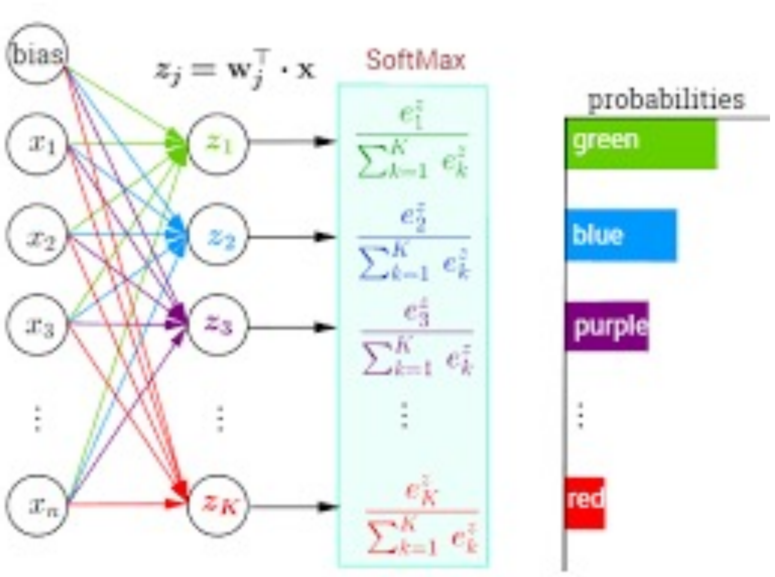
- Koike-Akino, T., "**Perspective of Statistical Learning for Nonlinear Equalization in Coherent Optical Communications**", *Signal Processing in Photonic Communications (SPPCom)*, DOI: 10.1364/SPPCOM.2014.ST2D.2, July 2014.
- Koike-Akino, T., Millar, D.S., Parsons, K., Kojima, K., "**Fiber Nonlinearity Equalization with Multi-Label Deep Learning Scalable to High-Order DP-QAM**", *Signal Processing in Photonic Communications (SPPCom)*, DOI: 10.1364/SPPCOM.2018.SpM4G.1, July 2018.
- Koike-Akino, T., Wang, Y., Millar, D.S., Kojima, K., Parsons, K., "**Neural Turbo Equalization to Mitigate Fiber Nonlinearity**", *European Conference on Optical Communication (ECOC)*, DOI: 10.1049/cp.2019.0803, September 2019.
- Koike-Akino, T., Wang, Y., Millar, D.S., Kojima, K., Parsons, K., "**Neural Turbo Equalization: Deep Learning for Fiber-Optic Nonlinearity Compensation**", *Journal of Lightwave Technology*, DOI: 10.1109/JLT.2020.2976479, March 2020.





Cross-Entropy Loss Function: Nonbinary to Binary

- Multi-class **single-label** cross-entropy: for non-binary coding
  - Conversion is slow since 2^n training is required per single word event
  - For high-order dual-polarization (DP)-QAM, it does not work well



DP-4QAM → DP-1024QAM

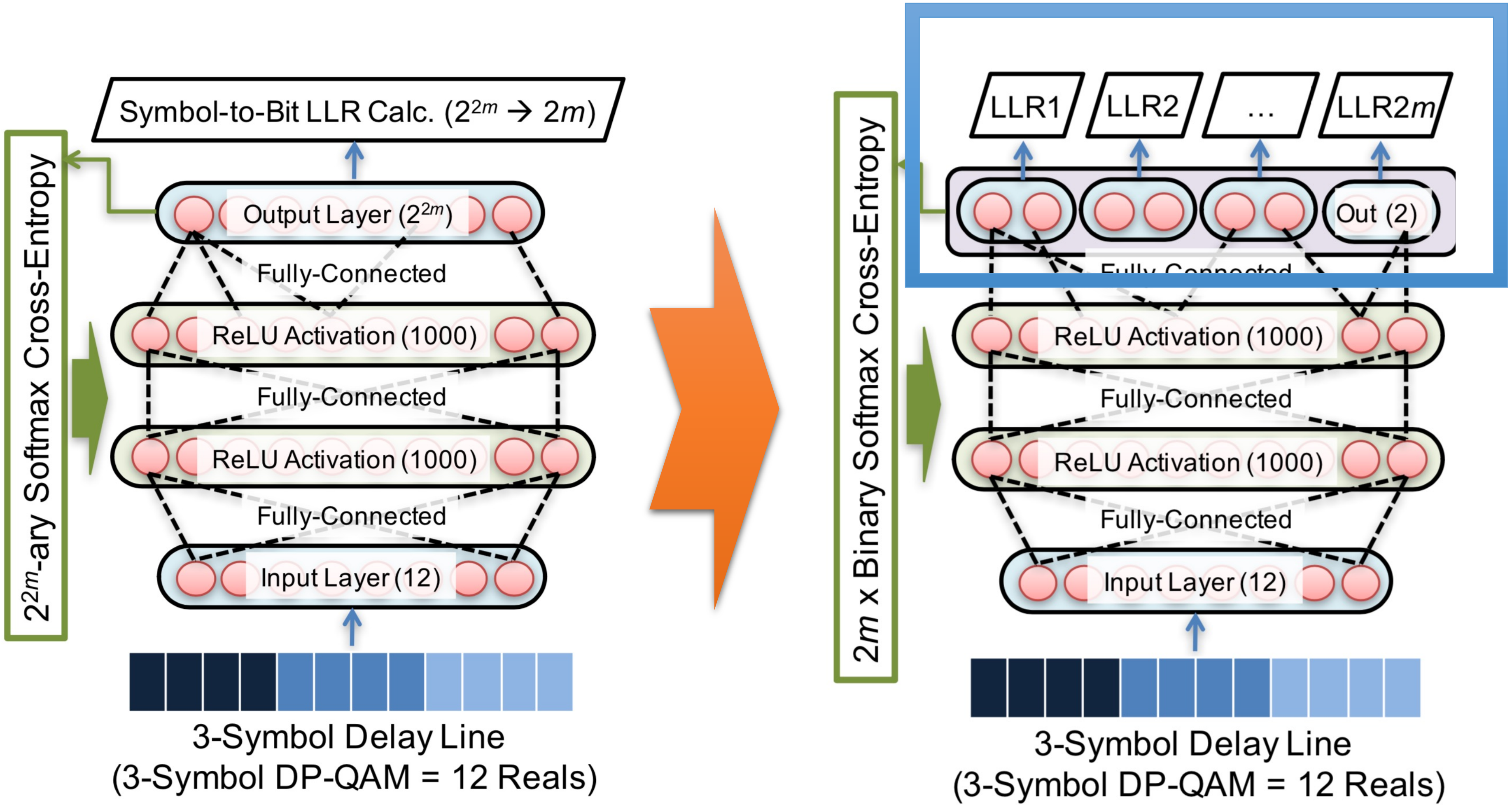
4^2 classes → 1024^2 classes!

20 binary classes

- Two-class **multi-label** cross-entropy: for binary coding
  - Multiple sigmoid cross entropy corresponds to bit-wise **LLR** (log-likelihood ratio)
  - DNN output can be directly fed back to soft-decision FEC decoder
  - Scalable to any high-order DP-QAM

Multi-Label DNN Architecture

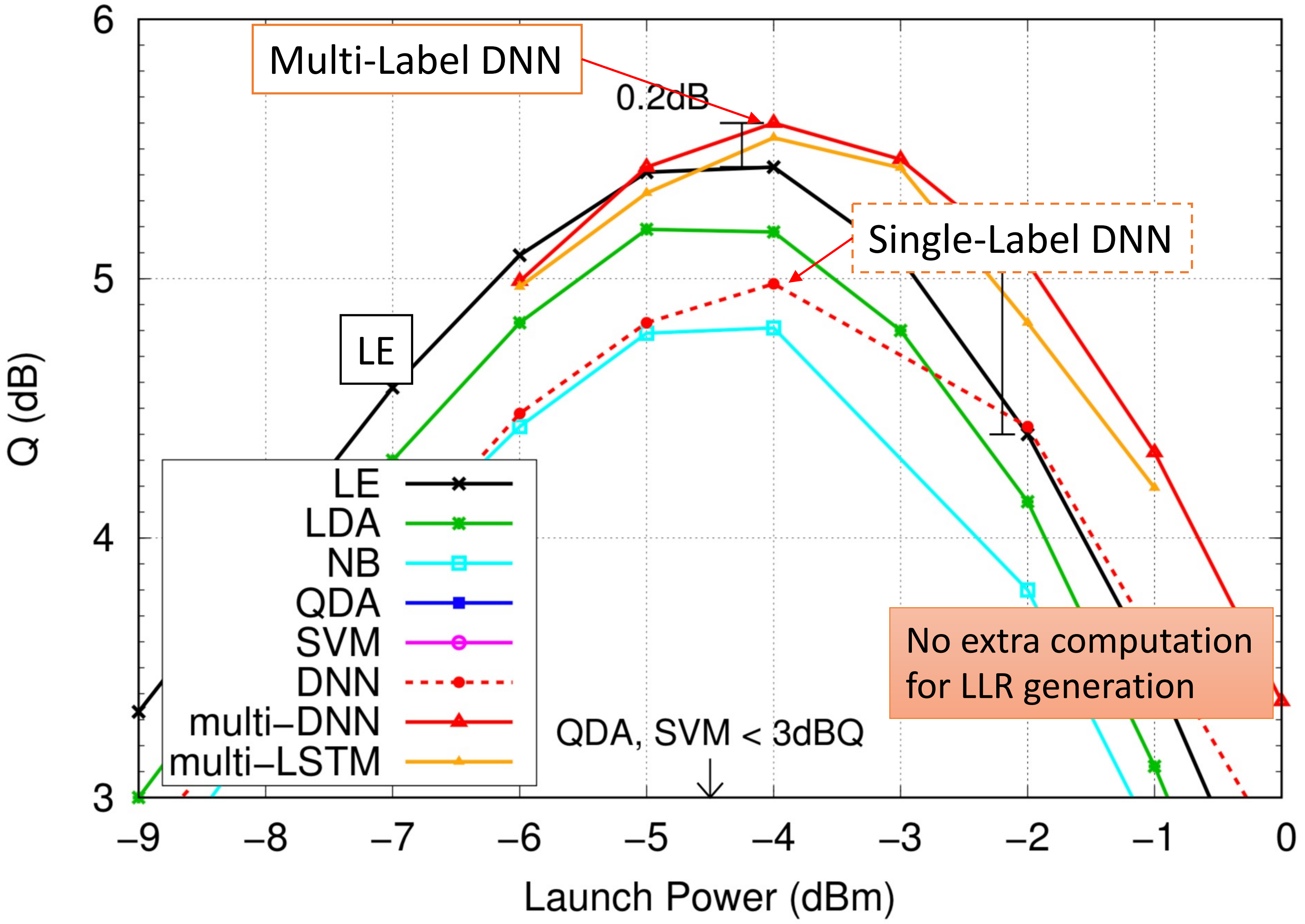
- Proposed DNN uses *n*-label sigmoid cross-entropy for *n*-bit modulation



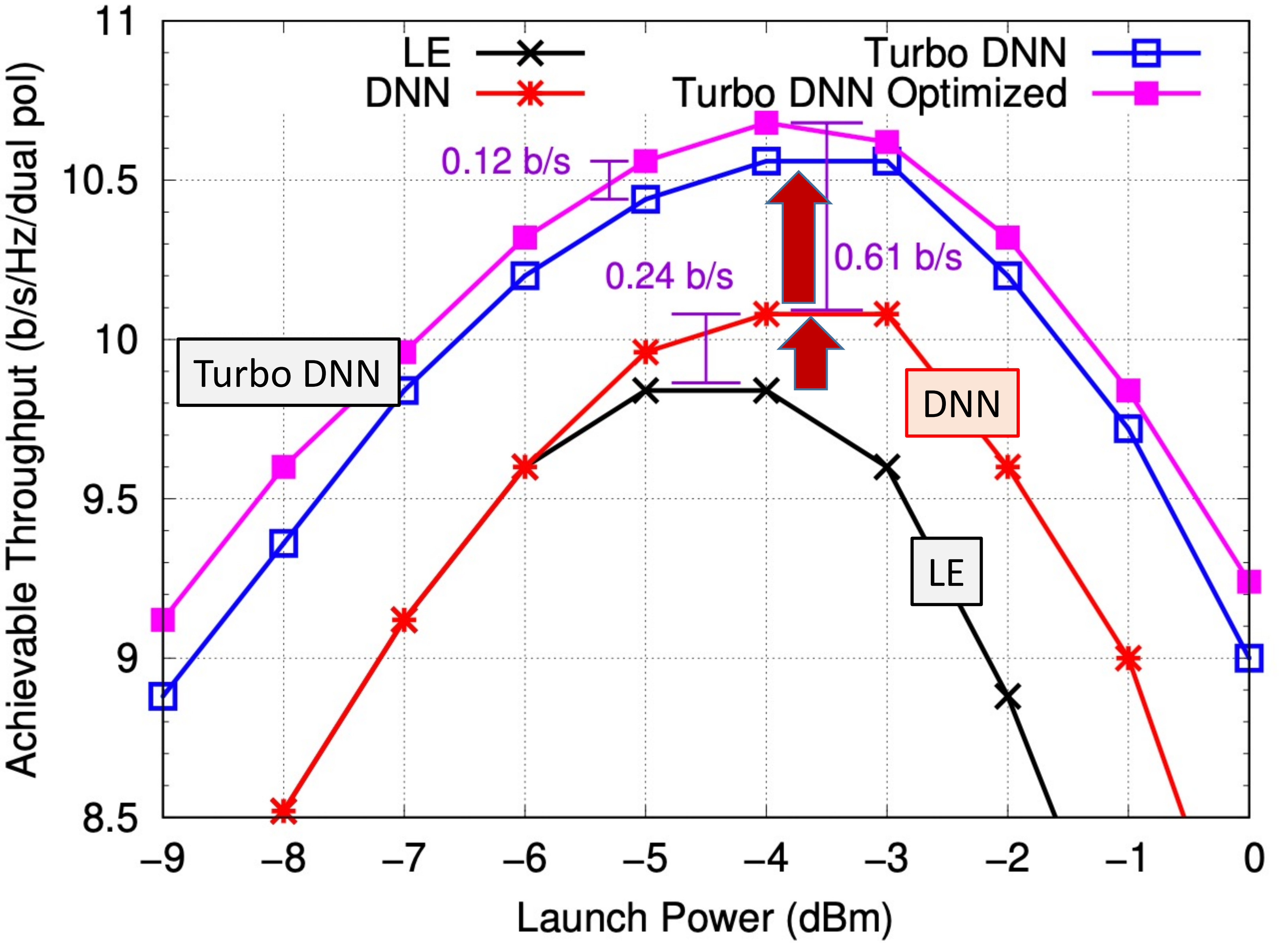
Single Label DNN (Nonbinary) | Multi-Label DNN (Binary)

Multi-Label DNN vs Single-Label DNN (DP-64QAM)

- Binary cross-entropy (BCE) performs better



DNN-TEQ Performance Evaluations (DP-64QAM)



MITSUBISHI ELECTRIC

Changes for the Better