Generative Digital Twins For High Performance System Design

Chris Cheng
Priyank Kashyap
Yongjin Choi

Hewlett Packard Enterprise

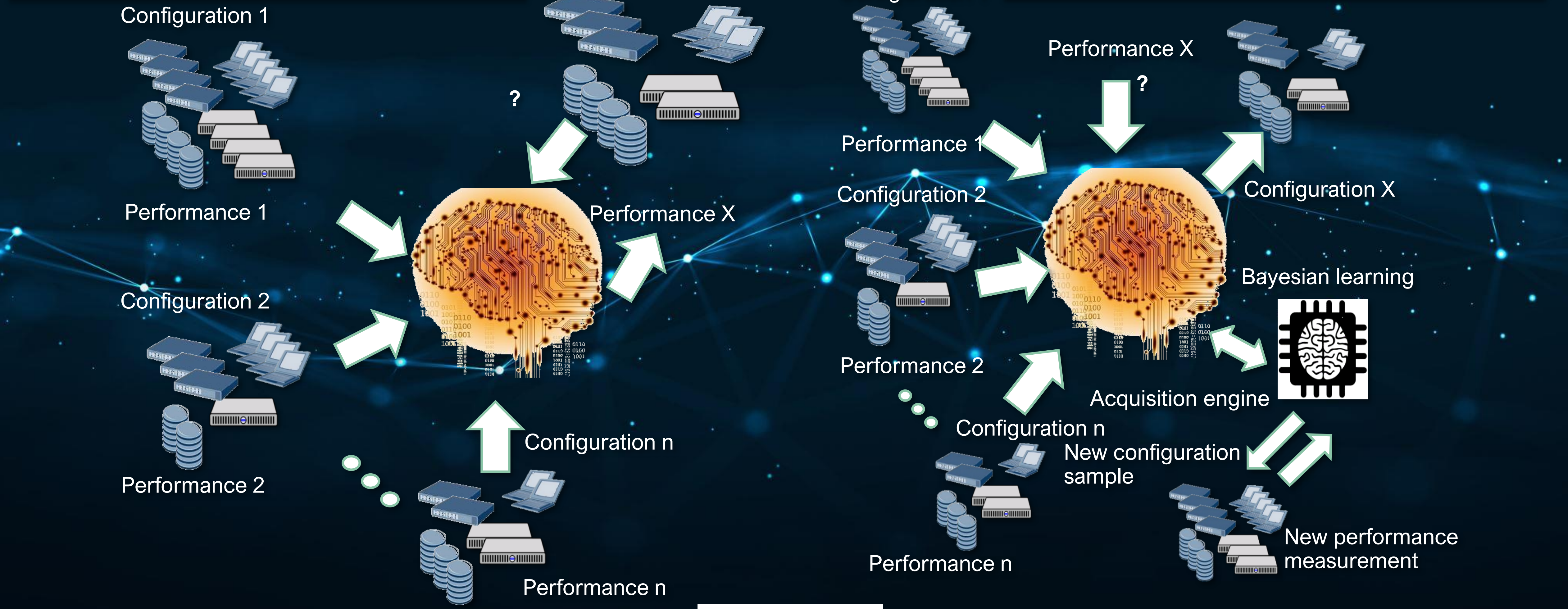
Outline

- Introduction to generative models and digital twins
- Generative Adversarial Networks (GANs)
- Conditional GAN for PAM4 SerDes
- Conditional GAN for advanced packaging power to thermal analysis
- Bayesian optimization of high-speed SerDes receivers
- Conclusion and future work

GENERATIVE MODELS AS DIGITAL TWINS

Discriminative Models
Conditional prob of output given an input

Generative Surrogate Models
Joint distribution of input/output

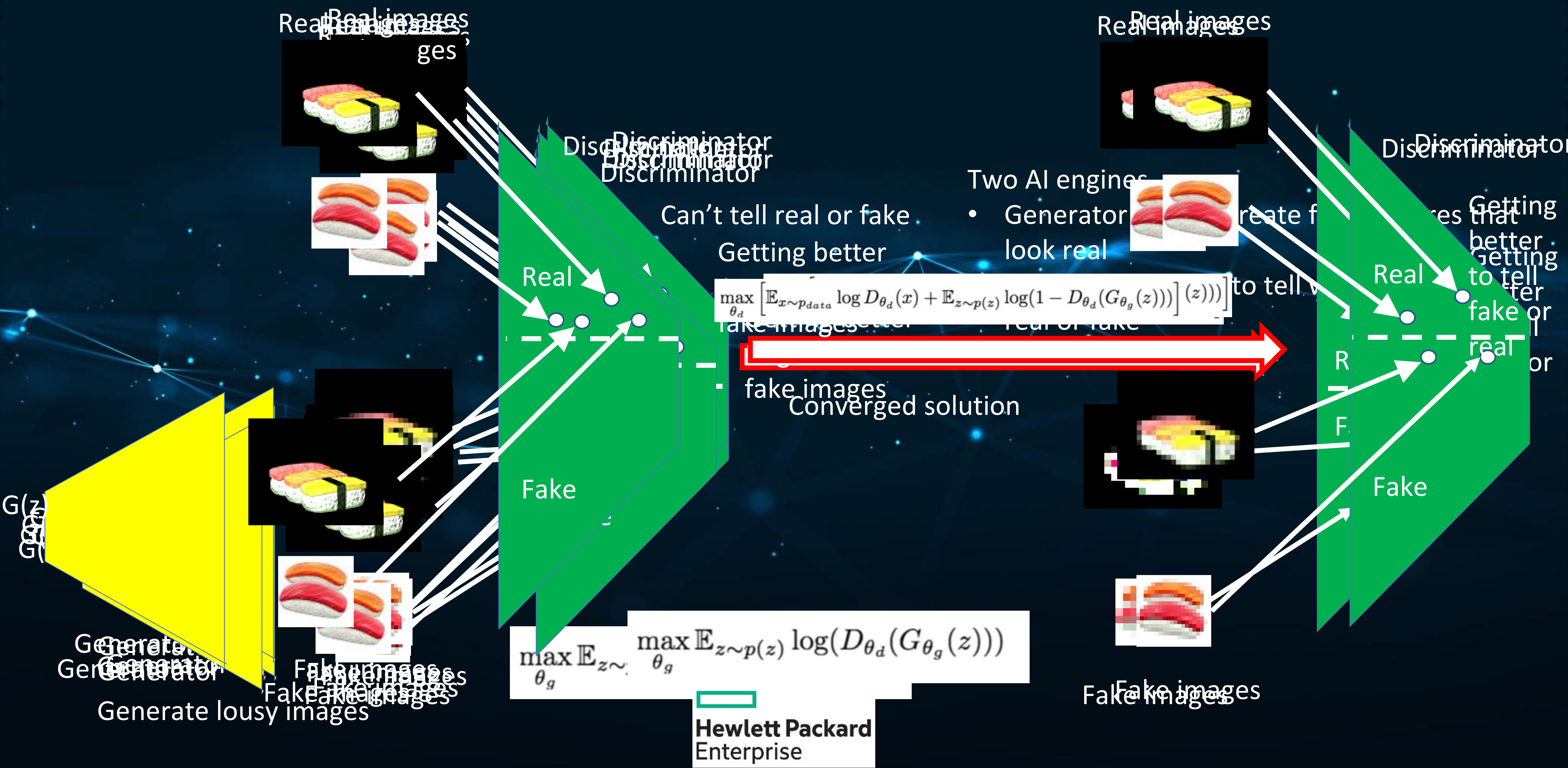


Hewlett Packard Enterprise

Generative Digital Twins

- Digital twins are computational models that cover the solution space within a targeted design limit
- Asymmetrical training vs. inference speed
  - Large dataset and high computational demand to train a generative model
  - Lightweight and fast computation for inference
- Realtime prediction of SI or multi-physics (power/thermal) systems
- Allow dynamic performance tuning based on changing input condition
- Generative models may have invertible solutions, i.e., given a desirable output, what is the most likelihood input condition
  - Can be constrained by power, space etc

Generative Adversarial Network



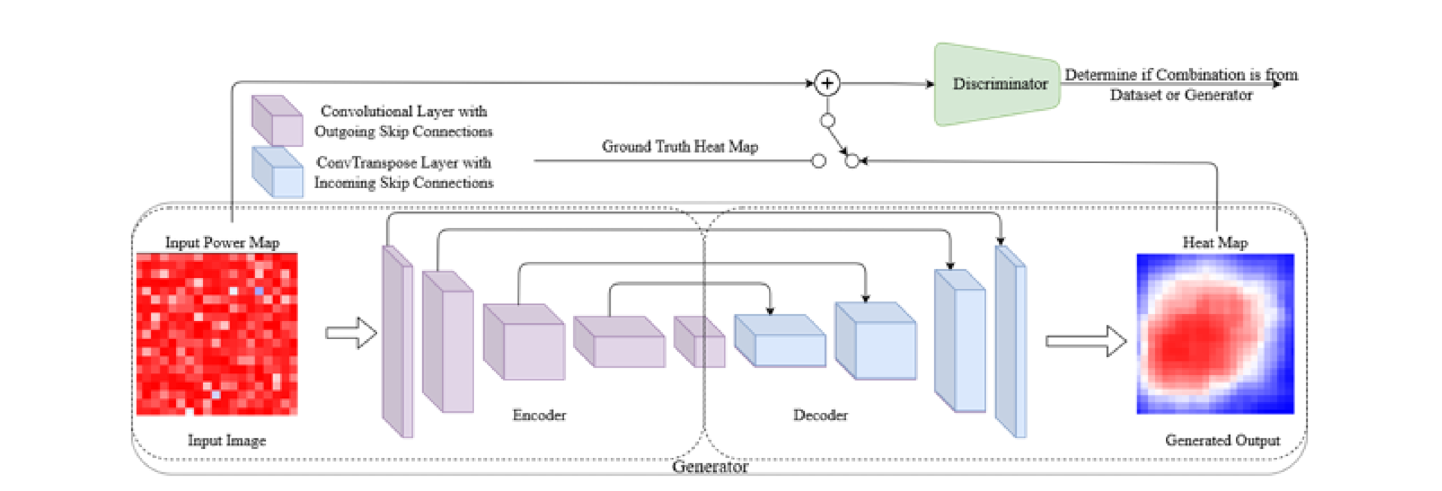
$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

$$\max_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$

Generator Example

CGAN Engine For Power To Thermal Analysis

- The generator is given the input power map to predict the corresponding heat map
  - Uses convolutional layers with skip connections on the encoder*
  - Decoder reconstructs from the hidden (latent) space
    - Contains Convolutional Transpose layers to upsample
- Skip connections ensure that no information is lost and that gradients are stable during the training phase



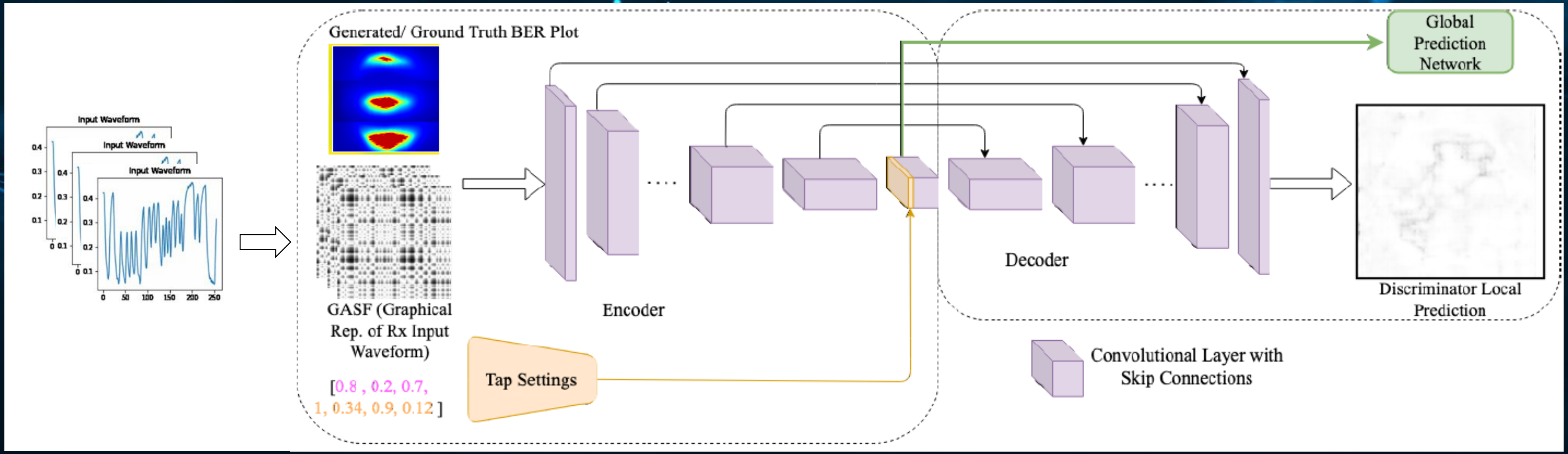
Kashyap et al: International 3D System Integration Conference 3DIC 2023

* P. Isola, J. Zhu, T. Zhou, and A. A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Discriminator Example

Conditional GAN For High Speed SerDes

- Discriminator is a U-Net architecture that predicts both a full pixel map (at decoder output) and a single true/false prediction (at the bottleneck) for a given input*
- Takes the input GASF and either the ground truth BER plot or generated BER plot
- Predicts whether the concatenated image is from the dataset or generator using two levels of prediction



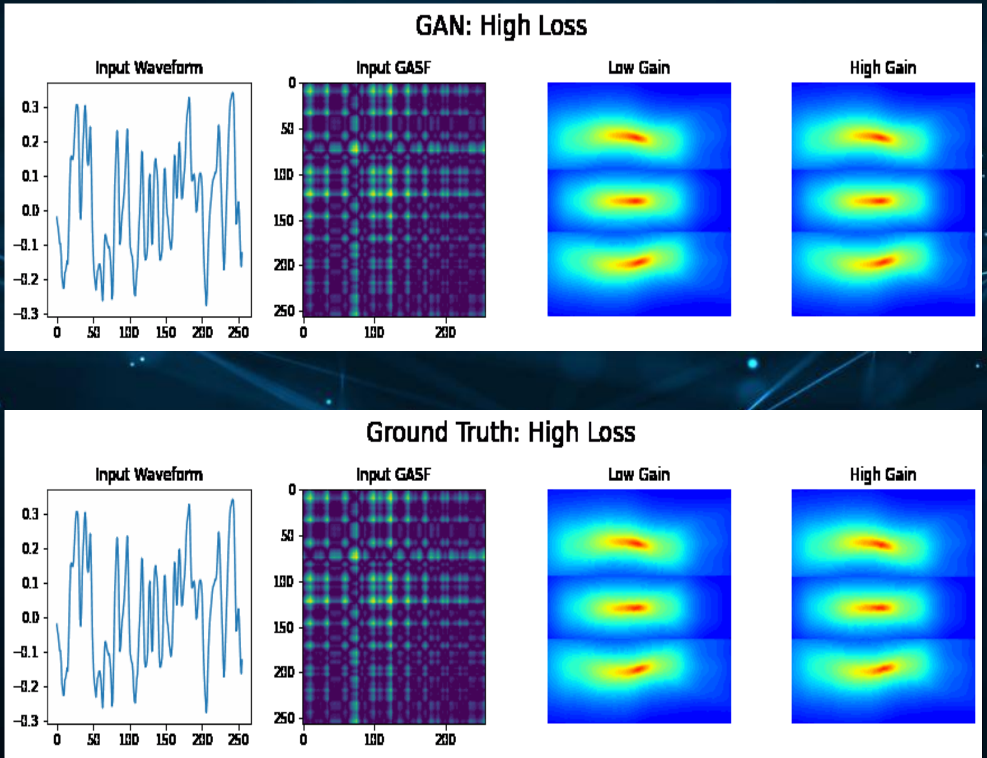
Choi Et al; DesignCon 2023

* E. Schonfeld, B. Schiele, and A. Khoreva. 2020. A U-Net based discriminator for generative adversarial networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8207–8216.

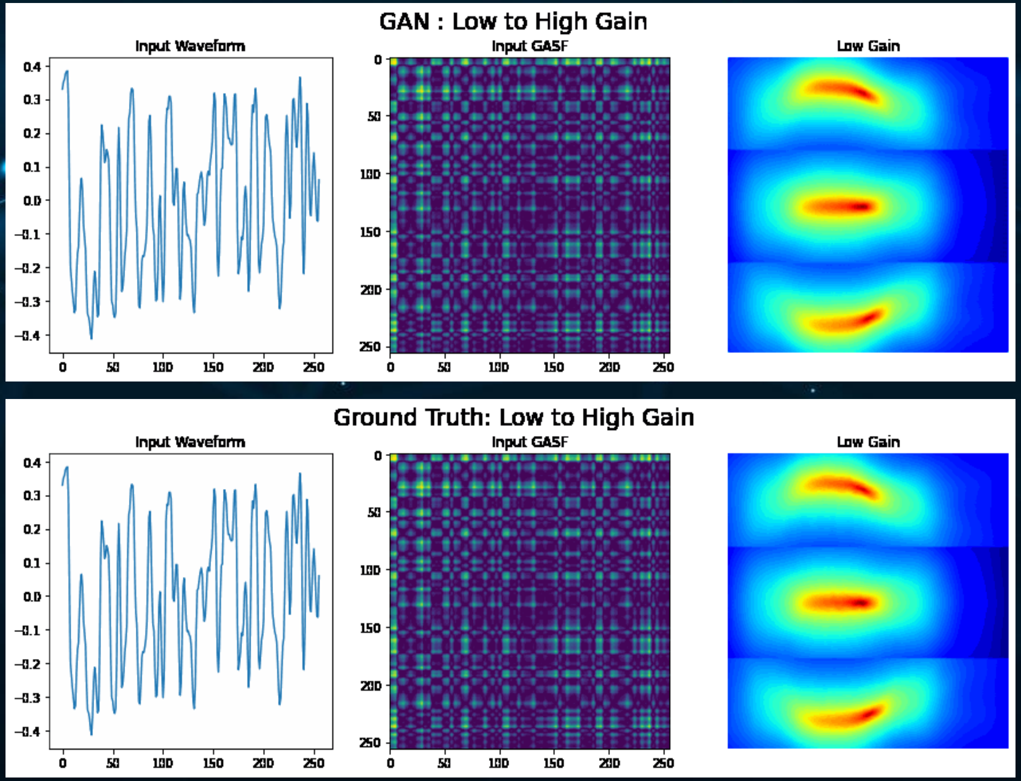
PAM4 Digital Twins

- cGAN engine can predict unseen input and equalizer condition

Interpolate Between Channel (Input) Conditions



Interpolate Between Gain (system) Conditions

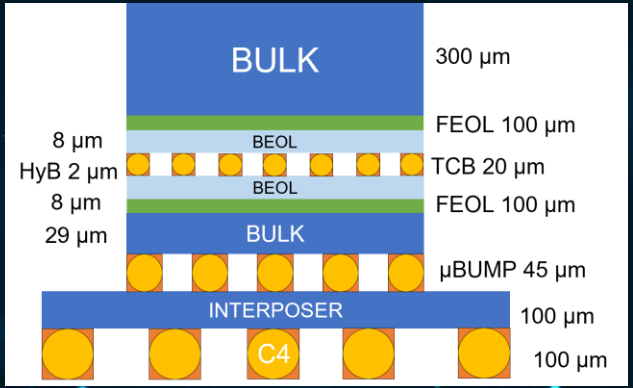


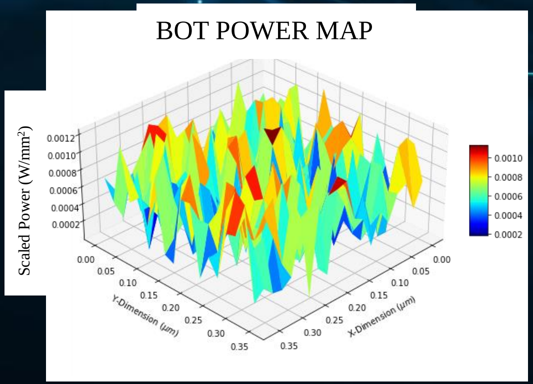
Choi Et al; 2023 DesignCon

Hewlett Packard Enterprise

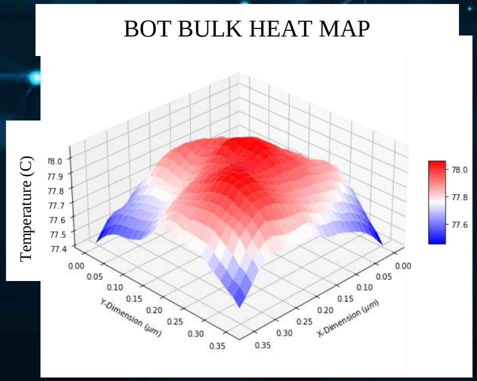
3DIC Power To Thermal Analysis

3DIC





Training power grid

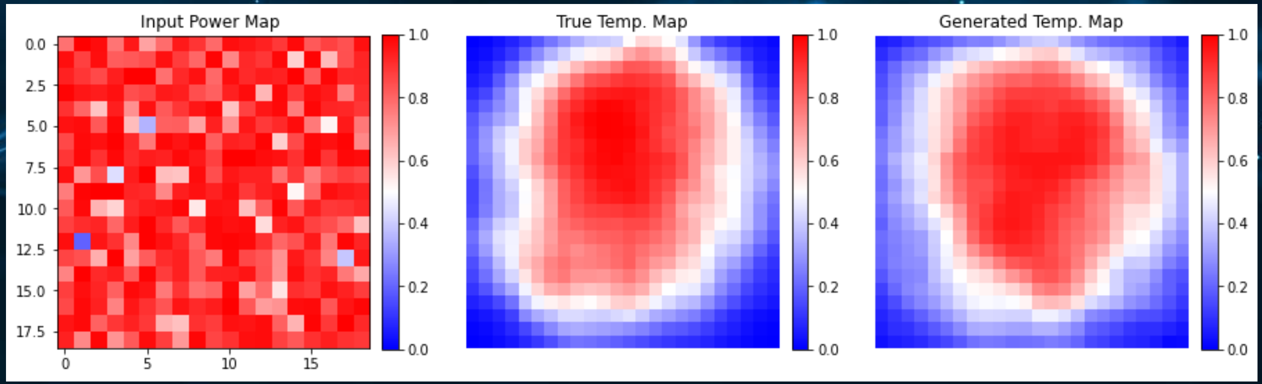


Training bulk heat map

Power Map

Heat Map

Generated Heat Map

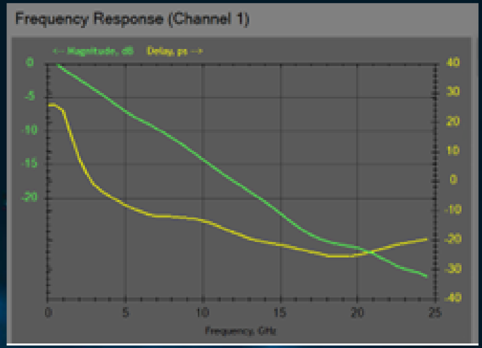


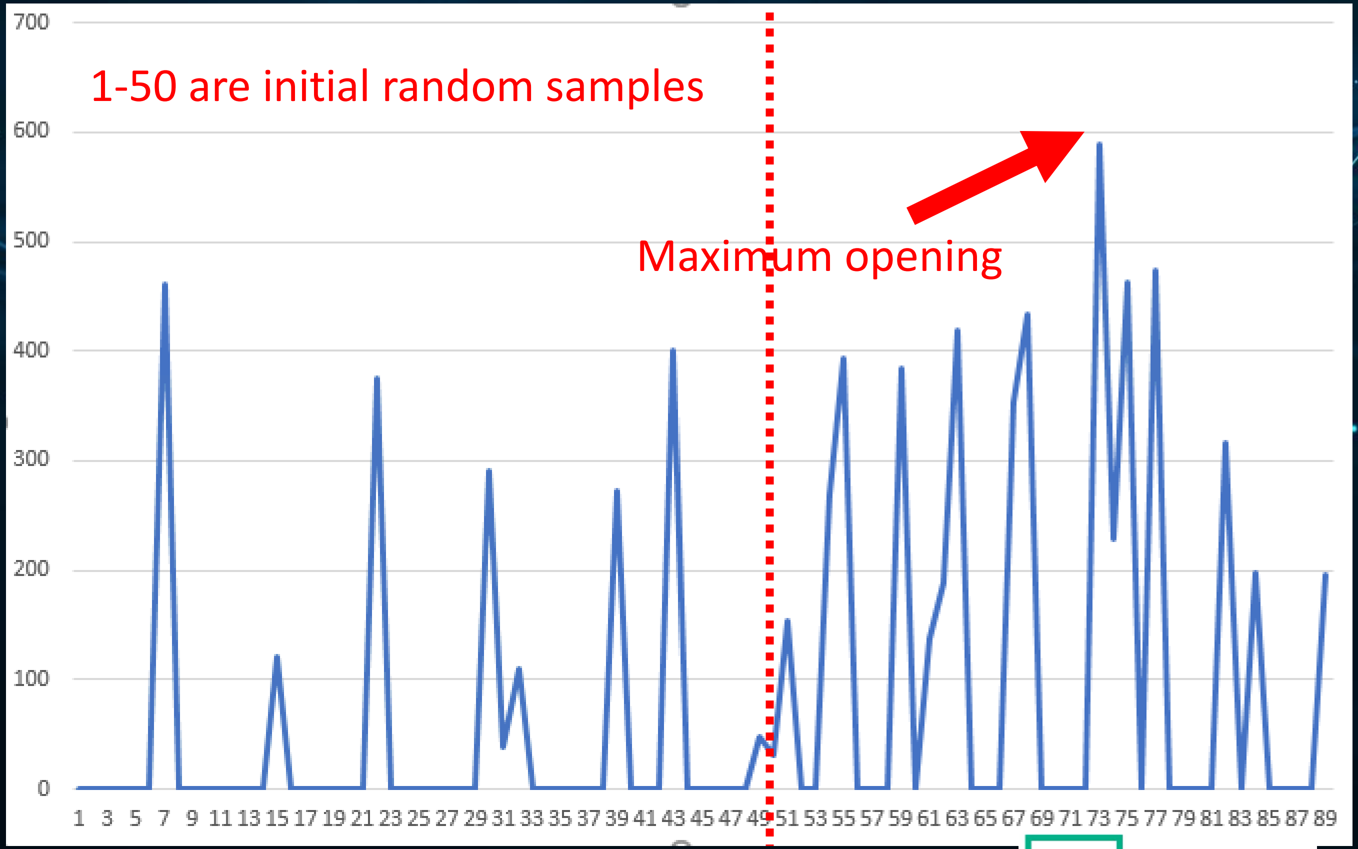
Kashyap Et al: International 3D System Integration Conference 3DIC 2023

Hewlett Packard Enterprise

Bayesian Learning for PCIe Gen 5 I/O Optimization

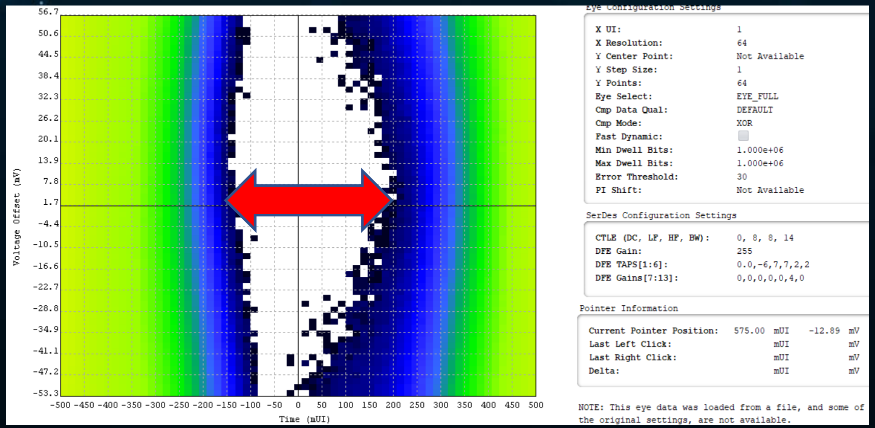
- System performance taps (up to 14) are set by Bayesian optimizer





max: 588 at 8th iteration (each iterations 3 parallel samples)

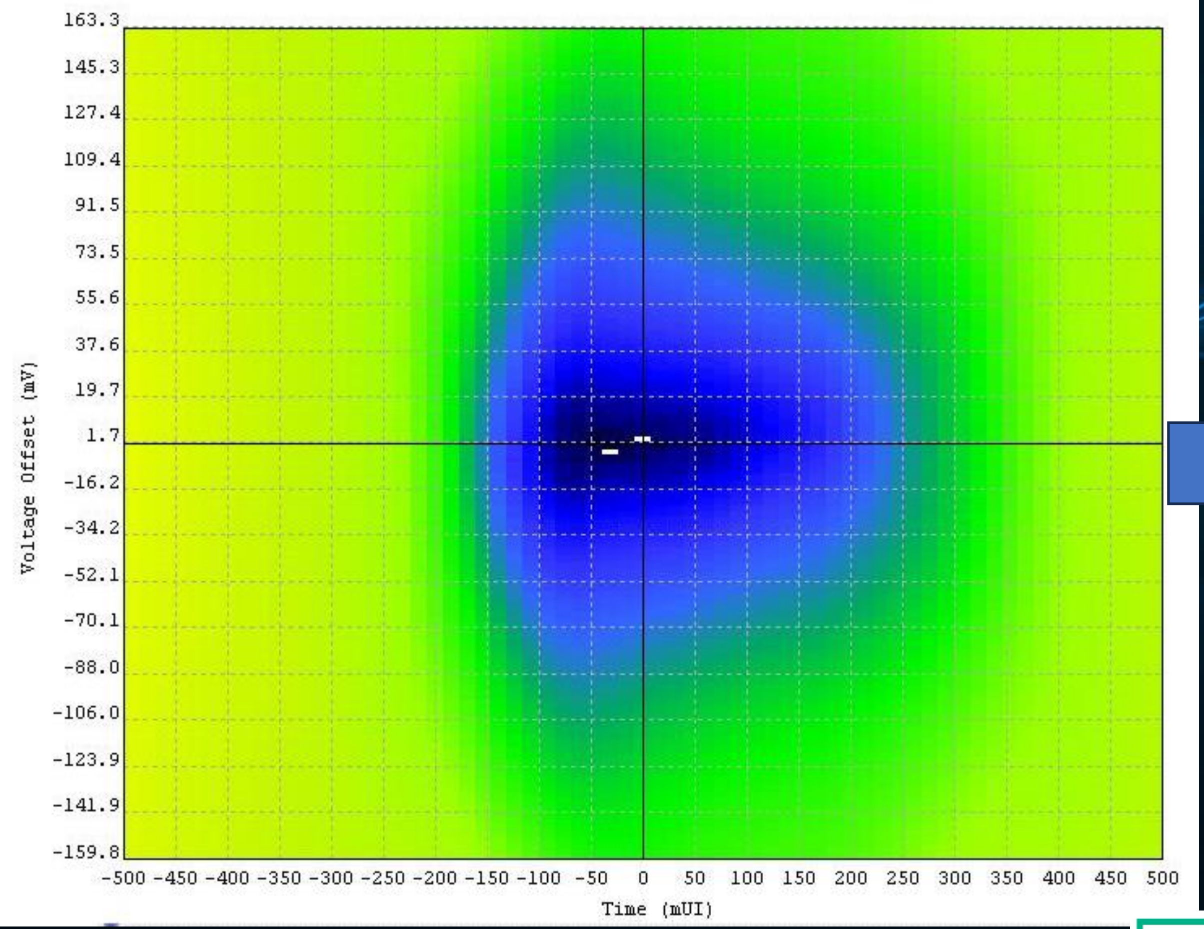
vs Auto tuning: 640



Bayesian optimization vs auto-tune PCIe Gen 5 I/O

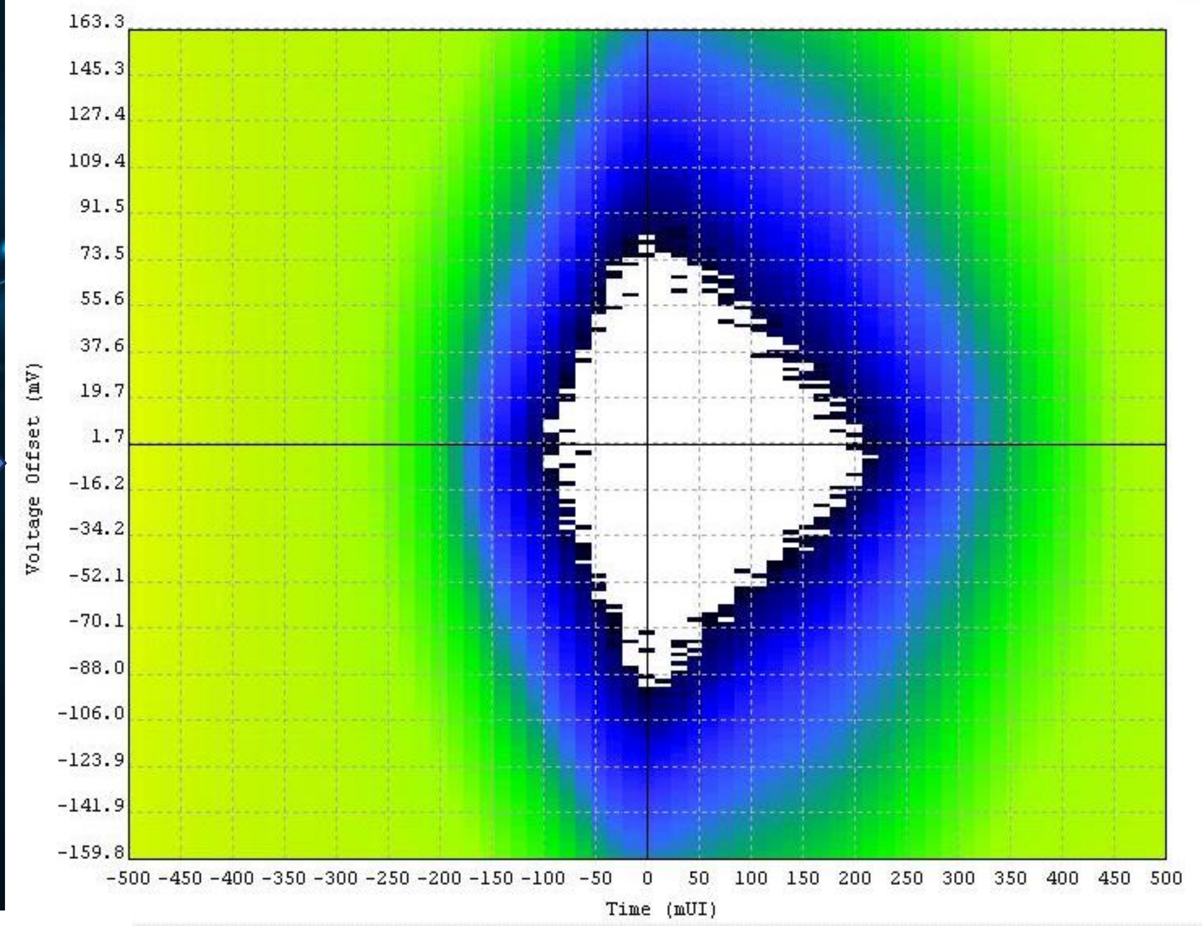
Auto tune

Eye Width: 0 mUI   Eye Height: 0 mV



Bayesian optimized

Eye Width: 287 mUI   Eye Height: 165 mV



Conclusion and future work

- cGANs can produce realistic results that can cover unseen input or output states within a design boundary
- We have moved on to transformer-based generator architectures for time series inputs (Under review)
- Application of Stable Diffusion will be our subsequent investigation if there are sufficient computing resources
- Both input and tap condition can scale to 100’s of features in a real design
- Digital twins can model the input and tap conditions in real time
  - It will open possibility of dynamic optimization of the system with the digital twin model as a parallel shadow system without disturbing the real system at runtime

Live demo of PAM4 CGAN and NRZ CGAN

Thank you

Hewlett Packard Enterprise