EXPLORING EARLY AI APPLICATIONS ON Q.ANT'S PHOTONIC NATIVE PROCESSING SERVERS

Executive Summary

As traditional Complementary Metal-Oxide-Semiconductor (CMOS) technology reaches its physical density scaling limits, performance gains are increasingly dependent on scaling out chip area, which drives unsustainable power consumption in modern data centers. This whitepaper introduces *Q.ANT's Native Processing Units (NPUs)*, a paradigm shift in computing that utilizes light instead of electricity to perform high-speed, energy-efficient operations. By modulating laser light through Photonic Integrated Circuits (PICs), Q.ANT enables "one-shot" optical computations that are particularly optimized for artificial intelligence and scientific simulations.

Author: Utz Bacher, Vice President Software at Q.ANT

Technology Overview: Photonic Computing

Photonic computing uses light instead of electricity for data processing. This technology enables very fast and energy efficient computations for workloads like artificial intelligence (AI). The photonic integrated circuit modulates laser light based on input data. Depending on the design of the PIC, interference of the light paths produces the result of a compute operation, which is captured by a detector. The optical operation is performed in one shot as light traverses the chip and modulation of the light is performed. While the optical operation is analog, a digital chip operated by a host system controls conversion between the digital and the analog domain, as well as modulation and detectors to perform the computation.
Light propagates with minimal attenuation inside the PIC. The modulation of the light is controlled through applying voltage to modulators (typically Mach Zehnder interferometers, or MZIs) and shifts the phase of the laser light signal. For this process, only minimal currents flow and, as a result, the PIC uses minimal electrical power for its operation.

Electro-Optical Modulators

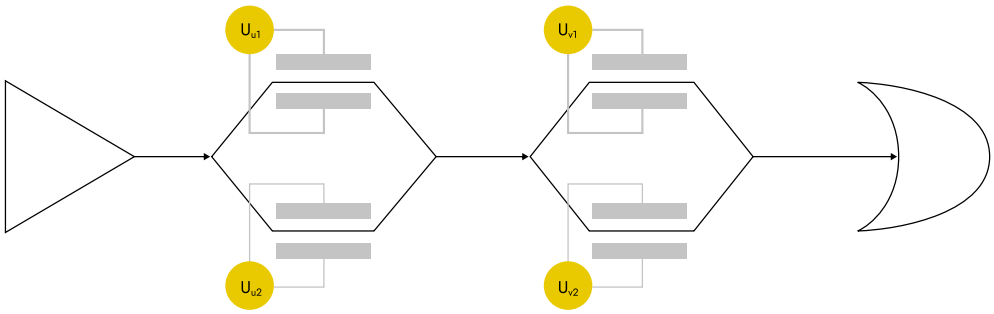

Figure 1: Schematic diagram of the components of a photonic computing system.
A product u*v is computed through according voltages applied to MZIs

Source: Q.ANT, 2026

The Challenge: The CMOS Wall

In contrast, CMOS technology is reaching density scaling limits; the operation performance per chip area is not improving significantly anymore. Instead, performance is only gained through scaling out the chip area, which increases power consumption. However, power consumption is a critical factor in many data center environments as grid capacity and power generation are scaling constraints for today's data centers.

Q.ANT Product Ecosystem

Q.ANT builds NPUs: PCIe (Peripheral Component Interconnect Express) accelerators comprising a PIC, device memory, and a digital electronic component controlling computation of data on the PIC. Several NPUs in a host server system shape a Native Processing Server (NPS). Q.ANT already offers the second generation of servers: after a prototype ("generation 0") demonstrating feasibility of the technology, a much higher performing NPU ("generation 1") was built, using clock cycles in the MHz range. The current version ("generation 2") operates at GHz speed and is able to run several compute operations in parallel. Future versions will further enhance computational performance.

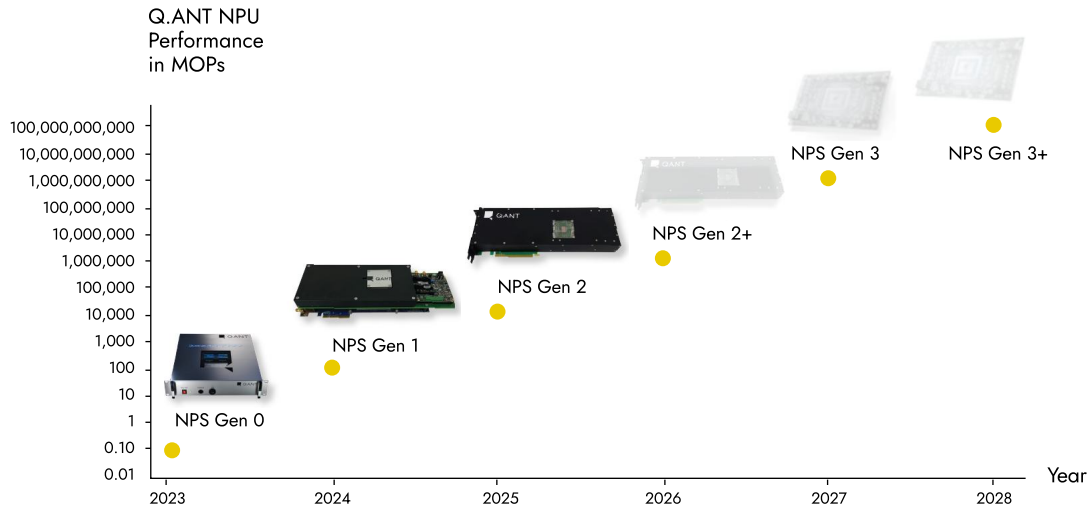

Figure 2: Native Processing Unit (NPU) Product Roadmap

Source: Q.ANT, 2026

The operations offered by Q.ANT NPUs include multiplications in several ways, including matrix multiplications performed in optics. In addition, non-linear operations are possible through appropriately controlling the modulation on the PIC, leading to a weighted sinusoidal function of a vector of input data. Future versions will introduce more computationally complex optical operations, enabling new classes of algorithms.

Technological Use Cases

Photonic computing is particularly well suited to AI use cases. Q.ANT's PICs provide the fundamental optical operations required for these use cases, allowing large fractions of the necessary computations to be addressed. This enables a variety of potential use cases, ranging from AI inference and training (e.g. large language models and generative AI) and advanced image processing (i.e. computer vision) to physics and scientific simulations (e.g. computational fluid dynamics, molecular dynamics and physics-informed neural networks).

Many of these workloads have traditionally been based on linear matrix multiplications. A great deal of research and development work has gone into these simple operations. In addition, classical hardware is not very efficient or fast at performing more complex computations – quite in contrast to Q.ANT's photonic technology.
We observe that models become more efficient when more complex and expressive operations are used for AI methods. The increased compute density of complex photonic operations results in much smaller models (i.e. they require fewer parameters); meaning less data bandwidth is required in the end. In short, complex photonic operations enable efficient AI models with a comparable level of output quality.

The software stack provided by Q.ANT to build such use cases is comprised of several layers:

- **The driver and base interface** are the foundation and expose the Q.ANT hardware through low level interfaces. Native operations are accessed through drivers controlling the NPU hardware. Python and C APIs are provided for native photonic operations and some higher level, composite operations.
- **Standard libraries and frameworks** are extended to exploit Q.ANT hardware. Higher level frameworks (such as PyTorch) are extended to use the driver and base interface to use native optical operations. This does not require proprietary programming interfaces such as NVIDIA CUDA (Compute Unified Device Architecture); instead programming on the framework level provides an agnostic way to exploit Q.ANT hardware.
- **Applications** use standard libraries and frameworks (such as PyTorch) or directly use the basic toolkit interface to execute algorithms. These algorithms can be standard AI algorithms or use advanced neural network layers built from complex operations.

Q.ANT's software stack is expanding heavily on all layers to run larger and more complex workloads and algorithms from various AI and compute frameworks with better performance.

Early Application Experiments

Digit Recognition

One of the first experiments on Q.ANT hardware has been a machine learning classic: recognition of handwritten digits based on the MNIST dataset[1]. Comprising 28 x 28 pixel grayscale images, this dataset is commonly used as a "hello world" example for image classification, requiring models to categorize inputs into one of ten digit classes. Initial experiments were based on images scaled down to 12 x 12 pixels. Meanwhile, the experiment is based on the original 28 x 28 pixels.

[1] A description of the MNIST dataset can be found here: https://www.tensorflow.org/datasets/catalog/mnist

Pre-trained weights for a two-layer fully-connected neural network were used with 784 neurons as the input, the hidden layer having 133 nodes using ReLU (Rectified Linear Unit) as activation function, and the output layer consisting of 10 nodes with a softmax activation function.

The results of the experiment provide evidence of successful execution of a neural network on Q.ANT hardware, as the network's operations have been executed on an NPU. It also shows significant performance improvements through hardware iterations and software stack optimizations: the inference process took about 30 seconds on generation 0, while execution with the latest software version on a single generation 2 NPU is orders of magnitudes faster taking merely a millisecond for one image.

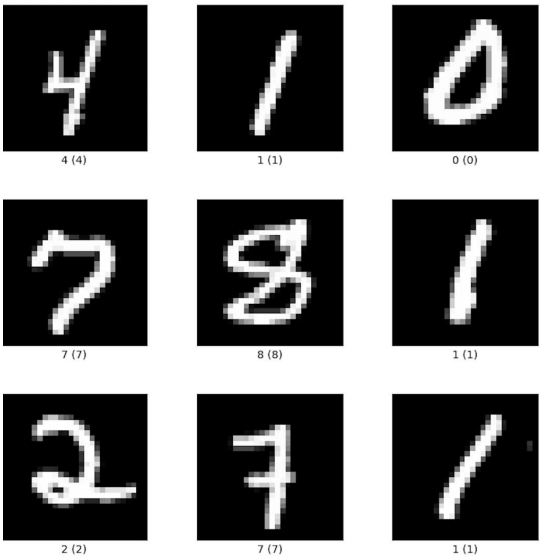

Figure 3: Exemplary representation of images and corresponding classes in MNIST dataset

Source: LeCun, Y., Cortes, C. and Burges, C.J.C. (1998) The MNIST Database of Handwritten Digits. New York, USA. Dataset license: CC-BY-SA-3.0.

Image Classification

Another conducted image classification experiment is based on a ResNet-18 convolutional neural network[2] of 5 layers with pre-trained weights. The model was tested on 224 x 224 pixel images classifying each into one of 1000 classes.

This image classification experiment based on the described pre-trained model uses the NPU for all linear (matrix) operations showing significant progress in hardware and software: from 82 seconds per classification on a single generation 1 NPU to 0.6 seconds on a single generation 2 NPU running on the latest software version.

[2] Detailed information on the used model can be found here: https://pytorch.org/hub/pytorch_vision_resnet/; https://arxiv.org/abs/1512.03385

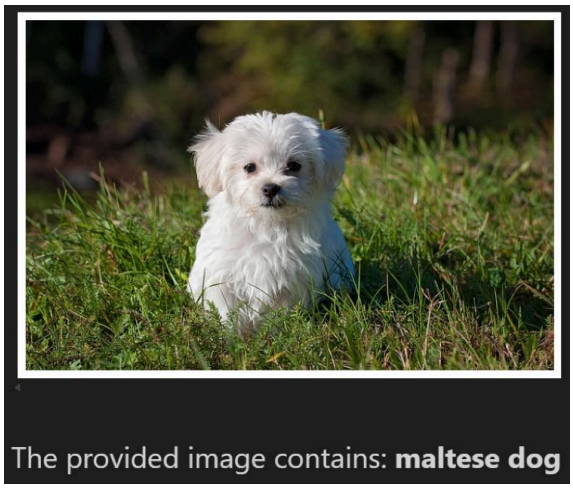

Figure 4: Exemplary result of an image classification experiment classifying animal species and objects

Source: Q.ANT, 2026

Image Segmentation

The next experiment demonstrates an image segmentation application on an NPU. This use case classifies the pixels of an input image into classes — in this case, detecting abnormalities in brain MRI scans based on shape features in the image. The utilized dataset consists of brain MRI scans and corresponding pixel segmentation masks.[3] The paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" by Ronneberger, Fischer and Brox[4] describes the details of the sophisticated, 23-layer U-net neural network executed on the NPU, which has been used in this experiment.
For this experiment, the image size has been scaled to 150 x 150 pixels. The segmentation process runs roughly 6.5 seconds on a single generation 2 NPU.

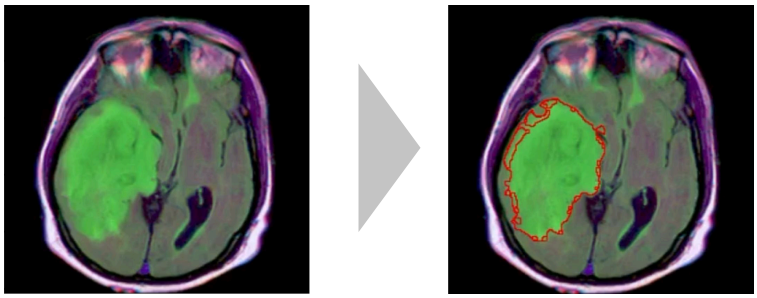

Figure 5: Exemplary result of an image segmentation example of genomic subtypes of lower-grade gliomas with shape features in CMR scans of human brains

Source: Own image based on dataset from Reddy, D., Saadat, N., Holcomb, J., Wagner, B., Truong, N., Bowerman, J., Hatanpaa, K., Patel, T., Pinho, M., Yu, F., Zhang, K., Lodhi, S., Madhuranthakam, A., Bangalore Yogananda, C. G., & Maldjian, J. (2026). The University of Texas Southwestern Glioma MRI dataset with molecular marker characterization and segmentations (UTSW-Glioma) (Version 1) [Data set]. The Cancer Imaging Archive. Dataset license: CC BY-NC-SA 4.0.

[3] Detailed information on the dataset can be found here: https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation

[4] The paper can be found here: https://arxiv.org/abs/1505.04597

Image Learning

Image learning is less of a practical experiment and more of a demonstration of a neural network's capabilities. In contrast to the previously described experiments, this use case includes the training of a neural network. The network trains to learn and predict the pixel content of an image. Training uses 128 x 128 pixel images as input data. The model's task is to predict the color (RGB values) for each pixel (X, Y coordinates). This experiment shows how well the model approximates the underlying data. It uses no specific dataset. Instead, random images, such as those recorded with a webcam, served as input.
This experiment focuses on a comparison between a classical MLP (multilayer perceptron) neural network with 5 layers and an advanced neural network utilizing learnable nonlinear operations, also with 5 layers. The advanced neural network learns a Fourier series in each node, increasing the model's computational complexity and power. This density allows a reduction of the number of required parameters, making the model significantly smaller. While the MLP neural network has more than 21,600 trainable parameters, the advanced neural network uses less than 11,500 trainable parameters. Despite the reduced number of parameters, the advanced neural network exploiting non-linearities performs much better — the training curves in figure 6 show how the advanced neural network's error decreases faster, demonstrating advantages over the MLP neural network in training speed and prediction quality.

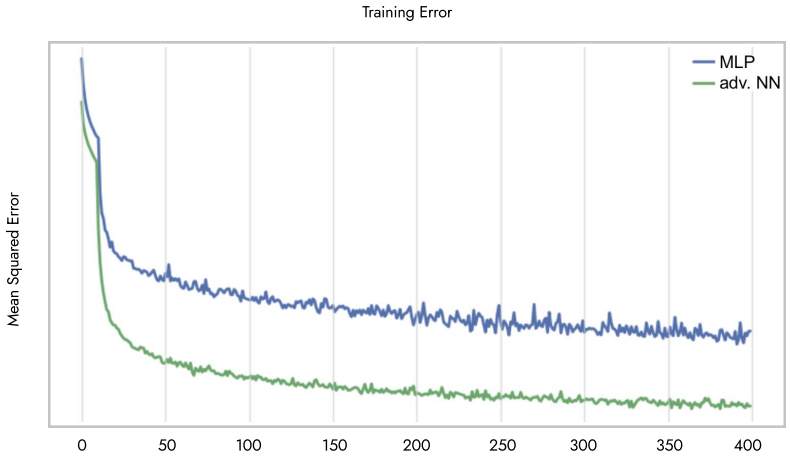

Figure 6: Comparison of the training error of the MLP and the advanced neural network across training epochs

Source: Q.ANT, 2026

The difference between the two networks is clearly visible in the quality of the predicted images (see figure 7): The advanced neural network reconstructs complex image patterns more accurately than the MLP leading to a predicted image that closely matches the ground truth (the image used for training).
This experiment shows the huge potential of using nonlinear operations for AI models, confirmed through an actual example running on Q.ANT's generation 2 NPU.

Ground truth

Multilayer perceptron neural network

Advanced neural network

Figure 7: Comparison of the ground truth with the images predicted by the MLP neural network and the advanced neural network

Source: Q.ANT, 2026. Original image by Günter Albers from stock.adobe.com.

Conclusion

Photonic computing is an emerging compute paradigm using light instead of electricity for data processing. Q.ANT's Native Processing Units are already used for various experiments with small AI workloads to demonstrate the capabilities of photonic computing today. These examples range from simple classification or segmentation tasks using neural networks with pre-trained weights to image learning using an advanced (non-standard) neural network. The hardware roadmap of Q.ANT shows a steady progression in performance and capabilities with every product generation. The software stack already enables many small standard AI use cases and is expanded for more complex workloads with better performance out of the box. The performance potential of Q.ANT's photonic computing technology can be amplified by running complex operations natively in optics. This multiplies the performance and energy-efficiency attributes of Q.ANT's products as compared to conventional computing technologies.

Contact

Q.ANT GmbH
Handwerkstr. 29
70565 Stuttgart
Germany

native-computing@qant.com
+49 711 25245-0
www.qant.com