Experimentelle Physik

Photonic non-von Neumann Processors

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Physik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Johannes Feldmann
aus Holtwick

2020

Dekan:	Prof. Dr. Gerhard Wilde
Erster Gutachter:	Prof. Dr. Wolfram H. P. Pernice
Zweiter Gutachter:	Prof. Dr. Gerhard Wilde
Tag der mündlichen Prüfung:	07.07.2020
Tag der Promotion:	

Contents

Abstract

In today's highly interconnected world with ever increasing speed of mobile networks and the simultaneous rise of machine learning and artificial intelligence, also the amount of data grows exponentially. Because conventional computers are reaching their limits in processing these large data volumes and the scaling of electronic circuits approaches its physical limits, new computing architectures must be explored to keep pace with the rapid developments in the modern information age.

In this thesis, three different approaches to tackling these challenges will be developed and experimentally demonstrated based on integrated photonics with phase-change materials. Integrated photonics provides a scalable platform to reliably control the flow of light on a chip, whereas phase-change materials, which change their ability to absorb light depending on their phase of matter, allow for an active and low energy tuning element in the photonic circuits. The basic building block of the combination of the two, a phase-change cell, is an optical waveguide with a phase-change material deposited on top. The intensity and the phase of light travelling down the waveguide can be altered by adjusting the composition of amorphous and crystalline fractions of the phase-change material via optical excitation through the waveguide. Because switching between the phases is a reversible and non-volatile process and the state of the material is preserved without further energy consumption, the integration of phase-change materials with photonics holds promise for low power data processing.

In the first part of this thesis, the phase-change cell is employed for arithmetic in-memory computing in analogy to a traditional abacus. In-memory computing, which means that the calculations are directly carried out in the memory, is used to circumvent the von Neumann bottleneck, which describes the limitations of conventional computer architectures due to the necessity of data transport between the physically separated memory and arithmetic unit. Exploiting the multilevel capabilities of phase-change materials, the fundamental calculation operations addition, subtraction, multiplication and division are implemented directly in base ten. The basic arithmetic unit is scaled to larger photonic circuits in a waveguide crossing structure, enabling random-access of the phase-change cells and operation with multiple digits.

The second part of the thesis focusses on implementing a scalable all-optical neural network. Neural networks are computational structures that mimic the behaviour of biological brains, which outperform conventional computers in terms of speed and energy efficiency by several orders of magnitude in cognitive tasks as speech- and pattern recognition. In a first step a single neuron,

the building block of neural networks, with synapses based on phase-change materials and a non-linear activation function is developed and experimentally applied to a basic pattern recognition task. Using a feedback loop the optical neuron can be trained (i.e. adjusting the synapses) in a supervised as well as in an unsupervised manner. In a second step a scalable architecture for combining many of the optical neurons and implementing multilayer networks is developed. A single layer of the proposed neural network architecture is fabricated and experimentally employed to recognize four different letters. Furthermore, the potential of larger neural networks and neurons with higher numbers of synapses is studied in simulations based on the experimental data.

In the third part of the thesis a photonic tensor core for executing matrix multiplications on a photonic chip is demonstrated. Matrix multiplications are the computationally expensive tasks in processing neural networks and the limiting factor for processing speed in conventional architectures. The presented photonic implementation is based on a waveguide crossing array with directional couplers and phase-change cells as the matrix elements. By avoiding resonant elements in the matrix unit, a high spectral range of wavelengths can be exploited for operating the multiplier with many input vectors in parallel based on wavelength-division multiplexing. Due to the parallelization and the high modulation and detection bandwidth available in the optical domain, the photonic tensor core holds promise for data-rates far exceeding existing electronic technologies.

List of Abbreviations

AI	Artificial intelligence
ALU	Arithmetic logic unit
ASIC	Application specific integrated circuit
CMOS	Complementary metal oxide semiconductor
CNN	Convolutional neural network
CPU	Central processing unit
DAC	Digital-to-analogue converter
DKS	Dissipative Kerr soliton
EBL	Electron-beam lithography
EDFA	Erbium-doped fibre amplifier
EOM	Electro-optic modulator
ER	Extinction ratio
FEM	Finite-element method
FPGA	Field programmable gate array
FSR	Free spectral range
GPU	Graphics processing unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ITO	Indium tin oxide
ITU	International Communication Union
MAC	Multiply-accumulate
MIBK	Methyl isobutyl ketone
NN	Neural network

PCM	Phase-change material
PIC	Photonic integrated circuit
PMMA	Poly methyl methacrylat
PTC	Photonic tensor core
PVD	Physical vapour deposition
ReLU	Rectified linear unit
ResNet	Residual neural network
STDP	Spike timing dependent plasticity
TPU	Tensor processing unit
VOA	Variable optical attenuator
WDM	Wavelength division multiplexing

1 Chapter 1. Introduction

In February 2011 the computer program 'Watson' developed by IBM competed in the famous quiz show 'Jeopardy!' with two record-winning candidates in three rounds and defeated his human competitors with a final score of $77.147 versus $24.000 and $21.600 [1]. 'Watson' was one of the first demonstrations of artificial intelligence that was capable of analysing natural language, which was special because it included understanding the meaning of words, linking different topics, distinguishing and pondering between ambiguities in the questions and quickly searching through vast amounts of information to find the correct answer, marking a major breakthrough of a rising technology [2–4].

Since 2011 much progress was made in the area of machine learning and artificial intelligence is considered a key enabling technology of the 21st century. In another, equally acknowledged but more recent competition between a machine and a human, 'AlphaGo', a computer program developed by the British company Google DeepMind lined up with Lee Sedol in the strategy board game 'Go'. 'AlphaGo' defeated Lee Sedol, who was considered the best 'Go' player in the world, with 4:1 in 2016. Because its complexity and the vast number of moves, a game of 'Go' can, other than for example chess, not deterministically be calculated with standard algorithms so that 'AlphaGo' was based on machine learning algorithms using deep neural networks training itself in self-play games and learning from human expert games [5].

Both examples show the huge potential of artificial intelligence (AI), which is nowadays widely applied in various areas of everyday life. Mobile phones process spoken language and are capable of face recognition [6]. Internet search engines distinguish between relevant and irrelevant information for a given search query within milliseconds and classify images by their content. More recently, AI lies at the heart of autonomous driving and enables new technologies in medical diagnostics [7, 8]. All these tasks have in common that they handle vast amounts of data bringing conventional computers and processors to their limits so that for example speech and face recognition tasks on mobile phones are often processed outside the device in cloud-based systems [9]. Applications using artificial intelligence currently generate huge amounts of data in the range of

80 exabyte per year which is expected to increase by more than an order of magnitude to 845 exabytes by 2025 [10], data volumes impossible to handle with conventional computers. Whereas shrinking integrated electronic circuits and fitting ever more transistors on a single chip using advanced fabrication techniques could hold up with the increasing demand for fast data processing for a long time, Moore's law stating that the number of transistors on a microprocessor doubles every two years is slowing down [11–13]. Equally important, the constancy of the power density while decreasing the transistor size (and increasing its speed) often referred to as Dennard scaling [14] is reaching its limits because of increased tunnelling (leakage) currents through the gates of the transistors, drastically downgrading the energy efficiency of smaller electronic circuits [15].

To overcome the challenges of handling the ever-increasing amounts of data in the modern information age while maintaining a reasonable power budget, new approaches to computing must be envisioned and new technologies developed. In the cognitive tasks present in the area of artificial intelligence, as for example speech- and face recognition or detecting and classifying objects in images, human brains (and biological brains in general) outperform classical computers by orders of magnitude in speed and energy efficiency, being an inspiring model for new developments to be based on. As an example, compared to a real mouse a personal computer simulates a mouse-scale cortex with 2.5 million neurons 9000 times slower while consuming 40 000 times more power, showing the huge potential of building brain inspired processors [16]. The key to the superior performance of biological brains compared to conventional computers lies in fundamentally different ways of processing the data. Conventional processors, which easily outperform human brains in arithmetic tasks, are based on the so called von Neumann architecture that operates in a serial way (command after command) and has its memory storing data and the actual program physically separated from its central processing unit (CPU) leading to a data transfer bottleneck limiting the speed of the computation. In contrast, a biological brain processes data in a highly parallel way without separation of processor and memory. Coming back to the competition between 'AlphaGo' and Lee Sedol this leads to a power consumption of 77 000 W for the computer versus 20 W for the human brain, clearly pointing towards hardware implementations of processors inspired by the brain and alternative approaches to computing.

Although CPUs and graphics processing units (GPUs) have greatly improved and are heavily employed in machine learning as hardware accelerators speeding up matrix multiplications, which are the key operation in many implementations of cognitive algorithms like neural networks, they still suffer from the downsides of the von Neumann architecture, driving the research on more specialized hardware as for example the tensor processing unit (TPU) from Google, which is an application specific integrated circuit (ASIC) specialized on executing matrix multiplications.

More recently, also optical approaches for applications in AI are emerging. Fast and long-distance data transfer is carried out in the optical domain for many years and is the foundation of the fast global internet because it enables data transport at the speed of light with low loss, high modulation speeds and parallelization (multiplexing). To increase computational speed of

supercomputers, optical links are nowadays also employed for communication between electronic chips and even in inter-chip data transfer, massively increasing the data rates [17, 18]. To date, most of the computation is still carried out in the electrical domain raising a need for energy and time consuming electro-optic conversions. All-optical approaches can potentially circumvent these conversions and hold promise for faster and more efficient data processing overcoming the limited bandwidth of electronics. An example of a photonic integrated circuit implementing an all-optical neural network was published in 2017 by Shen et al. [19] demonstrating vowel recognition on a silicon photonics chip, indicating the potential of optical computing at high speeds of up to 100 GHz compared to electronic processors operating at clock speeds of a few GHz. A downside of this approach is the use of thermal heaters for tuning the optical elements (Mach-Zehnder interferometers), resulting in an increased power budget. One way to conquer the constant power consumption by the heaters is to combine the photonic structures with non-volatile phase-change materials, leading to the platform of phase-change photonics, which is the basis of this thesis. Phase-change materials stand out due to the fact that they significantly change their optical and electrical properties based on their phase of matter and preserve their current state without further energy consumption, a feature which is widely used in optical data storage on DVDs or Blu-Ray RE [20–22].

In this thesis photonic integrated circuits (PICs) are combined with a phase-change material (PCM) adding an active component that can be used for data processing and computing. Based on previous research investigating the basic properties of the fundamental building block of phase-change photonics (an optical waveguide with evanescently coupled phase-change material), its optical properties, switching between its phases, endurance and energy consumption [23, 23–25], this thesis is focused on exploring and experimentally implementing unconventional all-optical approaches to computing that tackle the challenges of the increasing amounts of data in light of artificial intelligence and machine learning.

In a first step, a basic all-optical arithmetic unit similar to an abacus is developed, showing the potential of in-memory computing approaches and low power all-optical processing with phase-change materials. Other than conventional computers that operate binary, the proposed all-optical abacus is capable of working directly in base ten or other arbitrary bases at GHz speeds and a way of scaling from a single cell to a larger network is shown based on a waveguide crossing array.

The second part describes and experimentally demonstrates an all-optical implementation of a neural network capable of simple pattern recognition tasks. The proposed architecture is scalable to multiple layers and massively parallelizes the computation. Besides being trained by an external supervisor, the individual neurons of the network have the capability of self-learning, enabling the system to extract unknown features without input from the outside.

In the last part of the thesis a photonic tensor core is presented, a hardware accelerator for matrix multiplications capable of high-speed processing and compute densities achieved by paral-

lelization potentially outperforming current state-of-the art electronic implementations by orders of magnitude.

Chapter 2.
2 Theoretic foundations and motivation

In this chapter the background information for this thesis is explained in detail. After introducing the conventional processor architecture and comparing it to non-von Neumann processors – especially in-memory and neuromorphic computing – the concept of artificial neural networks, its applications and significance are outlined. In the last parts the experimental background of this work, integrated nanophotonic circuits and phase-change materials, are covered.

2.1. Non-von Neumann computing

Given the increasing amount of data in the modern information age [26], the demand for faster and at the same time more efficient computers and processors is rapidly growing. Opposed to this, Moore's law describing that the transistor density on an electronic chip doubles every two years [12] is slowing down, raising the need for alternative approaches to signal processing and computing. In the following, the conventional computer architecture with its strengths and weaknesses will be described, serving as a motivation for this work and leading to the development of alternative approaches as in-memory [27, 28] and neuromorphic computing [19, 29–34].

2.1.1. Conventional processors

The conventional processor architecture is the so called von Neumann architecture named after the mathematician John von Neumann, who developed the foundations of modern computers in the 1940s [35]. This architecture essentially consists of a central processing unit (CPU), a bus-system, memory and an input/output (I/O) unit. The main components are the CPU being the heart of the arithmetic processing and the memory, in which the program instructions and data are stored. These two entities are physically separated and connected by a bus-system shuffling the data between CPU and memory and to the I/O unit. Both these components are fast on their own and work with high precision but the data transfer between them leads to the so called von Neumann bottleneck limiting the actual processing speeds [36]. The execution of

a program is therefore limited by the transfer rate for instructions and data between memory and CPU. Another processor type, the Harvard architecture, widens the bottleneck by having separate memories for the program and the data. This way, loading instructions and data can be executed simultaneously and the data transfer is separated on two different paths [37, 38]. The Harvard architecture is for example widely employed in electronic microcontrollers such as the Atmel Atmega 128 [37, 39].

Conventional computers work in a serial way, a program is executed step by step. Despite making it easier to write deterministic programs, this limits the efficiency especially for tasks including huge amounts of data as present in cognitive tasks like pattern and speech recognition. Parallelism can usually only be achieved by physically replicating the processor employing multiple cores [27, 40]. Examples for electronic high-performance multi core processors are based on the Haswell or Skylake architecture from Intel with up to 18 and 28 parallel cores [41, 42].

As mentioned before, processors employed for artificial intelligence have to perform computationally very expensive tasks. The main operation that stands out for all cognitive tasks are matrix multiplications, which consist of many addition and multiplication operations. Graphics processing units (GPUs) achieve a certain degree of parallelization by exploiting many parallel arithmetic logic units (ALUs), which are the basic building blocks of conventional CPUs. However, GPUs are still based on the conventional architecture and therefore suffer from the von Neumann bottleneck.

To minimize the communication between ALUs and memory, special purpose processors have been developed. Two prominent examples are the TPU from Google [43] and Microsoft's Brainwave project based on a high-performance field programmable gate array (FPGA) [44]. The TPU is an application specific integrated circuit (ASIC) with its heart being a so-called systolic array. This architecture circumvents the von Neumann bottleneck by arranging individual ALUs in an array of size 256×256 that directly pass their computational output to the neighbouring ALUs, effectively implementing a matrix multiplication and minimising the number of accesses to the memory [45]. In Microsoft's FPGA the deep learning algorithms are directly programmed into the hardware and can be reprogrammed almost in real time.

2.1.2. Unconventional processors

To circumvent the von Neumann bottleneck, i.e. removing the need to transfer data between memory and processor, and benefit from massively parallel data processing, radically new approaches must be envisioned. One route of reaching high parallelism is quantum computing, taking advantage of the superposition of physical states arising from quantum mechanics potentially outperforming classical computers on computationally expensive tasks as for example factorization by several orders of magnitude [46]. However, quantum computing is still in its infancy and has only recently proven to beat conventional architectures on an especially designed

problem [47] and is based on completely new paradigms and technological platforms.

A promising alternative and classical approach to overcome the limitations of the von Neumann bottleneck that is based on existing technology and established fabrication procedures (i.e. complementary metal oxide semiconductor (CMOS)) is in-memory computing [27, 28]. Herein, data storage and computation are achieved at the same physical location [48, 49]. So called memcomputing devices that employ memory circuit elements for computation [50, 51] can for example achieve massively parallel processing in memristive networks [52] or binary logic [53, 54] and rely on the fact that the states of the individual memory elements depend on their history – information about previous calculations is stored in them and influences future processing thus removing the need to transfer data from and to an external memory.

A special kind of in-memory computing build for artificial intelligence is called neuromorphic computing, gaining more and more attention in the research of the past years [19, 55, 56]. Neuromorphic processors are built following the example of biological brains and are particularly useful for cognitive tasks as pattern and speech recognition. Mimicking the highly parallel structure of the brain, neuromorphic hardware architectures can potentially outperform conventional computers in classification tasks by several orders of magnitude and are already heavily applied behind the scenes on mobile phones [6]. They lie at the heart of modern day challenges as autonomous driving [6, 8] and video-stream analysis and can also be used in medical diagnostics [7, 57]. Electronic hardware implementations of neuromorphic processors are for example SpiNNaker (Manchester University) [58] and TrueNorth (IBM), implementing spiking neural networks with spiking rates similar to biological neurons (kHz). TrueNorth exploits the scalability of the CMOS platform and the implementation of one million neurons with 256 million synapses has already been demonstrated [59]. Further neuromorphic architectures are HICANN from Heidelberg University [60] or Neurogrid from Stanford University [16].

Recently, also photonic processors are more and more considered to be a promising technology for neuromorphic computing due to two main reasons. Firstly, high-speed photonic links readily overcome the data-transfer problem present in electronic circuits that arises from a high power consumption for charging and discharging electronic interconnects. Secondly, photonic circuits enable multiplexing signals in the same optical circuit potentially leading to a high degree of parallelism and therefore very high computational densities. A variety of different approaches to photonic neuromorphic computing are currently investigated such as reservoir computing, multiwavelengths- or coherent neural networks [61]. Photonic reservoirs map the input data to a higher dimensional space allowing for a much simpler classification of the data using linear regressions [62, 63]. Coherent photonic networks exploit fully optical matrix multiplications with high modulation speeds and high power efficiency using coherent light [19] and multiwavelengths approaches exploit wavelength multiplexing capabilities to gain highly parallelized processing [64].

It should be noted that brain-inspired software can also be developed on conventional hardware based on the von Neumann architecture but the term 'neuromorphic' in this work is related to

physical implementations of neuromorphic processors in hardware only.

2.2. Neural networks

Opposite to the outstanding performance and accuracy of conventional computer architectures in arithmetic processing, their ability to process huge amounts of data like video streams used for steering autonomous vehicles, speech processing and image analysis especially on mobile devices is rather limited [6]. Clear evidence for this is that to date most of the speech and pattern recognition tasks used in everyday life are still not performed on local computer chips [6] as for example in mobile phones but are externally analysed through cloud services [65]. When it comes to classifying objects on different scales in an image or video stream, the human brain is the most efficient tool to do so. As an example, Kuzum et al. estimated that simulating five seconds of brain activity on a (conventional) supercomputer takes 500 s and consumes 1.4×10^6 W of power [56], whereas the average power consumption of the human brain is only approximately 20 W promising advantages of several orders of magnitude in speed and energy efficiency for hardware implementations of neuromorphic processors for cognitive tasks. Up to now most of the neural network implementations are software-based and suffer therefore from the downsides of conventional hardware. However, huge efforts have been made optimizing the network structure so that they can beat human performance in some of the standard classification databases.

The main concept behind neuromorphic computers are neural networks. The following sections explain the general layout of such structures, different training techniques and their applications.

2.2.1. Layout and operation principle

From the outside a neural network (NN) can be viewed as a black box with several inputs taking information from the real world regarding a certain problem and several outputs revealing the result of the computation. The input to a neural network can for example be an image and the output can be information about objects that have been identified in the image. In light of autonomous driving, traffic signs, road users and road markings can be extracted from a camera input stream.

A NN is typically arranged in a layered structure with an input layer taking the physical input, several hidden layers and an output layer from which the results can be read (see Fig. 2.1a). Each layer is composed of a number of neurons (see Fig. 2.1b), which are the smallest building blocks of neural networks. The individual neurons receive n inputs $(x_1, x_2, \ldots, x_n)$ that are first summed and then sent to the activation unit. The synapses $(w_{1j}, w_{2j}, \ldots, w_{nj}$; j is the neuron index) are the connections to the previous layer and weigh the signals coming to the neuron. The weighted sum is calculated as $a = \sum_{i=1}^{n} w_{ij} x_i$ and is called activation energy. This activation energy is fed to the activation function $f(a) = y_i$ that decides whether an output signal is generated or not. If the summed input signal exceeds a certain threshold, the neuron generates an output signal and

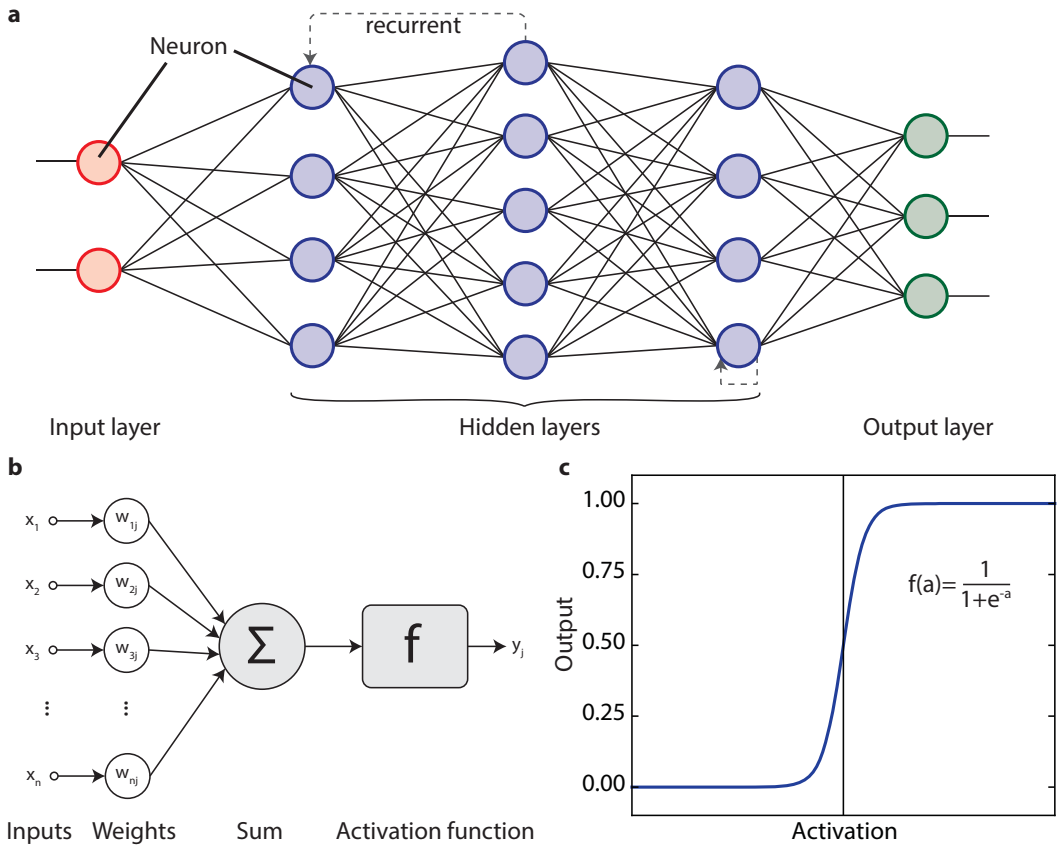

Figure 2.1.: Layout of a neural network. **a)** Layered structure of a neural network. Via several hidden layers an input layer that receives the input data is connected to an output layer that gives the computational result. Each layer consists of a number of neurons that take inputs from the previous layer's neurons and distribute their output to the next layer. Recurrent connections are used for time dependent inputs. **b)** Basic principle of a neuron. The input vector x is multiplied by a weight vector w and summed up. The weighted sum is sent to the activation function f that generates the output signal. The index j denotes the position of the neuron within a layer. **c)** A typical activation function is the sigmoid function that is 0 below a certain threshold and 1 if the activation (weighted sum of the inputs) exceeds the threshold.

sends it to the next layer of neurons. Mathematically the threshold function is often resembled as the sigmoid function of the form (see Fig. 2.1c)

$$f_{\text{sigmoid}}(a) = \frac{1}{1+e^{-a}}. \tag{2.1}$$

Below a certain value of the activation energy no output is generated, above the threshold the output is limited to a maximum value. The sigmoid function is advantageous for software implementations of neural networks as it is continuously differentiable, a property desirable for training of the network (see Sec. 2.2.2). Other common activation functions are the hyperbolic tangent, softmax (often used for output neurons) or the swish function [66–69]. In the last years the rectified linear unit (ReLU) function, which was initially introduced to neural networks by

Hahnloser et al. with a biological motivation, became more popular as it was shown that training of deep neural networks is more successful with the ReLU-function than with the classical sigmoid function [70]. The ReLU function is defined as

$$f_{\mathrm{ReLU}}(a) = \max(0, a) . \tag{2.2}$$

Depending on how the connections between the neurons are implemented, different types of networks can be identified. In feedforward networks the information is processed layer by layer from input to output and the information only travels in one direction. In recurrent neural networks instead, neurons can also have connections to previous layers or to neurons of the same layer. This way also temporal patterns in the input data can be observed which is especially helpful for speech recognition and any time dependent problems.

2.2.2. Learning techniques

Each neural network must be trained in order to find suitable weights to accomplish a certain task. A neural network structure can hold hundreds of layers and several millions of parameters (the weights/synapses), making the training a computationally expensive optimization problem [6]. In the past years a lot of efforts have gone into reducing the number of parameters of a NN while keeping its prediction accuracy. Using for example residual neural networks (ResNets) the number of layers can be increased by implementing short-cuts between layers (residual connections). In the training process unused layers can this way be switched-off tuning only one parameter [71].

Generally, two different learning techniques can be distinguished: supervised and unsupervised learning. Supervised learning is applicable in all situations where a set of training data with pairs of input and output patterns is available. It is for example employed for classification tasks or digit recognition with datasets of pre-classified images on hand. The most common algorithm for training neural networks is the backpropagation algorithm [72, 73] but besides this many other iterative solvers such as genetic algorithms can be applied [74]. All these algorithms have in common that they try to find the global minimum in the solution space. Minimizing the prediction error is therefore not a trivial task given the huge number of parameters. Indeed the computational power needed for training neural networks has been doubling every 3.4 months in the past six years [75].

To achieve the best performance, the learning parameters have to be carefully tuned in order for the algorithm not getting stuck in a local minimum. In case of the backpropagation algorithm each training cycle consists of three steps. In a first step a training pattern (i.e. an image to be classified) is presented to the network and propagated through it (feed forward). In the second step the networks output is compared to the expected outcome and the deviation is considered the error of the prediction and given by a loss function. In the last step the error is propagated backwards through the network, while changing the weights slightly depending on their influence

on the error effectively reducing it for the presented input pattern. Because the change of the weight is calculated via the derivative of the loss function with respect to the weight parameters, it is also called a gradient descent method. These three steps are repeated with a certain training set until the desired accuracy of the prediction is achieved.

If no training data is available and an unknown input stream has to be analysed, unsupervised learning techniques can be exploited in which the neural network trains itself and adapts to repeating patterns in the data. This can be accomplished employing a certain learning rule as for example spike timing dependent plasticity (STDP) that is related to the biologically inspired paradigm by D. Hebb: 'neurons that fire together, wire together' [76]. All weights that contribute to triggering an output pulse in a neuron are strengthened, whereas the others are weakened. This way the neural network can adapt to certain repeating patterns in the input data.

2.3. Nanophotonic circuits

The processor architectures discussed in this thesis are based on the nanophotonics platform. Nanophotonic circuits are the optical counterpart to electronic integrated circuits. Instead of guiding electrons through copper wires on semiconductor chips, photons are guided through optical waveguides. As the miniaturization of electronics has significantly slowed down because of the increasing heat dissipation and leakage current when reducing the element size especially of transistors and their interconnects [40, 77], integrated nanophotonic circuits gain more and more interest. Similar to electronic circuits, which can be built from discrete components like resistors, transistors and capacitors on a breadboard, free space optics consisting of lenses, filters and optical detectors can be shrunk to the nanoscale on a photonic chip. Besides the miniaturization nanophotonic circuits also offer a significantly improved stability and are less sensitive to accidental misalignment compared to free space optics [78–80]. This is especially promising for information processing and quantum optics but also serves as a platform for new approaches to computing as described in this work, benefiting from the high speed and bandwidth of optical circuits [40, 81, 82]. Compared to electronic circuits that suffer from heat accumulation because of the electrical resistance of the connections, optical waveguides are transparent for light and therefore do not heat up. Wavelength multiplexing techniques, which allow to use the same optical waveguide to be shared by several signals at the same time without interference, inherently offer the ability of highly parallel signal processing and computation [83–85]. Multiplexing based on different modes or polarizations add additional degrees of freedom to furthermore increase the capacity of optical interconnects [86]. As already mentioned in Sec. 2.1.1, a limiting factor for conventional electronics is the data transfer between processor and memory. High speed data transfer over long distances is well established in the optical domain employing optical fibres transmitting data at the speed of light being the foundation of high-speed internet. More recently, optical interconnects are also considered for inter-chip [18] and even intra-chip [17] communica-

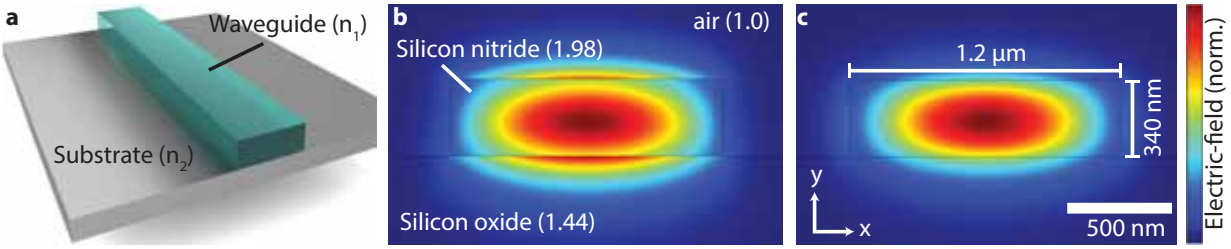

Figure 2.2.: Dielectric waveguides. **a)** Ridge waveguide. Light is guided in the higher refractive index medium. **b)** FEM simulation of a silicon nitride waveguide at a wavelength of 750 nm — TM-like mode. **c)** TE-like mode.

tion. Before processing the optically transmitted data in a computer, the optical information has to be converted to an electronic signal. By operating directly in the optical domain, nanophotonic circuits and processors are a promising way of circumventing this data transfer bottleneck and removing the need for time and energy consuming conversions between the electrical and optical domain.

The biggest disadvantage of photonic integrated circuits in comparison to electronic circuits is the footprint of the devices. Conventional on-chip waveguides are limited in their size [87, 88], the minimum radius of bends on-chip is determined by the wavelength, geometry and refractive index of the material used [88]. A typical width for on-chip waveguides in Si_3N_4 is for example around 1 µm with a height of 340 nm when operated at telecom wavelengths (1.55 µm) with minimum bend radii of 20 µm − 30 µm. By using higher refractive index materials as for example silicon, the size can be reduced to smaller dimensions (approx. 500 nm × 220 nm and bend radii down to 5 µm). Further miniaturization can be achieved by using shorter wavelengths for example in the visible range, mainly limited by the transparency windows of the materials employed.

A way to overcome the diffraction limit and building devices with even smaller footprint are plasmonic waveguides [88]. Typical implementations rely on surface-plasmon polaritons [89] that are quantized oscillations of the charge carrier density and are localized to a metal-dielectric interface. However, to date these structures are too lossy to be implemented in larger nanophotonic circuits. With advanced fabrication techniques also a stacked architecture with several waveguide layers can be envisaged [90–92] making use of the third dimension. This way also losses and crosstalk as caused by waveguide crossings on a chip can effectively be decreased.

Nanophotonic circuits are based on dielectric waveguides that enable confinement and guidance of light on a chip analogous to electronic wires in the electric domain. Fig. 2.2a) shows an integrated ridge-type waveguide. The main principle for waveguiding is a local variation of the dielectric constant to which the picture of ray optics offers an intuitive understanding based on total internal reflection. Light propagating in medium 1 with a refractive index of n_1 is, when incident at an angle θ_{in}, reflected at the interface to a second medium with refractive index of n_2 according to Snell's law $n_1 \sin(\Theta_{\text{in}}) = n_2 \sin(\Theta_{\text{out}})$ under the angle θ_{out}. The critical

angle $\theta_{\text{crit.}} = \arcsin\left(n_1/n_2\right)$ defines the minimum angle for which total internal reflection occurs so that the light is fully reflected into medium 1. This way light can be confined and guided inside a waveguide with higher refractive index than its surrounding media. For a more detailed understanding of the properties of the guided light the electromagnetic theory based on Maxwell's equations needs to be considered [93]. As the solution to Maxwell's equations for most of the geometries is not a trivial problem, optimization algorithms for example based on the finite-element method (FEM) can be employed to calculate the field distribution of a certain waveguide geometry.

Fig. 2.2b) and c) show the simulated modes of the silicon nitride ridge-type waveguide as fabricated throughout this work (simulated with COMSOL Multiphysics and FEM). Two types of modes can be discriminated based on the distribution of the electric field, that are the transverse-magnetic(TM)-like modes (see Fig. 2.2b) and the transverse-electric(TE)-like mode (see Fig. 2.2c). For TM-like modes the magnetic field component in the direction of propagation is almost 0, whereas on the opposite for the TE-like mode the electric field component in this direction vanishes. In Fig. 2.2 it can be seen that the normalized electric field is continuous in y-direction for the TE-like mode and continuous in the x-direction for the TM-like mode. The mode profile for a certain waveguide at a fixed wavelength is solely determined by its geometry and the refractive indices of the media involved. As can be seen in the mode simulations, the electric field is confined in the centre of the silicon nitride but a significant part also leaks out into the air. This can be exploited for coupling with functional structures as for example phase-change materials.

2.4. Phase-change materials

In order to exploit the nanophotonic platform for computing, an active component in addition to the passive waveguide structures is necessary to tailor the amplitude and phase of the propagating electromagnetic waves. In this work, phase-change materials (PCMs) are used as such an active tuning element. Phase-change materials are materials that significantly change their electrical and optical properties when undergoing a phase transition [21, 22, 94]. For example, the resistivity or the refractive index of typical phase-change materials changes by more than an order of magnitude upon transition between the amorphous and crystalline state. PCMs were originally discovered by Stanford R. Ovshinsky who observed electrical switching phenomena in the 1960s [95] and gave first explanations for the microscopic mechanisms. After initial attempts to developing a commercial product – an electronic memory based on PCMs – in the 1970s [96], finally the optical properties of phase-change materials lead to a huge commercial success of rewritable data storage as CDs, DVDs and later Blu-Ray RE [94, 97, 98]. All these storage media make use of the change in refractive index between the amorphous and crystalline state, which can be observed in a simple reflectivity measurement for example in a DVD drive. It is important

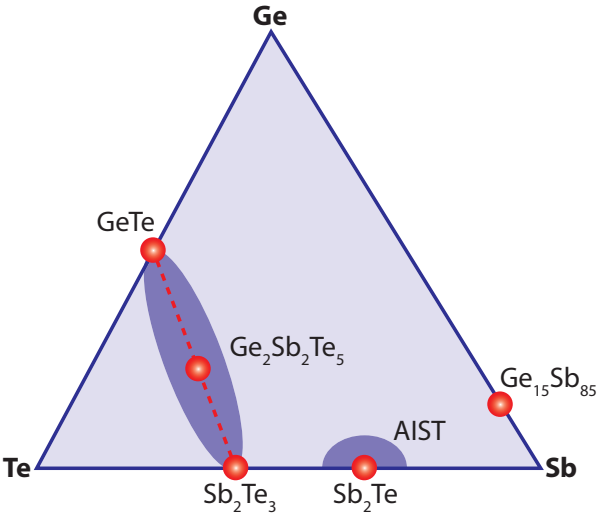

Figure 2.3.: Ternary phase-diagram of Ge:Sb:Te showing the most prominent phase-change materials. Particularly along the GeTe:Sb_2Te_3 line many PCMs are found.

to note that these materials are rewritable which means that PCMs can be cycled between the two states in a reversible manner [99]. Besides the strong optical and electrical contrast as well as a high cyclability, phase-change materials have to show a good stability and fast crystallization times [22, 98]. Fast switching speeds below one nanosecond have already been shown in nanophotonic applications [100].

In this work only non-volatile phase-change materials that preserve their state at room temperature for many years are considered, although in literature also volatile materials as vanadium-oxide that show a reversible transition of the optical properties upon excitation are often referred to as phase-change materials [101, 102].

2.4.1. Material types

Most phase-change materials belong to the chalcogenides that are chemical compounds of elements of the oxygen-family (i.e. O, S, Se, Te) and at least one metal or more electropositive element (i.e. As, Ge, Sb, In). They can be found in the ternary phase-diagram of Ge:Sb:Te [103] (see Fig. 2.3). Especially the pseudo-binary line from Sb_2Te_3 to GeTe reveals many PCMs [104, 105] as for example $Ge_2Sb_2Te_5$ (GST), that is also the phase-change material mainly employed throughout this work. Another compound that is exploited for rewritable data storage as DVD-RW is based on Sb_2Te in combination with silver and indium (i.e. $Ag_5In_5Sb_{60}Te_{30}$, also called AIST) [94, 98]. A third family of compounds can be considered as doped antimony (i.e. $Ge_{15}Sb_{85}$) and has been extensively studied due to its potentially fast switching properties [106]. More recent research has additionally lead to monoatomic phase-change materials, which solve issues with the exact stochiometry when downscaling the material volume [107] and also interfacial phase-

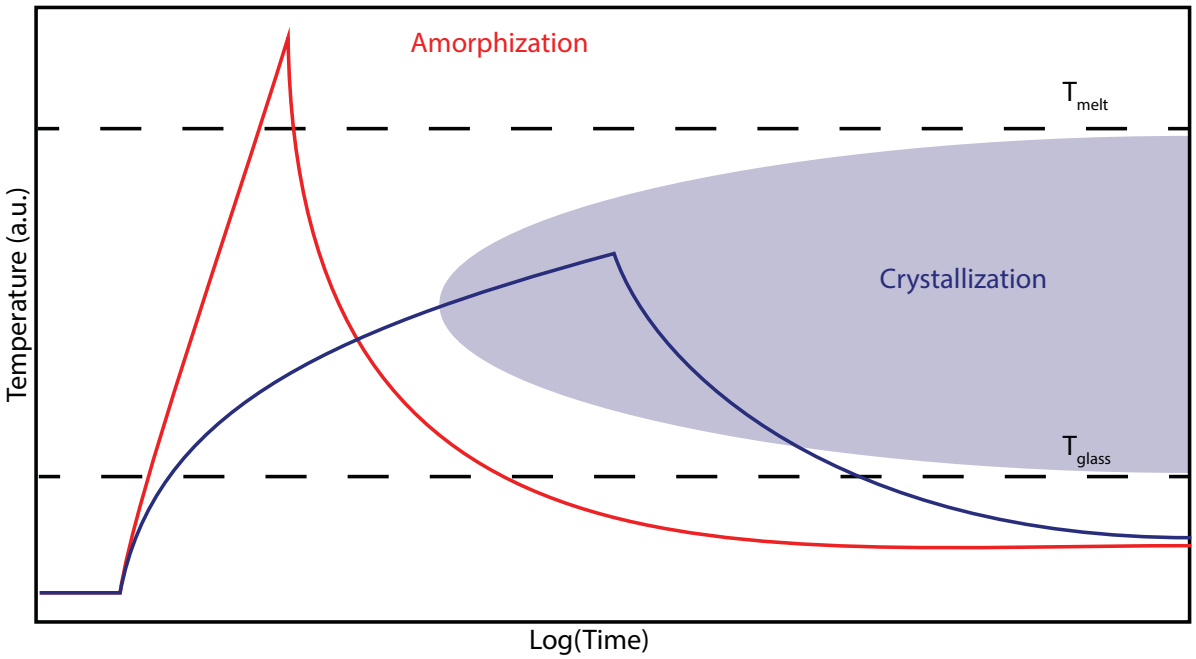

Figure 2.4.: Schematic illustration of the temperature-time dependence during the switching process. When rapidly cooled down after heating the PCM above its melting temperature, crystallization can be avoided and the unordered amorphous state is preserved (red). By heating the PCM above the glass transition temperature and keeping it at elevated temperatures allowing the atoms enough time to move, crystallization can be induced (blue).

change materials that combine different PCMs to artificial superlattices, effectively reducing the switching to one dimension and therefore decreasing the switching energy [108].

2.4.2. Switching mechanism

The phase transition itself can be induced optically as well as electrically by heating the material above the crystallization or melting temperature with optical or electrical pulses. Fig. 2.4 shows the process of the phase transition as a function of time. To amorphise the material from the crystalline state, it has to be heated above the melting temperature T_{melt}. Now the atomic long-range order is lost and the material can stay in the amorphous state if the melt is quickly quenched below the glass transition temperature. By elevating the temperature again between the glass transition temperature T_{glass} and below the melting temperature and allowing the atoms enough time to diffuse, the atoms relax back into the energetically preferred crystalline state. It is important to note that both states of the PCM are stable for many years at room temperature, making them ideal candidates for non-volatile memories as well as in-memory computing applications.

In order to explain the extraordinary difference of the material properties of PCMs between the amorphous and crystalline state, a closer look at the atomic order and the bonding mechanisms

involved must be taken. In the amorphous phase the material is dominated by covalent bonds and no long-range order is present [109]. Therefore, the electrons are highly localized leading to a low electrical conductivity and accordingly the PCM acts as an isolator. Upon switching to the crystalline state, the local atomic arrangement changes slightly and so-called resonant bonds are formed [20, 98]. Herein, electrons are shared between several atoms leading to a strong delocalisation opposed to the covalent bonds in the amorphous phase. The electronic polarizability strongly increases while the optical bandgap decreases, resulting in a vast change of the optical properties of the phase-change material. Because only small changes in the atomic arrangement are necessary, phase-change materials offer fast switching times as described in more detail in Sec. 3.3.

Chapter 3.
3 Phase-change photonics

In this chapter the integration of phase-change materials with the nanophotonic platform will be described. PCMs add an active component to the passive photonic waveguides enabling reconfigurable optical circuits and therefore routing and computation. Especially the non-volatility of phase-change materials holds promise for low power data processing.

3.1. Waveguide coupled PCMs

As sketched in Fig. 3.1a) the PCM is deposited on top of a waveguide, so that the light travelling down the waveguide can evanescently couple to the PCM. For most of the experiments and all simulations, the phase-change material $Ge_2Sb_2Te_5$ (GST) is used with a refractive index at 1550 nm of

$$n_{\mathrm{aGST}} = 3.94 + i0.045 \qquad n_{\mathrm{cGST}} = 6.11 + i0.83. \tag{3.1}$$

In some of the experiments in Chap. 5 AIST is employed. All photonic structures are simulated and fabricated in silicon nitride with a waveguide width of 1.2 µm assuring single-mode operation (transverse-electric) at a wavelength of 1550 nm. The waveguide height is fixed by the silicon nitride layer thickness of the commercially purchased wafer and varies in the range of 325 nm to 343 nm. The optical mode simulations at a wavelength of 1550 nm (FEM in COMSOL Multiphysics) of a 2D cross section is shown in Fig. 3.1b) and c), considering a 10 nm GST film and a 10 nm indium tin oxide (ITO) protection layer on top of the waveguide. The ITO layer prevents oxidation of the GST. The film-thicknesses are chosen following previous work in [23] and offer good switching contrast and low insertion loss in the amorphous phase. With the PCM in the amorphous state the normalized electric field is still centred in the middle of the silicon nitride waveguide, almost unaffected by the PCM yielding an effective refractive index of $n_{\mathrm{eff,aGST}} = 1.60 + 0.002i$. The imaginary part of the refractive is low, leading to an optical

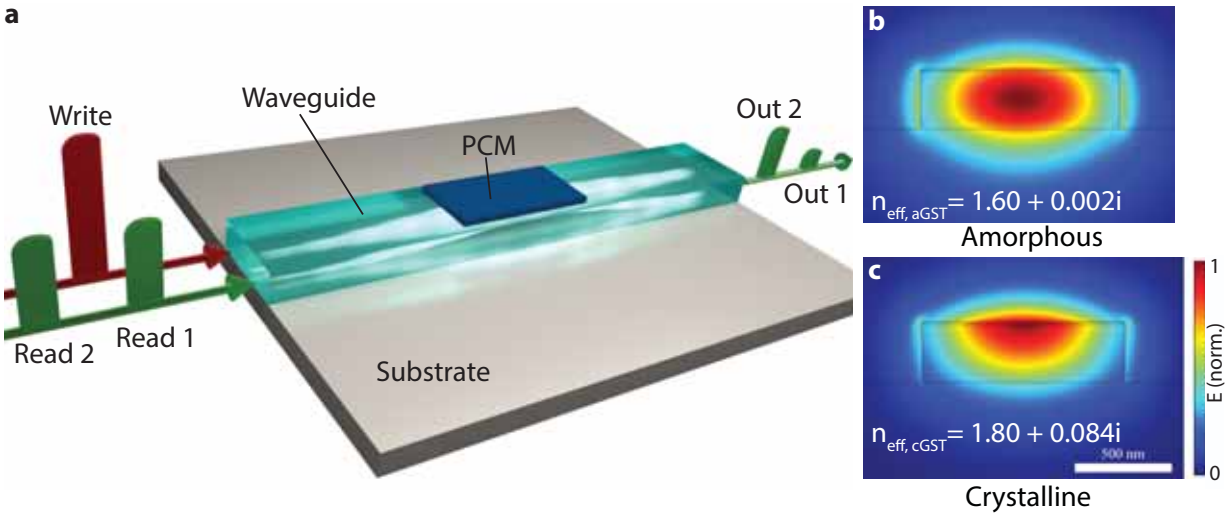

Figure 3.1.: Sketch of the basic phase-change photonic cell. a) The phase-change material is deposited on top of an integrated waveguide. With the PCM in the crystalline state, a first read pulse ('Read 1') is absorbed resulting in a low output pulse ('Out 1'). After amorphizing the material with a higher power write pulse ('Write'), a second read pulse is fully transmitted ('Out 2'). **b)** Mode simulation of a silicon nitride waveguide with 10 nm GST and 10 nm ITO on top at a wavelength of 1550 nm, with the PCM in the amorophous state. **c)** The same geometry as in b) but with the PCM in the crystalline state reveals a mode strongly attracted to the top surface of the waveguide.

absorption of 0.07 dB/µm at 1550 nm. On the contrary, with the GST in the crystalline state the optical mode is highly attracted by the GST and the effective refractive index of the mode increases to $n_{\mathrm{eff,cGST}} = 1.80 + 0.084i$ yielding an absorption coefficient of 2.96 dB/µm. It should especially be noted that the imaginary part of the refractive index changes by more than an order of magnitude, giving rise to a highly effective amplitude modulation.

The phase transition between the amorphous and crystalline state of the PCM can be reversibly induced with optical pulses through the same waveguide enabling all-optical operation of the waveguide coupled PCMs. Fig. 3.1a) shows the general operation principle. Assuming the GST to be in the crystalline state a low power probe pulse ('Read 1') is sent to the waveguide. As the light strongly couples to the PCM via the evanescent field, most of the light is absorbed ('Out 1'). By sending a higher power write pulse ('Write') the amorphous PCM is heated up above the melting temperature and rapidly quenched below the glass transition temperature such that amorphization is induced. A following probe pulse ('Read 2') now is mainly absorbed in the crystalline PCM and only a small amount of the light reaches the output ('Out 2'). In a similar way, by sending an optical pulse with appropriate energy lower than the amorphization pulse, the GST can be heated above the glass transition temperature resulting in crystallization of the GST. A more thorough description of the switching behaviour with optical pulses in the basic PCM-cell considering also the microscopic composition of amorphous and crystalline parts in the

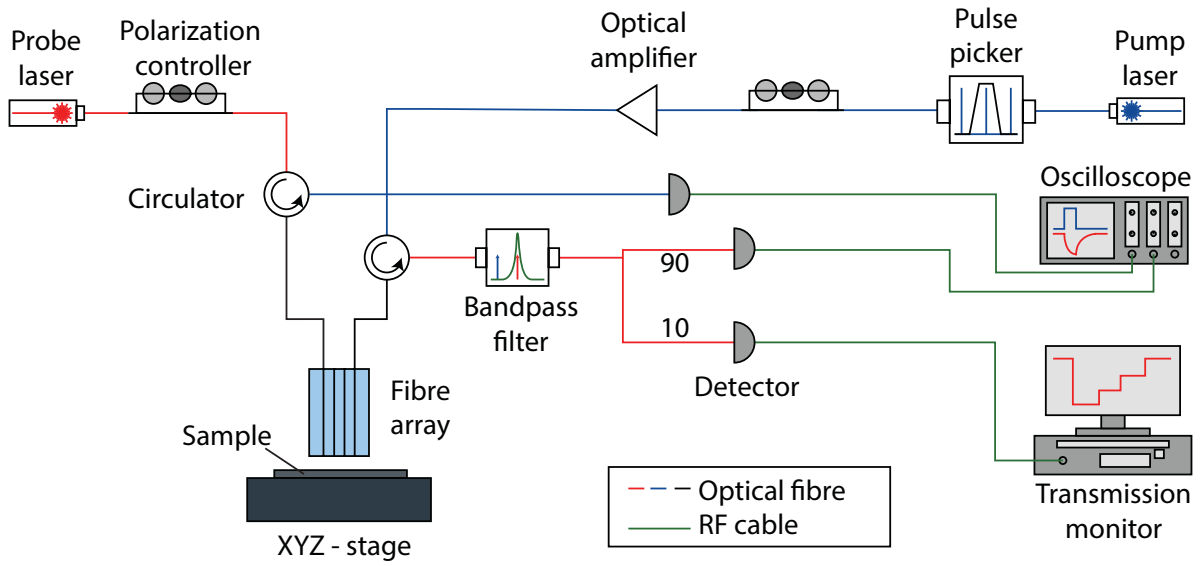

Figure 3.2.: Pump-probe setup for basic sample characterisation. Two counterpropagating optical paths are employed for the operation of a basic PCM-cell, to allow for separation of the two signals using optical circulators. The probe path is used for reading the PCM-state and consists of an optical transmission measurement. The pump path is used to generate optical pulses for switching. The transient behaviour of the cell during switching and dynamic effects are observed with an Oscilloscope ($> 1\,\mathrm{GHz}$), whereas the static properties are monitored with a slow detector ($\approx 200\,\mathrm{kHz}$).

PCM after switching can be found in [24].

3.2. Experimental techniques and measurement platform

The experimental setup to characterize and operate the basic phase-change cell is sketched in Fig. 3.2 and consists of two counterpropagating optical paths based on [23]. The first is the probe path (red) and resembles a simple optical transmission measurement. The light is generated by a wavelength-tuneable continuous wave laser (Santec, TSL 510) coupled to the chip via a fibre array mounted on top of a x-y-z-stage. After transmission through the on-chip waveguide, the probe light is filtered and split in two parts (90:10) and detected using two photodetectors. The first is a fast detector ($> 125\,\mathrm{MHz}$) visualized on an oscilloscope and enables the observation of the fast transient effects of the probe signal during a switching pulse. The second detector is a slow detector ($< 200\,\mathrm{kHz}$) that is used to measure the state of the PCM. For optically switching the phase-change material the counterpropagating optical path (blue) is used. The switching pulse is generated from a continuous-wave (CW) laser (New Focus, Telecom Test Laser 6427) and an electro-optic modulator (Lucent Technologies, 2623CS) that cuts short pulses out of the CW-light

controlled by an electrical pulse generator (HP, 8131A)[1]. After transmission through the on-chip device the optical pulse is recorded with a fast detector and displayed with an oscilloscope. To separate the two counterpropagating optical paths, circulators are employed before the fibre array and to further clean the probe signal from the pump light that is more than an order of magnitude higher, a tuneable bandpass filter provides sufficient suppression of the pump light. With fibre-based polarization controllers the polarization of the light can be optimized to provide efficient coupling to the TE-mode of the on-chip waveguides.

The fibre-array is polished under an angle of 8° allowing for efficient coupling to the chip. Bragg-grating couplers [110–112] are employed as input and output ports to the photonic structures with a coupling loss of approximately 7 dB − 9 dB and a bandwidth of approximately 30 nm. Grating couplers enable to integrate many devices on the same chip with the advantage that the waveguides do not have to be guided to the edge of the chip and (depending on their design) require only one lithographic step. It should be noted that the transmission of the grating couplers can be improved with negative angle couplers to about 4 dB per coupler and the bandwidth and coupling efficiency can be largely increased using 3D couplers fabricated via direct laser writing as for example total internal reflection [113, 114].

3.3. Switching events and comparison between GST and AIST

In all experiments in this work the switching of the phase-change material is carried out optically via pulses send through the waveguide (in-plane). Whereas most of the experiments are performed with GST, some experiments in Chap. 5 employ also AIST[2]. This section gives a short comparison of both materials regarding their properties in the basic PCM-cell. Fig. 3.3a) shows the thermo-optic response [24, 115, 116] of both a waveguide with GST and one with AIST for two different pulse lengths. The upper panel depicts the measured optical pulse exciting the PCM and the lower panel the change in the transmission of the probe light. The change in the transmission in this case is volatile as no switching of the PCM is induced and reflects the temperature dependent refractive index of the PCM leading to a decreasing transmission during the excitation. For both materials it is observed that shorter pulses lead to shorter cooling times meaning that the transmission settles faster to the initial level. This can be explained as follows: when an optical pulse arrives it is mainly absorbed in the PCM due to the non-zero imaginary part of the refractive index. For longer optical pulses the heat also diffuses to the surrounding waveguide and substrate. The cooling time mainly depends on the temperature gradient between the PCM and

[1] Note that for the experiments with short pulses a mode-locked picosecond laser from Pritel with a repetition rate of 40 MHz is employed. Using a custom-build pulse picker, individual pulses can be selected from the pulse train and used for switching the PCM.

[2] Note that GST and AIST are used because they are available for sputtering in the group of Prof. Bhaskaran at the University of Oxford where the PCM is deposited and have proven suitable for integration with photonic structures in previous work.

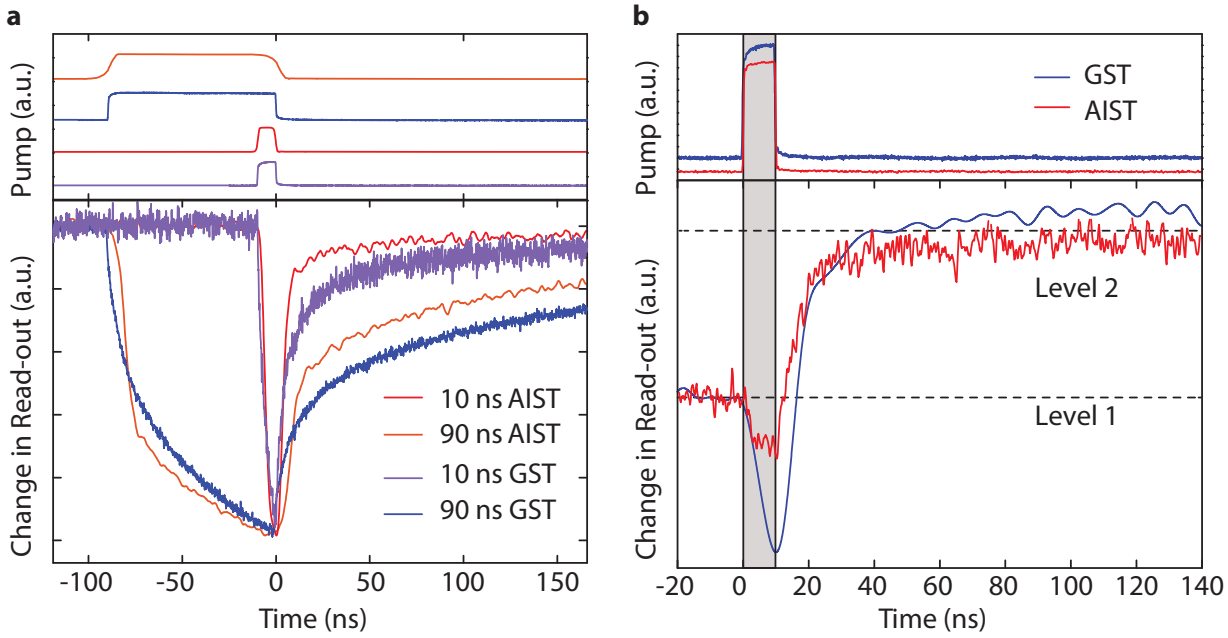

Figure 3.3.: Comparison of the basic PCM-cell with AIST and GST. a) Thermo-optic response of both materials (bottom) upon excitation with 10 ns and 90 ns optical pulses (top) without switching. **b)** Amorphization of both materials with a 10 ns pump pulse. GST-data taken from [115].

its environment. Because shorter pulses more specifically heat the PCM itself, shorter cooling times can be expected as seen in the experiment because the heat diffuses away faster [116]. The cooling constants for both materials are on the same order of magnitude. Fig. 3.3b) shows the time-dependence of the probe transmission during an optical pulse that induces amorphization for both materials. The pulse width is 10 ns, again leading to a similar behaviour for both GST and AIST. After the initial drop due to the thermo-optic effect, the transmission rises two a new level indicating the switching of the PCM. Both cells are settled after about 40 ns.

Two conclusions can be drawn from the data in Fig. 3.3. The first is that shorter optical pulses lead to shorter dead times of a PCM cell, which is important for fast signal processing. As will be shown in Chap. 5 switching with pulses of a width of a picosecond is possible leading to sub-nanosecond operation times while also drastically reducing the pulse energies. The second conclusion is that within the scope of this work no significant difference in the behaviour of the two phase-change materials AIST and GST can be observed. However, it should be noted that they differ in their crystallization process (GST is nucleation dominated whereas AIST is growth dominated [117, 118]) and a detailed analysis could reveal advantages and downsides of the specific materials for example in terms of endurance or switching energies. More details on the switching process of the photonic phase-change cells can be found in [24, 116].

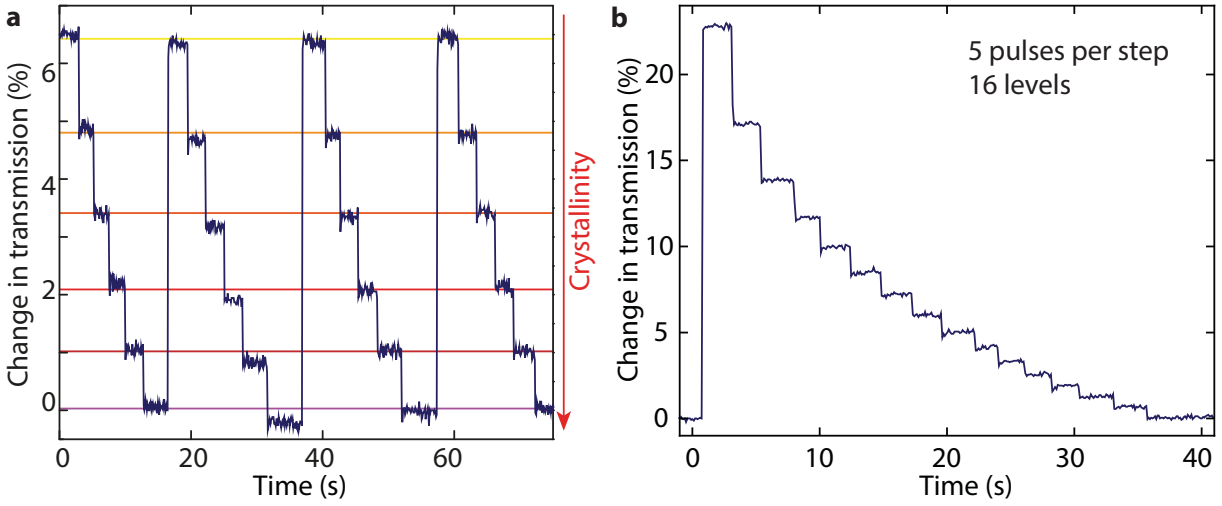

Figure 3.4.: Multilevel operation of a PCM-cell. **a)** A single PCM-cell is repeatedly switched through six different levels, leading to a change in the optical transmission. Each step is induced by a single pulse of a length of one picosecond. **b)** PCM-cells exhibiting sixteen individual levels. Each step is induced by a train of five consecutive pulses.

3.4. Multilevel operation

The non-volatility of the phase-change materials together with the large contrast in their properties between the phases of matter make them ideal for the application in optical data storage. Compared to conventional digital electronic memories they have an additional advantage as shown in the transmission measurement in Fig. 3.4. Besides the two phases fully crystalline and fully amorphous, the PCM can also be in almost arbitrary states in between, enabling a multilevel operation of the phase-change cell increasing the density of the memory [116, 119–122]. The multilevel capability arises from micro structuring of the phase-change material. Depending on the properties of the switching pulse only partial amorphization or partial crystallization can be induced, meaning that only a fraction of the material changes its state while the other fraction remains unaffected. An electromagnetic field travelling down the waveguide only senses the average phase-state of the whole PCM-patch, leading to intermediate states between fully amorphous or fully crystalline although the material itself can only exist in either of the states. Fig. 3.4a) shows the change in optical transmission[3] through a basic PCM-cell (3 µm AIST) with six different levels. Single optical pulses of a length of 1 ps have been used to induce switching leading to very fast operation. It should be noted that other than the timescale in the plot suggests because the individual pulses are sent manually, the switching events (transition between levels) take place on the order of nanoseconds. All levels can reproducibly be reached and are clearly distinguishable. Fig. 3.4b) further shows an example of a PCM-cell operated with 16 individ-

[3] The change in transmission ΔT is defined as $\Delta T = \left(T_{\max} - T_{\min}\right)/T_{\min}$ with $T_{\max}$ and $T_{\min}$ being the maximum and minimum transmission, respectively.

ual levels. By increasing the optical contrast between the bottom and the top level, more steps can be distinguished in between. In the example shown, a train of five consecutive pulses with 25 ns separation and a width of 1 ps are employed for each crystallization step. It is important to note, that the phase-change material shows an accumulative behaviour, meaning that each step of crystallization is induced by the exact same pulse (sequence) with the same energy. This property enables simple arithmetic as explained later in Chap. 5. X. Li et al. have demonstrated a 5-bit optical memory discriminating 34 levels with the standard phase-change cell as described in [122]. By even further increasing the overall optical contrast for example using photonic crystal cavities [123], micro resonators [124] or increasing the signal to noise ratio of the measurement even more levels can be exploited.

Compared to electronic implementations of multilevel memories using phase-change materials, photonic approaches are less prone to drift of the individual levels. The resistance states of phase-change random-access memory (PCRAM) drift due to structural relaxations in the material [125–128] making the readout of the memory over time more difficult. Additionally, in electronic phase-change memories an iterative program-and-verify method is used for programming [129, 130], involving several write and reset pulses until the correct state is reached. Using the photonic multilevel cell, reliable multilevel operation with single pulses is possible [25, 122], resulting in faster and more energy efficient operation.

3.5. Endurance

Another important characteristic as well for memories as for data processors is the endurance of the system. Fig. 3.5a) shows the measurement of a phase-change cell binary switched between the crystalline and the amorphous state for 50 000 times (after pre-cycling to get to a steady state [116]) with a stable contrast of 20%. Amorphization is induced by a single pulse (100 ns) with an energy of 430 pJ, whereas crystallization is achieved with 18 consecutive pulses with energies decreasing from 160 pJ to 114 pJ (100 ns pulse width). While the contrast is stable for 50 000 switching cycles, measuring the endurance for a million cycles becomes experimentally challenging as depicted in Fig. 3.5b). The red line depicts the optical switching contrast, whereas the blue line depicts the reference transmission through an unswitched PCM-cell. The pulse energy used for amorphization is again 430 pJ but crystallization is carried out by a pulse train consisting of 18 pulses with 150 pJ energy each. All pulses are of 100 ns length. The switching contrast decreases gradually from approximately 30% to below 15% over one million cycles. It is known that the PCM-cells need to be cycled a few times before they can be operated in a stable mode [116] achieving a steady state resulting in the initial drop of the optical contrast. The almost linear decrease of the contrast afterwards can be explained by the reference transmission measured through an independent device. Over the duration of the experiment the experimental setup shows a drift, especially the alignment between the on-chip grating couplers and the fibre array

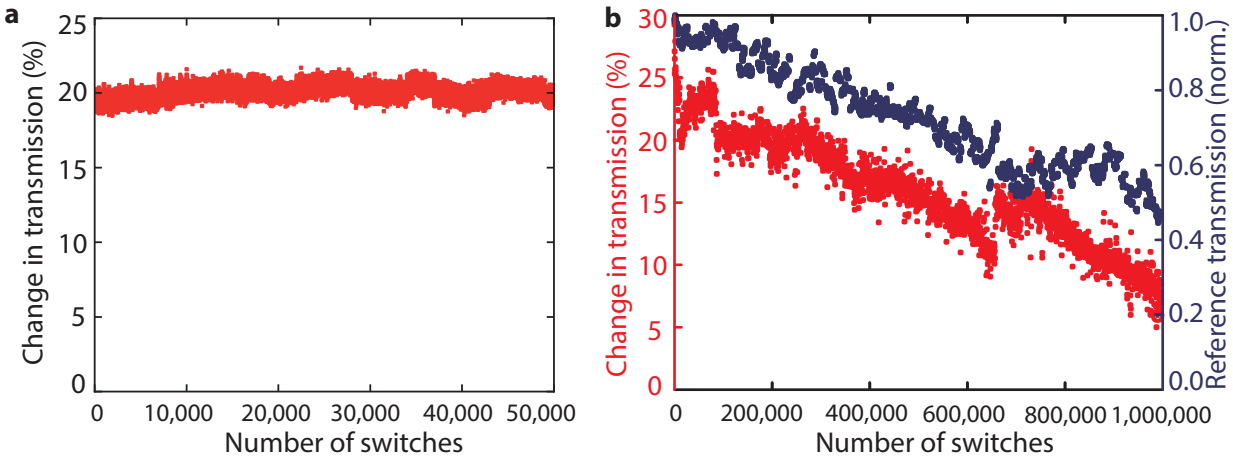

Figure 3.5.: Endurance measurement of a single PCM-cell. a) Constant change in transmission of about 20% over 50 000 cycles. **b)** Change in the optical transmission upon switching for one million cycles in comparison to the reference transmission measured through a separate device that was not switched. As the reference transmission decreases, also the pulse energy and therefore the optical contrast decreases [131].

on top gets displaced, leading to the decreasing reference transmission. Because the reference transmission decreases, also the energy of the optical pulse reaching the PCM-cell to be switched decreases, leading to the reduced optical contrast. The original transmission level could be reached again by realigning the chip after this experiment, also enabling a switching contrast of approximately 20% (note however, that stabilizing the setup for the time of the measurement ($> 24\,\mathrm{h}$) is difficult, so that the experiment is stopped after 1 million cycles).

These results indicate that the PCM-cells operated all-optically exhibit a good reversibility throughout a large number of cycles exceeding the prevailing endurance of other non-volatile memories as for example FLASH-memories with cyclability in the order of $10^4 - 10^5$ [132]. In other studies 10^{12} cycles for PCMs switched electrically have already been shown [56] and 10^{15} cycles are predicted from theoretical studies [133]. However, the search for new PCMs with even better properties in terms of endurance, energies and switching times is a vivid and ongoing field of research.

Chapter 4.
4 Fabrication

In this chapter the fabrication of the nanophotonic circuits is explained. A three-step lithographic process is carried out on a silicon nitride on silicon oxide platform. The chip-layout is created using the GDSHelpers library [134].

The typical techniques in nanofabrication to pattern photonic integrated circuits are photo lithography and electron-beam lithography (EBL) [135, 136], which enable feature sizes down to the nanometre scale. Photo lithography is usually employed in industrial production due to its faster exposure times compared to electron-beam lithography. However, photo lithography requires a photo-mask for exposure, which is typically fabricated using an EBL-process to achieve the high resolution necessary for nanophotonic circuits. The prototypical devices used in this thesis are therefore directly fabricated using electron-beam lithography.

In EBL a resist spin-coated on a sample is exposed by a confined electron beam following a pre-defined pattern. The electron beam induces a chemical reaction in the resist, which can for example be linking polymer chains (negative-tone resist) or breaking them up (positive-tone resist). A subsequent development step is carried out by treating the resist with a certain chemical that dissolves the non-exposed (negative-tone resist) or the exposed (positive-tone resist) resist. The remaining mask can then be transferred to the substrate in an etching step or used to deposit a new material in the defined areas with a subsequent lift-off process.

The photonic structures fabricated in this work are prepared from a silicon nitride wafer (commercially purchased from Rogue Valley Microdevices) with a layer stack of 343 nm Si_3N_4 on top of 3330 nm silicon oxide on a 525 µm silicon substrate (note that for the experiments in Chap. 7 the silicon nitride film thickness is 325 nm, due to variation in the purchased wafers). The layer stack allows for fabrication of single mode waveguides with a width of 1.2 µm.

Fig. 4.1 illustrates the fabrication process consisting of three major steps. In a first step of lithography, windows for the deposition of alignment markers are exposed in the positive-tone electron beam resist poly methyl methacrylat (PMMA) 950k 4.5. After development in a mixture

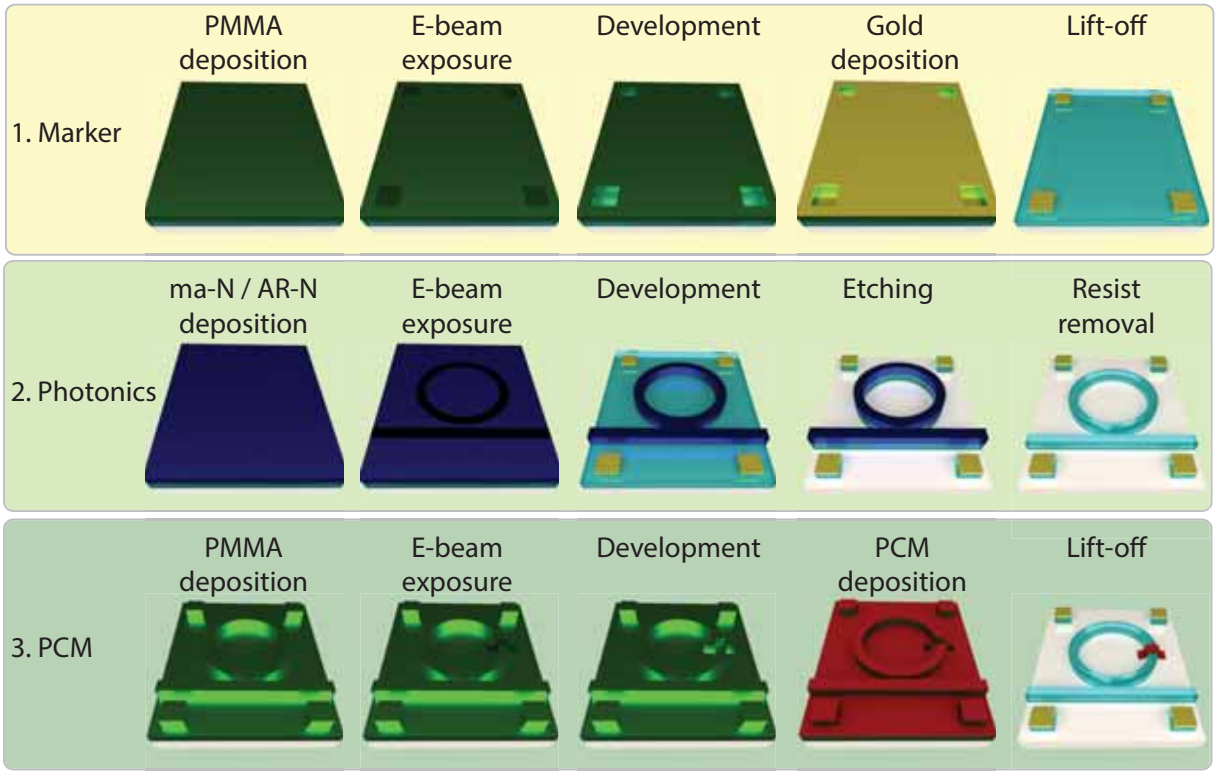

Figure 4.1.: Fabrication process. The fabrication consists of three major steps. In the first step gold markers for alignment purposes are deposited using the positive-tone electron beam resist PMMA (green) and a lift-off process. The second step of lithography defines the photonic structures in the negative-tone resist ma-N (blue). After development, an etch mask is obtained that is transferred to the substrate. In the last step, again using PMMA, windows for the PCM (red) deposition are opened and aligned to the photonic waveguides using the gold markers. A subsequent lift-off reveals the final sample.

of isopropanol and methyl isobutyl ketone (MIBK), 7 nm chromium and 70 nm gold are deposited using physical vapour deposition (PVD). A subsequent lift-off process in acetone removes the PMMA and leaves the alignment markers on the surface of the sample. These are used to precisely position the electron beam in the following lithographic steps.

In the second step the photonic structures are defined. The pattern representing waveguides, grating couplers and other photonic components are inscribed in the negative tone resist ma-N 2403 (note that for the samples in Chap. 7 AR-N 7520.12 is used). The mask is developed with MF-319 and transferred into silicon nitride via a reactive ion etching process based on CHF_3 and oxygen. The remaining resist on top of the photonic structures is removed in an oxygen plasma.

In the last lithographic step windows for the deposition of the phase-change material are opened in PMMA analogous to the process for the markers. Using the alignment markers, the PCM-windows are precisely aligned to the photonic structures. 10 nm of the PCM (GST or AIST) are sputter deposited with a 10 nm layer of ITO to prevent oxidation of the PCM. The resist is again

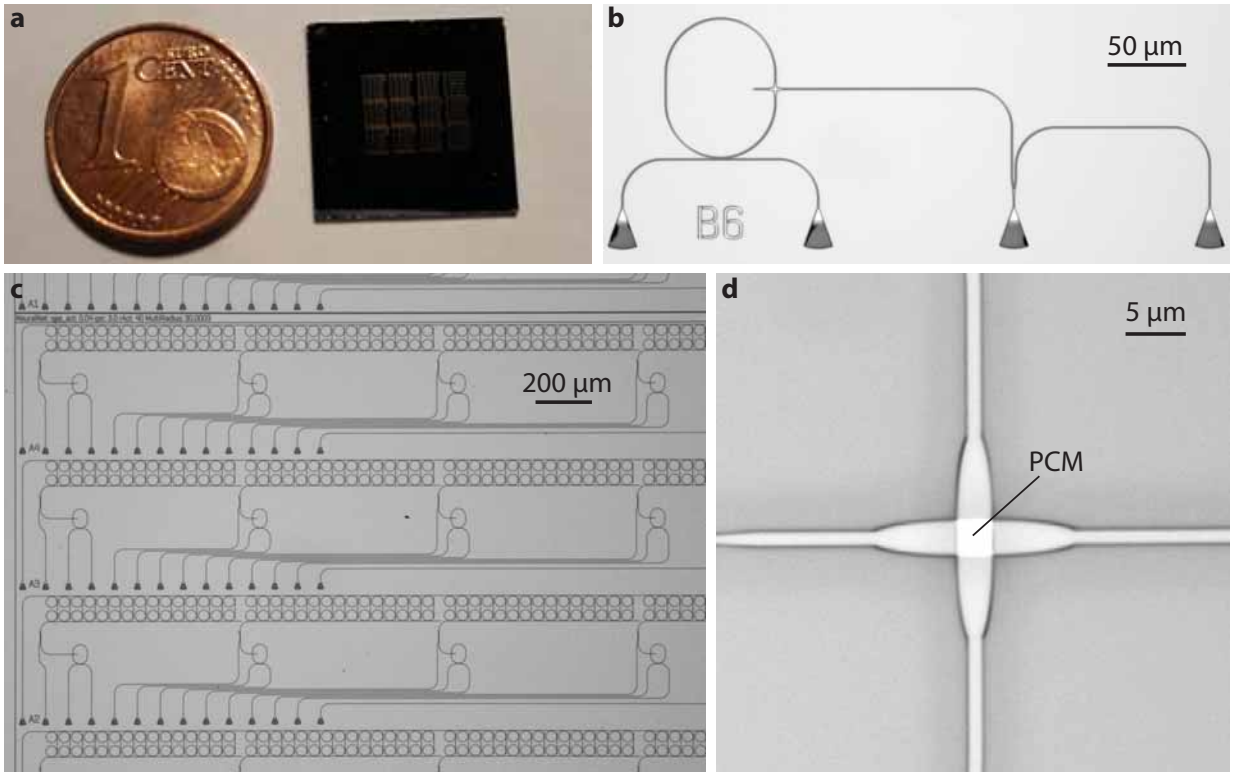

Figure 4.2.: Micrographs of fabricated on-chip devices. a) Nanophotonic chip in comparison to a one cent coin. **b)** Optical micrograph of a single photonic device with grating couplers (four triangular structures at the bottom) as input and output ports for coupling light. **c)** Several photonic circuits fitted on a small area of the chip. **d)** Close up of a waveguide crossing with PCM deposited on the intersection showing precise alignment between the waveguide and the PCM.

lifted-off with acetone.

Fig. 4.2 shows optical micrographs of a fabricated photonic chip with different photonic structures. Fig. 4.2d) proofs the precise alignment between the waveguides and the deposited PCM. The waveguides fabricated with the developed electron-beam lithographic process and subsequent reactive ion etching have been shown to exhibit very low loss of down to 21 dB/m [137]. More details on the exact fabrication procedure are given in App. A.1.

Chapter 5.
5 All-optical abacus

Based on the phase-change photonics platform described in the previous chapters, prototypical non-von Neumann processors are developed and described in Chap. 5 to Chap. 7. Whereas Chap. 5 is based on arithmetic operations, Chap. 6 describes a small-scale neuromorphic processor capable of basic pattern recognition tasks and in Chap. 7 a photonic tensor core able to perform matrix multiplications at high speeds in parallel is implemented.

This chapter focusses on basic arithmetic that is established using a single PCM-cell capable of addition, subtraction, multiplication and division in arbitrary bases in analogy to an abacus. The concept is in a second step expanded in a waveguide crossing array with individually addressable cells employing a two-pulse switching scheme. The presented results are based on [131].

5.1. Single cell operation

The fundament of a conventional processor is its ability to carry out basic arithmetic tasks with high precision and repeatability. This chapter explains and experimentally demonstrates the basic unit of a first approach to a phase-change photonics processor carrying out addition, subtraction, multiplication and division using light.

5.1.1. Basic arithmetic

The basis for the arithmetic processing unit developed in this work is counting optical pulses with a phase-change material. A similar approach was already demonstrated with a free-space optical setup by Wright et al. [138] preparing the ground for integration in the nanophotonic platform offering a scalable architecture.

The operation principle of the arithmetic unit is similar to that of a traditional abacus [139] where several rods, each representing a certain place value, hold ten beads representing the digits. In the photonic counterpart the rods are individual phase-change cells and the beads correspond to a certain state of crystallization. By inducing quanta of crystallization in the PCM, the optical

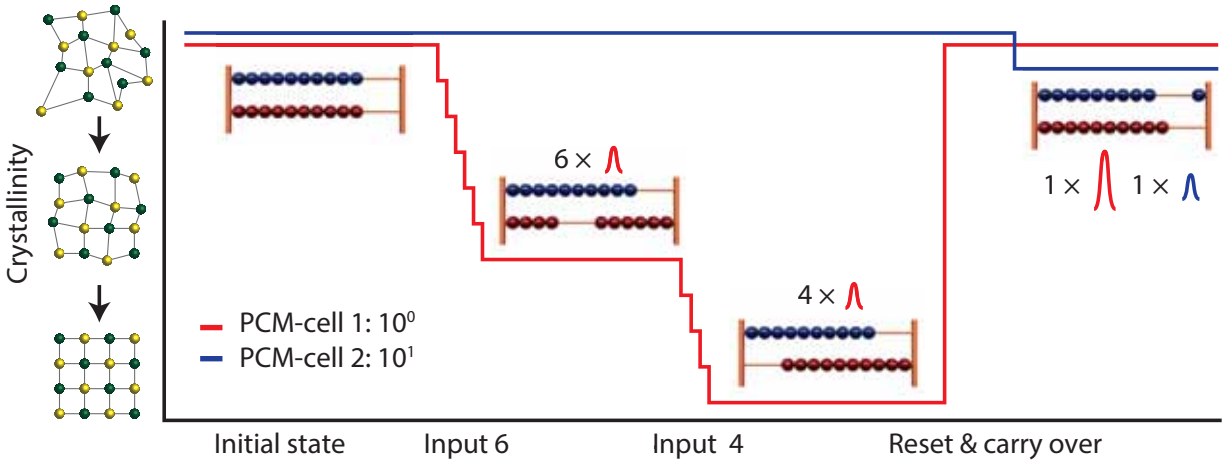

Figure 5.1.: Illustration of the operation of the basic arithmetic unit in base ten. Two PCM-cells are employed to carry out the addition $6 + 4 = 10$. The first cell stores the ones and the second cell stores the tens. The operation principle follows the addition with an abacus. The first addend ('6') is induced by six optical pulses inducing crystallization in the first PCM cell. After adding the second addend ('4') the cell is fully crystalline and a carry-over has to be executed, meaning that the first cell is reset to its initial amorphous state and a single step of crystallization is induced in the second cell representing '10' (adapted from [131]).

transmission can be tuned stepwise and utilized for the arithmetic operations. Fig. 5.1 sketches the summation '$6 + 4 = 10$' employing two phase-change cells that represent the ones (red) and the tens (blue).

Starting with both cells in the unordered amorphous state, six identical optical pulses (representing the first addend) are sent to the first cell inducing six steps of crystallization. This is analogous to shifting six beads on the row representing the lowest place value of an abacus to the right as indicated in the image. Now the second addend ('4') is sent to the cell again using the same optical pulses leaving the PCM in the crystalline ground state. This situation corresponds to all beads of one row shifted to the right so that a carry-over must be executed in which all beads are shifted back to their initial position on the left and one bead of the next higher rod is shifted to the right (blue). The same operation is carried out with the optical abacus. The first PCM-cell (red) is amorphized to its initial state using a higher power optical pulse, and a single pulse is sent to the second cell inducing one step of crystallization. The calculation result can now be read from the phase state of the different PCM-cells (again analogous to an abacus). The first PCM-cell is in the amorphous state representing zero ($0 \cdot 10^0$) and the second cell is left in the first crystalline state representing ten ($1 \cdot 10^1$) revealing the result $0 + 10 = 10$.

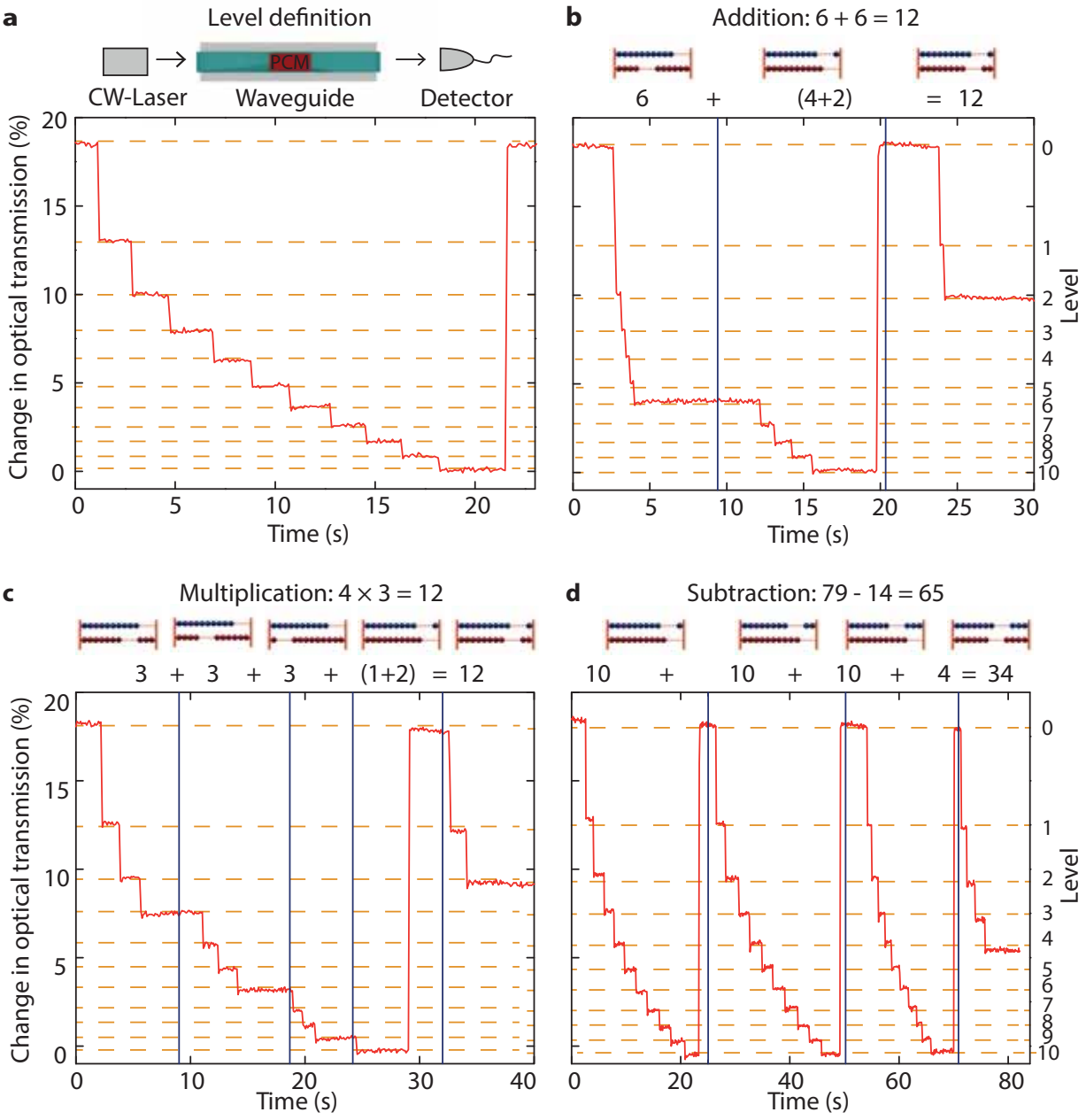

Figure 5.2.: Arithmetic operations in base ten. a) Defining the transmission levels. Before operating the phase-change cell, the base for the calculation must be chosen and the transmission levels defined accordingly. Eleven levels are needed for operation in base ten.**b)** Addition: Analogous to a traditional abacus, the summation $6 + 6 = 12$ is performed. By sending the first six optical pulses (first addend) the transmission reaches level 6. When inserting the second addend, after four pulses the crystalline base state is reached, and a carryover is performed. This includes resetting the PCM-cell and storing the carry-over in a second cell (not shown). The result 12 can be found from one carry-over (1×10) and the transmission state (level 2) of the PCM-cell. **c)** Multiplication is performed as sequential addition by reducing the task $4 \times 3 = 12$ to the summation $3 + 3 + 3 + 3 = 12$. **d)** Subtraction: Using the nine's complement method the subtraction $79 - 14 = 65$ can be reduced to the addition $20 + 14 = 34$. The minuend and the difference are replaced by their nine's complement (see more details in the main text). The examples are chosen to include at least one carry-over to emphasize the analogy to an abacus. All data is taken from [131].

When performing arithmetic tasks with the phase-change cell in the experiment, first the base for the calculation must be defined and the transmission levels found accordingly. Fig. 5.2 shows the definition of the eleven levels necessary for calculations in base ten. Ten levels are needed to represent the digits $0-9$ and one additional level is used to detect if a carryover has to be performed. The crystallization steps are induced by five consecutive picosecond pulses with 12 pJ energy each. The number of pulses is chosen to induce enough contrast to be able to clearly distinguish all levels. The amorphization (reset) is realized with a train of ten pulses with 19 pJ per pulse.

As an example for addition the summation $6+6=12$ is performed on the photonic abacus (Fig. 5.2b). Starting from the amorphous phase (highest optical transmission), in a first step six pulses are sent to the PCM cell inducing six crystallization steps. Now the second addend is sent to the cell but after the first five pulses the fully crystalline level ten is reached. Again, similar to the situation with an abacus when all beads of a rod are slid to the left, a carryover has to be performed. With an abacus this means shifting all ten beads back to their initial position on the left and moving one bead of the next higher rod to the right. The same is executed with the PCM-cell. The crystalline PCM is reset to its initial amorphous state and one crystallization step is induced in a second cell representing the tens (not shown here, see Sec. 5.2 for more details on the carryover). After the reset, the remaining two pulses are sent to the cell and the result of the calculation can be read from the optical transmission measurements. The cell representing the first place value is in state two and one carryover was performed revealing the correct result $2+10=12$. Although the timescales of the plots suggest a slow operation time because the cell is operated manually, the switching process can take place in the sub-ns regime offering GHz processing speeds [116, 124]. The waiting time between the individual steps is chosen for demonstration purposes and the whole calculation can be processed in less than 12 ns in principle.

Because multiplication can be viewed as sequential addition, it can be executed in the same way using the same PCM cells. In Fig. 5.2c) the execution of $4\cdot 3=12$, which is first converted to $3+3+3+3=12$, is demonstrated. Analogous to the previous example the addends are input to the cell step by step. When adding the fourth addend the crystalline level is reached again and the carryover is performed. After resetting the PCM-cell the remaining two pulses induce crystallisation steps leaving the cell on level two. From this value combined with one carryover the correct result is found.

Turning to subtraction the concept of the nine's complement has to be introduced, which can be used to reduce also the subtraction to an addition task [140]. An exemplary calculation is shown in Fig. 5.2d) in which the task $79-14=65$ is performed. To convert the problem to a pure addition, the nine's complement of the minuend 79 is added to the subtrahend 14. The nine's complement of the resulting value is the solution to the initial subtraction. In general, the nine's complement is calculated by subtracting the input value from 10^N-1 with N being the number of digits. For the present example that leads to the addition $20+14=34$, with 20 being

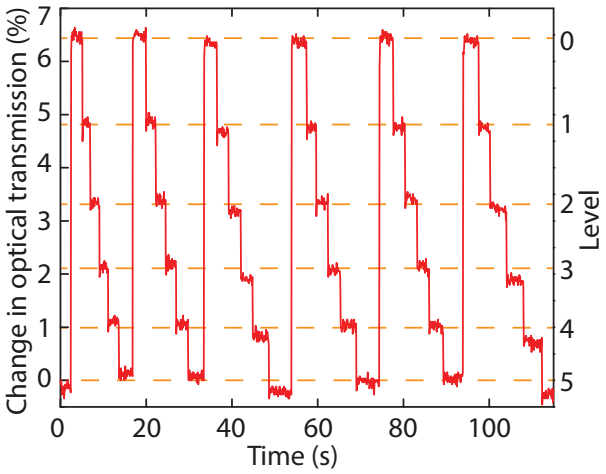

Figure 5.3.: Operation with single pulses. Phase-change cell repeatedly cycled through six levels for six times showing multilevel operation with single optical picosecond pulses (data from [131]).

the nine's complement of 79 and 34 being the complement of 65 – the correct result of the initial task. In the photonic implementation (Fig. 5.2d)) at first twenty pulses for the minuend are input including two carryovers. Next the remaining 14 pulses are sent leaving the PCM cell in state four and adding another carryover. Summing everything up $(3 \cdot 10 + 4)$ gives 34 and therefore the nine's complement of the correct result is obtained. Similar to viewing multiplication as sequential addition, also division can be viewed as repeated subtraction and can therefore also be computed with the same concept in the photonic cells.

The examples shown in this section are executed directly in base ten in opposition to conventional computers that use binary arithmetic increasing the computational power and density. It should also be noted that by using phase-change materials the calculation result is stored in a non-volatile way not requiring any energy to preserve the state. Circumventing the memory bottleneck, the next computation can be executed in the same cell without the need to move the data to the memory.

5.1.2. Single pulses

To reduce the pulse energies used to perform the calculations, all operations can also be performed with single optical pulses per crystallisation step. This way the contrast between the individual levels is decreased reducing the number of clearly distinguishable levels and therefore the maximum base a device can be operated in.

Fig. 5.3 shows six cycles over six levels of crystallisation induced in a PCM-cell with a single picosecond pulse of 8 pJ energy per crystallisation step. Amorphisation is achieved with five consecutive pulses of 14 pJ per pulse, enabling low energy and high-speed computation.

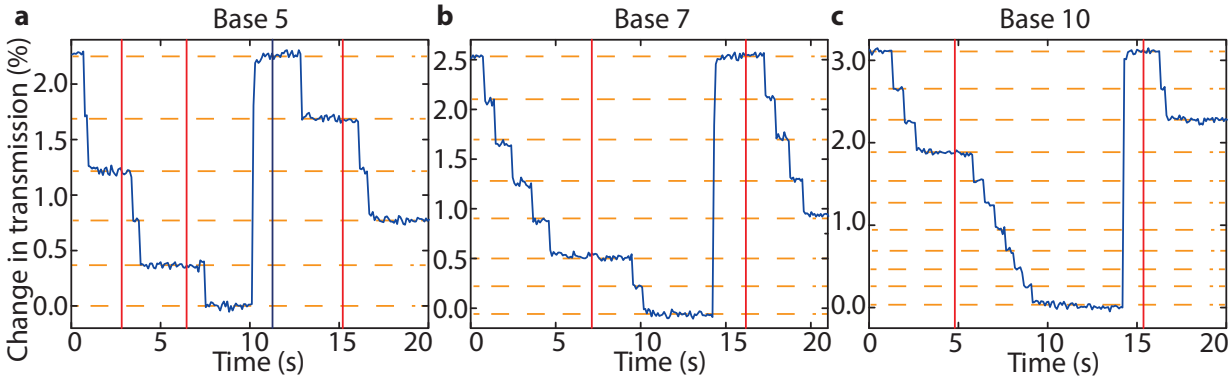

Figure 5.4.: Arithmetic in different bases. **a)** The multiplication 4×2 calculated in base five. **b)** Addition of $5 + 6$ in base seven and **c)** $3 + 9$ in base ten [131].

5.1.3. Different bases

Because the number of levels between the crystalline and amorphous state can be almost arbitrarily chosen only limited by the ability to distinguish them in an optical transmission measurement, the base of the arithmetic operations can be chosen suitably for the actual arithmetic problem. Fig. 5.4 shows three examples of calculations performed with the same PCM cell (2 µm of AIST) in different bases. With a fixed crystalline and amorphous state, the amount of levels can be chosen via the pulse energy or the amount of pulses per step. For example, a device operated with single pulses in base ten can be used in base five simply by sending two pulses per step. If single pulse operation is required, the pulse energies can be tuned in order to induce more or less crystallization per step. The limitations for the number of different levels are mainly the available optical transmission contrast and the sensitivity and noise of the photodetector. The first can for example be increased by using longer PCM-cells or higher PCM-film thicknesses. All operations in the different bases shown in Fig. 5.4 are performed with the same PCM-cell underlining its versatility.

5.2. Two-digit arithmetic with carryover

In the examples shown so far the carryover to a second cell was not physically stored in a second PCM-cell. Fig. 5.5 depicts the experimental setup employed to achieve an automatic carryover with two PCM-cells. As explained in Sec. 3.2, the optical setup is again based on a probe- and a pump-path. The probe-path consists of a continuous transmission measurement for monitoring the phase state of both cells. Therefore, the light from a CW-laser is split in two equal parts and measured with two photodetectors after passing through the PCM-cells.

The pump-path induces the crystallization steps and carries out the calculation. With a 90 : 10 splitter the pump pulse is divided in two parts. The higher power pulse is guided to the first cell representing the ones and the lower power pulse to the second element representing tens.

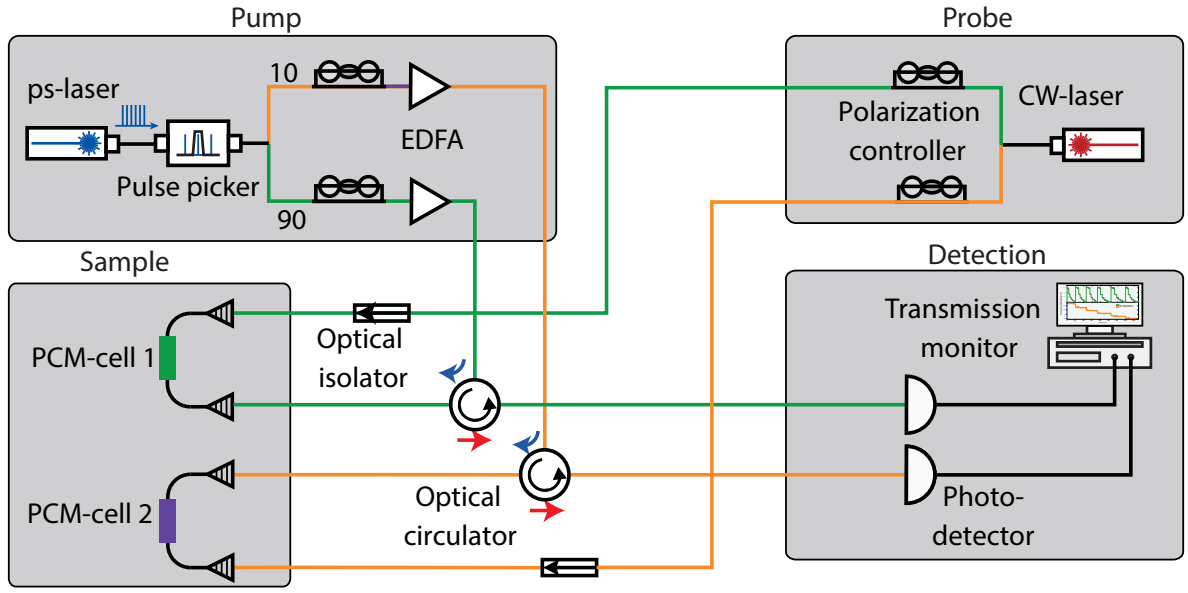

Figure 5.5.: Setup for arithmetic with automatic carryover using two PCM-cells. The optical part of the setup is divided in a pump- and a probe-path. The probe-path is used to continuously detect the state of the PCMs, and the pump-path induces the switching of the cells. To achieve the automatic carryover, the pump light is split so that 90% of the light are guided to the first PCM-cell representing ones and 10% to the second cell representing tens. The pulse energies are chosen such that a single crystallization pulse in the first cell is not enough to switch the second cell but the amorphization pulse when resetting the first cell (to perform the carryover) induces a single crystallization step in the second cell.

The pulse energies are chosen in such a way that the crystallization pulses for the first PCM-cell (90% of the initial pulse) leave the second cell unaffected[1]. When a carryover of the first cell is performed the higher power amorphization pulse is sent and resets the phase-change material to its initial amorphous state. The 10% of the pulse that are split to the second cell are now enough to induce a crystallisation step in the PCM representing the carryover to the next higher place value.

Fig. 5.6a) demonstrates the experimental results achieving automatic carryover employing two PCM-cells working as a basic counter. The change in transmission is shown as a function of time for both cells. After sending five pulses to the first cell reaching its crystalline ground state and leaving the second cell unaffected, the reset to the initial amorphous state is performed by sending a high-power pulse to the first cell. The pulse energy split to the second cell is enough to induce a single crystallization step. This process is repeated five times resembling a simple counter till 24 in base five. In order to demonstrate that the optical abacus can do more than counting pulses and is a powerful tool for arithmetic, an example of a two-digit subtraction making use of the automatic carryover is presented in Fig. 5.6b). According to the rule of converting the task

[1] It should be noted that the exact relation of the two powers can be tuned using the erbium-doped fibre amplifiers (EDFAs) in both paths.

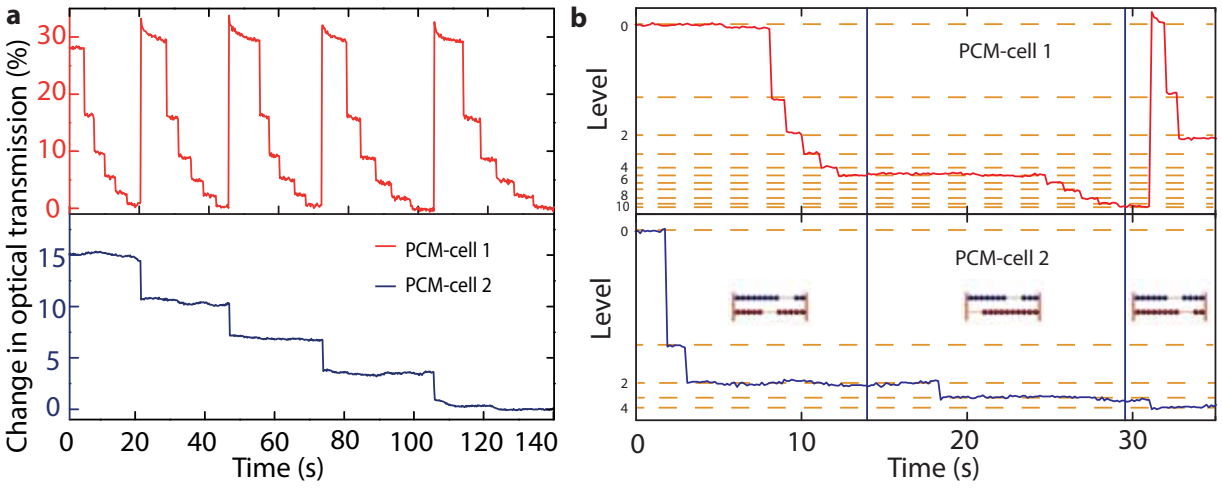

Figure 5.6.: Performing the automatic carryover to a second PCM-cell. a) Counting optical pulses with two PCM-cells. Every five pulses to the first cell a reset is initiated storing the carryover in a second PCM-cell by inducing a crystallization step. **b)** Processing the two-digit subtraction $74 - 17 = 75$ with two PCM cells and automatic carryover (adapted from [131]).

$74-17 = 57$ to an addition using the nine's complement method it is reduced to $25+17 = 42$ (the example is chosen to include two digits and a carry-over to the second cell). In a first step the first addend 25 is input. Therefore, two crystallization steps are induced in the second PCM-cell representing the tens (blue) and after that five steps in the first PCM-cell representing the ones (red). In the second step the level of crystallization of the second cell is increased by one before adding the remaining seven pulses to the first cell. When reaching the fully crystalline state in the first cell after sending five pulses, the amorphization pulse to reset the phase-change material is sent. From Fig. 5.6b) it can be seen that at the same time a crystallization step is induced in the second cell. After adding the remaining two pulses to cell one, the final result of the calculation is encoded in the two cells. Cell two remains in state four representing 40 and cell one remains in state two adding up to 42. By employing even more phase-change cells, the platform can be extended to operate with higher numbers (more digits).

Compared to a simple counter that would have needed 42 optical pulses, the operation only took 15 optical pulses because the higher digits can be directly accessed. The process could be further sped up by setting the first addend directly using only one pulse per digit following the process shown in [116] by starting from the crystalline level and choosing the amorphization pulse energy to reveal the correct transmission state or using a double-pulse method as shown in [122] to arbitrarily reach all transmission levels independently of the initial level by manipulating the pulse shape. Following this approach, the number of pulses necessary to perform the calculation could be reduced to ten. The results shown in this section are carried out with two independent on-chip devices that are combined off-chip with the 90:10 splitter. In the next sections a waveguide crossing structure will be developed to address larger arrays of PCM-cells individually on chip.

5.3. Crossing design

When scaling a single PCM-cell to larger networks an important structure is the crossing of two waveguides to gain more flexibility in routing light signals in an integrated photonic circuit. The requirements that have to be fulfilled by such a waveguide structure are low insertion loss, meaning that no light gets lost (scattered out) and low crosstalk, meaning that all the light is transmitted over the crossing and not coupled to the intersecting waveguide, while maintaining a small footprint. The design of a waveguide crossing element as employed throughout this work is shown in Fig. 5.7a). The incoming waveguides are elliptically tapered to the crossing section and symmetrically tapered back to the initial waveguide width at the outputs. The elliptical tapering is chosen to expand the mode and avoid the excitation of higher order and radiation modes [141]. The waveguide width is tapered from the initial width of 1.2 µm to 3.0 µm at the intersection point with a taper length of 7 µm and therefore an overall footprint of $17 \times 17\,\mu m^2$. As indicated in Fig. 5.7a), the crossing parameters are first estimated using 3D-FEM (COMSOL Multiphysics) and 3D-FDTD (Meep [142]). A short pulse is launched in the left waveguide (source) and the electric field after a short time of propagation is shown. It can be seen that most of the light is transmitted across the intersection without crosstalk to the vertical waveguide.

Fig. 5.7b) and c) show optical micrographs of the fabricated devices used to determine insertion loss and crosstalk of the specific crossing design. In order to estimate the insertion loss, devices with different numbers of crossings are fabricated and the optical loss in a transmission measurement is plotted as a function of the number of crossings in Fig. 5.7d). The reference output (see Fig. 5.7b) is used to account for variations in the input power between the measurements of different devices. From the slope of the linear fit in Fig. 5.7d) the insertion loss of the waveguide crossing is obtained as (0.226 ± 0.003) dB/crossing (measured at a wavelength of 1550 nm)[2]. The crossing design is chosen as a compromise between device footprint, ease of fabrication in a single lithographic step and low insertion loss. Other approaches comprising crossing arrays exploiting Bloch waves or more sophisticated tapering exhibit an insertion loss below 0.03 dB [143, 144]. Using the device shown in Fig. 5.7c), the crosstalk to the intersecting waveguide is measured to be below −50 dB, limited by the sensitivity of the photodetector.

5.4. Two-pulse switching

The low loss waveguide crossing developed in the previous section can now be used to build larger waveguide crossing arrays combining several PCM-cells and reducing the addressing overhead by employing a two-pulse scheme as shown in the sketch in Fig. 5.8a). All PCM-cells can be individually accessed by sending two optical pulses to perpendicular waveguides that intersect at

[2] A similar measurement employing a standard waveguide crossing without tapering shows an insertion loss of (0.94 ± 0.03) dB/crossing.

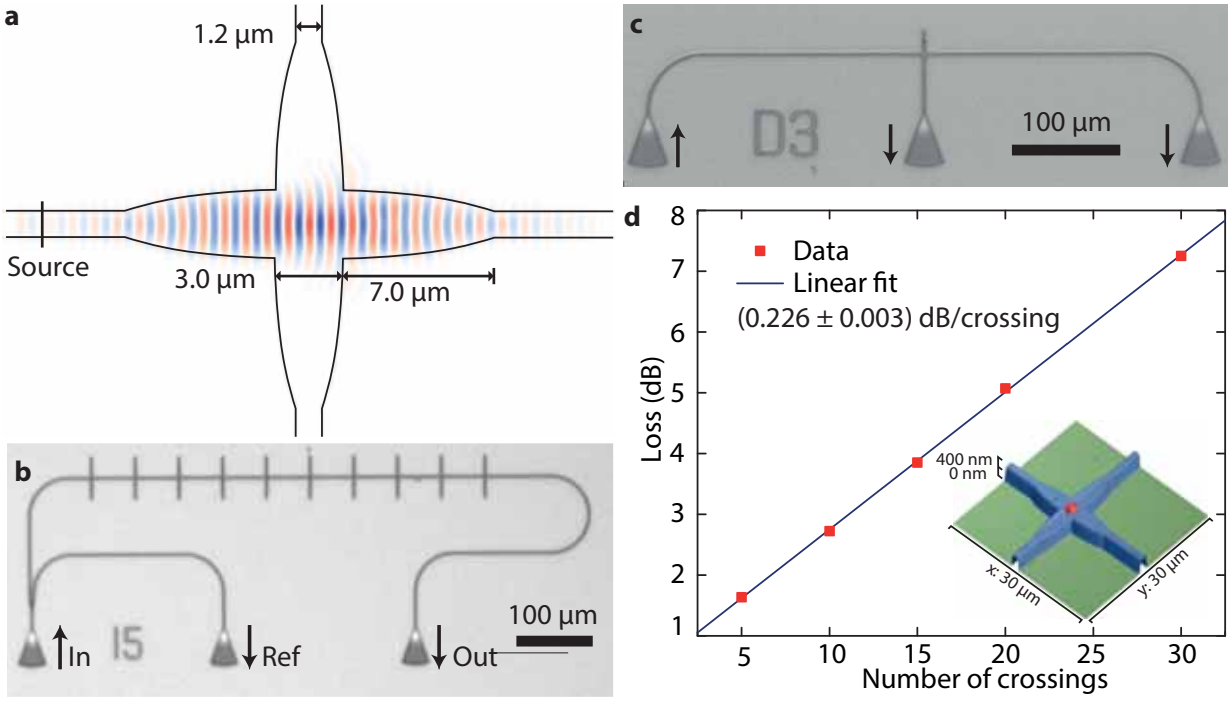

Figure 5.7.: Waveguide crossing. **a)** Design of the waveguide crossing using elliptical tapering. The transmission of a short optical pulse launched in the left waveguide simulated using 3D-FDTD shows low crosstalk. The overall footprint of the crossing is $17 \times 17\,\mu m^2$. **b)** and **c)** Optical micrographs of the fabricated devices to test crosstalk and insertion loss of the waveguide crossing. **d)** Transmission loss as a function of the number of crossings (see b)). The slope of the linear fit reveals an insertion loss of 0.23 dB/crossing. The inset shows a false-coloured atomic-force microscopy image of the fabricated crossing (adapted from [131]).

the selected cell. The pulse energies are chosen such that only the overlapping pulses can induce a phase transition in the PCM. This way it is possible to readout and switch $N \times N$ PCM-cells with only $N + N$ waveguides reminiscent of an electronic crossbar memory. The PCM is deposited in the crossing area (see Fig. 5.8b). By sending two pulses on perpendicular waveguides and carefully tuning the pulse energies, switching of only the selected PCM-cell at the intersection point can be achieved. The pulse energies have to be chosen in such a way that a single pulse does not affect the state of the PCM but two pulses together induce crystallization or amorphization.

5.4.1. Basic characteristics of the PCM-crossing

A scanning-electron micrograph of a waveguide crossing with a phase-change material (GST) deposited on top is shown in Fig. 5.8b) proofing good alignment between the photonic waveguide crossing and the phase-change material. An important parameter when switching the PCM-cell via two perpendicular waveguides is the delay between the arrival times of the two pulses at the waveguide crossing. Both pulses will heat up the PCM but if the time separation between the two allows for the PCM to cool down before the second pulse arrives, no switching will be induced.

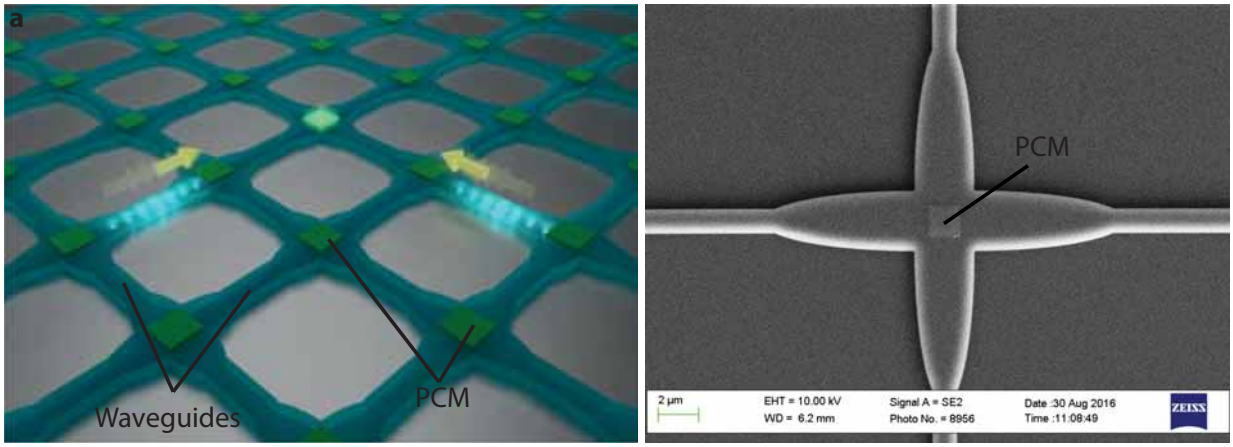

Figure 5.8.: Waveguide crossing array. a) Sketch of a waveguide crossing array illustrating the two-pulse addressing scheme. The PCM-cells deposited on top of the waveguide crossings can be individually addressed using two perpendicular pulses. **b)** Scanning electron micrograph of a fabricated waveguide crossing with PCM deposited in the middle [131].

This operation therefore represents a logical AND.

Fig. 5.9a) shows this relation in more detail. In this experiment optical pulses with a width of approximately one picosecond are used and sent to the PCM-cell via the perpendicular input waveguides with a variable delay (more details on the experimental setup are given in App. A.2). As expected, when the pulses arrive at the same time the contrast in the optical transmission is the highest and if the delay between the pulses is larger than 150 ps, no change in the optical transmission is observed. The data confirms that a single pulse does not excite the PCM to switch its phase of matter. The maximum contrast which can be achieved can be tuned by the size of the PCM-element. As depicted in Fig. 5.9a) the largest change in optical transmission is gained if the complete crossing section ($3\,\mu m \times 3\,\mu m$) is covered. To proof that the two-pulse switching scheme does not prevent the ability of multilevel operation, Fig. 5.9b) shows amorphization in a single step and stepwise recrystallization over eleven levels enabling arithmetic operations in base ten.

5.4.2. Optical random-access in a photonic crossbar array

Combining the basic crossing structure to larger arrays leads to an elementary random-access memory (see Fig. 5.10a) similar to electronic crossbar memories. The memory consists of a 3×3 waveguide crossing array with the PCM elements deposited on each crossing. Exploiting the two-pulse switching scheme described in the previous section, a two-dimensional addressing of the memory cells is enabled. By selecting the corresponding row and column waveguides the element on their intersection can be addressed individually, as the phase-state of the phase-change material is not altered by a single pulse but only by the two overlapping pulses. In Fig. 5.10b) it is experimentally demonstrated that only the selected PCM-cell is switched. Therefore, the

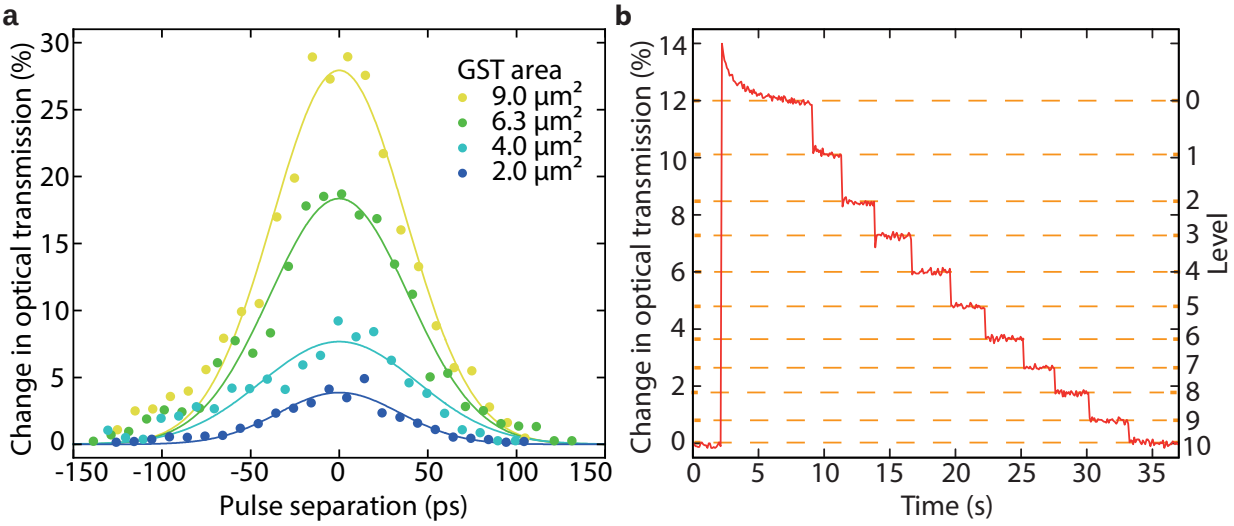

Figure 5.9.: Two-pulse switching. a) Optical transmission contrast as a function of the pulse separation for different sizes of the GST patch on the crossing. **b)** Multilevel operation using the two-pulse switching scheme. The initial decay after amorphization is introduced by the probe power that induces crystallization if it is chosen too high [145].

optical transmission measurements for the rows with the elements '$c+d$' (red) and '$a+b$' (blue) (see inset of Fig. 5.10b) are simultaneously monitored, while sending optical pulses for switching element 'd' with the two-pulse switching scheme. From the fact that the transmission through cells '$a+b$' remains constant during the switching process it can be concluded that only cell 'd' is switched.

The maximum size of such an array is limited by the maximum pulse energy that does not switch a PCM-cell alone and the absorption of light along several PCM-cells and can be estimated in the following way: with the maximum power without switching a PCM-cell with a single pulse P_0 and the loss per crossing and PCM α, the maximum power of a single pulse arriving at the Nth element in a row can be calculated to be

$$P_{\mathrm{N}} = P_0 \cdot (1-\alpha)^{(N-1)}\,, \tag{5.1}$$

so that the total power at the intersection with the combined perpendicular pulses is

$$P_{\mathrm{NN}} = 2 \cdot P_0 \cdot (1-\alpha)^{(N-1)}\,. \tag{5.2}$$

In order to switch the PCM cell, the relation $P_{\mathrm{NN}} > P_0$ has to be fulfilled leading to the expression

$$\alpha < 1 - \sqrt[N-1]{\frac{1}{2}} \tag{5.3}$$

for the loss per crossing. For $N = 3$ as in the experiment this for example leads to $\alpha < 0.3$

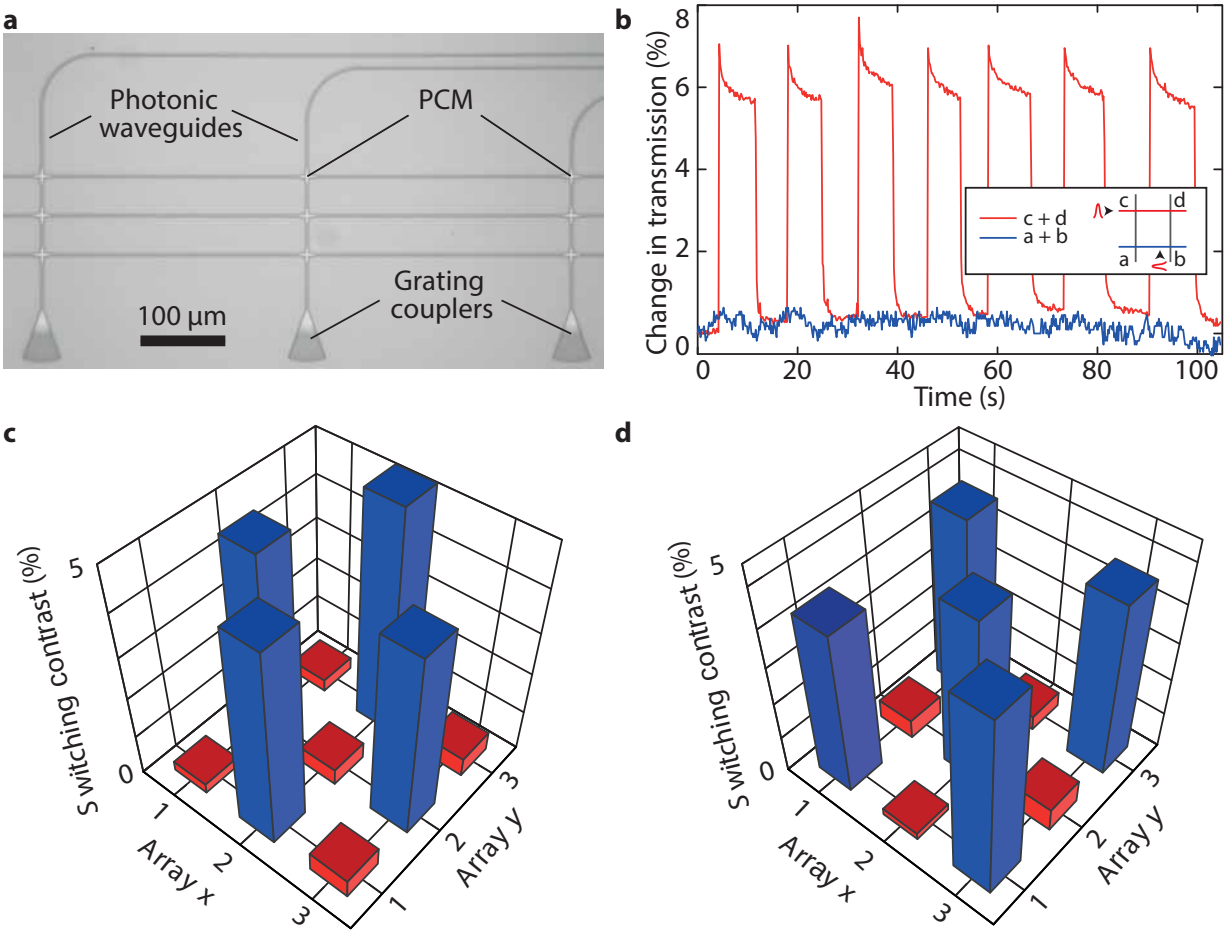

Figure 5.10.: Random-access with a photonic waveguide crossing array. **a)** Optical micrograph of a crossing array with nine PCM-cells. **b)** Independently switching cell '*d*' in a 2×2 array. **c)** and **d)** Readout of two complementary states in a 3×3 photonic waveguide array (adapted from [131]).

or 1.5 dB and for $N = 10$ to $\alpha < 0.074$ or 0.33 dB limiting the maximum contrast that can be achieved within a single cell. It should be noted that for a given loss coefficient α a matrix size of $2N \times 2N$ can be accessed by sending the pulses from all four sides of the waveguide array so that the longest optical path to a specific element involves passing N crossings. Several of these arrays can then be combined to obtain a larger photonic memory. The downside of this approach is that a destructive multilevel readout as explained in the next section has to be used to read the state of a certain memory element. This is due to the fact that a single probe pulse that is sent to only one waveguide is always influenced by all PCM-cells along its way.

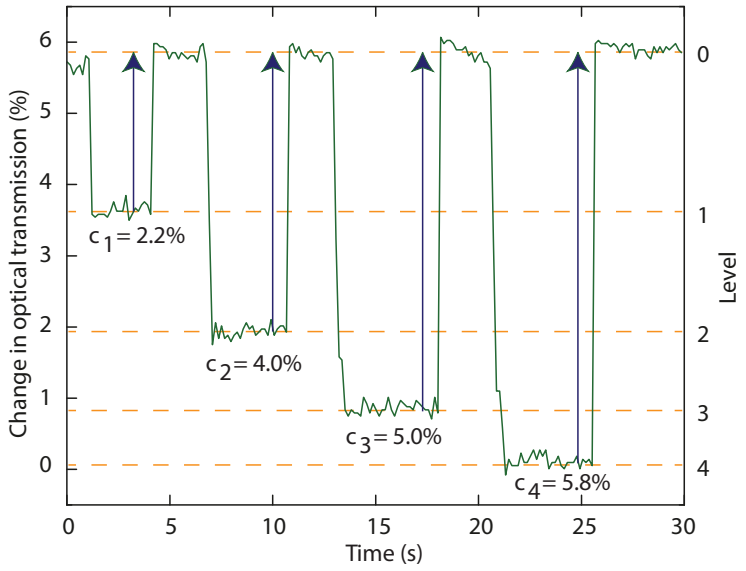

Figure 5.11.: Destructive multilevel readout. For reading the state of a PCM-cell in the waveguide array, a reset is performed by sending an amorphization pulse. The change in the optical transmission reveals the state of the PCM-cell. As the information of the cell is deleted during the amorphization, a subsequent write pulse must be employed to store the value in the memory element again.

5.4.3. Destructive multilevel readout

For reading the cells of the waveguide crossing array, a destructive method is used, meaning that the information in the specific cell is erased during readout. The same addressing scheme as for writing the cell is also employed for reading, with the difference that the pulse energies are not chosen to set a specific level but to fully amorphise the phase-change material. The readout principle is shown in Fig. 5.11, where the transmission is plotted as a function of time. Four different levels are programmed successively and reset (blue arrows) to the fully amorphous state to demonstrate the readout-scheme. As can be seen, the change in the optical transmission depends on the amount of crystallization (c_1-c_4) induced for a specific level and can therefore be used as a readout method. Because the information is erased in this process, the previous state has to be rewritten after readout. However, it should be noted that improved structures making use of wavelength-division multiplexing can be used to enable fully scalable all-optical photonic memories with non-destructive readout [119].

5.5. Conclusions

In this chapter an abacus-like all-optical basic calculation unit for arithmetic processing was demonstrated that provides arithmetic in arbitrary bases. Other than conventional electronic processors, which operate with binary data, the photonic abacus is based on multilevel operation

and can therefore readily carry out calculations in bases that suit the arithmetic problem. The principle of the calculation unit is based on counting light pulses and storing them in different states of crystallization of a phase-change material. Because of the non-volatility of the PCM, the all-optical abacus provides an energy efficient storage of the calculation results and more importantly circumvents the von Neumann bottleneck by processing the data directly in the memory.

Employing pulses of one picosecond width, pulse energies below 20 pJ and switching times in the nanosecond regime, the optical abacus holds promise for fast and efficient arithmetic units operating at GHz speeds. Using a novel two-pulse switching scheme in a waveguide crossing array individual addressing of arbitrary PCM-cells in the array is achieved enabling scalability to larger networks of PCM-cells. This way a photonic random-access memory was implemented using a destructive readout scheme.

Chapter 6.
6 All-optical neural network

Complementary to the previous chapter describing arithmetic processing based on counting optical pulses, in this chapter an all-optical neuromorphic architecture inspired by biological brains will be discussed. Although traditional computers are much faster and more precise in arithmetic calculations than humans, brains succeed over them by orders of magnitude in speed and energy efficiency when dealing with large datasets and the extraction of features out of these [56]. The main reason for this is a fundamentally different approach to handle the data. As detailed in Sec. 2.1.1, in a conventional computer memory and processor are physically separated and commands are executed one after the other. Human brains instead consist of billions of neurons with even more interconnects (synapses) in which the information is stored and processed in a highly parallel way. The 'memory' is represented by the synapses and therefore directly part of the processing unit without architectural separation between memory and processor.

In this chapter, the development of the single building block of a neural network – a neuron – on the phase-change photonics platform will be described, its functionality experimentally validated and the different learning rules as described in Sec. 2.2.2 will be applied. Having implemented a single neuron, a way of connecting it to larger neural networks is detailed in Sec. 6.2. After describing the principles on which the scalability of the architecture relies, an experimental example of a neural network consisting of four neurons and sixty synapses capable of simple pattern recognition tasks will be demonstrated. In the last section, the possibilities of larger neural networks with more synapses per neuron are explored via simulations based on the experimental results. The experiments presented in this chapter are based on [146].

6.1. Single neurons

As previously stated, a neuron is the basic unit of a neural network. In its abstracted form inspired from biological neurons it consists of three main parts: the synapses acting as the inputs to a neuron and weighting the incoming signals coming from previous neurons in the network, the

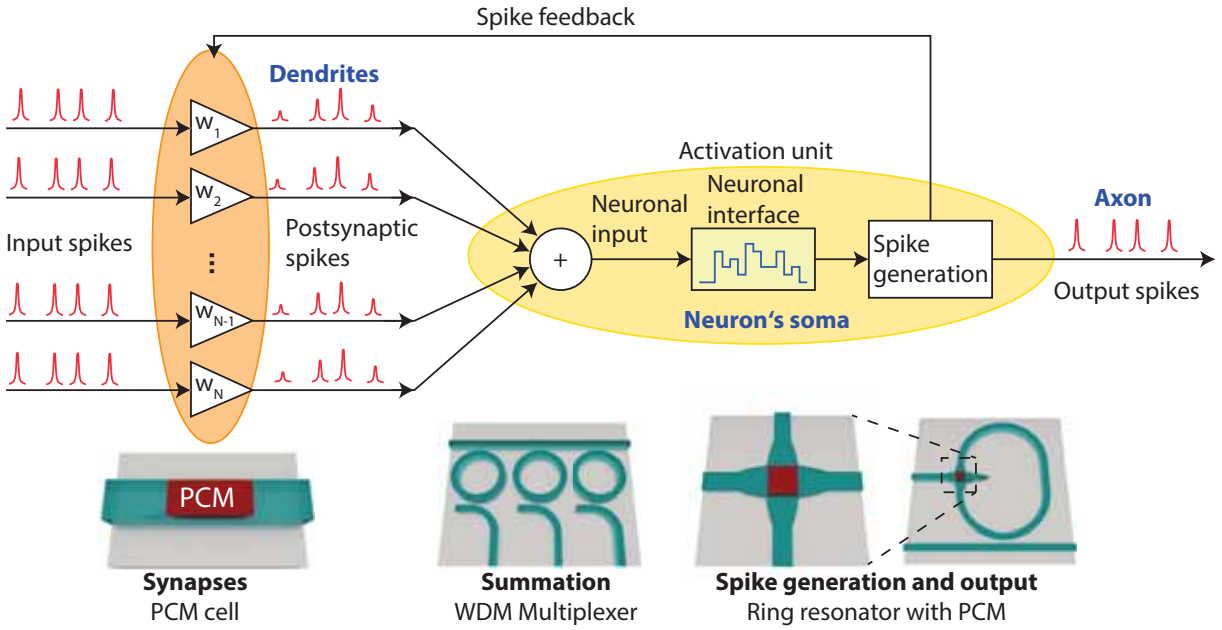

Figure 6.1.: Schematic of an artificial neuron. Different input signals from previous neurons are weighed in the synapses and the postsynaptic spikes are summed in the neuron's soma that then decides if an output spike is generated and sent to the next neuron via the axon. The output signal is also guided back to the synapses to impose a feedback to implement a learning rule. The waveguide structures below sketch the photonic implementation of the individual parts of the neuron.

summation of the weighted input signals and finally the activation unit, which decides whether an output signal is generated and sent to the next layer of neurons.

6.1.1. Concept of the artificial neuron

The schematic in Fig. 6.1 depicts the underlying principles of a single artificial neuron in comparison to its biological counterpart. The input spikes, which can be either output signals from previous neurons or encoded input data, are weighted in the synapses (orange ellipse) with the weights w_1 to w_N. These are factors that the input spikes are multiplied with and represent the strength of a connection between two neurons. The dendrites lead the weighted postsynaptic spikes to the neuron's soma. Here, the incoming signals are first summed up and then compared to a certain threshold and the decision if an output spike is generated is taken (activation unit). Potential output spikes are both guided to the next layer of neurons via the axon and back to the synapses to enable the possibility of implementing a learning rule and update the connections between the neurons.

6.1.2. Weighting mechanism – the synapses

In the all-optical neuron presented in this work, the weighting of the input signals is implemented by a basic phase-change cell that intrinsically offers the desired features of a synapse via its variable absorption of light in its different states. As shown in Fig. 6.1, the phase-change material is deposited on a waveguide corresponding to a dendrite that connects two neurons. By tuning the absorption of the phase-change material, the incoming signal can be weighted from the amorphous state, in which all light is transmitted ($w = 1.0$), gradually down to the fully crystalline state implementing the weight $w = 0.0$. As will be seen later, in the synapses and therefore in the phase-change material the information about features that the neural network is able to extract are stored after a suitable training process.

Opposed to the most software-based implementations of artificial neurons, the hardware implementation described in this work is restricted to only positive weights between 0.0 and 1.0, limiting the capabilities of the neural network and restricting standard training algorithms. However, the capability of pattern extraction and self-learning is still preserved and additionally a correlation detection between patterns is enabled, as will be shown in the simulations at the end of this chapter. The lack of negative weighting prohibits cancellation of two input signals to the neuron. In a more complex photonic implementation of the synapse for example taking advantage also of the phase of the light and therefore allowing for constructive and destructive interference, also negative weights could be achieved [147]

It should be noted that throughout this chapter a weight of 1.0 always corresponds to the amorphous state where almost all of the light is transmitted. Other than what might be intuitively assumed, in the fully crystalline state ($w = 0.0$) some light is still transmitted although much less than compared to the amorphous state. However, this has an influence on the overall activation energy after the summation as will be discussed later.

6.1.3. Summing the weighted inputs – the multiplexer

After weighting the input signals they must be summed to retrieve the activation energy that is fed into the activation unit. To do this, all input spikes coming from different waveguides are added up in a single waveguide using the wavelength division multiplexing (WDM) technique [64, 148]. By using different wavelengths for each input interference effects can be avoided, removing the need to also control the phase of the incoming electromagnetic waves. WDM is a technique in which multiple optical signals on different wavelengths are combined into a single optical fibre or waveguide and is heavily applied in telecommunication networks to distribute different signals and transport them on the same physical channel.

Fig. 6.2a) shows the basic principle of a wavelength multiplexer based on optical ring resonators. Ring resonators are chosen because they allow for low-loss filtering of wavelengths and have a compact footprint compared to other multiplexers as arrayed waveguide gratings or multiplexers

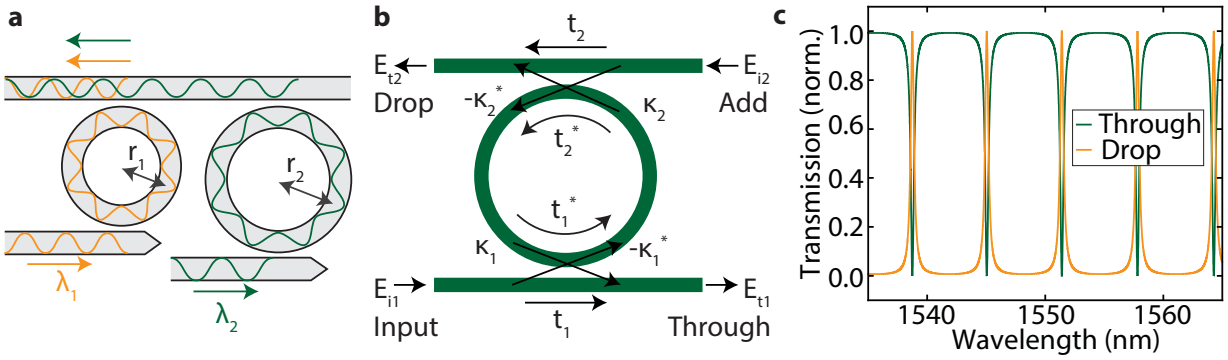

Figure 6.2.: Wavelength division multiplexing with ring resonators. a) Using ring resonators with different radii, different wavelengths can be combined into a common waveguide. **b)** Sketch of an optical micro-ring resonator in add-drop configuration with the individual coupling parameters. **c)** Typical through- and drop-port spectrum of a critically coupled resonator assuming no loss in the waveguides, $t_1 = 0.92$ and a radius of 30 µm. On resonance, all the light from the input-port is transferred to the drop-port, whereas off-resonance all the light is transmitted.

based on Bragg couplers. Additionally, they mainly rely on only two design parameters and offer a way of tuning the extinction ratio of the resonances via the coupling efficiency. A ring resonator acts like an optical filter. If a feeding waveguide is brought in near vicinity to the resonator waveguide, the incoming light can couple to the ring if it fulfils the resonance condition $2\pi r n_{\text{eff}} = m\lambda$, with the radius r, the effective refractive index of the propagating mode n_{eff}, the wavelength λ and m being a positive integer number. Thus, only if the optical pathlength of the resonator is equal to a multiple of the wavelength of the propagating light, it can interfere constructively inside the ring resonator. If a second waveguide is added to the configuration a so called add-drop resonator is obtained. Now the incoming light can couple to the second waveguide via the ring resonator. As only wavelengths satisfying the resonance condition will couple to the second waveguide, this structure can be used for wavelength division multiplexing by employing resonators with different radii as depicted in Fig. 6.2a).

Besides this basic concept of ring resonators based on the resonance condition giving a natural and qualitative understanding, a more fundamental derivation is necessary to also make quantitative predictions about what fraction of the incoming light is coupled to the different ports of the add-drop resonator. Starting from a coupling matrix based on the variables defined in Fig. 6.2b) and following [149], the complex mode amplitude for the through-port can be derived as

$$E_{\text{t1}} = \frac{t_1 - t_2^* \alpha e^{i\Theta}}{1 - t_1^* t_2^* \alpha e^{i\Theta}}. \tag{6.1}$$

Here, t_1 and t_2 denote the coupling coefficient from the input- to the through-port and from the add- to the drop-port, respectively (the $*$ indicates the complex conjugate). α is the loss factor with $\alpha = 1$ denoting no loss inside the resonator and $\Theta = 4\pi^2 n_{\text{eff}} r/\lambda$, establishing the

relation to the resonator geometry. Similarly, the mode amplitude obtained at the drop-port can be calculated to be

$$E_{\mathrm{t2}} = \frac{-\kappa_1^* \kappa_2 \sqrt{\alpha} e^{i\Theta/2}}{1 - t_1^* t_2^* \alpha e^{i\Theta}}. \tag{6.2}$$

Fig. 6.2c) shows a typical through-port ($P_{\mathrm{through}} = |E_{\mathrm{t1}}|^2$) and drop-port ($P_{\mathrm{drop}} = |E_{\mathrm{t2}}|^2$) spectrum of a ring resonator assuming the waveguide geometry used throughout this work, a radius of 30 µm and no propagation loss. It can be seen that for wavelengths fulfilling the resonance condition all the light from the input-port is transferred to the drop-port. The light off-resonance remains unaffected and is transmitted to the through-port.

Knowing the relation between coupling efficiency, propagation loss and the transmission factors t_1 and t_2, it can be concluded that the maximum coupling efficiency is achieved in a symmetric configuration with equal input and output gaps between waveguides and the resonator, if the propagation loss is zero. If the waveguide loss instead cannot be neglected, the maximum drop-port transmission on resonance is found from (6.2) and leads to the condition

$$\alpha = \left| \frac{t_1}{t_2} \right| \tag{6.3}$$

for the coupling parameters and therefore the gaps between input and output waveguides and the ring resonator. More details on the exact derivation can for example be found in [149].

By adding several ring resonators with different radii to the same waveguide, all resonance wavelengths can be combined on the common drop waveguide. In the context of the neural network this technique can be used to sum all the weighted input powers coming from previous neurons on a single waveguide.

When designing a WDM-filter out of ring resonators it is important to control the spacing $\Delta\lambda$ between the resonance wavelengths of the individual resonators in order to obtain non-overlapping resonances and minimized crosstalk between the multiplexer channels. A simple relation between the change of the ring resonator radius Δr and the wavelength shift $\Delta\lambda$ for a given resonance wavelength of the first resonator λ_0 and its radius r_0 can be obtained from the resonance condition (assuming a negligible change in the effective refractive indices for the given wavelength shift):

$$\Delta r = \frac{\Delta\lambda}{\lambda_0} r_0. \tag{6.4}$$

To account for fabrication offsets in the designed multiplexers, several ring resonators in add-drop configuration are fabricated and the experimental wavelength shift as a function of the designed (based on (6.4)) shift is evaluated. Fig. 6.3a) shows measured transmission spectra for fifteen ring resonators with a designed wavelength spacing of 0.5 nm. It can be seen that the resonances have little overlap in their minima and are equidistantly spaced. Due to fabrication imperfections, the actual wavelengths spacing differs from the designed shift and is determined to be $\Delta\lambda =$

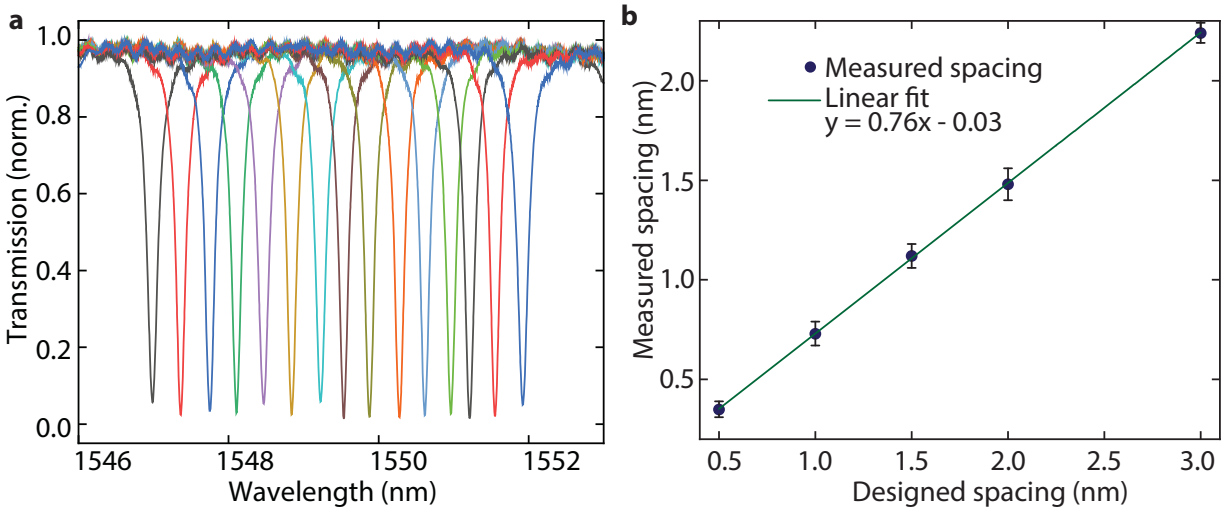

Figure 6.3.: Characterisation of the wavelength division multiplexer. a) Through-port transmission spectra for fifteen individual ring resonators showing non-overlapping and equidistant resonances. The ring radius was constantly increased from left to right to achieve a wavelength spacing of 0.35 nm. The radius of the left most resonator is 30 µm. **b)** Measured wavelength spacing for different designed spacings of the resonance wavelengths. Each datapoint is obtained by measuring fifteen resonators as shown in a).

(0.35 ± 0.04) nm. By measuring the optical spectra as in Fig. 6.3a) and extracting the wavelength shift for different designed spacings, the relation between measured and designed spacing can be obtained (see Fig. 6.3b). From the linear fit the parameter to be used when fabricating a wavelengths multiplexer for a specific distance between two channels can be extracted. The small standard deviation (error bars) obtained for each datapoint underline the reliable fabrication of several ring resonators with different radii.

6.1.4. Activation unit

After weighting and summing up the input vectors, the signal is fed to an activation unit. This part of the neuron decides if an output pulse is emitted or not based on the comparison of the weighted input signal with a certain threshold power. As explained in Sec. 2.2.1, the activation functions can be of different forms like for example a step or sigmoid function. The activation function implemented in this work is exploiting the switching threshold of a PCM-cell and a ring resonator as shown in the scanning-electron micrograph of a fabricated device in Fig. 6.4a). The weighted input light is guided to the PCM on a crossing in a ring resonator. A bus waveguide passing the ring on the bottom is employed as neuron output and connected to the next layer of neurons.

The activation unit works as follows: the PCM on the ring is initially in the crystalline state and a probe pulse is sent through the bus waveguide on resonance so that it is coupled to the resonator and no light exits the output waveguide. If now the weighted input power that is guided

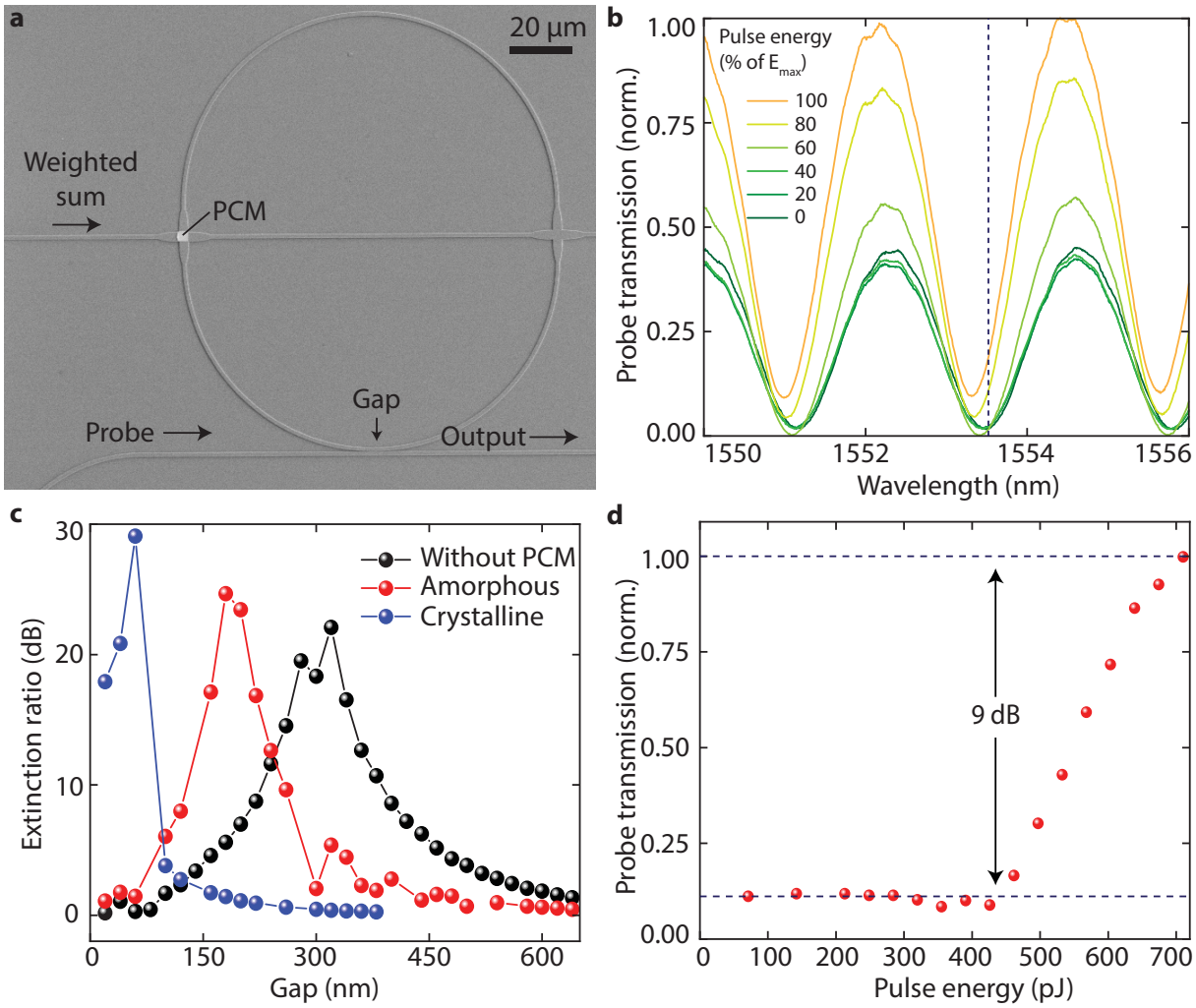

Figure 6.4.: Activation unit. **a)** Scanning-electron micrograph of the activation unit comprised of a ring resonator with bus waveguide (probe) and a waveguide crossing with a PCM on top. Note that the second intersection of the waveguide with the ring on the right is solely used as a test port and is not related to the function of the unit. **b)** Optical spectra of the ring resonator obtained from a transmission measurement through the probe path after sending different pulse energies to the PCM starting from the crystalline state. Only above 40% of the maximum pulse energy (720 pJ), a significant change in the spectra is visible due to amorphization of the cell. Both resonance wavelength and resonance depth change upon the phase transition. **c)** Extinction ratio as a function of the gap between probe waveguide and resonator. The lower the optical loss in the resonator, the larger the gap to achieve critical coupling (highest extinction ratio). **d)** Output transmission of the activation unit at a fixed wavelength (indicated by the dashed line in b) as a function of the pulse energy. The transmission stays almost constant at a low level till approximately 420 pJ. Above this energy amorphization is induced leading to a linear increase of the probe transmission resembling the rectified linear unit function (adapted from [146]).

to the crossing in the ring exceeds the switching threshold of the PCM, it is partly amorphized. As a consequence, the resonance condition of the resonator changes because the real part of the refractive index of the PCM changes, and therefore the optical pathlength of the ring. Now the resonance wavelength of the ring is shifted and the probe pulse does not longer couple to the ring but is transmitted to the output. In addition to the shift of the resonance wavelength, also the change of the imaginary part of the refractive index of the PCM plays an important role in generating the activation function. The extinction ratio, a measure for the depth of the resonance dip, and the point of critical coupling (with the lowest transmission to the through-port) of a certain ring resonator configuration strongly depend on the propagation loss inside the ring and the gap to the bus waveguide. Because the gap is fixed, the extinction ratio alters when changing the waveguide loss.

Both effects can be observed in the measured spectra in Fig. 6.4b). The plot shows the transmission spectra (measured using the evanescently coupled probe waveguide) of the activation ring resonator after supplying different input energies (weighted sum) to the PCM on the intersection. Before sending a different pulse energy, the PCM state is crystallized to its ground state. Firstly, it can be seen that only above a pulse energy of approximately 40% of the maximum energy the spectra start to change significantly. At this point the optical pulses start amorphizing the phase-change material. Secondly, with increasing degree of amorphization (higher pulse energies) the resonance wavelength slightly shifts to smaller wavelengths. This is expected, as the optical path length of the resonator decreases with decreasing real part of the refractive index. At the same time the extinction ratio decreases because the waveguide loss decreases as shown in Fig. 6.4c). Here the extinction ratio (ER) in dB is plotted[1] as a function of the gap between resonator and waveguide before the PCM deposition, with amorphous PCM (as deposited) and after crystallizing the PCM on a hotplate at 230 °C for 15 min. For a ring resonator with only one bus waveguide the point of critical coupling (highest extinction ratio) is given for $\alpha = |t|$, with α being the loss factor and t the transmission coefficient (see Sec. 6.1.3) [149]. The lower the propagation loss ($\alpha \to 1$) the higher the transmission coefficient and therefore the larger the gap between resonator and waveguide. This relation is also reflected in the graph in Fig. 6.4c): the lowest propagation loss is obtained before deposition of the PCM (as also the amorphous PCM absorbs a small fraction of the light), leading to critical coupling at approximately 300 nm. With the PCM in the crystalline state the highest extinction ratio of 30 dB is found at a gap of only 70 nm, revealing the desired working point of the activation unit. As the transmission of the probe pulse should be low in the initial (crystalline) state, the ring resonator configuration is chosen to be critically coupled. By amorphizing the PCM now, the extinction ratio is significantly reduced as can be seen from the plot for the amorphous PCM in Fig. 6.4c), contributing strongly to the activation function obtained in Fig. 6.4d).

[1] The extinction ratio is calculated as $\mathrm{ER} = 10 \cdot \log_{10}(\frac{P_{\max}}{P_{\min}})$, with $P_{\max}$ being the maximum power (between the resonances) and $P_{\min}$ the minimum transmitted power (on resonance).

Fig. 6.4d) shows the transmission taken from the spectra at a fixed wavelength (indicated by the dashed line in Fig. 6.4b) revealing a curve similar to the ReLU-function. The output of the neuron is very low for pulse energies below 420 pJ and then increases linearly for higher energies by up to one order of magnitude (9 dB). The maximum pulse energy is chosen in order to achieve sufficient optical contrast without damaging the PCM. The contrast could be further improved by optimizing the ring resonator configuration by designing it to be closer to critical coupling in the initial crystalline state (so that no light is transmitted). After probing the output of the neuron, the activation unit must be reset to the crystalline state in case it was switched. From Fig. 6.4b) it can be concluded that the insertion loss of the activation unit is high, as only approximately 25% (−6 dB) of the probe pulse is transmitted to the output. This could further be improved by optimizing the resonator geometry. Reducing the optical loss of the waveguide crossing or using smaller PCM-patches to reduce the absorption can potentially lead to narrower resonances (higher quality factors) and improve the influence of the resonance shift.

6.1.5. Supervised learning

Having implemented all the essential elements comprising a single neuron, they can now be combined to a complete neurosynaptic system. In a first experimental step a four-input neuron is fabricated and operated in a supervised mode. Fig. 6.5a) shows an optical micrograph of the fabricated photonic neuron with four synapses. The input signal is encoded on different wavelengths, sent to the input grating couplers illustrated by the coloured arrows and weighted with the PCM-synapses. Exploiting the filtering capabilities of micro-ring resonators, the weighted input signals are combined on a single waveguide leading to the activation unit. The output spike is generated, as described in the previous section, if the PCM on the bigger resonator on the right is switched and the probe pulse transmitted to the output.

In the case of supervised learning the weights of the neuron are set in a training process by an external supervisor. In the given example, the neuron is supposed to differentiate between simple pixel patterns consisting of four black or white pixels as shown on the bottom of Fig. 6.5b) and c). The four pixels are encoded in four optical pulses that are sent to the four inputs of the neuron. Fig. 6.5b) shows an example of a neuron that is able to recognize the simple pattern ‘1010’, meaning that a pulse is sent to the first and the third input of the neuron. In this case training of the neuron simply consists of making the first and third PCM transmissive (amorphous) and keep the other weights crystalline. This is achieved by sending optical pulses for switching (off-resonance) to the corresponding PCM-cells. In the given example both input pulses of the ‘1010’ pattern are fully transmitted to the activation unit and the combined pulse powers are high enough to induce switching of the PCM on the activation ring leading to an output pulse as depicted in Fig. 6.5b). It should be noted that a single pulse (e.g. ‘1000’ or ‘0010’) is not enough to switch the activation PCM. For the other patterns shown on the bottom, part of the incoming

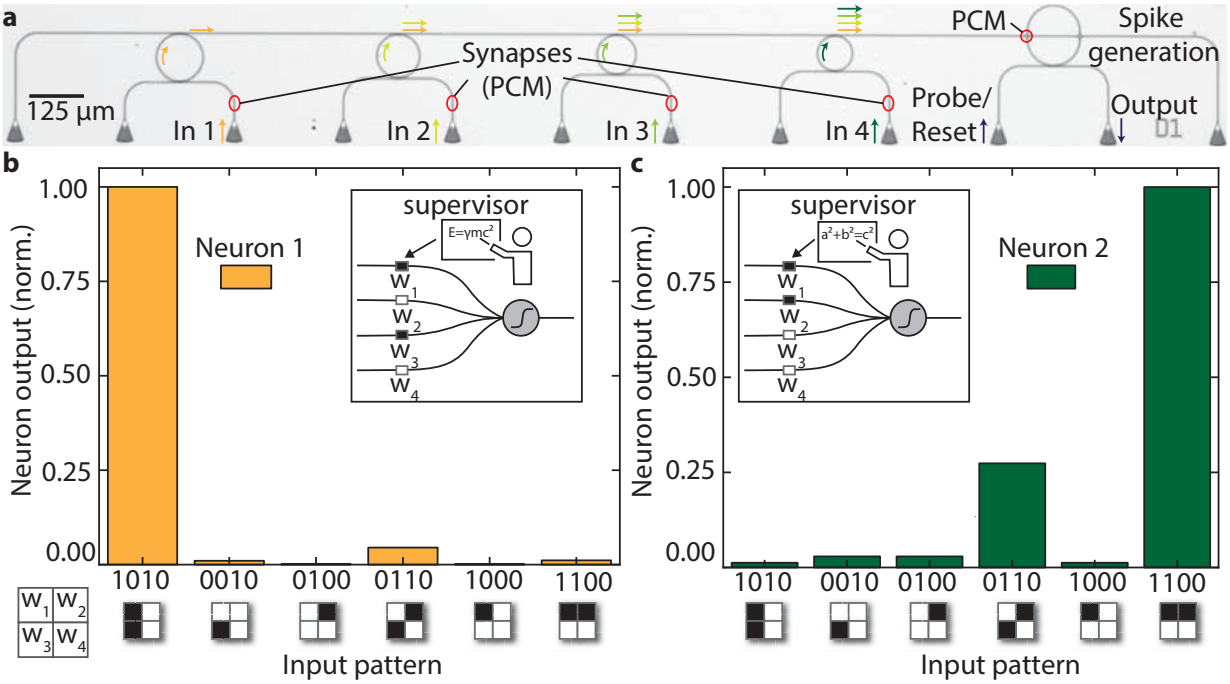

Figure 6.5.: Single all-optical neuron. a) Optical micrograph of a fabricated photonic neuron with four inputs. The four inputs on different wavelengths (colour coded) are combined using micro resonators and guided to the PCM of the activation unit. Using the probe pulse and the bigger activation ring resonator a spike is generated if the PCM is switched. **b)** and **c)** Output of two neurons trained to recognize different patterns. After setting the synapses (weights/PCM) by an external supervisor, the neurons only respond to the specified pattern (adapted from [146]).

light is absorbed in a crystalline PCM-weight and the energy arriving at the activation unit is not enough to trigger an output pulse.

Similarly, a second neuron is trained to recognize the pattern '1100' (see Fig. 6.5c), illustrating the successful implementation of an all-optical neuron. Other than electronic implementations on conventional computers that have to carry out the weighting, summation and activation function over several clock cycles including loading the weights from a separated memory, the optical implementation presented in this work performs the complete operation in parallel in a single time step.

Besides the strong output signal obtained for the correct pattern, the second neuron also reacts to the pattern '0110'. The reason for this is that because of imperfections in the fabrication and also the experimental setup, not all optical pulses that arrive at the neuron and the activation unit have the exact same power. In this case the power on input two of the neuron contributes most to the activation energy and as this input belongs to the learned pattern and its synapse is transmissive, also the pattern '0110' with its additional energy on input three induces switching of the activation unit.

Although this malfunction can be easily resolved by a better fine tuning of the pulse energies

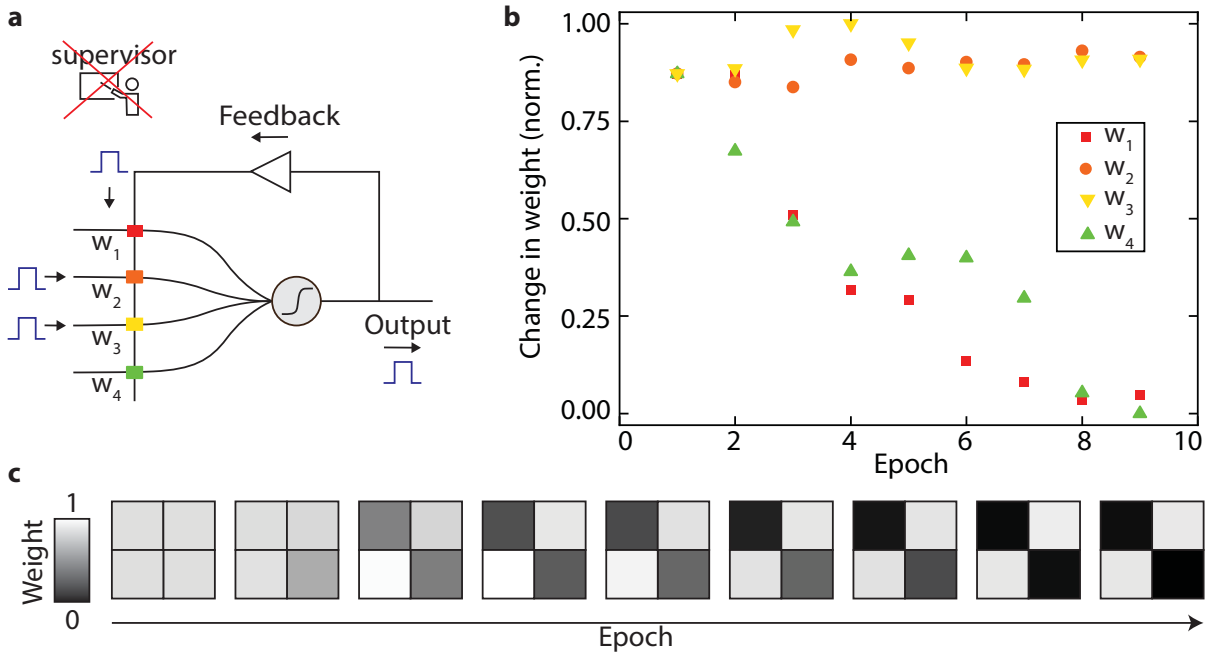

Figure 6.6.: Unsupervised learning. **a)** Schematic of the implementation of unsupervised learning in the optical neuron. Part of the optical output pulse is guided back to the synapses via a feedback loop updating the weights following the Hebbian learning rule. **b)** Experimental demonstration of unsupervised learning. The four synaptic weights of a single neuron are monitored over several epochs while repeatedly showing the pattern '0110'. While the two weights corresponding to the input pattern stay transmissive, the two unrelated weights continuously decrease until the neuron has fully adapted to the pattern. **c)** Evolution of the weights over time. Starting in the amorphous (transmissive) state, only the two synapses not contributing to the output spike are crystallized (black) over several epochs (adapted from [146]).

and also using more input signals instead of only four, it still demonstrates a disadvantage of this neuron implementation that only allows for positive weights [0..1]. All patterns that include the pulse pattern the neuron is trained to recognize, like for example '1111' or '1110' will also trigger an output pulse. In an implementation including negative weights, the input signals would cancel out (for example through destructive interference) so that solely the trained pattern is recognized.

6.1.6. Unsupervised learning

Supervised learning as described in the previous section can only be applied if a training set of data is present with known output for certain input values (see Sec. 2.2.2). These datasets can then be used to optimize the weights of the neural network to fulfil a certain task for example using the error backpropagation algorithm. If such a dataset is not available, for example because the pattern to be found in a dataset is not known before, unsupervised learning methods can be applied in which the weights of the neural network are automatically adjusted without the help of an external supervisor.

Fig. 6.6a) illustrates how this can be implemented with the optical neuron developed in this work. The general structure of the neuron is the same as in the supervised case, only that a part of the output spike is guided back to the weights of the neuron in a feedback loop. This setup mimics the spike timing dependent plasticity (STDP) that is observed in biological neurons in a very basic way [76, 150, 151]. In STDP connections between neurons are strengthened and weakened based on the time difference between their input and output spikes. If a subsequent neuron fires right after the previous neuron, it is likely that the previous neuron contributed to triggering the output spike of the second neuron and therefore their connection is strengthened. This rule is often simplified as the Hebbian learning rule: 'neurons that fire together, wire together' [76]. The weight is adjusted depending on the time delay between two spiking events of both neurons. Negative time delay (the second neuron spikes before the first neuron) leads to weight depression, whereas positive time delay leads to a potentiation [152]. The strength of the induced change of the weight depends on the absolute time difference between two events and is maximized for small delays.

In the photonic implementation with the feedback loop, part of the output pulse is fed back to the weighting elements and the input pulse length is chosen long enough to overlap with the feedback pulse. The pulse energies are chosen in such a way, that the combined input and feedback pulse lead to amorphization of the PCM and therefore strengthening of the connection. A single feedback pulse on the other hand induces crystallisation and weakening of the synapse based on the same principles as employed for the two-pulse switching scheme in Sec. 5.4. With this configuration all inputs that contributed to triggering the output spike are enhanced while the others are softened. Fig. 6.6b) and c) show the experimental results obtained from a single neuron with four inputs operated with the feedback loop. All weights are initially set to the amorphous state, leading to high transmission and a pulse pattern '0110' is repeatedly sent to the neuron. The two incoming pulses are fully transmitted to the activation unit and induce the generation of the output pulse that is partly transferred back to the synapses via the feedback loop. The weights of the two contributing pulses (w_2 and w_3) stay amorphous as they feel the input and the feedback pulse. The inactive inputs are partly crystallised by the feedback pulse, so that the input pattern is transferred to the weights of the neuron after several epochs. Now the neuron will only spike if the correct input pattern is shown similar to the case of the supervised neuron. Fig. 6.6c) visualizes the evolution of the synapses according to the input pattern. Starting from four equal amorphous weights the input pattern is transferred to the synapses after nine epochs. Employing this learning rule, repeating patterns in an unknown data stream can be extracted as shown in more detail in Sec. 6.4.1.

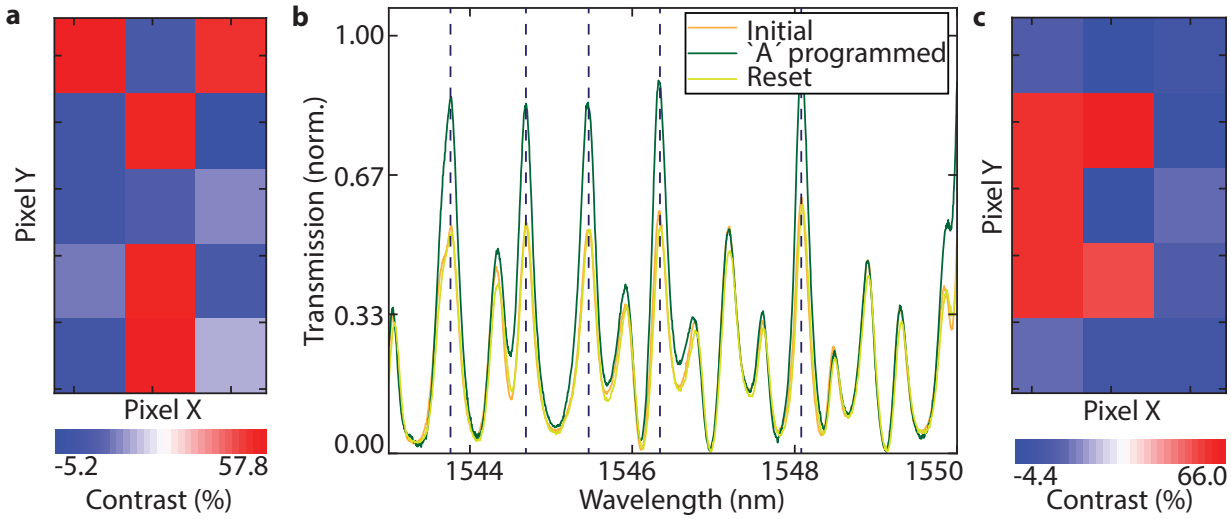

Figure 6.7.: Reconfiguration of weights. **a)** Letter 'A' encoded in fifteen PCM-cells (synapses) with a maximum switching contrast up to 60%. **b)** Optical transmission spectra of an optical neuron with the fifteen synapses (represented by the pixels) shown in a), each peak corresponds to one PCM. From the initial states, the five red pixels (complementary to the pixels comprising the letter) are inscribed in the peaks marked with the dashed lines. To demonstrate reconfigurability the PCM cells are then reset and the initial spectrum is obtained again. **c)** The same synapses are programmed to the pattern '3' (adapted from [146]).

6.1.7. Reconfigurable weights

An important advantage of using phase-change materials as the weighting elements in neural networks besides their non-volatility is the reversibility of the phase transition over many cycles. This means that the neurons can potentially be reprogrammed during the supervised training process so that the optimization can be performed directly in the photonic circuit. During unsupervised learning the neurons have the capability of adapting to changing patterns in the data as shown in the previous section.

Fig. 6.7 shows the reconfigurability of the weights in more detail on an example neuron with fifteen synapses (PCM-cells). In Fig. 6.7a) the neuron is trained to recognize the letter 'A', meaning that the five red pixels are amorphized. The three curves in Fig. 6.7b) correspond to the optical spectra in the initial state, after programming 'A' and after resetting the weights to the initial state. The wavelengths representing the individual pixels can clearly be distinguished and the initial spectrum (orange) matches the spectrum after the reset (light green) well. The five peaks in the spectrum representing the 'A' indicated by the dashed lines show a contrast in transmission of about 60%. After the reset, the same optical neuron was re-programmed to recognize the number '3' (see Fig. 6.7c), illustrating that the neuron successfully learned a new pattern demonstrating its reconfigurability.

6.1.8. Experimental Setup

Fig. 6.8 shows the experimental setup used to characterize the function of a single neuron as described in the previous sections. The top panel shows the sketch of an on-chip neuron with four inputs as discussed before. The optical setup is comprised of three main parts, which are reading the state of the PCM-weights (green), setting the weights and generating input patterns (orange) and the operation of the activation unit (blue). Reading of the weights is simply achieved in a transmission measurement. Light from a CW-laser (Santec, TSL 510) is equally split in four parts and sent to the corresponding input ports. Because the input couplers are polarization sensitive, the polarization for each path is controlled individually. Circulators prior to each port help separating the probe light (green) and the counterpropagating light for setting the weights and generating the input patterns (orange). The transmitted probe light indicating the state of the PCM-weights is read with the photodetectors $D0 - D3$ (New Focus, Model 2011).

Input patterns are generated by first combining the light from four individual CW-lasers operated at different wavelengths, which are multiplexed on a single fibre. The amplitude of each signal is controlled via the laser output power. Now a pulse of 200 ns width is cut from the signal with an electro-optic modulator (EOM) (Lucent Technologies, 2623CS) controlled by an electrical pulse generator (Agilent, HP 8131A) and amplified using and EDFA (Pritel). The optical pulse is then demultiplexed again and sent to the corresponding inputs of the optical neuron. The same path can also be used to set the state of the PCMs serving as the weights by simply using higher energy pulses. Note that the EOM is polarization sensitive and therefore an additional polarization controller for each optical path has to be added before the EOM.

The probe pulses for the activation unit (blue) are obtained similar to the input pulses using another CW-laser and an EOM and are monitored on detector $D4$. By sending higher power pulses on resonance, the PCM on the ring can be reset to its crystalline state after every cycle. In case of unsupervised learning, the outcoming probe pulse is picked up, amplified off-chip and sent back to the chip as feedback pulse to the ports indicated by $F1 - F4$.

6.2. Multilayer networks

In the previous section the photonic implementation of the basic building block of neural networks, the neuron, was described and experimentally demonstrated. The following sections present a pathway to scaling from a single neuron to a complete multilayer all-optical neural network ending with the demonstration of a neural network consisting of four neurons and sixty synapses capable of recognizing simple images.

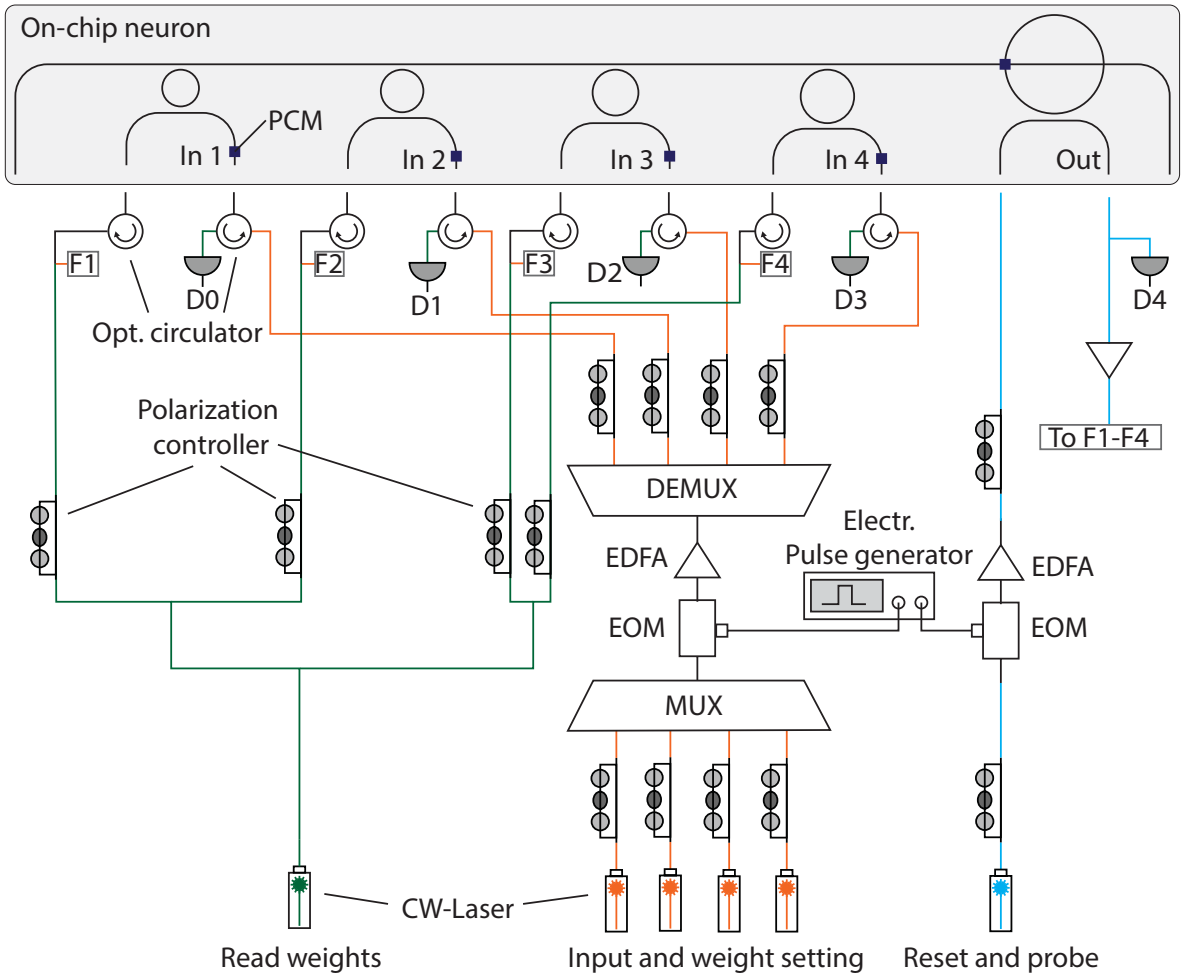

Figure 6.8.: Setup for the operation of a single neuron with four inputs. The setup consists of three main optical paths. The first is used to read the state of the weights (green) and consists of a basic transmission measurement using a CW-laser, polarization controllers and detectors ($D0 - D3$). The second path is used to input the patterns and program the weights (orange). The pulses to represent the input patterns or used for setting the weights are cut from four CW-lasers and an EOM. The pulses are amplified using an EDFA and demultiplexed for distribution to the different input ports. The third optical path is used to probe and reset the state of the activation unit (blue).

6.2.1. Scalability and photonic implementation

A critical point of all photonic circuits used for data processing is their scalability. Similar to scaling an electronic circuit from a single transistor to a full processor, a way to scale the single optical neuron to a large neural network needs to be developed.

As described in Sec. 2.2.1 in more detail, a neural network generally consists of an input layer, an output layer and several hidden layers in between (see Fig. 6.9). Each layer holds a certain number of neurons connected with the previous layer and emits its outputs to the next

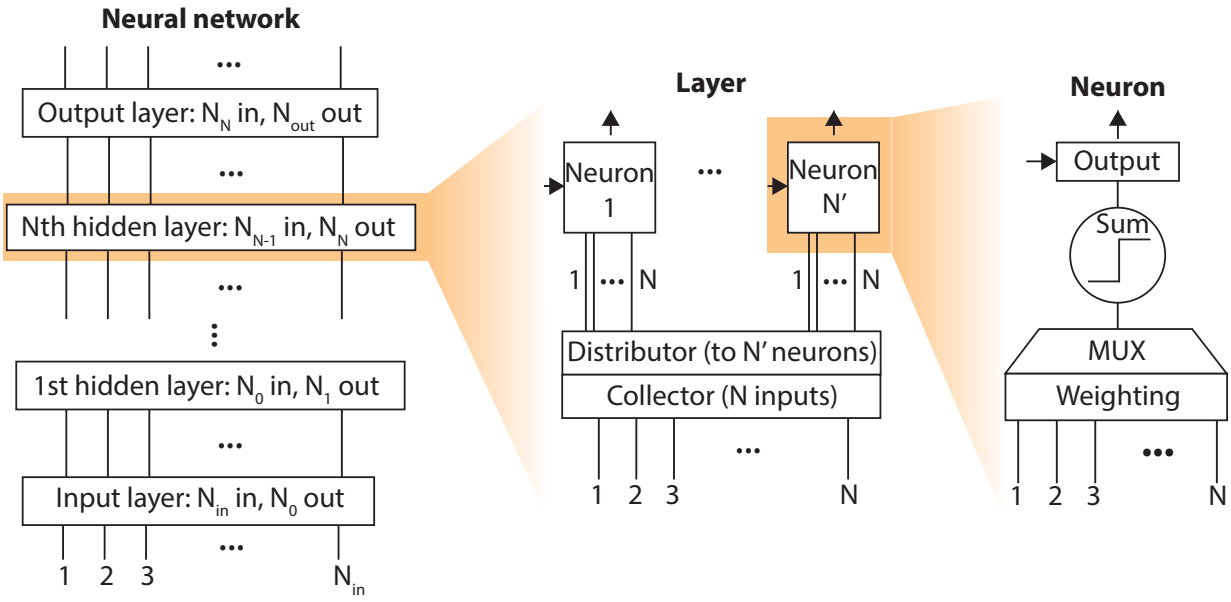

Figure 6.9.: General structure of the photonic multilayer network. As the classical neural network, the photonic network consists of an input layer an output layer and several hidden layers. On each layer the inputs from the previous layer are first collected (Collector) and then equally distributed (Distributor) to the individual neurons of the network. A single neuron consists of the weighting unit, a multiplexer to combine the inputs and the activation function that decides if an output pulse is generated.

layer. The layers fulfil the collection of the input data from the previous layer (Collector) and equally distribute the information to its neurons (Distributor), leading to a fully connected neural network. The neurons themselves process the data and are implemented as described in Sec. 6.1.

The photonic implementation of a single layer is illustrated in Fig. 6.10 and heavily relies on multiplexing and demultiplexing the signals from previous neurons with ring resonators. The neurons in the previous layer are designed to emit light on different wavelengths $(\lambda_1, \ .., \lambda_N)$, defined by the optical pathlength of the activation ring so that they can be multiplexed on a single waveguide without interference and minimized loss. The collector consists of several ring resonators in an add-drop configuration with a common drop waveguide. The configuration of the input and output gap of the resonators is designed to transmit all the light to the drop-port and the combined signal is then passed to the distributor that equally splits the power to the neurons on the current layer. This is achieved in the reverse configuration of the collector, again with ring resonators in add-drop configuration. As every neuron of the previous layer has to be connected to every neuron on the current layer, the power on every wavelength has to be split equally between the neurons. This is achieved by adjusting the coupling efficiency to the drop-port of the resonators (varying the gap between resonator and waveguide) following the equation $c = \frac{1}{N'+1-i}$ with N' being the total number of neurons in the layer and i the index of the neuron. Assuming a layer with three neurons, the first neuron would in this case receive $\frac{1}{3+1-1} = \frac{1}{3}$ of the

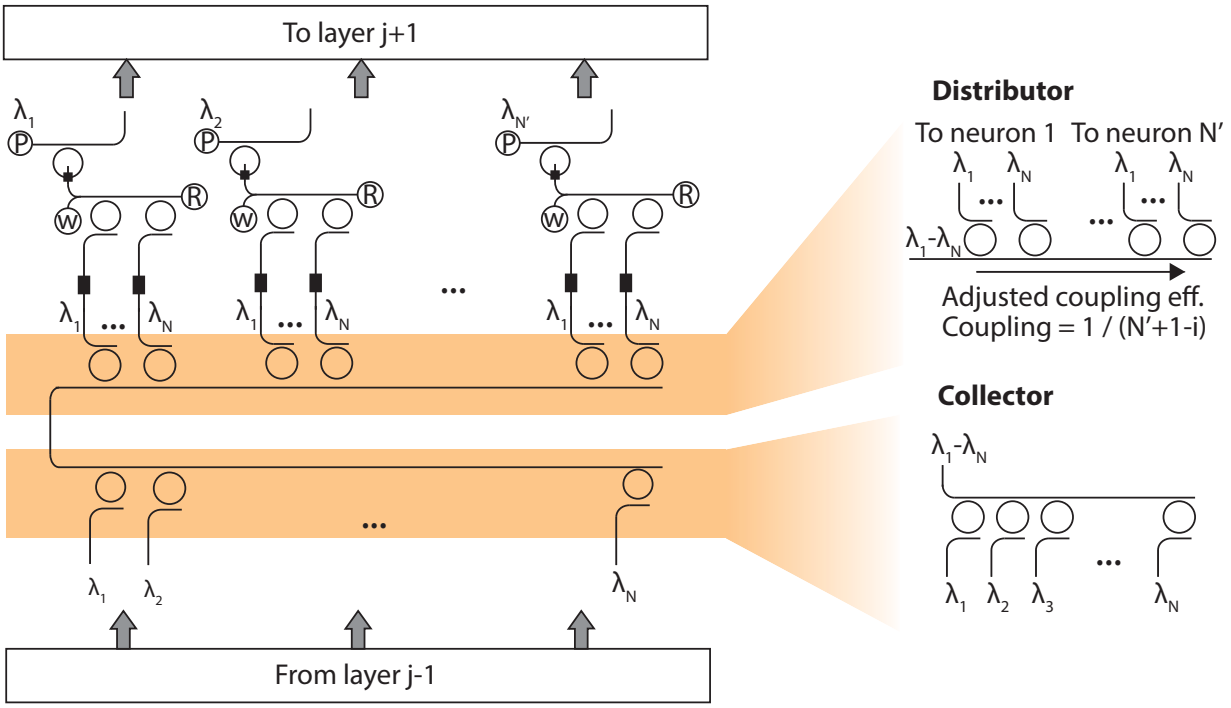

Figure 6.10.: Schematic of the photonic implementation of the multilayer neural network. The input signals from the previous layer on different wavelengths are combined on a single waveguide using a multiplexer based on ring resonators (Collector). Now the signals are distributed to the individual neurons of the layer again using ring resonators in the add-drop configuration. To split the signal equally to all the neurons, the coupling efficiency to the drop-port of the resonators is adjusted as shown on the right (Distributor). The ports used for probing the activation function, setting the weights and resetting the activation PCM are labelled with 'P', 'w' and 'R', respectively.

incoming light, the second neuron $\frac{1}{2}$ of the remaining light and the last neuron all the light that is left ($\frac{1}{1}$). The correct coupling efficiencies can be obtained by carefully designing the input- and drop-port gaps as described in more detail in Sec. 6.2.3. After passing through the neurons, the output pulses are again generated on different wavelengths, so that several layers can be stacked on top of each other.

This layered architecture has an important advantage against networks in which the same light pulses are travelling through the full network. The probe power is sent to each neuron individually (ports 'P' in Fig. 6.10) serving as the input for the next layer. This design on the one hand clearly separates the individual layers preventing mixing of signals during the propagation of the information through the whole network and on the other hand solves the problem of loss when building a many layer network. As each layer has its own inputs, the power levels can be preserved throughout the entire network.

In order to operate the proposed optical neural network, individual optical access to the PCM-cells is necessary so that each weight can be trained without influencing the other weights. In

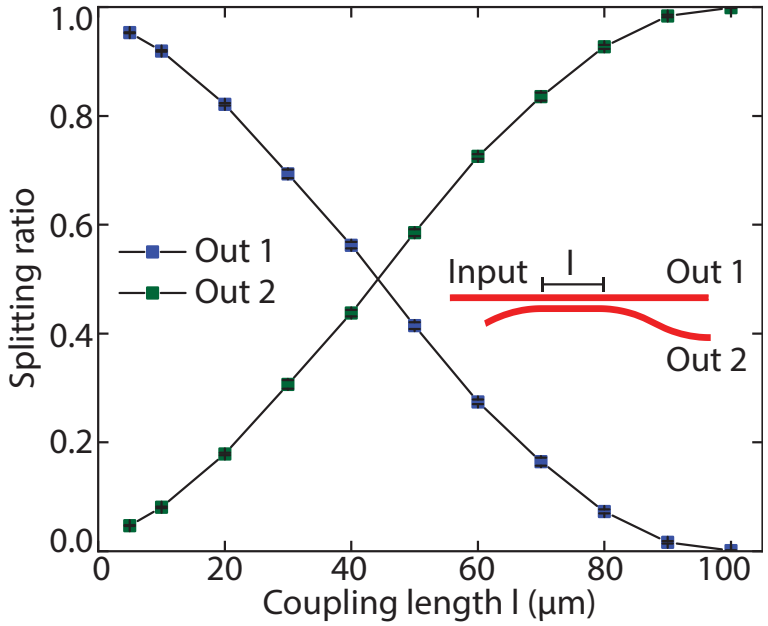

Figure 6.11.: Directional coupler. Splitting ratio between the two output ports of the directional coupler depicted in the inset as a function of the coupling length l measured at a wavelength of 1550 nm.

the presented network this is achieved through the input ports marked 'w' in Fig. 6.10. Light sent through these ports propagates backwards compared to the flow of the input data. Each neuron has one 'weight-input' port so that by selecting the corresponding port and the wavelength corresponding to a certain weight each cell can individually be addressed. Because the waveguide employed for setting the weights is the same as the waveguide leading the weighted sum of a neuron to its activation unit, a splitter is implemented to guide 80% of the light to the activation unit and only 20% of the power from the weight-input port to the PCM cells. This is achieved using directional couplers detailed in the next section. The splitting ratio is implemented to allow for lower energy operation of the neuron after setting the weights and ensuring that enough power reaches the activation unit for triggering an output pulse. In future designs, the optical switching of the weights could be replaced by electronic switching of the PCM making the weight input port and the directional coupler redundant. To reset the activation unit after a potential switching event, the ports labelled 'R' are employed.

6.2.2. Directional couplers

Directional couplers are optical components used to split light between two waveguides. This is achieved by bringing the waveguides in close vicinity exploiting evanescent coupling of the modes. In the proposed photonic implementation of a neural network they are employed to allow for accessing the individual PCM-cells of the neuron.

The directional coupler design and its performance is shown in Fig. 6.11. The splitting ratio between the two output ports ($P_i/\left(P_1+P_2\right)$ with $i \in 1,2$) is plotted as a function of the coupling length together with the standard deviation over five devices per datapoint as the error bars. The gap between the two waveguides is 500 nm allowing for a fine control of the splitting ratio, as the coupling to the second port gets stronger for smaller gaps. The experimental data shows the expected $\sin^2$ behaviour of the splitting as function of the coupling length l [93]. To achieve the 80 : 20 splitting ratio as used in the neural network explained in the previous section, a coupling length of 24 µm is used.

6.2.3. Multiplexer and tuning of coupling efficiency

The correct distribution of the input signals to the individual neurons and between the layers is a crucial step in the implementation of a neural network. The design chosen in this work relies on wavelength multiplexing with ring resonators and has its advantages especially in the fact that no waveguide crossings are needed to interconnect different neurons and crosstalk between signals is therefore avoided. The two main issues that have to be solved when employing ring resonators for multiplexing in the proposed architecture are, firstly, achieving a fixed and reproducible spacing between the resonance wavelengths of the individual rings, and secondly, controlling the amount of light coupled to the drop-port of the resonator. The first was already addressed in Sec. 6.1.3 describing wavelength multiplexing for a single neuron. To align the wavelengths of the neural network to the off-chip setup a resonance spacing of 100 GHz (approx. 800 pm) according to the International Communication Union (ITU) grid is chosen.

The second issue in designing the ring resonators for the optical neuron is the coupling efficiency to the drop-port. In case of the collector ideally all light from the previous layer of the network is transferred to the common drop waveguide so that a maximum coupling of light to the drop-port is desired. From the theory explained in more detail in Sec. 6.1.3, the relation $A = |t_1|/|t_2|$ with the propagation loss factor in the ring A and the transmission factors t_1 and t_2 is found. t_1 and t_2 describe the transmission from the input to the through-port and from the add- to the drop-port, respectively. For a lossless resonator with $A = 1$ this would lead to a symmetric configuration with equal input and output gap. Since for a real device scattering loss must be considered, ring resonators with different configurations are fabricated and the coupling efficiencies experimentally validated. Fig. 6.12a) shows the transmission to the drop-port with a fixed input gap ($a = 400$ nm) for a varying output gap. From (6.3) the best configuration for $A < 1$ is expected for $t_1 > t_2$, meaning larger output gaps compared to the input. The plot shows the best transmission for an almost symmetric configuration, indicating that the propagation loss in the fabricated devices is low so that a symmetric configuration is further used in the neural network implementation when a maximum coupling of light to the drop-port is necessary. To further optimize the drop-port transmission, the maximum transmission for ring resonators in a symmetric configuration

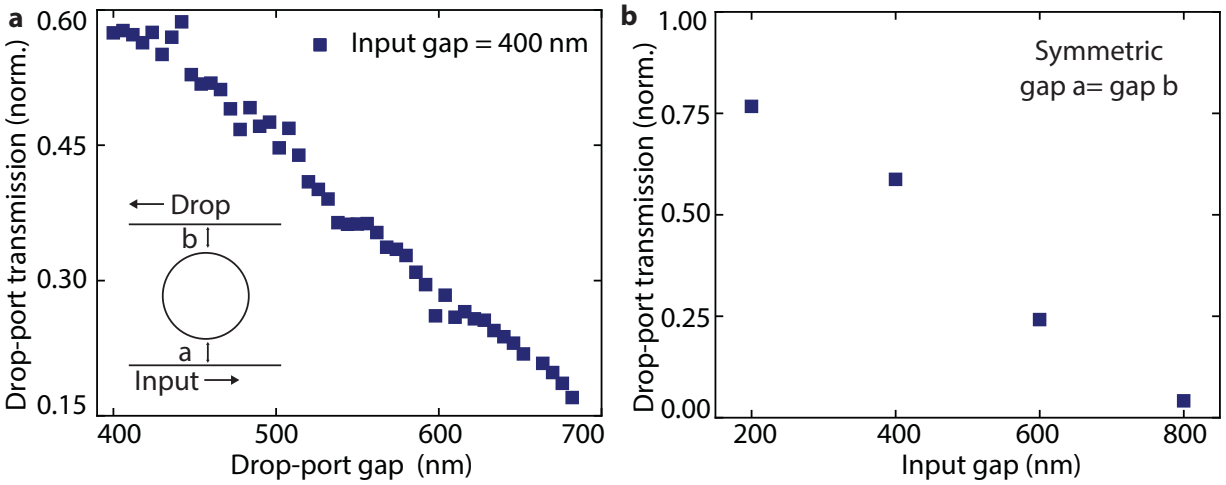

Figure 6.12.: Ring resonators in add-drop configuration. **a)** Drop-port transmission on resonance for a fixed input gap $a = 400\,\text{nm}$ as a function of the drop-port gap b. The maximum transmission to the drop-port is achieved for the almost symmetric configuration. **b)** Maximum drop-port transmission for different ring resonators with symmetric coupling gaps. The transmission decreases with increasing coupling gap.

as a function of the gap is shown in Fig. 6.12b). For coupling gaps of $a = 200\,\text{nm}$, a drop-port transmission of 75% is obtained and it can be concluded that smaller gaps lead to higher coupling efficiencies so that a gap of $a = 100\,\text{nm}$ is chosen for the ring resonators in the collector of the neural network.

In case of the distributor, a variable coupling to the drop-port is necessary which is achieved by carefully configuring input- and drop-port gap so that equal amount of the light from the collector is sent to all neurons on the layer. To find the correct configurations, devices with varying input and output gap are fabricated and a map with the corresponding coupling efficiencies and insertion loss is obtained (see Fig. 6.13). The ratio of light transmitted to the drop-port in relation to the through-port on resonance is indicated by the black contour lines and is defined as $P_{\text{drop}}/\left(P_{\text{drop}} + P_{\text{through}}\right)$, with P_{drop} being the power measured at the drop-port and P_{through} being the through-port power. The colours depict the loss of the whole add-drop resonator (the sum of the through- and drop-port power divided by the through-port power off-resonance). Similar to what was derived in the previous part, a line with the maximum coupling can be drawn with slightly larger drop-port gaps compared to the input gap, which is expected for a loss factor of $A < 1$. Importantly it should be noted that for the same splitting ratio configurations with significantly differing loss can be found. Whereas a configuration of 0.3 µm and 0.6 µm reveals a loss of more than 90% and a splitting ratio of 0.9, the same ratio is obtained in the configuration 0.1 µm and 0.1 µm with a loss below 20%. This is important as a combination of splitting ratio and remaining light (determined by the loss) must be found in order to guarantee that an equal amount of the incoming light from the collector is transferred to the individual neurons.

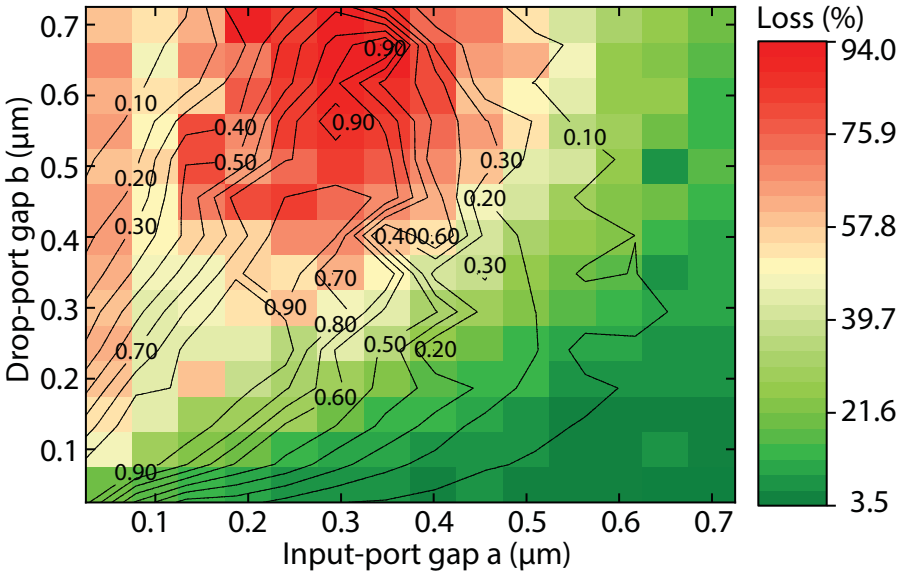

Figure 6.13.: Coupling efficiencies and loss of add-drop resonators in different configurations. Each square corresponds to one device with the specific input and output gaps. The colourmap depicts the insertion loss of the device and the contour lines the ratio of the power transmitted to the drop- and through-port ($P_{\text{drop}}/\left(P_{\text{drop}}+P_{\text{through}}\right)$). Because low loss is generally preferable, the green region from the diagonal to the lower right corner contains the most useful configurations.

Tab. 6.1 shows an exemplary set of input and output gaps that is chosen for a neural network layer with four neurons. The desired coupling is calculated according to the formula $c_i = \frac{1}{N+1-i}$ assuming a lossless system. The resulting transmission in the last column gives the fraction of a certain input pulse that is transmitted to each neuron taking into account the experimental values (splitting ratio and loss) depicted in Fig. 6.13. It is calculated as

$$t_i = c_i^m \cdot \prod_{j=1}^{i} \left(1-l_j\right) \cdot \left(1-c_{j-1}^m\right), \text{ with } \; c_0^m = 0. \tag{6.5}$$

From Tab. 6.1 it can be seen that each neuron therefore receives $15\%-16\%$ of the incoming light, well satisfying the condition of equal power distribution. Because of the loss introduced by the multiplexer, only $4 \cdot 15\% = 60\%$ of the light from the previous layer is preserved in this case.

6.2.4. A multi-neuron network

Based on the scaling path, device characteristics and parameters obtained from the individual elements of a photonic neural network determined in the previous sections, a neural network consisting of four neurons with 15 synapses each is constructed and experimentally tested on a pattern recognition task. Fig. 6.14a) shows an optical micrograph of the fabricated all-optical

Neuron position i	Desired coupling c_i	Gap a (nm)	Gap b (nm)	Measured coupling c_i^m	Loss l_i	Resulting transmission to neuron t_i
1	0.25	400	100	0.165	0.093	0.15
2	0.33	350	100	0.228	0.109	0.15
3	0.50	350	150	0.373	0.152	0.16
4	1.00	200	150	0.823	0.290	0.16

Table 6.1.: Exemplary resonator configuration for a layer of four neurons. Derived parameters from the data in Fig. 6.13 to achieve an equal splitting of light to the four neurons. As the insertion loss of the resonators must be considered, the measured coupling loss differs from the theoretically obtained optimal desired coupling coefficients. The results show that each neuron receives a fraction of approximately 15% of the incoming signal.

neural network implementing one layer of the proposed scalable architecture. The device consists of sixty synapses and more than 120 ring resonators for multiplexing the signals. With the input port at the left most grating coupler, the signal is distributed to the individual neurons through the upper row of ring resonators.

Fig. 6.14b) shows a magnified micrograph of a single neuron within the layer. The triangular structures on the bottom row are the grating couplers used as input and output ports for the network. Through the left port the input patterns are sent to the optical neurons. The ports labelled W_i are used to program the weights of the individual neurons and guide the pulses through the directional couplers in the backward direction to the specific synapse based on the wavelength of the switching pulse. To sustain the functionality of the activation unit, only 20% of the optical pulse is transferred to the cross-port of the directional coupler as described in more detail in previous sections. The pulses to probe the state of the activation units are sent to the ports P_i and the output is read from ports O_i. Fig. 6.14c) and d) show close-ups of the PCM-synapses (GST) between the distributor and the combining multiplexer rings and the activation unit, respectively. The length of the phase-change material patches employed as weighting elements is 3 µm. The images show a precise alignment of the PCM in respect to the photonic waveguides.

The fabricated all-optical neural network can now be tested with a basic image recognition task similar to the single neurons before. The pixel patterns consisting of fifteen pixels each (see Fig. 6.15a) are again encoded in optical pulses on different wavelengths (1..15). As illustrated in Fig. 6.15b), each pixel corresponds to one resonator pair in each neuron. The pattern 'A' therefore consists of pulses on the wavelengths $1, 3, 5, 8$ and 14 (note that for simplifying the optical setup, the complementary pattern – white pixels – is used). If it is presented to the neuron, the power on each of the wavelengths is equally split to the four neurons via the distributor. By training each neuron to one of the patterns, for example amorphizing the weights $1, 3, 5, 8$ and 14 for neuron one to recognize 'A', all neurons will only give an output pulse if the correct input is shown. Fig. 6.15c) shows the change in the output spike intensity, defined as the optical contrast between the probe pulse amplitude before and after showing a certain pattern, for all four neurons in dependence

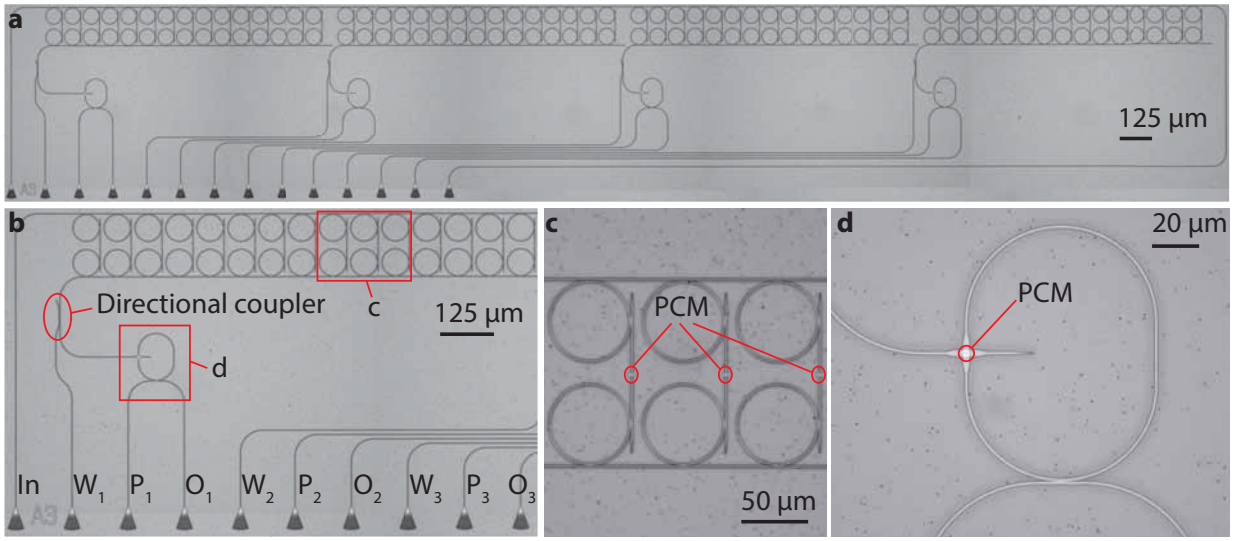

Figure 6.14.: Optical micrograph of a photonic neural network with four neurons. a) Overview of the complete network. **b)** Magnified image of a single neuron within the network. The signal coming from the input port is first (partly) distributed to the neuron using the upper row of resonators (distributor) and then weighted via the PCM-synapses. The second row of resonators combines the weighted input signal via WDM and the bus waveguide leads to the bigger activation unit ring resonator. **c)** Close-up of the resonator array comprising the distributor and the PCM-synapses. Each PCM patch has a length of 3 µm. **d)** Micrograph of the activation unit with the PCM deposited on the waveguide crossing.

on the presented pattern. It can clearly be seen that the neural network correctly distinguishes between the individual patterns and every neuron recognizes only one of the patterns.

Again, it is important to note that the complete calculation is executed in a single timestep. The weighting of the input pattern with all-four neurons is performed simultaneously showing the huge potential of photonic neural networks. By combining several of the layers according to the system architecture presented in this chapter, larger neural networks can be obtained. As the proposed neural network design does not suffer from crosstalk between individual layers because the input signal only travels until the next activation unit and the output spikes are generated from additional independent light-sources in every layer, the design presents a scalable photonic architecture for the hardware implementation of neural networks.

6.2.5. Experimental setup

The experimental setup used to operate the optical neural network is depicted in Fig. 6.16. The input patterns (green path) are, similar to the operation of the single neuron, generated from several CW-lasers. The signals on different wavelengths are multiplexed on a single fibre and modulated with an EOM and a pulse generator. After subsequent amplification of the pulse using an EDFA, the input pattern is sent to the input port of the neuron. Additionally, one of the lasers comprising the pattern is also used to read out the states of the PCM cells. Therefore,

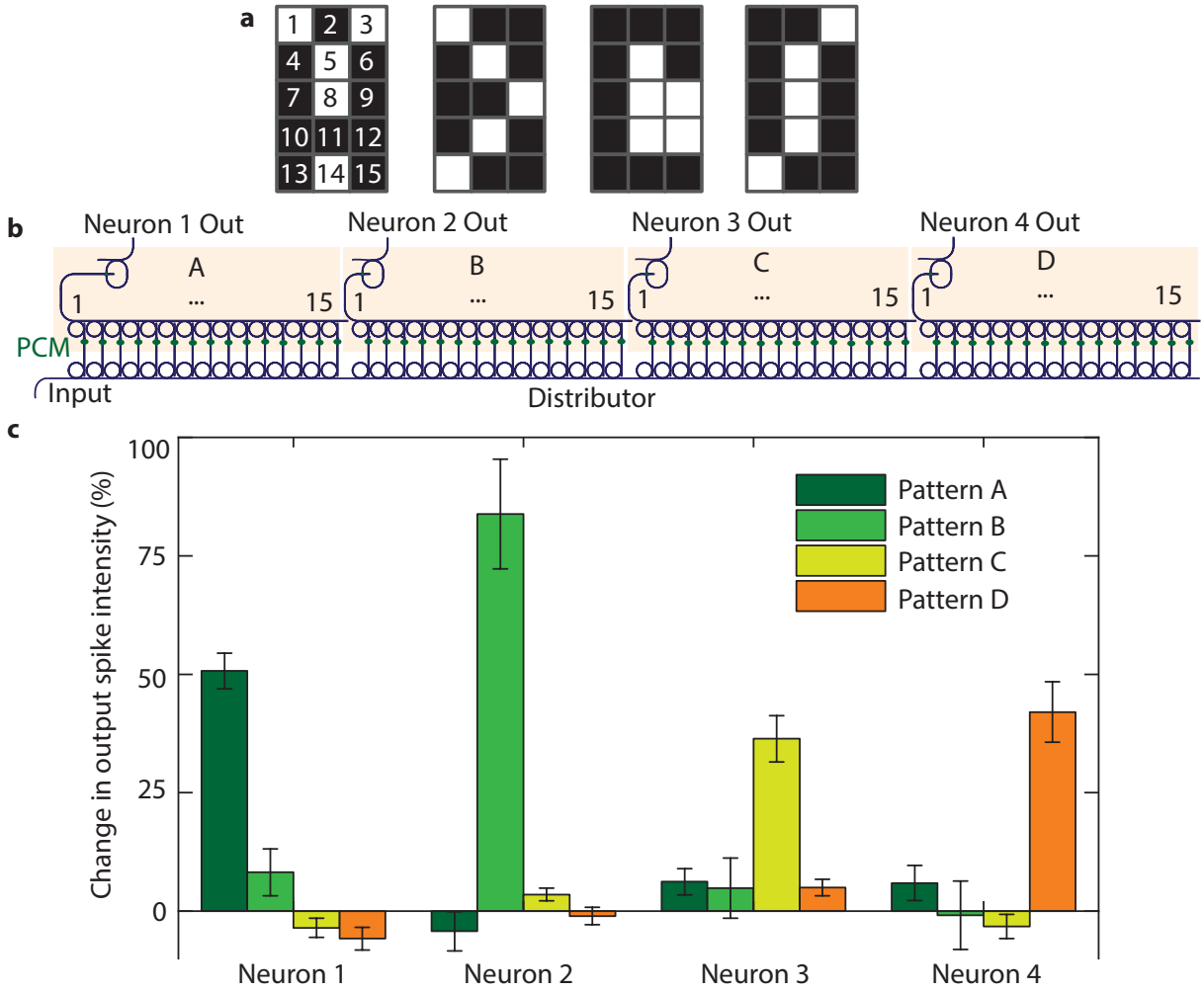

Figure 6.15.: Pattern recognition with the all-optical neural network. a) The four pixel patterns ('A', 'B', 'C', 'D') recognized by the neural network. Each pattern consists of 15 pixels and is encoded in optical pulses on different wavelengths (1..15). **b)** Schematic representation of the optical neural network shown in Fig. 6.14a). Each pixel of the pattern in a) corresponds to a pair of resonators (1..15) in each neuron. **c)** Change in the output spike intensity for all four neurons responding to the different patterns. Each neuron successfully recognizes only one of the letters.

part of the light is split and guided around the pulse generation circuit and by selecting the correct wavelengths corresponding to the synapse to be read and selecting the detector at port $W1 - W4$ depending on the neuron to be read all PCM-cells can be monitored.

To program the weights (orange path) pulses of 200 ns length are generated from a CW-laser. These are split into four equal parts using a 1×4 splitter and then sent to the different weight inputs ($W1 - W4$) of the neural network. To adjust a specific synapse, first the neuron is selected turning on or off the VOAs and then the specific PCM-cell can be chosen adjusting the wavelength of the optical pulse according to the corresponding ring resonator.

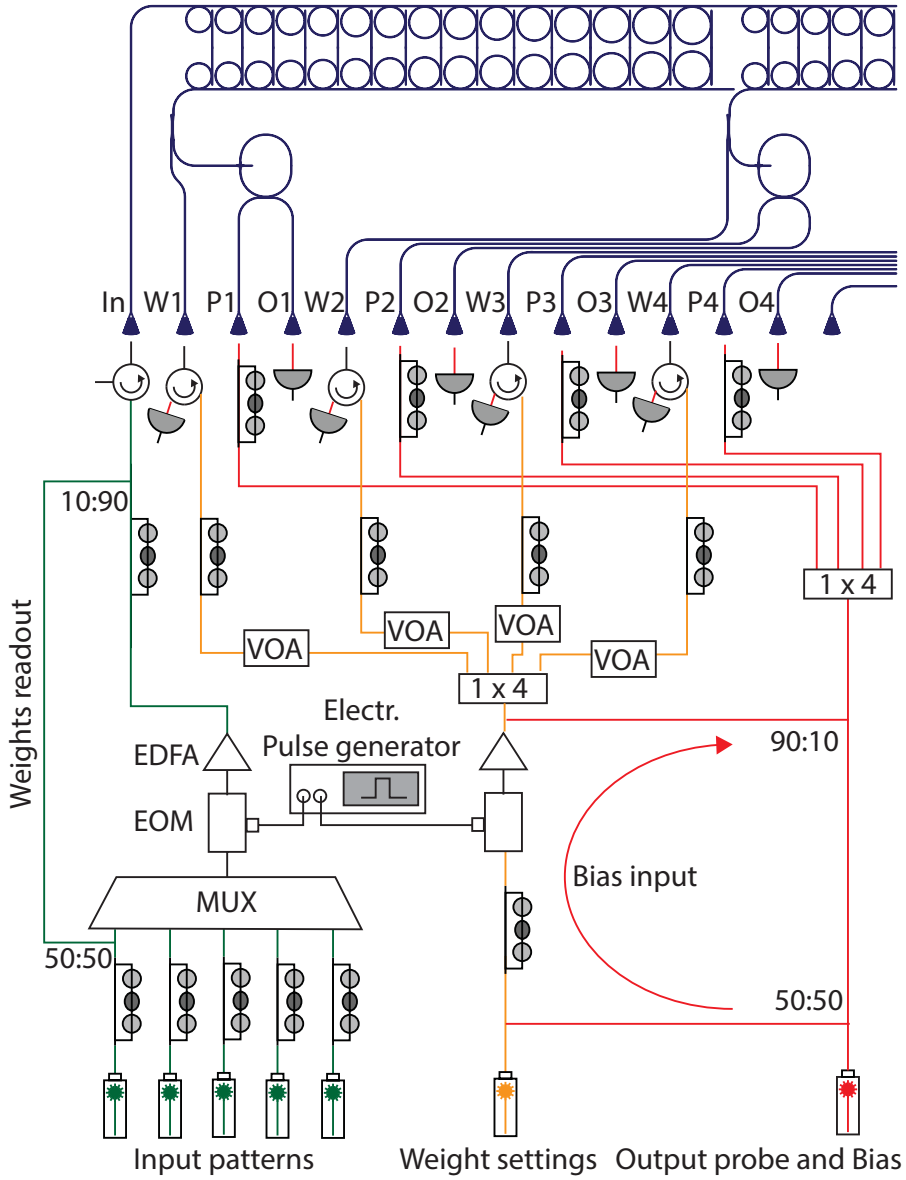

Figure 6.16.: Experimental setup for operating the optical neural network. Similar to the operation of the single neuron, the input patterns are encoded with CW-lasers on different wavelengths (green) that are combined with a MUX. An electrical pulse generator modulates an EOM to cut out a 200 ns pulse pattern. One of the lasers is additionally used to read the weights of the neurons. To program the weights (orange) an additional laser with EOM, pulse generator and EDFA generate the switching pulses. Via variable optical attenuators (VOAs) the light is selectively guided to the weight input ports $W1 - W4$. The output probe and bias path (red) consists of a CW-laser that continuously measures the state of the activation units with the four photodetectors at the output ports $O1 - O4$. To bias the activation unit when showing a pattern to the neuron, a bias pulse can be generated using the components of the weight setting path.

The state of the activation unit is continuously monitored using a CW-laser split in four equal parts and sent to the probe ports $P1 - P4$. The output is detected with four photodetectors at

the output ports $O1 - O4$. Making use of the same pulse generation path as used for the weight setting, also a bias and reset pulse can be generated and sent to the activation units. A bias pulse can be used to reduce the energy needed at the activation unit to switch the PCM on the crossing. The reset pulse is necessary to crystallize the neuronal PCM after the activation unit has switched.

6.3. Power and energy considerations

The energy consumption of the demonstrated neural network can best be evaluated together with an explanation of the temporal sequence of operation. Because the phase-change material enables non-volatile storage of the weights, no continuous power supply is necessary to maintain the state of the neural network and the only power consumption is generated from sending the input pulses and switching the PCM-cells.

The neural network is operated in the following way: in a first step the network is trained, which includes switching all its PCM-cells working as synapses to the desired state. The pulse energy needed to induce a full amorphization step in a PCM-cell of 3 µm length is measured to be 4.7 nJ inside the weight input waveguide for a 200 ns pulse. Because of the subsequent directional coupler that only transmits 20% of the optical pulse to the neuron and a loss of approximately 30% for the ring resonator in a symmetric add-drop configuration, the energy arriving at the PCM-synapse is calculated to be $4.7\,\mathrm{nJ} \cdot 0.2 \cdot 0.7 \approx 660\,\mathrm{pJ}$. This the maximum energy needed to set a PCM-synapse to the maximal state of transmission and has to be partly applied once for every synapse depending on the desired state of the synapse. After setting all weights of the network, the PCM-cells keep their state without further energy consumption and can therefore be considered a minor factor in the overall power budget.

In the feed forward step of the neural network first the input pattern is sent to the neuron, then a probe pulse to generate the output signal and in a last step the activation units that have been switched must be recrystallized to the initial state. The pulse energy of one pulse of the pulse pattern inside the input waveguide is measured to be approximately 700 pJ. This energy is then equally distributed to the four neurons on the layer using the on-chip distributor. According to the experimentally obtained parameters (see Tab. 6.1) the power at the PCM-synapse after passing the first ring resonator (distributor) is $700\,\mathrm{pJ} \cdot 0.907 \cdot 0.165 = 105\,\mathrm{pJ}$, where 0.165 is the splitting ratio to the first neuron and 0.907 the loss factor. The second ring resonator in a symmetric configuration combining the weighted inputs onto the waveguide leading to the activation unit multiplies another loss factor of 0.7 that includes the loss of the second resonator in a symmetric configuration. As the directional coupler before the activation unit only transmits 80% of the light, the power arriving at the neuronal PCM on the crossing is approximately $0.7 \cdot 0.8 \cdot 105\,\mathrm{pJ} \approx 60\,\mathrm{pJ}$. Assuming one of the patterns used in the experiments that consists of five optical pulses and that the neuron was trained to recognize this pattern (meaning the corresponding PCM-synapses are

in the amorphous state) the total power arriving at the activation unit is approximately 300 pJ. Two points can be concluded from this estimation. The first is that the 105 pJ of a single pulse of the input pattern is well below the switching threshold of 660 pJ derived for the switching pulse at the synapse, so that the input pattern can not alter the state of the synapses. The second is that the energy of 300 pJ obtained from the full input pattern at the activation unit is not enough to fully amorphise the neuronal PCM, so that a bias input is needed to increase the overall power level at the activation unit (see Sec. 6.2.5). However, it should be noted that by using more inputs per neuron the power at the activation unit would increase making the bias input obsolete.

Returning the activation unit to its initial state is achieved by sending approximately five pulses with decreasing energy between 300 pJ and 100 pJ to the neuronal PCM leading to an overall energy for the reset of around 1.0 nJ. Considering a whole feed forward step for one neuron consisting of sending the input pattern and re-initialization, the energy adds up to $5 \cdot 105\,\mathrm{pJ} + 1000\,\mathrm{pJ} = 1.5\,\mathrm{nJ}$. The full neural network presented consisting of four neurons consumes about $5 \cdot 700\,\mathrm{pJ} + 4 \cdot 1000\,\mathrm{pJ} = 7.5\,\mathrm{nJ}$ per cycle. The difference between the two values for a single neuron and the full network is the result of the loss in the distributor and depends on the specific structure of the neural network. Because the exact pulse energies can vary a lot depending on the fabrication (overlap of the resonances) and drift in the off-chip setup, an estimation of the variation of the pulse energies can be found from the error bars obtained in the operation of the network in Fig. 6.15. These error bars show a variation in change of the output spike intensity of the neurons of up to 12%. Because these variations are mainly introduced by a variation of the pulse energies arriving at the activation unit that result in different amounts of amorphization, it can be concluded that also the pulse energies vary in the range of $10\% - 15\%$.

However, it has to be noted that the pulse energies can be considerably reduced by using shorter optical pulses. As demonstrated in Chap. 5 and also in [116] the energies for switching the phase-change material can be reduced to around 20 pJ by using picosecond pulses, gaining a factor of 30. Optimizing the photonic circuit (for example the multiplexer losses) and introducing electronic switching of weights making the directional coupler obsolete can further improve the energy efficiency.

6.4. Simulations

Having demonstrated the all-optical neural network on a network with up to four neurons and sixty synapses limited by constraints of the optical setup, in this section the potential of larger neurons with higher numbers of inputs and larger networks is explored. Therefore, simulations based on the experimentally obtained parameters of the optical neuron are carried out.

All simulations presented in this section are based on the activation function obtained in Sec. 6.1.4 and the linear regression depicted in Fig. 6.17. Therefore, the function is divided in a constant part below 60% of the maximum pulse energy and a linearly increasing function

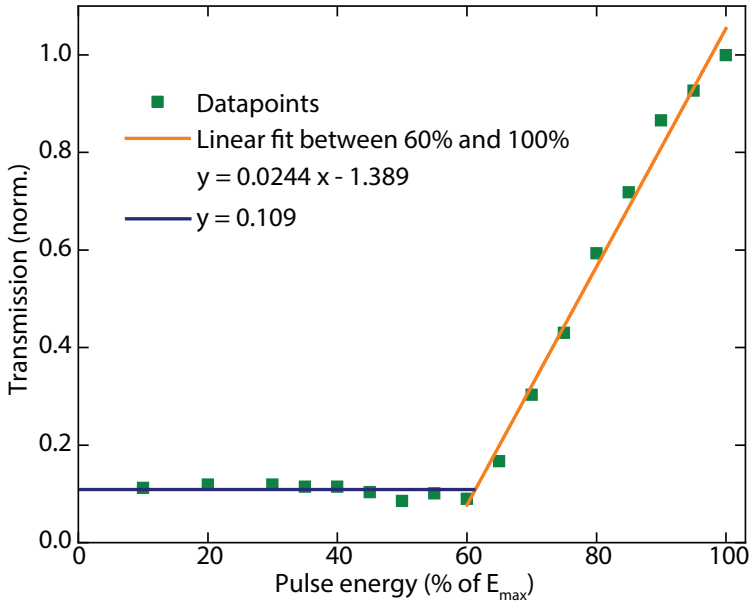

Figure 6.17.: Experimentally obtained activation function with linear regression. To perform simulations of the optical neural network, two linear fits are used to represent the activation function. The first covers the range from $0\% - 60\%$ and the second the linear increase above 60% of the maximum pulse energy.

above 60% resembling a ReLU function. The behaviour of a single neuron is emulated in a python script and expanded to more synapses. Individual neurons are combined to achieve larger neural networks.

6.4.1. Unsupervised learning for pattern recognition

To further investigate the performance of unsupervised learning, a single neuron in the following simulations consists of 4096 synapses connected to one activation unit to recognize a 64×64 pixel image. Similar to the experiments, each pixel of the image is mapped to one input of the neuron. Weight depression is achieved in ten equal steps of crystallization comparable to the unsupervised learning experiment shown in Sec. 6.1.6 and potentiation occurs in a single step also following the experiment.

To validate the simulation model, in a first step a task similar to the experimental unsupervised learning demonstration with the four-pixel image is carried out in which a single neuron is supposed to autonomously adapt to the pattern shown in Fig. 6.18a) on the left. The pattern is a black and white pixel image of size 64×64 ('WWU'). In a first training run all synapses are initially in the amorphous state (depicted by white pixels in epoch 0) and either the 'WWU'-pattern or a random noise pattern (newly generated in every epoch) is shown to the neuron in a random sequence. The ratio for the appearance of either of the patterns is $50 : 50$ and the noise

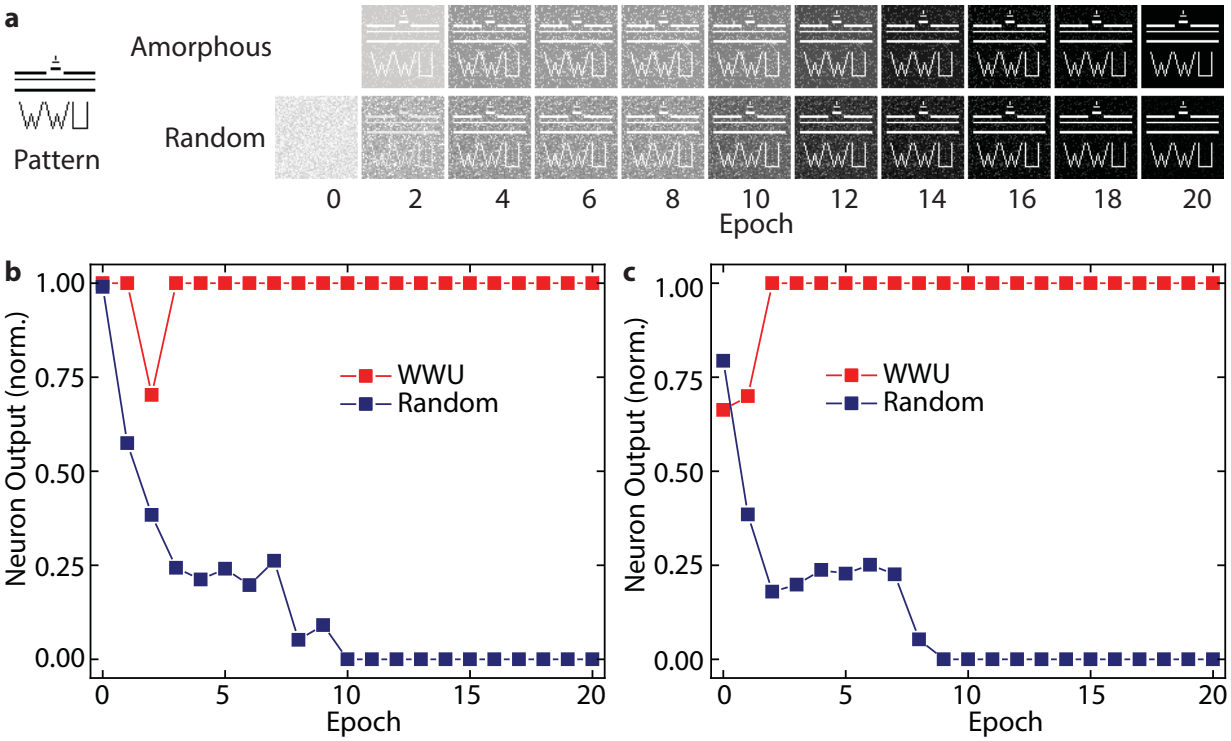

Figure 6.18.: Unsupervised learning of a single pattern with 4096 inputs. a) Evolution of the synapses of the neuron over 20 epochs starting from fully amorphous synapses and randomly initialized weights. The neuron is shown the pattern on the left or a noise pattern. The synapses fully adapt to the repeated pattern over 20 epochs. **b)** and **c)** show the output of the neuron for the 'WWU' pattern and a random pattern over time for amorphous starting weights and random starting weights.

pattern consists of 4096 randomly generated values $(0-255)$. In each epoch the neuron follows the learning rule via the feedback loop that was also experimentally implemented and from the evolution of the weights it can be seen that the neuron fully adapts to the repeatedly appearing 'WWU'-pattern after 20 epochs.

The neuron output for amorphous starting weights after each epoch is shown in Fig. 6.18b). In the beginning both patterns ('WWU' and noise) generate an output pulse because all synapses are transmissive. If an output pulse is generated, all inputs contributing to the output spike are amorphized and all the others are crystallized by one step. As the 'WWU' pattern is repeatedly shown, some of the input synapses contribute more often to the output spike than the others and are therefore more likely to stay in the highly transmissive amorphous state. The strength of all other synapses is gradually decreased over time leading to the full adaption of the neuron to the 'WWU' pattern after 20 epochs.

Fig. 6.18c) shows the output of the neuron when initialized with random weights before the training process. Because randomly some of the synapses that belong to the 'WWU' pattern are initialized in the crystalline or partly crystalline state, the output signal for both patterns is less

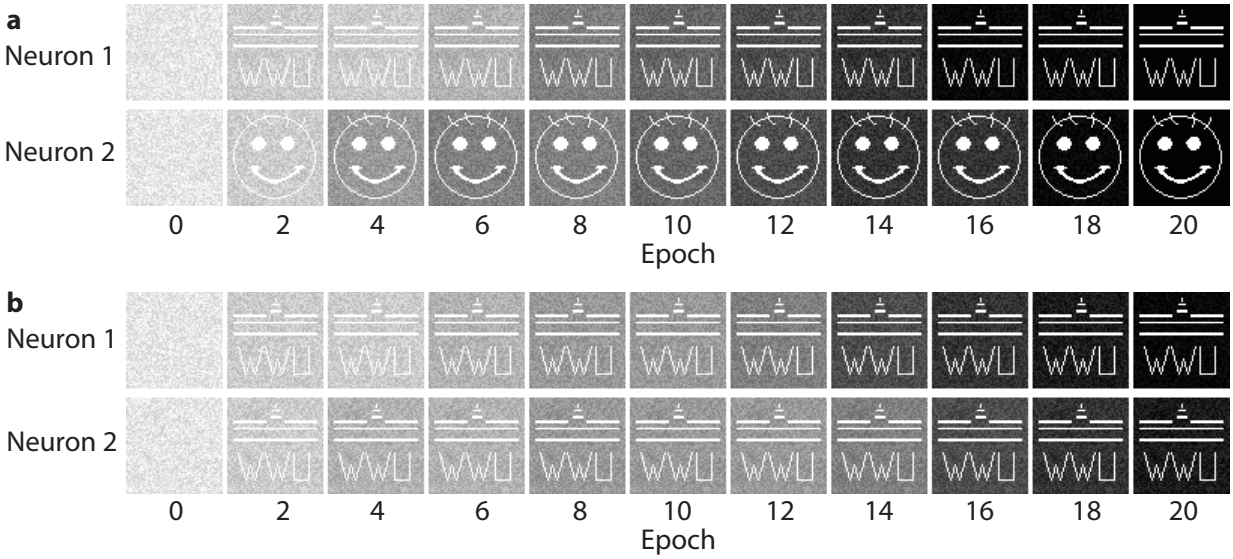

Figure 6.19.: Unsupervised learning with two neurons. **a)** Both neurons successfully recognize one of the patterns 'WWU' and 'smiley' after the training process. **b)** Neuron one and two both adapt to the same pattern illustrating a known problem of unsupervised learning called co-specialization.

than 1.0 in the beginning and the output for the random pattern in this case is even higher than for the 'WWU' pattern. Still the neuron is able to extract the repeated pattern and its synapses adapt to the 'WWU' pattern over time.

In the next step a two-neuron system with 4096 synapses each is simulated to recognize two different patterns, a 'smiley' pattern and the 'WWU' pattern. In Fig. 6.19a) the development of the synapses of both neurons is shown, illustrating that both successfully adapt to different patterns. During the training process in 20% of the cases the 'smiley' or 'WWU' pattern is shown and in the remaining 80% a random noise pattern. The noise is necessary to depress wrongly potentiated synapses over time. In Fig. 6.19b), with the same simulation setup but different random initialization, both neurons adapt to the same pattern. This is a common problem for unsupervised learning techniques and is called co-specialization. The result of the learning process depends on the starting point of the synapses and the sequence in which the patterns are shown. To prevent co-specialization, inhibitory connections between the neurons can be introduced for example implementing a 'winner-takes-all' rule meaning that in each epoch only the neuron with the highest output is allowed to update its weights.

In a last step the inhibitory connections between the neurons are also implemented in simulation to build a neural network based on the photonic neurons that is capable of recognizing ten digits (0..9) after an unsupervised learning process. Again, each neuron consists of 4096 synapses that are initialized in the fully amorphous state. Fig. 6.20a) shows the evolution of the weights over 200 epochs illustrating that each digit was learned only once because of the 'winner-takes-all'

rule. If several neurons have the same output (which is the highest) only the first of the neurons is allowed to update its weights. The neurons successfully extract the individual patterns from the data stream and the network is able to recognize all digits after the training process without intervention of an external supervisor. It can be seen in Fig. 6.20a) as well, that after ten epochs the neurons $6-10$ are still in their initial amorphous state. This is due to the inhibitory connections, because to none of the patterns shown in the first 10 epochs they gave the highest output signal so that their weights were not yet updated. The higher specialized the other neurons already are the more likely it will be for the neurons with still amorphous synapses to have the highest output and start specializing to a certain pattern.

An interesting side effect of the photonic neurons only allowing positive weights is an inherent capability of detecting correlations between different patterns. Fig. 6.20b) and c) depict the evolution of the output of neuron one that adapted to digit '5' and neuron three that adapted to '1' over 200 epochs. For neuron one the response to the other digits is not fully supressed which is caused by the fact that no negative weights are available to cancel out the signal reaching the activation unit. Therefore, the network also detects similarities between the different patterns and the height of the output of each neuron as response to a certain pattern is directly related to the overlap with the pattern that the neuron is trained to recognize. The digit '5' for example shares many pixels with the digit '6' or '8', which therefore result in the second and third highest output signal from neuron one. On the other hand, as depicted in Fig. 6.20c), the pattern resembling the '1' has only negligible overlap with the other patterns. Besides recognizing the ten digits, the neural network is also capable of detecting correlations and similarities between the different patterns.

6.4.2. Hidden layer network for language identification

Based on the pathway to scaling the proposed neural network architecture to multilayer networks, simulations of a three-layer network as shown in Fig. 6.21a) are carried out. Hidden layers increase the number of parameters (synapses) of the neural network and therefore improve the possibility to adapt to more complex input patterns. When designing the number of hidden layers in a network it has to be considered that more layers also greatly increase the training time of a neural network.

The simulated network consists of five input neurons, three hidden-layer neurons and two output neurons and will be trained to identify if a given text is written in English or German language. The network structure with one hidden layer of three neurons is chosen to show the potential of small networks that can potentially be implemented with the experimental setup employed in this work in future. The vowels 'a', 'e', 'i', 'o', 'u' in a given text are first counted in a pre-processing step and divided by the total number of characters to get the fractions of each vowel. These are the five input values for the neural network. Because these fractions

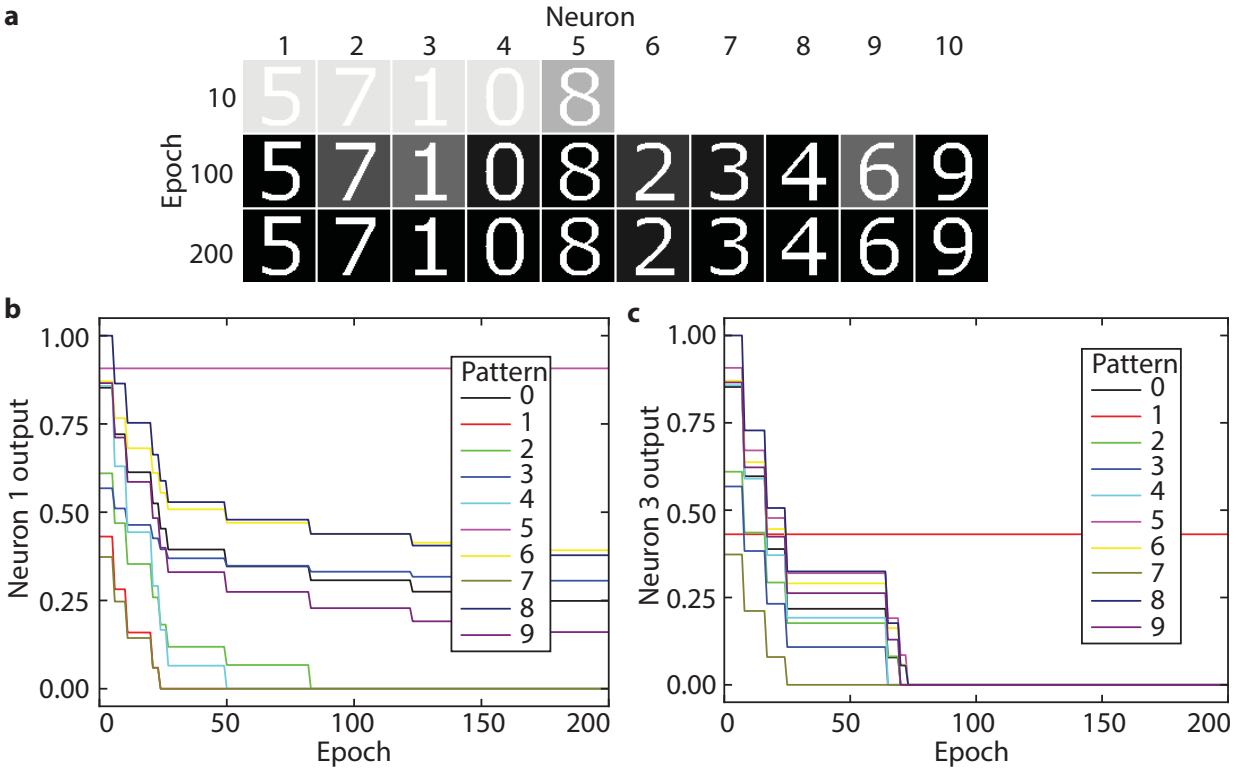

Figure 6.20.: Digit recognition with a network consisting of ten neurons and 4096 synapses. a) Every neuron specializes to one of the digits in an unsupervised learning process over 200 epochs with inhibitory connections based on the 'winner-takes-all' rule. **b)** and **c)** The output of neuron 1 and 3 over time in response to all digits. Neuron one recognizes the digit '5'. As this pattern is similar to the patterns representing '6' and '8', these digits result in the second highest output of the neuron. Neuron three on the other hand is specialized to the digit '1', which has low overlap with other digits. The height of the neuron output in response to different patterns is therefore a measure of the similarity between the patterns.

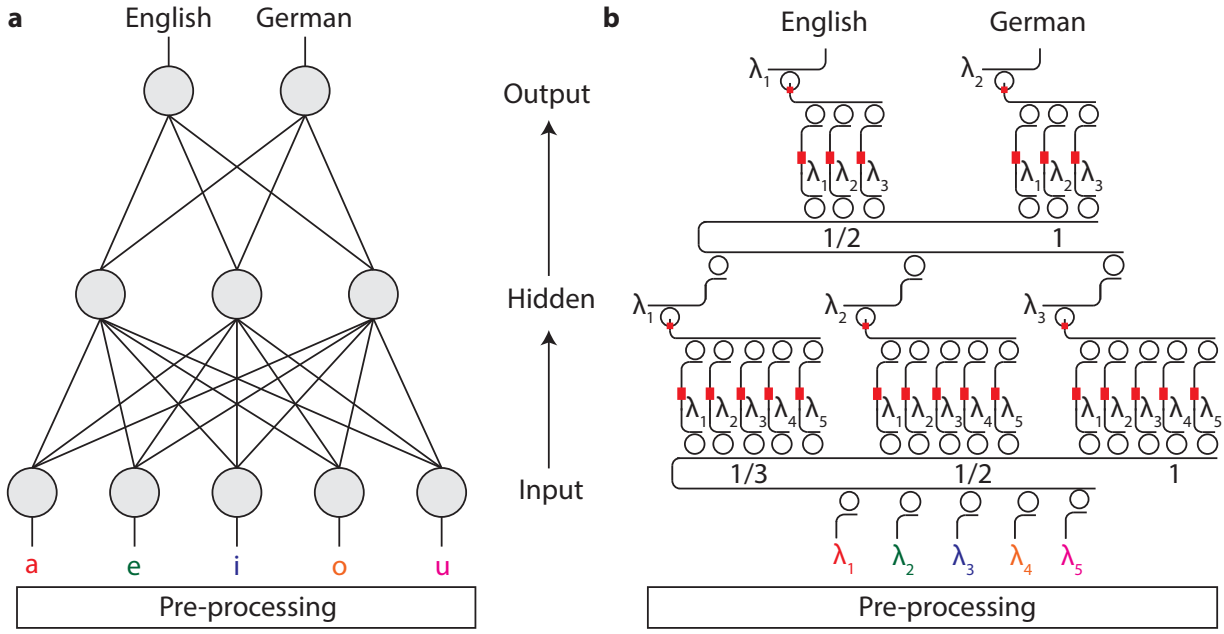

Figure 6.21.: Multilayer network for language identification. **a)** Layout structure consisting of five input neurons, three hidden layer neurons and two output neurons. The five input neurons receive the fractions of the vowels 'a', 'e', 'i', 'o' and 'u' of a certain text (pre-processing) as input. If it is an English text, the first output neuron will fire, for a German text the second output neuron fires. **b)** Translation of the general structure into a photonic circuit based on the architecture proposed in this work. The input signals are encoded on five different wavelengths.

are different for different languages, the network is able to distinguish between languages if the network is trained correctly. Fig. 6.21b) shows the photonic translation of the general network based on the design proposed in this chapter. The input signal is again encoded on the amplitude of different wavelengths that are sent to the five inputs of the network.

To test the network, short input texts with a length up to 150 words are processed and presented to the network. The text input is taken from [153] and [154]. The achieved accuracy in relation to the number of words in the input text is shown in Fig. 6.22. The accuracy for single words is below 50% because not all words include vowels so that both output neurons do not respond. Already above 35 words the accuracy reaches 90%. For texts with more than 150 words the accuracy is above 99.6%. This simulation shows that a small neural network consisting of only a few neurons can already solve complex tasks such as language identification, underlining again the potential of specialized hardware for cognitive computation.

6.5. Conclusions

In this chapter at first a single all-optical neuron was implemented capable of basic pattern recognition. The input signals were weighted using phase-change materials and summed up

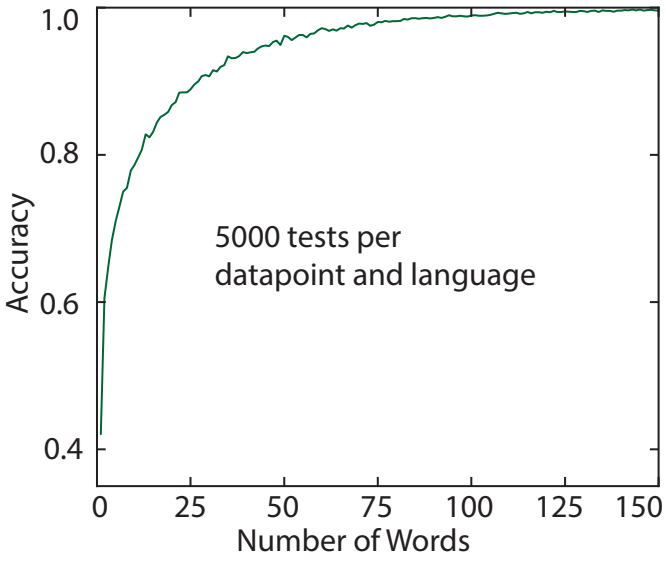

Figure 6.22.: Accuracy of the language identification network as a function of the number of words. The network is presented 5000 text inputs per datapoint and the accuracy is calculated. Above a word count of 35 the accuracy already increases above 90%.

exploiting WDM. The activation unit is comprised of a PCM embedded in a ring resonator that effectively implements a ReLU activation function. The optical neuron works solely in the optical domain and can be operated in a supervised and unsupervised mode. In a second step a scalable architecture was developed that allows to combine the single neurons in layers and connect these layers to deep neural networks. A single layer of this structure comprising sixty synapses and more than 120 optical components is fabricated and successfully applied to pattern recognition of four different letters. The neural network carries out a full matrix multiplication in a single time step and can be operated at GHz rates. To investigate the potential of neural networks with more neurons and synapses, simulations based on the experimental behaviour of the single neurons were carried out in a last step showing potential for digit recognition. A small network using a hidden layer is used to demonstrate the capability of small networks even for complex tasks as language identification.

The presented architecture shows a pathway to scaling from single optical neurons to deep (many-layer) neural networks with phase-change materials as passive weighting elements that non-volatily store the information of the learned patterns.

Chapter 7.

7 Photonic accelerator for convolutional neural networks

Besides building all-optical neuromorphic processors for cognitive tasks mimicking biological brains, a step in between is developing hardware accelerators that take over specific tasks of an electronic processor to increase its performance – preferably a computationally expensive part of the processing that has to be carried out many times in a similar form. In case of neuromorphic computation for neural networks, a very common mathematical operation are matrix multiplications that consist of many multiply-accumulate (MAC) operations of the form $a+b\cdot c$. For example, calculating the weighted sum for a given input in each layer of a neural network (see Sec. 2.2) can be viewed as a matrix multiplication if the weights of all neurons are written in form of a matrix w and the input pattern as the vector v. The resulting vector $o = w \cdot v$ holds the weighted sum (activation energy) for all neurons in the layer that is fed to the individual activation units. In case of the all-optical neural network demonstrated in Chap. 6 the matrix multiplication is performed by the collector and distributor in combination with the PCM-cells.

A lot of effort is currently put into the development of electronic hardware accelerators for matrix multiplications [43] that aim to speed up cognitive tasks. In conventional computers parallelization is achieved through using ever more processor cores, still limited by the von Neumann bottleneck and the data transfer between processor and memory. More recent electronic implementations employ electronic crossbar memories with phase-change materials to efficiently perform such operations [28] in a computational memory.

When implementing a hardware accelerator for matrix multiplications on the platform of phase-change photonics, two main constraints must be considered. First, to gain processing speed a passive operation of the device in a simple transmission measurement is desirable, keeping the switching of PCM-cells to a minimum. Second, as photonic circuits cannot compete with electronic circuits in terms of device footprint, which in the end defines the production cost of a processor, small matrix sizes and a high degree of parallelization are preferable. The first condition is

fulfilled by pretrained neural networks in which the matrix elements (the weights) are fixed and can be stored in the phase-change cell in a non-volatile fashion so that only the input patterns need to be changed. The second constraint is fulfilled by a special class of neural networks which are convolutional neural networks (CNNs) that usually require matrix sizes below 1000×1000 *elements [155]. Parallelization can be achieved by WDM as will be explained in Sec. 7.2.3.*

7.1. Convolutional neural networks

Convolutional neural networks [71, 156] are the currently best performing neural networks on image classification tasks. In the well-known ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [157], which is a competition on image classification in computer vision based on the public ImageNet database [158], CNNs such as AlexNet (2012) [159], GoogleLeNet (2014) [155] or ResNet (2015) [71] have proven to achieve the lowest error rates in the past years. The competition consists of detecting objects ('dog', 'house', 'car',...) in a subset of images of the ImageNet dataset, localizing them (drawing a bounding box) and assigning a general label classifying the objects in the image. CNNs gain their strength from so called convolution layers that extract features of the input image in a pre-processing step, making them easier to detect for a subsequent classical neural network. The convolution operation consists of convolving the input image with a certain image filter (kernel), which is typically of size 3×3 or 5×5, and is the computationally expensive task when calculating CNNs. The relatively small kernels can be combined to a filter matrix resulting in a small matrix suitable to be implemented with photonic structures. The convolution operation can be mapped to matrix multiplications and efficiently performed in a highly parallel way using phase-change photonics as described in the following.

7.2. Photonic tensor core

The heart of the photonic tensor core (PTC) build in this work is based on a waveguide crossbar array as depicted in Fig. 7.1 that implements a fixed weight matrix with phase-change materials and is employed for parallel matrix-vector multiplications. The input vectors are modulated on the optical spectrum of a coherent single soliton frequency comb, which is also implemented on a silicon nitride chip, providing a compact and high bandwidth source with a fixed wavelength spacing. By multiplexing several input vectors, very high computational speeds competitive with current electronic implementations are demonstrated.

The general concept is based on ideas and previous experiments by Carlos Rios and Nathan Youngblood from the University of Oxford [160], who implemented a first matrix multiplier on the phase-change photonics platform on a small scale. Throughout this thesis, a scalable and highly parallelized implementation was designed and experimentally realized in close relation with Nathan Youngblood.

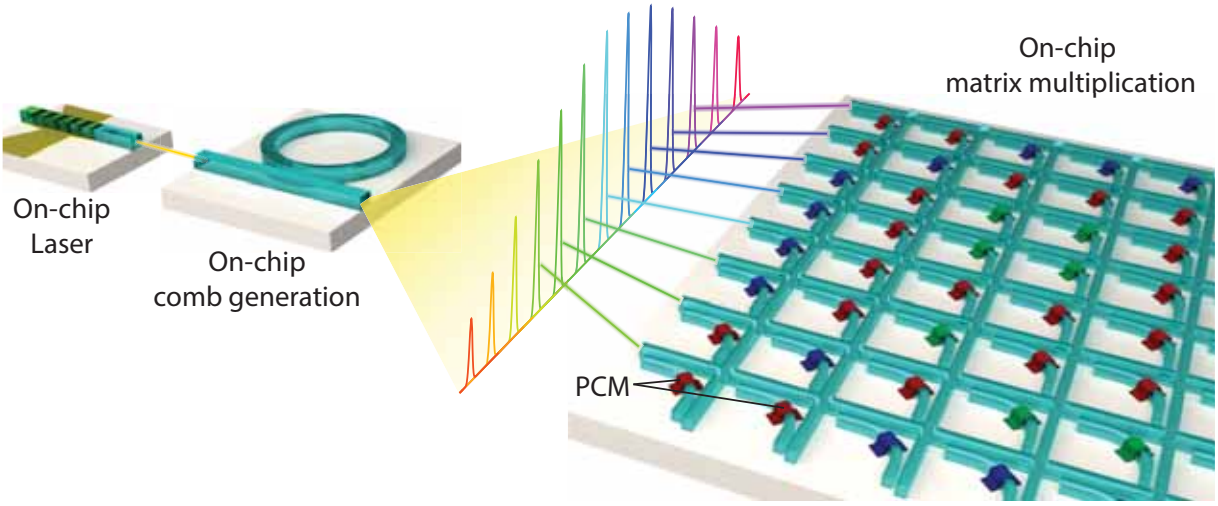

Figure 7.1.: Basic principle of the proposed photonic matrix multiplier. A high power CW-laser generates a frequency comb in an on-chip micro resonator. The comb lines are modulated to represent the input vector of the matrix multiplication. The matrix is implemented in a waveguide crossing structure employing the PCM as matrix elements.

7.2.1. Architecture and basic operation principle

Matrix multiplications consist of scalar multiplications and subsequent addition of products. The addition can optically be performed by incoherent addition of optical powers, similar to the implementation of the summation in the optical neuron. The multiplication $a \cdot b = c$ is based on the transmission through a phase-change cell as illustrated in Fig. 7.2a). The first factor a is encoded in the amplitude P_{in} of the incoming light and the second factor b is inscribed in the attenuation coefficient of the PCM. From the measured output power P_{out} the multiplication result can be extracted. The same principle is employed in the sketched photonic device in Fig. 7.2c) to carry out the matrix-vector multiplication in Fig. 7.2b). To perform the matrix multiplication $A^T \cdot x = y$, the input vector x is encoded in the amplitude of light with different wavelengths λ_1 to λ_i and sent to the different horizontal waveguides corresponding to the rows of the vector. The matrix is encoded in the transmission factors of the PCM-cells in the matrix. The input light is equally distributed to each cell of the matrix by designing the splitting ratios of the directional couplers accordingly. The horizontal directional couplers guide the light to the columns of the matrix and can be calculated analogue to the splitting ratios in the optical neural network as $1/(j+1-k)$ with j being the total number of columns and k the column index. The vertical couplers provide equal splitting between the rows so that the same amount of light from each input row arrives at a certain output and the splitting ratios are derived as $1/l$ with l being the row index. The resulting vector y can be derived from the output powers in the columns, that basically represent the sum of the scalar products between the input vector and the transmission factors defined by the phase-change cells in each column multiplied by $1/(ij)$. By encoding the

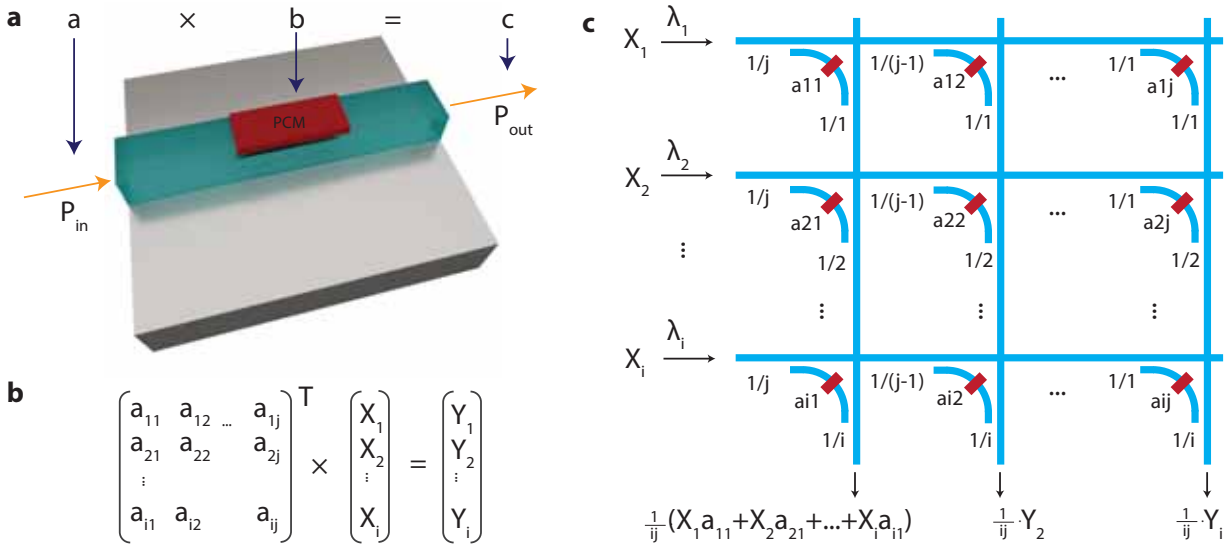

Figure 7.2.: Photonic matrix multiplication. a) The transmission measurement of a basic phase-change cell can be viewed as a scalar multiplication. The input light of power P_{in} is multiplied with the absorption coefficient of the PCM b resulting in the output power P_{out}. **b)** Matrix multiplication performed in the waveguide array in **c)**. The PCM-cells (red) inside the waveguide array resemble the matrix A. Using directional couplers the input vector x, encoded on different wavelength, is distributed to the PCM-cells resulting in a transmission at the vertical output waveguides proportional to the matrix-vector multiplication result.

entries of a vector on different wavelengths, the summation of the individual scalar products is not disturbed by interference effects.

However, it is important to note that the measured output power does not directly reveal the correct output value but requires a post-processing step. To make this evident, a single scalar multiplication needs to be explained in more detail. To calculate the product $a \cdot b = c$, the first factor ($a \in [0,1]$) is encoded in the amplitude of the input light so that the input power is given by $P_{\text{in}} = a \cdot P_0$, with a fixed standard input power P_0. The second factor, $b \in [0,1]$, is encoded in the loss factor of a phase-change cell l_{PCM} ($l_{\text{PCM}} \in \left[l_{\text{cryst}}, l_{\text{amorph}}\right]$, with l_{cryst} and l_{amorph} being the loss factors in the crystalline and amorphous state, respectively). The result c is then calculated from the transmitted optical power P_{out} measured at a photodetector. As the light is not fully absorbed even in the fully crystalline state ($l_{\text{cryst}} > 0$), the output power always depends on the input power, even if the factor b is set to 0. The output power is given by $P_{\text{out}} = a \cdot P_0 \cdot l_{\text{PCM}}$. For $b = 0$ (and therefore $l_{\text{PCM}} = \left(l_{\text{cryst}} + l_{\text{amorph}}\right)/2$) and $a = 1$ this results in an output power $P_1 = 1 \cdot P_0 \cdot \left(l_{\text{cryst}} + l_{\text{amorph}}\right)/2$ and $P_2 = 0.5 \cdot P_0 \cdot \left(l_{\text{cryst}} + l_{\text{amorph}}\right)/2$ in case of $a = 0.5$. Because $P_1 \neq P_2$ this reveals an offset problem present due to the non-zero loss factor of the PCM. To correct for this offset a reference value has to be subtracted from the output power that depends

on the input power leading to the term

$$c = \left(P_{\text{out}} - P_{\text{ref}}\right) / \Delta P_{\text{max}}. \tag{7.1}$$

$P_{\text{ref}} = a \cdot P_0 \cdot l_{\text{PCM}}^{\text{ref}}$ is the power transmitted if the phase-change material is in the reference state $l_{\text{PCM}}^{\text{ref}}$, which in the given example is $l_{\text{PCM}}^{\text{ref}} = \left(l_{\text{cryst}} + l_{\text{amorph}}\right)/2$.[1] ΔP_{max} is a normalization factor that is fixed and given by the maximum power difference between the reference state and the fully amorphous state: $\Delta P_{\text{max}} = P_0 \cdot l_{\text{amorph}} - P_0 \cdot l_{\text{PCM}}^{\text{ref}}$. It should be noted that the reference power P_{ref} depends on the input power $a \cdot P_0$ and therefore is a scalar multiplication by itself. Plugging the previous expressions into the equation for c this finally leads to

$$c = a \cdot \frac{l_{\text{PCM}} - l_{\text{PCM}}^{\text{ref}}}{l_{\text{amorph}} - l_{\text{PCM}}^{\text{ref}}} \tag{7.2}$$

and

$$b = \frac{l_{\text{PCM}} - l_{\text{PCM}}^{\text{ref}}}{l_{\text{amorph}} - l_{\text{PCM}}^{\text{ref}}}. \tag{7.3}$$

When going from scalar multiplication to dot products and matrix multiplications equation (7.1) still holds, only that P_{ref} now also becomes a dot product that depends on the input vector. In case of a dot product $c = \vec{a} \cdot \vec{b}$ with m-dimensional vectors $\vec{a}$ and $\vec{b}$, P_{out} will be given by

$$P_{\text{out}} = a_1 \cdot P_0 \cdot l_{\text{PCM}}^{1} + \ldots + a_m \cdot P_0 \cdot l_{\text{PCM}}^{m} = P_0 \sum_{i=1}^{m} a_i \cdot l_{\text{PCM}}^{i} \tag{7.4}$$

and similarly

$$P_{\text{ref}} = a_1 \cdot P_0 \cdot l_{\text{PCM}}^{\text{ref},1} + \ldots + a_m \cdot P_0 \cdot l_{\text{PCM}}^{\text{ref},m} = P_0 \sum_{i=1}^{m} a_i \cdot l_{\text{PCM}}^{\text{ref},i}. \tag{7.5}$$

This means that for every multiplication, two optical measurements have to be carried out as will be explained in more detail in Sec. 7.4.3.

7.2.2. Equally distributing the light to all matrix cells

In order to achieve equal distribution of the input vectors to all columns of the matrix and also ensuring equal fractions of all inputs to the different rows in the measured output power, the splitting ratios of the directional couplers in the matrix have to be chosen carefully. In Sec. 7.2.1 for simplicity reasons a lossless photonic circuit was assumed to calculate the splitting ratios of the vertical and horizontal directional couplers. In an actual fabricated photonic device, the insertion loss of the directional couplers l_{DC} as well as the loss of the waveguide crossings l_{C} must

[1] Note that the reference state can in generally be any value in the range $l_{\text{cryst}} \leq l_{\text{PCM}}^{\text{ref}} \leq l_{\text{amorph}}$ depending on the given problem. If only positive matrix values are required, the reference state can for example be set to l_{cryst} to exploit the full dynamic range of the PCM.

be taken into account when evaluating the splitting ratios for equal splitting of the light. With the row and column indices k and i the through-port transmission of the vertical couplers can recursively be calculated to be (see more details in App. A.3)

$$t_{k+1} = \frac{1}{l_c l_{\mathrm{DC}} \left(1 - t_k\right) + 1} \tag{7.6}$$

with $l_{\mathrm{DC}} = l_{\mathrm{C}} = 1$ meaning no loss. A similar relation is found for the through-port transmission of the horizontal directional couplers:

$$t_{i-1} = \frac{1}{l_c l_{\mathrm{DC}} \left(1 - t_i\right) + 1}. \tag{7.7}$$

Equivalent to the calibration of directional couplers in the optical neural network (Sec. 6.2.2), calibration devices are fabricated to measure the splitting ratios with a slightly different design (see the inset of Fig. 7.3a) to the one previously presented. Fig. 7.3a) shows the splitting ratio[2] between through- and cross-port of the directional couplers as a function of the coupling length, well following the expected $\sin^2$ behaviour. With the coupling gap of 200 nm the whole range of splitting ratios ($c \in [0, 1]$) can be achieved within a coupling length of maximal 63 µm. Because of the tapering region introduced to reduce the insertion loss of the directional couplers, the splitting ratio for a coupling length equal to zero is already about 15% as the waveguides are brought together closely in a single point, raising the need to use coupling lengths above the first maximum of the curve at around 30 µm to achieve smaller splitting ratios. This can be improved in further designs by varying the gap between the two waveguides with zero coupling length to obtain smaller coupling fractions with short directional couplers. The insertion loss of the directional couplers is estimated in a measurement concatenating several of the couplers as shown in the inset of Fig. 7.3b. The linear regression reveals an insertion loss of 0.14 dB. The insertion loss of the crossings is 0.23 dB (see Sec. 5.3).

An important factor in the design of the photonic matrix is the wavelength dependence of the directional couplers setting a constraint to the maximum wavelength range that can be exploited for processing. Fig. 7.4 shows the cross-port splitting ratio of the directional couplers as a function of the coupling length for wavelengths in the range from 1500 nm to 1619 nm. It can be seen that in the range to the first maximum of the spectra at around 30 µm, which in principle holds all necessary splitting ratios in the range from 0.0 to 1.0, all curves show similar ratios with a deviation of up to 2.4 dB for small coupling lengths. The maximum splitting ratio used in the experiments is indicated by the dashed vertical line and shows a wavelength dependent difference in the splitting ratio of 2.2 dB. Although the directional couplers already enable to use a broad wavelength range larger than 100 nm, more sophisticated designs can exhibit an even larger range with more uniform wavelength response as shown for example in [161].

[2] The splitting ratio is calculated as $P_{\mathrm{cross}} / \left(P_{\mathrm{through}} + P_{\mathrm{cross}}\right)$.

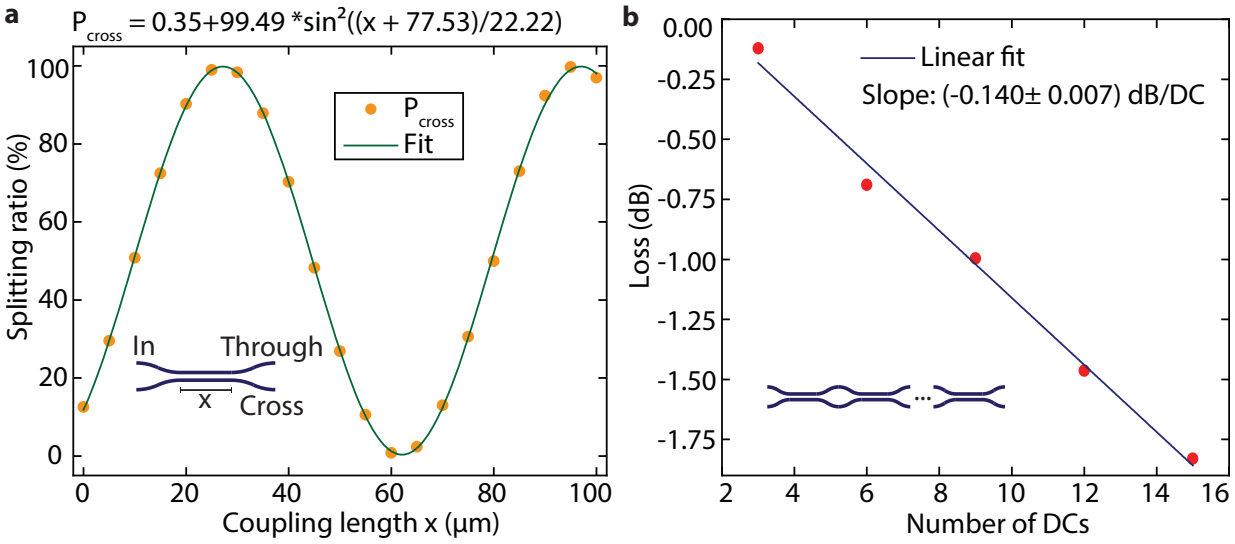

Figure 7.3.: Directional coupler calibration. a) Cross-port transmission as a function of the coupling length x. The inset shows the symmetric coupler design that is used. **b)** Measurement of the insertion loss of the directional couplers (DCs) at a wavelength of 1550 nm, by concatenating them as shown in the inset. From the linear regression a loss of 0.14 dB per coupler is obtained.

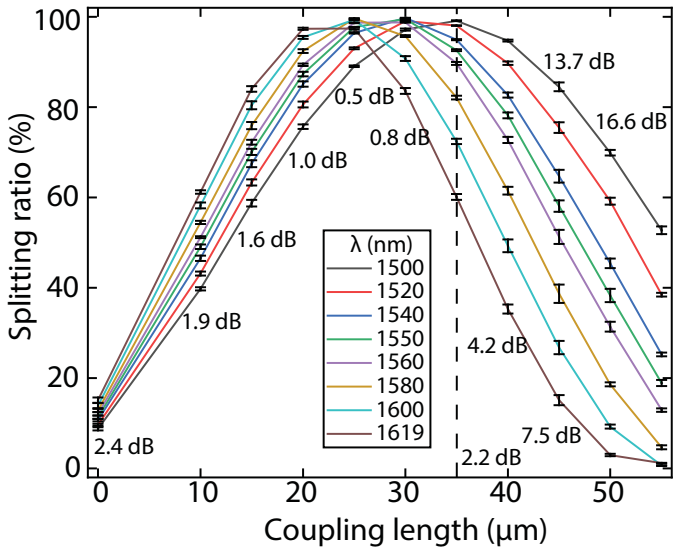

Figure 7.4.: Wavelength dependence of the directional couplers in the range from 1500 nm **to** 1619 nm**.** The dashed line indicates the largest coupling length used in the experiments, with a variation in the splitting ratios of 2.2 dB between the wavelengths.

7.2.3. Multiplexing input vectors

A huge advantage of the presented implementation of a photonic matrix multiplier is the capability of multiplexing many input vectors carrying out many multiplications in parallel. Fig. 7.5 illustrates the basic principle of multiplexing four input vectors in a 4×4 matrix, which is ex-

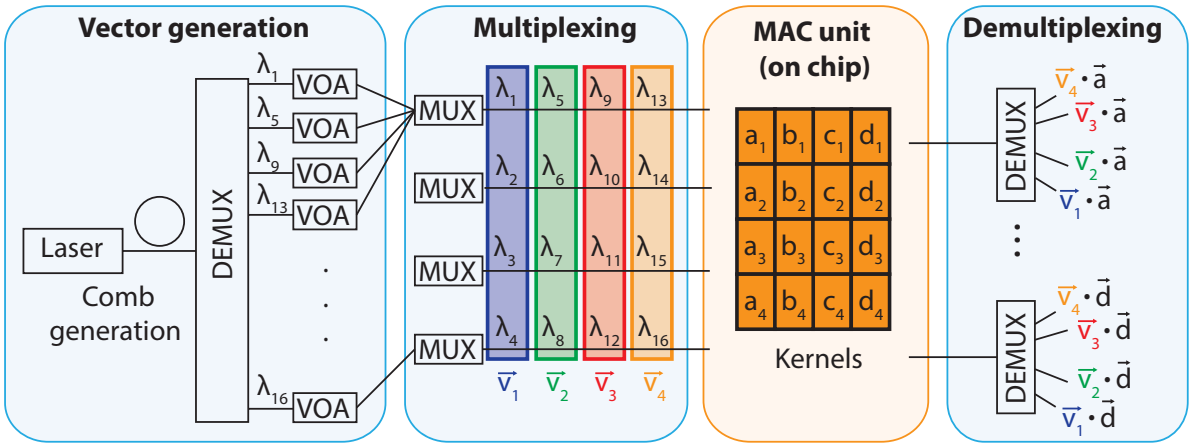

Figure 7.5.: Multiplexing input vectors. The wavelengths source needed to modulate the input vectors is a frequency comb generated using an on-chip microring resonator. The comb lines are demultiplexed and individually modulated. By combining the signals on wavelengths of different vectors corresponding to the same row via multiplexing, they can be simultaneously fed to the on-chip matrix. After passing through the matrix, the wavelengths are demultiplexed according to the input vectors.

perimentally implemented in Sec. 7.4, exploiting wavelength division multiplexing. Each vector is encoded on four different wavelengths generated by a coherent frequency comb (see Sec. 7.2.4). The 16 individual wavelengths λ_1 to λ_{16} are then demultiplexed and their amplitudes independently modulated with variable optical attenuators (VOAs) according to their corresponding input vector. The first vector for example consists of the wavelengths λ_1 to λ_4 and the second of λ_5 to λ_8. After merging all wavelengths corresponding to the different rows of the matrix via multiplexing (e.g. $\lambda_1, \lambda_5, \lambda_9, \lambda_{13}$ to row 1), they are sent to the inputs of the matrix. At the matrix output the wavelengths are demultiplexed according to the vectors revealing the multiplication results. This way the same matrix can be used to perform four multiplications in parallel and more depending on the number of available wavelengths in the frequency comb spectrum, offering a huge advantage compared to electronic implementations and significantly increasing the computation density as will be shown in more detail in Sec. 7.4.7.

7.2.4. Frequency combs

A frequency comb is a light source consisting of equidistantly spaced discrete frequency lines and is for example used for metrology, spectroscopy and optical communications [162–165]. While first demonstrations of frequency combs were achieved with big table-top systems, research in the past years led to the realization of comb generation with fully chip-integrated lasers and micro resonators in platforms compatible with CMOS fabrication [166] making frequency combs more and more attractive also for photonic data processing.

Chip-based frequency combs can be generated using continuous wave lasers with powers below

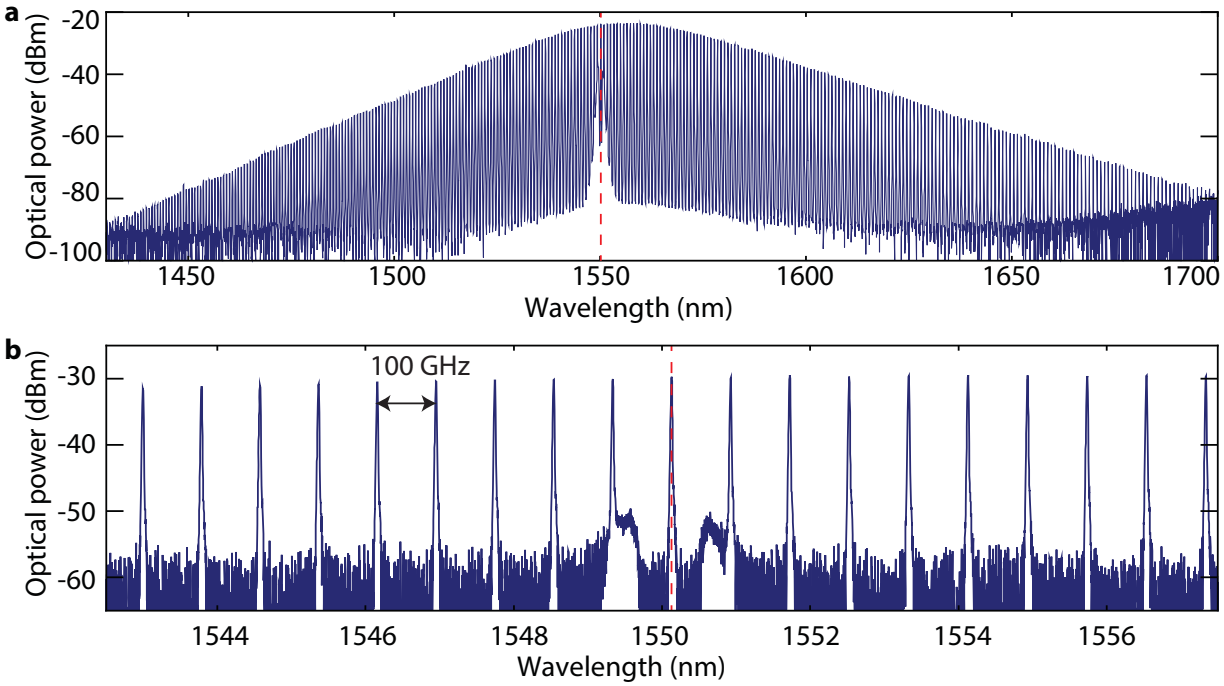

Figure 7.6.: Optical spectrum of the on-chip frequency comb used in this work. a) Full spectrum of the comb spanning over 270 nm in the single soliton state. **b)** Close-up of the frequency spectrum around the pump wavelength at 1550 nm showing equally spaced frequency lines with a spacing of 100 GHz. The red dashed line indicates the pump wavelength.

1 W and rely on dissipative Kerr solitons (DKSs) [162]. Solitons are optical fields that preserve their field profile during propagation. DKS can be generated in ring resonators and rely on a balance of gain and loss of optical power as well as nonlinearity and dispersion. The resonators are designed so that the nonlinearity compensates for the dispersion and a single soliton can circulate within the cavity. Based on the four-wave mixing principle, a frequency comb can form inside the resonator with a frequency spacing equal to the free spectral range of the resonator [167, 168], which in this work is 100 GHz.

Fig. 7.6 shows the optical spectrum of the frequency comb in the single soliton state[3] employed in this work. It spans a spectral range of over 270 nm showing approximately 340 frequency lines. The pump wavelength of 1550 nm on which the ring resonator is pumped with around 1 W is highlighted with the red dashed line. Figure 50b) shows a close-up of nineteen comb lines around the pump line. The frequencies are equally spaced with a spacing of about 800 pm (100 GHz). The pump line is filtered and attenuated to the power level of the other lines so that it can also be used in the operation of the photonic matrix. The microresonator used for the comb generation has a quality factor of around 10^7 and is fabricated in Si_3N_4 with a photonic Damascene process [170].

[3] To prepare the single soliton state a standard tuning technique is employed as described in [169].

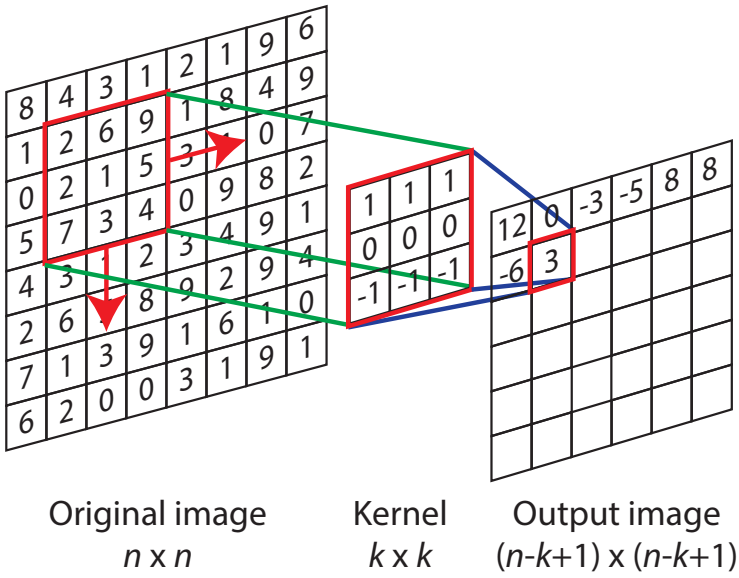

Figure 7.7.: Principle of a convolution operation. An image filter (kernel) of size $k \times k$ is moved over the original input image of size $n \times n$ and the pixelwise product between kernel and image area reveals the convolution between image and kernel.

7.2.5. Convolution operation

The photonic tensor core is used in this work to perform convolution operations, which are the computationally expensive tasks in convolutional neural networks and can be reduced to matrix multiplications. Convolution operations are also common in image processing to apply filters for blurring, edge detection or sharpening of images.

Fig. 7.7 explains the convolution operation between an input image and a kernel as processed on the convolution layers of a CNN in detail. The image and the kernel (filter) are convolved by shifting the kernel step by step over the image as indicated by the red arrows and calculating the sum of the pixelwise products between kernel and the corresponding area of the input image (with the numbers representing the colours of the pixels). In the example depicted in Fig. 7.7 calculating the dot product between the part of the image highlighted by the red square with the kernel this leads to

$2 \cdot 1 + 6 \cdot 1 + 9 \cdot 1 + 2 \cdot 0 + 1 \cdot 0 + 5 \cdot 0 + 7 \cdot (-1) + 3 \cdot (-1) + 4 \cdot (-1) = 3.$

The filter is shifted pixel by pixel over the image leading to a total number of $(n-k+1) \cdot (n-k+1)$ dot products per convolution operation, making it a computationally expensive problem and the critical bottleneck in processing convolutional neural networks. This problem is especially well suited for the photonic tensor core as in a fully trained CNN the kernels are fixed and of small size (typically 3×3 or 5×5). The values in a kernel are learned in a training process and are then only once programmed to the PCM-cells in the photonic matrix.

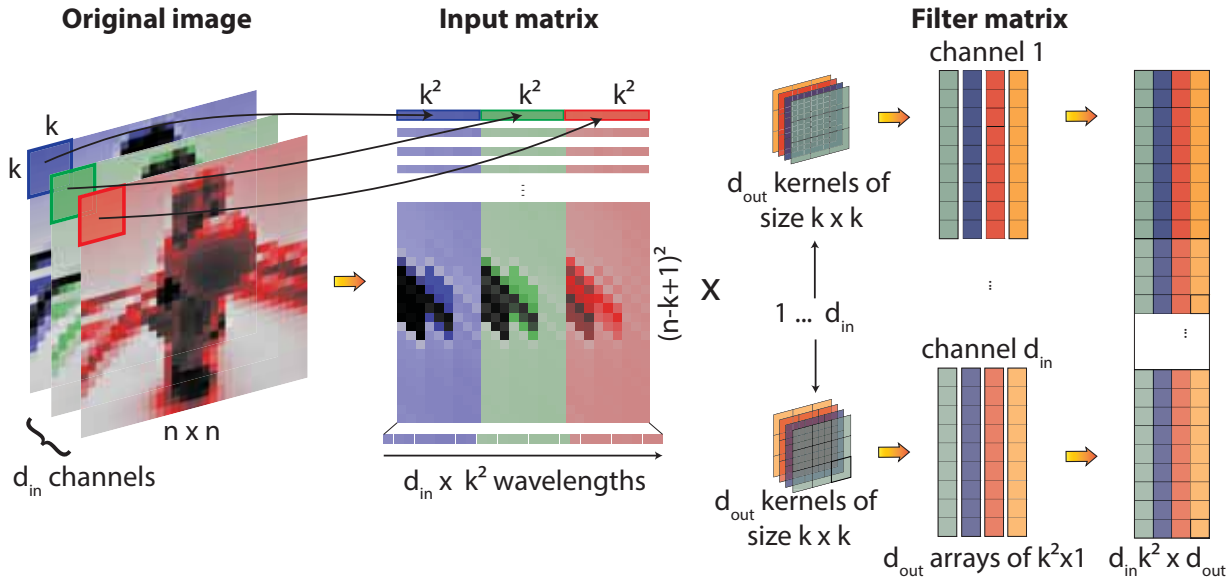

Figure 7.8.: Construction of input and output matrix in a CNN. The original input image with its d_{in} channels is converted to an input matrix by stacking the pixels of a filter volume into a row of the input matrix for every filter position. Similarly, the filter matrix is constructed by stacking the kernels into the columns of the matrix. Each column of the matrix holds one kernel.

7.2.6. Construction of the input and kernel matrix

In the actual implementation of a convolutional neural network several channels are convolved with several kernels simultaneously. The d_{in} channels can for example be the separated red, green and blue (RGB) pixel-values of an input image as shown in Fig. 7.8 on the left. To construct the input matrix from the original image of size $n \times n$, the pixels of all channels corresponding to one kernel area $(k \times k)$ are stacked into the rows of the input matrix leading to a matrix size of $d_{\text{in}} \times k^2 \times (n-k+1)^2$. Each row of the matrix holds one vector that can be processed with the photonic matrix. In a similar way the d_{out} kernels are combined to a kernel matrix that is then programmed into the PCM-cells of the actual photonic matrix. As shown in Fig. 7.8 on the right the individual kernels for all channels are stacked into the columns of the filter matrix, resulting in a matrix size of $d_{\text{in}} \times k^2 \times d_{\text{out}}$. Each kernel represents a certain image filter, which is learned throughout the training process of the convolutional neural network. The whole convolution operation can this way be mapped to a sequence of matrix-vector multiplications in which only the input vectors have to be changed suitable for the photonic tensor core.

7.3. Experimental setup

In the experiments demonstrating the photonic tensor core two setups have to be distinguished. The first is used for matrix vector multiplication with single vectors per time step demonstrating

the basic concept. The second is a more advanced version of the setup capable of processing multiple vectors per time step demonstrating the strength in parallelization of the proposed architecture. In both cases only the photonic matrix with the phase-change materials and the frequency comb generation are implemented on-chip. The setup for the frequency comb generation is provided by the group of Prof. Dr. Kippenberg at EPFL (Lausanne, Switzerland).

7.3.1. Single matrix vector multiplication

The experimental setup can be divided in four main parts (see Fig. 7.9). The first part is used to generate the wavelengths needed to represent the input vector. This is achieved with a soliton microcomb (see Sec. 7.2.4) generated with a tunable continuous-wave laser in the telecom band (40 mW) amplified to approximately 1 W with an EDFA. In an optimized switching process a single soliton state is prepared [169] in a high-Q ring resonator, leading to a very stable and broadband frequency comb with a wavelengths spacing of 100 GHz determined by the free spectral range (FSR) of the resonator. With a second EDFA the comb is amplified to account for optical loss in the setup.

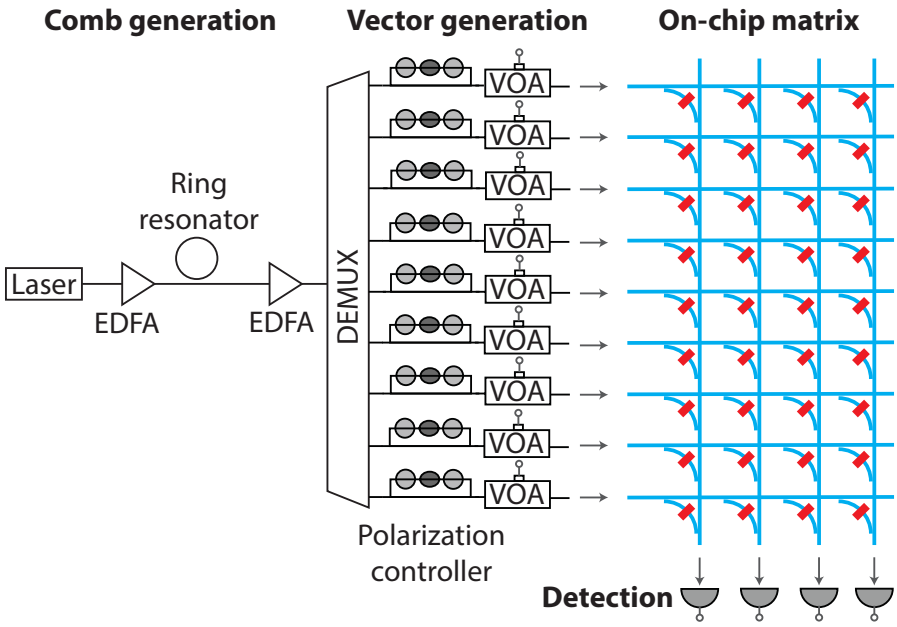

Figure 7.9.: Optical setup for matrix vector multiplication. To multiply single vectors with a 9×4 matrix, nine wavelengths of the frequency comb spectrum generated with a micro resonator are separated using a fibre-based demultiplexer. After adjusting the polarization and modulating the amplitudes of the comb lines with a variable optical attenuator VOA, the signals are sent to the input couplers of the on-chip matrix. The multiplication result is measured with four photodetectors.

The second part is the generation of the input vector, for which the different comb lines are demultiplexed and individually amplitude-modulated with VOAs to represent the entries of the

input vector. The modulators are controlled via a computer and a custom-made digital-to-analogue converter (DAC) conversion board that is interfaced with an Arduino microcontroller. Using polarization controllers, the coupling to the on-chip grating couplers is optimized. The wavelength channels $27 - 35$ of the ITU grid corresponding to the wavelength range 1549.32 nm - 1555.75 nm are used in this experiment.

The third part of the setup is the actual photonic matrix that is implemented on-chip as described in Sec. 7.2 and fabricated in a three-step electron-beam lithography process in silicon nitride (see Chap. 4). The matrix acts passively in the experiment and holds the filter matrix stored in the PCM-cells (red). The PCM employed is GST (3 µm length).

The last part is the detection of the output vector accomplished with standard photodetectors (New Focus, Model 2011) off-chip.

7.3.2. Multiplexing input vectors

Fig. 7.10 shows the experimental setup used for matrix multiplication with four input vectors. Similar to the previous section, a soliton microcomb is demultiplexed into the sixteen wavelength

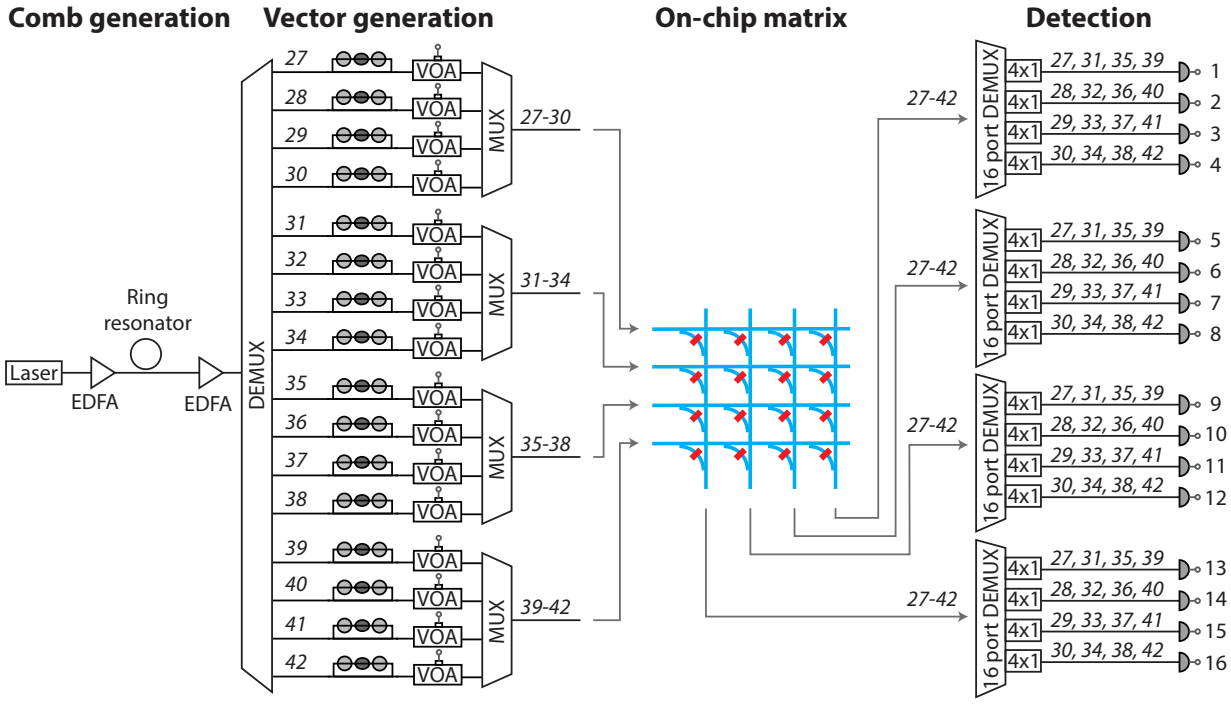

Figure 7.10.: Setup for multiplexing four input vectors in a 4×4 matrix. To parallelize matrix vector multiplications, wavelengths division multiplexing is employed. Each of the four vectors is encoded on four wavelengths (depicted by the channel numbers $27 - 42$ of the ITU grid). The entries of the vectors corresponding to the same row are then multiplexed together and sent to the on-chip matrix. At the matrix output the signals are demultiplexed according to the input vectors and the results can be read from sixteen photodetectors.

channels 27 to 42 of the ITU grid. Because the photonic matrix used is of size 4×4, one vector is represented by four wavelengths and therefore four individual vectors are obtained. Each wavelength is modulated with a VOA again controlled by a custom-made DAC board and a computer. The wavelengths corresponding to the same row index of the vectors are multiplexed together and sent to the corresponding matrix input (for example channels $27 - 30$ represent the first entries of the four vectors and are sent to row one of the matrix). After passing through the matrix, each output contains all sixteen wavelengths. These are demultiplexed and then combined again corresponding to the input vectors as shown in Fig. 7.10. The results are read from sixteen photodetectors (New Focus, Model 2011). In this way $16 \cdot 4 = 64$ multiply-accumulate operations can be processed in a single time step.

7.4. Experimental results of the photonic tensor core

The experiments demonstrating the potential of the photonic tenor core are divided in four parts. In the first part the general principal is investigated using a 9×4 matrix for multiplications with a single vector per step. The second part tackles the challenge of the electronic post-processing resulting from the necessity to subtract a reference value that relies on the input vector as explained in Sec. 7.2.1 and shows a way to resolve it. The next section addresses the multiplexing capabilities of the photonic matrix with the experimental demonstration of 64 MAC operations per time step, whereas the last part sheds light on the speed and energy efficiencies of the proposed architecture and gives a future projection for a fully integrated implementation.

7.4.1. Programming the matrix elements

Before doing an actual matrix multiplication, the matrix elements must be inscribed in the PCM-cells. Fig. 7.11a) shows an optical micrograph of a fabricated 4×4 matrix. The input vectors and the multiplication results are coupled on and off the chip via 3D-printed total internal reflection couplers (see Fig. 7.11c). These couplers are optimized for broadband transmission and the variation of their peak transmission in the range from 1480 nm to 1640 nm is below 0.25 dB [113, 114] making them ideal for coupling a broad range of lines of the frequency comb to the chip.

Fig. 7.11b) shows a scanning-electron micrograph of a PCM-patch deposited on top of a waveguide crossing with good alignment to the photonic structures. The waveguide crossing is needed to efficiently program the state of the PCM in a matrix cell as illustrated in more detail in Fig. 7.11d). To be able to conveniently address all cells independently, three additional grating couplers are added to each unit of the matrix. The two alignment grating couplers are used to align the fibre array to the ports in a simple transmission measurement. The third grating coupler in the middle ('Switching In') is used to send the optical pulse for switching the PCM-cell on the crossing. This additional coupler is necessary to prevent unintended switching of neighbouring

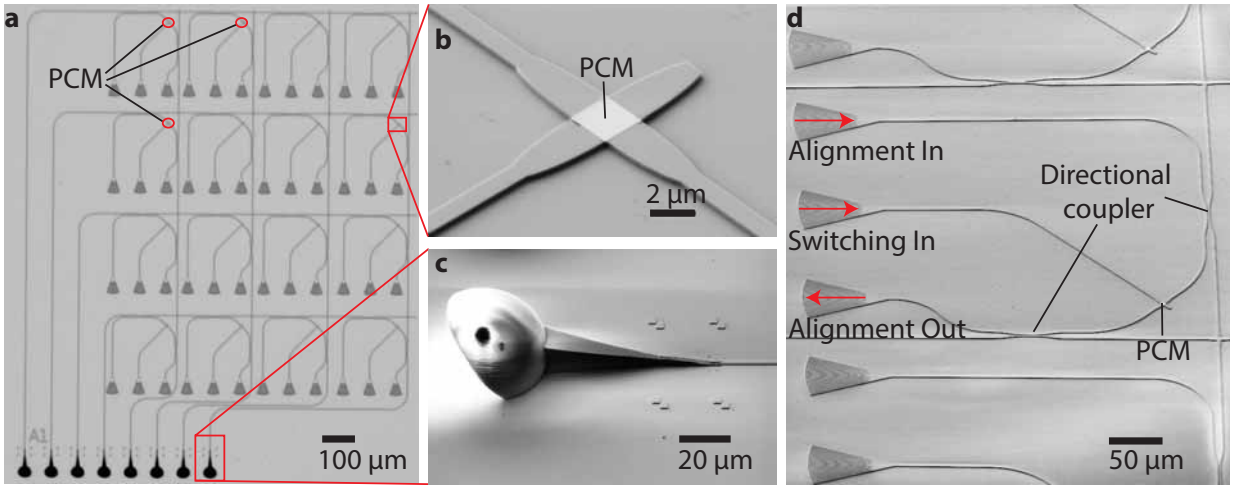

Figure 7.11.: Micrographs of a fabricated 4×4 matrix device. **a)** Optical micrograph of a photonic 4×4 matrix. **b)** Scanning-electron micrograph with the phase-change material deposited on top, showing good alignment between waveguides and the PCM. **c)** Scanning-electron micrograph of a 3D printed total internal reflection coupler used as input and output couplers. **d)** Scanning-electron micrograph of a single matrix cell highlighting the alignment and switching inputs for programming the PCM.

cells. Depending on the splitting ratios of the directional couplers in a matrix cell only a part of a potential switching pulse sent to one of the alignment couplers arrives at the selected PCM-cell, the other part travels further on to the next cell and could displace its value. Using the additional perpendicular input, only the intended matrix element is programmed.

Note also that for a similar reason all directional couplers are chosen to have less than 95% splitting ratio in order to always be able to align to the cell employing the alignment couplers. Especially the cross-port coupling ratios of the upper right cell of the matrix would otherwise be 100%, making it impossible to do the alignment with the two couplers inside a matrix cell.

The optical pulses employed for switching the matrix elements have a width of 200 ns and a maximum pulse energy up of approximately 800 pJ. This way a reversible contrast of the optical transmission of up to 70% is achieved. The matrix elements are set by manually sending optical pulses till the desired transmission state is reached with the necessary accuracy. A more sophisticated switching process using a double-step pulse as shown in [122] would enable single shot programming of the matrix elements. It should also be noted that the switching energies can be considerably reduced to below 20 pJ by utilizing short optical pulses in the picosecond range. Alternatively, the photonic matrix can be potentially operated in a mixed mode electro-optic version in future, by switching the PCM with electrical pulses enabling easier integration with conventional electronics [171, 172].

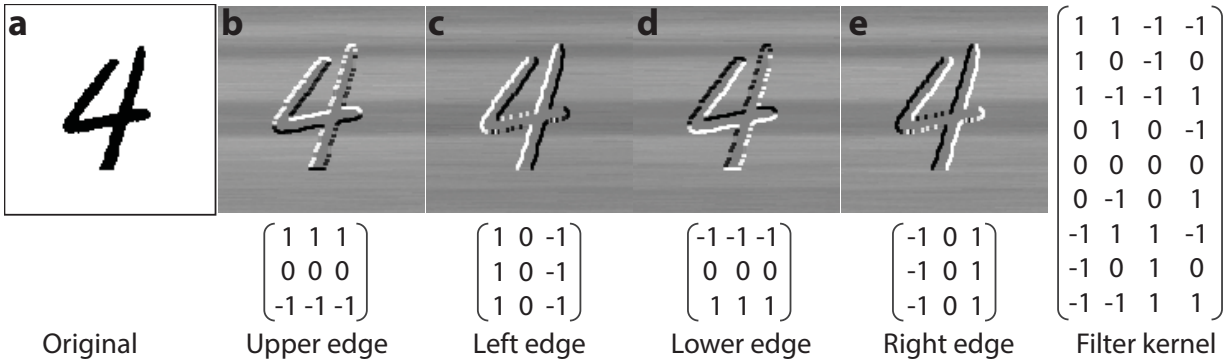

Figure 7.12.: Experimental edge detection with one vector per timestep. **a)** The original image shows a handwritten '4' [173] and consist of 128×128 pixels. **b)-e)** Convolution results highlighting different edges of the input image. The full filter kernel programmed into the matrix is shown on the right.

7.4.2. Single matrix vector multiplications

In a first experiment the convolution operation between an input image and an image kernel is performed with one input vector per time step. The matrix size used in this experiment is 9×4 leaving space for four kernels of size 3×3. The input image is of size 128×128 pixels in 8-bit greyscale (single channel). To better visualize the results of the convolution operation the input kernels applied are common edge detection filters as depicted in Fig. 7.12. In a convolutional neural network, however, the values of a kernel are usually learned throughout a training process and do therefore not necessarily represent an obvious pattern. Corresponding to Sec. 7.2.6, the filter matrix constructed from the four kernels shown underneath the images is shown in Fig. 7.12 on the right.

Fig. 7.12 shows the experimental results of the convolution operation. Because four kernels (and therefore four matrix columns) are used, the number of output images is also four. Each of the output images highlights a different edge of the input image showing a handwritten '4' (Fig. 7.12a) proofing the correct operation of the matrix. Fig. 7.12b) for example highlights the upper edges whereas panel c) highlights the left edges of the input image. To achieve the final output image as shown in the figure, every experimental output value is offset by +0.5. The values below 0 are set to 0 (representing black pixels) and the values above 1 are set to 1 (representing white pixels), which is the standard procedure as commonly used in image processing software to improve the contrast of the edge detection.

Because all kernels are encoded in the same matrix, all output images are obtained simultaneously leading to a total of more than 63 000 inner-product operations. The variation in the grey background of the output images is due to a fluctuation of the input power over the time of processing the image (about 4 mins in this configuration), because the reference values that need to be subtracted from the experimentally obtained values are calculated from the measurement

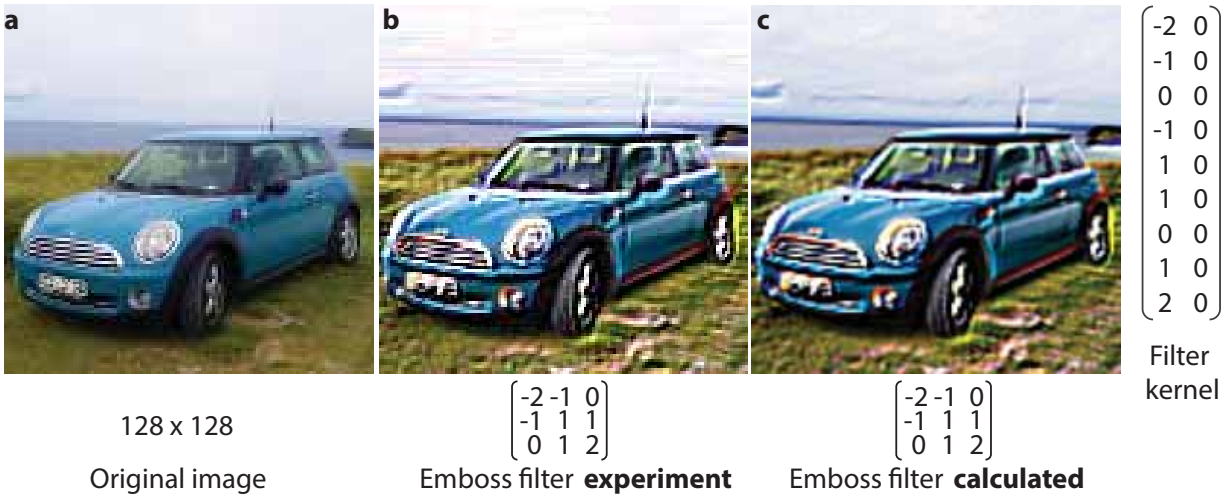

Figure 7.13.: Experimental convolution without electronic post-processing. **a)** Original image of size 128 × 128. **b)** Experimental convolution operation applying an emboss filter. **c)** Calculated result for applying the emboss filter on the original image. The filter kernel is shown on the right.

of the matrix in the reference state that is taken before setting the actual matrix entries corresponding to the kernels. The reference values therefore are fixed and do not reflect the variation of the input power.

7.4.3. Matrix multiplication without electrical post-processing

In the previous section an electrical post-processing step was required to subtract a reference value and obtain the actual result of the inner product. In a real application this would hinder the photonic matrix to reveal its full potential, because for every optical product also an electrical product has to be calculated. The system would therefore still suffer from the limitations of current electronic systems. To avoid the electronic processing the reference value can also be calculated optically by leaving one column of the matrix in a reference state. The reference state is given by the loss-factors ($l_{\mathrm{PCM}}^{\mathrm{Ref}}$) as described in Sec. 7.2.1. Using a balanced detection between a kernel column and the reference, the output value can directly be read from the detector.

To demonstrate this experimentally one column of a 9 × 2 matrix is set to the reference state of the PCM (see the right column of the filter kernel shown in Fig. 7.13), which is in this case the middle between the crystalline and the amorphous state. The maximum contrast between the two states is chosen to be 65%, therefore the reference transmission for all elements is 32.5% above the crystalline ground state. It is important that the transmission for all elements in both columns that correspond to the same entry of a kernel (same row index in the filter matrix) cancel out exactly in order for the balanced detection to work when they are both in the reference state.

With the reference column set to the correct values, the filter column can be adjusted. In

the example shown in Fig. 7.13 an emboss filter is applied to the input image that consists of five different values $(-2, -1, 0, 1, 2)$. These values are projected to the PCM-states between fully crystalline and the amorphous state so that -2 corresponds to the crystalline state, $+2$ to the amorphous state and 0 to the reference state. In terms of switching contrast from the initial crystalline state this leads to the contrast values 0%, 16.25%, 32.5%, 48.75% and 65% for the five different kernel entries.

The result of the convolution between the original image (Fig. 7.13a) and the emboss filter is shown in Fig. 7.13b). The comparison with the electronically calculated convolution with an image manipulation software shows very good agreement and proofs that no electronic post-processing is needed to perform the optical convolution operation. It should be noted that in this example all three colour-channels are treated individually and one after the other. The same filter was applied to the red, green and blue channel and the obtained images recombined in the end. This operation could be executed in parallel by multiplexing the input vectors of the three image channels as explained in the next section.

7.4.4. Parallelizing the convolution operation

The power of the photonic tensor core for matrix multiplications presented in this work stems substantially from the capability of wavelength division multiplexing (WDM) for parallelizing the processing. As the filter matrix stays constant during the whole convolution operation, only the input vector that is modulated on the lines of the frequency comb needs to be changed. Because the photonic matrix employs no resonant elements as for example ring resonators, a wide range of the optical spectrum can be used to encode multiple vectors.

The experimental demonstration of multiplexing four vectors per time step is shown in Fig. 7.14 with three different input images. According to Sec. 7.2.3, a 4×4 matrix with a kernel size of 2×2 is used and four input vectors consisting of sixteen wavelengths are processed in parallel. In every time step four pixels of all four output images are obtained, leading to a speed-up of a factor of four compared to single vector operation. The image kernels applied are again constructed for edge detection, as can for example clearly be observed at the edges of the bricks in the first image. Because the post-processing in this experiment was done electronically, the horizontal lines again are caused by variations in the input power during the computation. Additionally, a weak diagonal sub-patterning can be observed, which is caused by a slight misconfiguration between the four photodetectors that result in the four pixels per time step. Because the detectors are slightly offset to each other, the repeated sub-patterning occurs but could be resolved by fine-tuning the experimental setup.

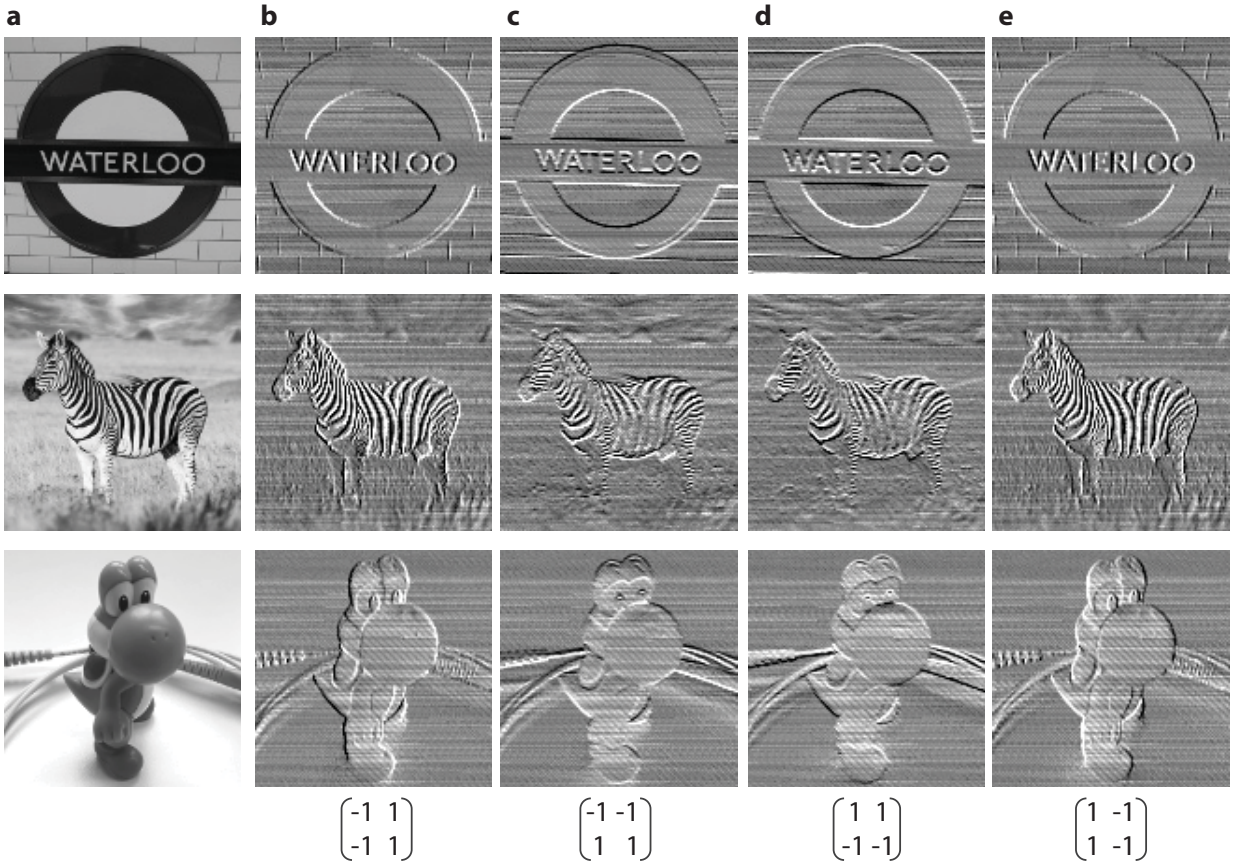

Figure 7.14.: Parallel convolution processing with four vectors and a 4×4 matrix. **a)** Original images. **b) - e)** Experimental results for edge detection with 2×2 kernels and four input vectors per time step.

7.4.5. Accuracy measurements

An important metric for computational hardware is its accuracy. In order to characterize the accuracy of the proposed photonic tensor core, 100 000 inner products are calculated with a fixed matrix and varying vectors with nine entries. Fig. 7.15a) shows the measured calculation result as a function of the expected outcome scaled down to the range $[0, 1]$. The inset shows the histogram of the data revealing a standard deviation of 0.008 with a mean value of 0.002. Fig. 7.15b) shows the calculated standard deviation for a fixed resolution of the input vectors from two to eight bits. The intersection with the blue line depicting 0.008 reveals a resolution of the photonic matrix of 5 bit.

7.4.6. Projections to the future

The performance of the photonic tensor core demonstrated so far is limited by the available electronic setup used to operate the system. Especially the clock speed was limited by the custom

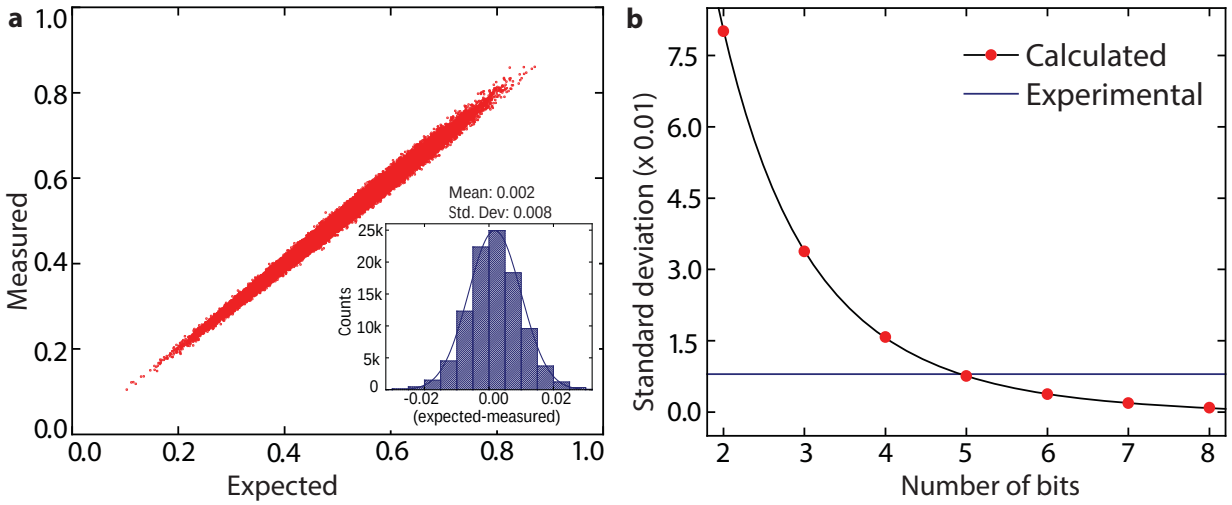

Figure 7.15.: Accuracy of the photonic tensor core. a) Measured multiplication result of an inner product of a 9×1 vector with a 9×1 matrix for 100 000 inner products. The result is scaled to the range $[0, 1]$. The inset shows the histogram of the data with a standard deviation of 0.008. **b)** Calculated standard deviation with a fixed precision given by the number of bits. The intersection with the experimental standard deviation (blue line) reveals a resolution of 5 bit for the photonic tensor core.

DAC-board and the Arduino microcontroller to approximately 1 kHz. Because the operation of the photonic matrix is based on a passive transmission measurement, the speed is in general only limited by the modulator and detector bandwidth.

Fig. 7.16a) shows the experimental setup used to characterize the frequency response of the photonic matrix to estimate the achievable processing speed. Nine wavelengths generated by the soliton microcomb also used for the convolution experiments are modulated separately with a 20 GHz EOM (Optilab, IMC-1550-20-PM). A network analyzer (Keysight, 18 GHz E5063A) generates the signal for modulation and receives the signal from a 12 GHz photodiode (New Focus, Model 1554b) employed to measure the optical output after passing through the matrix. The data in Fig. 7.16b) shows the measured frequency response up to 14 GHz and is obtained by subtracting the performance of the off-chip components (the modulator output is directly connected to the detector) from the result of the whole system including the on-chip matrix. Each line is measured separately. As expected, no loss due to the matrix is seen in the frequency response plot confirmed by the flat response. The maximum frequency is limited by the weakest link in the setup, which is the photodetector with a 3 dB bandwidth of 12 GHz. The open eye-diagram (Fig. 7.16b) further confirms that modulation at 13.5 GHz is possible without any loss.

Considering a modulation and detection speed of 14 GHz while multiplexing four vectors in a 9x4 matrix a processing rate of more than 2 TMAC/s ($4 \times 9 \times 4 \times 14$ GHz) can be achieved. However, this estimation is limited by the available experimental setup in this work, especially the

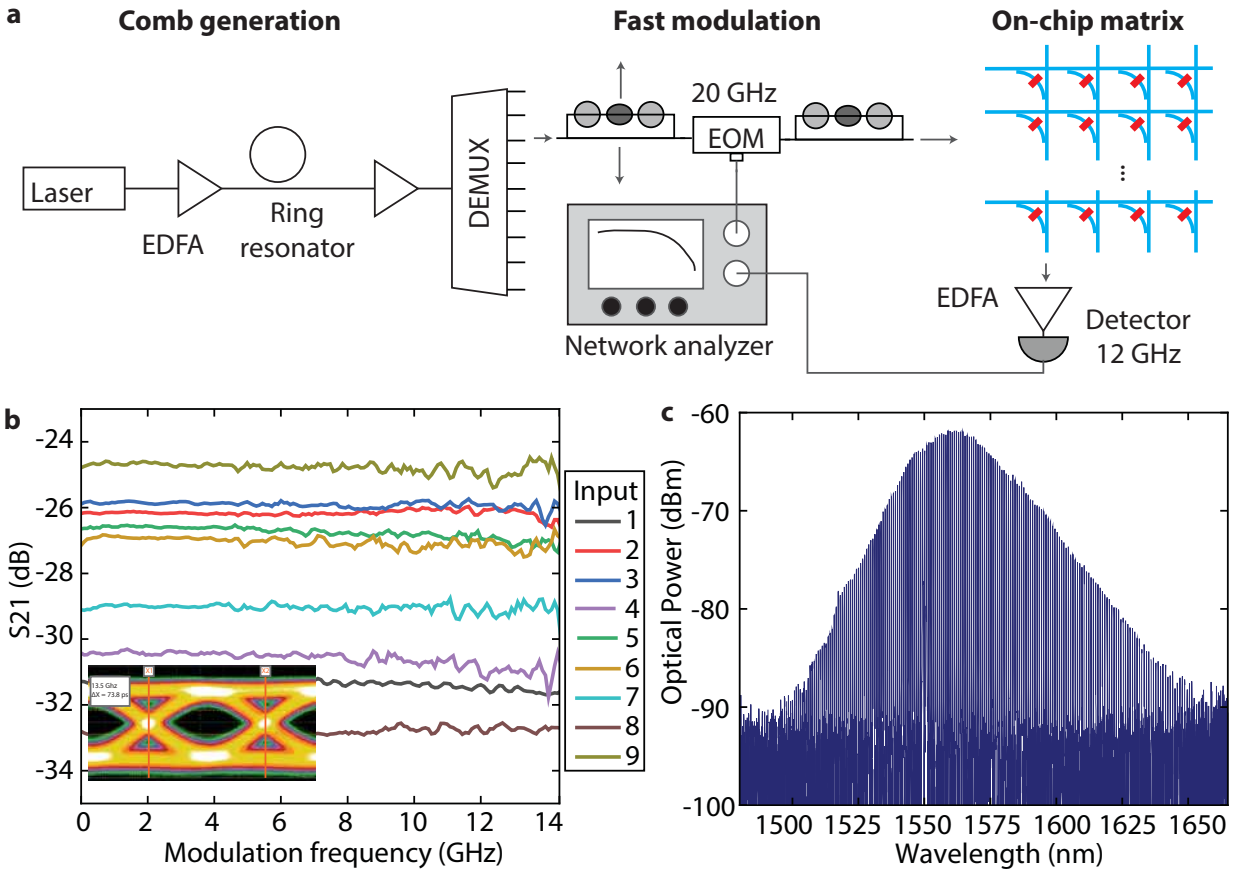

Figure 7.16.: Fast modulation. **a)** Setup used to modulate the input vectors with up to 14 GHz. The lines of the frequency comb are individually modulated with an EOM and a network analyzer and sent to the matrix input. A photodetector with 12 GHz bandwidth feeds the electronic signal back to the network analyzer. **b)** Flat frequency response for the nine inputs of the matrix up to 14 GHz. The inset shows an eye-diagram at 13.5 GHz. **c)** Frequency comb measured after transmission through the device. Using the broadband couplers as inputs, a frequency range of almost 150 nm can be exploited.

off-chip multiplexers limiting the number of wavelengths that could be used and the bandwidth of detectors and modulators.

When estimating the total number of wavelengths available for processing, two main factors have to be considered. On the one hand the number of comb lines the frequency comb can generate and on the other hand the wavelength dependent parts of the photonic matrix, which are in this case the directional couplers and the input and output couplers used to couple the light on and off the chip. It is especially important to analyse the wavelength dependence of the directional couplers because they define how well the equal splitting of light to the individual matrix elements can be achieved in a certain spectral range. As the normalization factor used to replicate the result of an inner product is an absolute value that usually depends on the matrix element with the lowest transmission, a good signal to noise ratio for all matrix elements can only be guaranteed for relatively small differences between the transmission of the matrix elements for all parallel wavelengths. In Sec. 7.2.2 it was concluded that a spectral range of at least 120 nm can be exploited for the input vector generation and even more is possible with dedicated coupler designs.

Fig. 7.16c) in addition shows the optical spectrum of the soliton microcomb after transmission through the on-chip matrix and the total internal reflection couplers, indicating that a range of more than 150 nm is available for the matrix processing. With a spacing of 100 GHz this leads to more than 187 wavelengths.

Considering current state of the art photonics with modulation and detection speeds up to 100 GHz the photonic tensor core therefore holds promise for upscaling to the PetaMAC/s ($= 1 \times 10^{15}$ MAC/s) range within a single matrix, when further decreasing the spacing of the comb lines to 50 GHz and moderately scaling the matrix to 50×50 elements surpassing any state-of-the-art electronic matrix multiplier. The matrix size is mainly limited by the optical loss inside the matrix due to the fan-out, waveguide crossings and directional couplers. By decreasing these loss factors, for example by using multilayer platforms with negligible waveguide crossing loss [90, 92], matrix sizes of up to 1000×1000 holding a million parameters on a single chip can be envisioned. If additionally using several incoherent frequency combs and encoding a complete vector on a single wavelength, the number of vectors to be processed in parallel is solely determined by the number of comb lines available in the spectral range (i.e 187 as previously estimated), pointing towards ExaMAC/s ($= 1 \times 10^{18}$ MAC/s) operation on a single chip.

7.4.7. Comparison of the computational power with state-of-the-art hardware accelerators

When comparing different approaches of hardware accelerators, suitable metrics have to be defined first. Two important figures of merit in this case are the compute density [174] and the energy

needed per MAC operation. The compute density is defined as

$$\text{compute density} = \frac{\text{speed (MACs/s)}}{\text{area per MAC unit (mm}^2\text{)}}. \tag{7.8}$$

The size of the unit cell of the photonic matrix in this work is $285\,\mu\text{m} \times 354\,\mu\text{m}$, yielding a compute density of $14\,\text{GHz} \cdot 4/(285\,\mu\text{m} \times 354\,\mu\text{m}) = 0.55\,\text{TMAC}/(\text{s} \cdot \text{mm}^2)$ when considering a 14 GHz modulation speed and multiplexing of four vectors. However, it should be noted that the unit cell can be considerably reduced in two ways. First, the implemented photonic matrix exploits additional grating couplers in each cell for optically switching the PCM (see Sec. 7.4.1), which can be replaced by electronic microheaters [171] reducing the cell size significantly. Second, the minimum bend radius of the silicon nitride waveguides ($> 30\,\mu\text{m}$) prevents further downscaling of the matrix unit cell. By employing the higher index platform silicon-on-insulator, state of the art in photonic foundries with bend radii down to 5 µm, the cell size can potentially be reduced down to $30 \times 30\,\mu\text{m}^2$. The size is now defined by the maximum length of the directional couplers and the waveguide crossings inside a single matrix unit. The longest directional coupler needed achieves a 50% splitting of the light to the cross-port and can be build in SOI with approximately 10 µm length for a 400 nm strip waveguide design and a 100 nm gap [175]. The 100%[4] splitter in the presented photonic architecture that is needed to guide the remaining light to the last column of the matrix is used for simplicity reasons in the design but can be replaced by a simple waveguide bend transferring the full optical power to the last column of the matrix. A low loss crossing with less than 0.1 dB insertion loss in the telecom band can be as compact as $9 \times 9\,\mu\text{m}^2$ [143], justifying an overall cell size of the matrix of $30 \times 30\,\mu\text{m}^2$. Because this unit cell considers the longest directional coupler needed for the complete matrix, the overall average matrix size could further be decreased by not sticking to a fixed grid but reducing the sizes of cells with shorter directional couplers. Assuming a multiplexing scheme with 64 parallel vectors [179, 180] and a modulation and detection speed of 100 GHz [181, 182], this leads to a compute density of $100\,\text{GHz} \cdot 64/30\,\mu\text{m}/30\,\mu\text{m} = 7100\,\text{TMAC}/(\text{s} \cdot \text{mm}^2)$. Tab. 7.1 shows a comparison with state-of-the art digital and analogue hardware accelerators revealing a compute density that is more than four orders of magnitude above comparable architectures.

In order to calculate the energy consumption of the photonic matrix per MAC-operation the equation

$$E_{\text{MAC}} \geq \frac{1}{N^2} \frac{2h\nu}{\eta} 2^{2N_b} \tag{7.9}$$

based on [174] can be employed, estimating the energy needed to overcome shot noise assuming a fixed precision N_b. N is the size of the matrix, $h\nu = 1.28 \times 10^{-19}\,\text{J}$ the photon energy at a wavelength of 1550 nm and η the combined quantum efficiency of the detector and the optical loss in the system. The optical loss, which is the highest contribution to η, is estimated in the

[4] In fact, a 95% splitter is used to be able to align a fibre array for switching the PCM (see Sec. 7.4.1)

Technology	Clock speed	Compute density TMAC/(smm^2)	Energy pJ/MAC	Precision bit	Reference
Photonic matrix (SiN)	14 GHz	0.56	206	5	This work
Photonic matrix (SOI)	100 GHz	7100	1.6[a)] 0.025[a)]	8 5	Future projection
Photonic coherent DNN	100 GHz	5.5	0.038[b)]	3.5	[19]
Photonic hybrid laser NN	5 GHz	4.5	0.22[b)]	5.1	[176]
Haswell E5-2699 v3 (CPU)	2.3 GHz	0.0039	110[c)]	8	[45]
Google TPU (ASIC)	700 MHz	0.14	0.43[c)]	8	[45]
Nvidia Titan V (GPU)	1.2 GHz	0.015 0.0075 0.0038	20[c)] 40[c)] 80[c)]	16 32 64	[177]
TrueNorth	2.5 kHz	0.00051	0.27[c)]	5	[59]
Neurogrid	40.1 kHz	0.0056	119[c)]	13	[16]
PCM crossbar array (electronic)	1 MHz	3.16[d)]	0.212	4	[28]

Table 7.1.: Comparison between the photonic matrix and other approaches. a) Upper boundary for assuming lossless crossings and directional couplers. **b)** Efficiency per MAC based on the dominating power consumption of the heaters (10 mW [19] and 5 mW [176]). The precision of 3.5 bit is estimated based on the experiments demonstrated with a 4×4 matrix. **c)** Energy efficiency is estimated by dividing the wall-plug power by the MAC rate for each processor. **d)** Estimation based on a unit cell assuming a 1024×1024 crossbar with eight analogue to digital converters of 50 µm × 300 µm each. (Table adapted from [178]).

following. For the experimental setup used in this work the optical loss from the first multiplexer to the detector after passing through the matrix is measured to be −45 dB for a 9×4 matrix in single vector mode and −51 dB for the 4×4 matrix when multiplexing four input vectors. Additional −20 dB must be added to account for the conversion efficiency of the frequency comb. Considering the insertion loss of the individual components of the setup (16-port MUX: −5 dB; Polarization controller and VOA: −1 dB; GST: −3 dB; 2× coupling to/off chip: −9 dB) and the matrices (Matrix (9×4): −15 dB; Matrix (4×4): −12 dB) the measured values are well explained. The difference between the two setups arises from the additional multiplexers needed at the output of the matrix for the parallelized version. In particular the combination of the four wavelengths belonging to one vector is accomplished with a 4×1 splitter which accounts for additional loss of −6 dB. Considering the measured loss, the energy per MAC is about

$$E_{MAC} \geq \frac{1}{N^2} \frac{2h\nu}{\eta} 2^{2N_b} = \frac{1}{4^2} \frac{2 \cdot 1.28 \times 10^{-19}\,\mathrm{J}}{10^{(-7.1)}} 2^{2 \cdot\ 5} \approx 206\,\mathrm{pJ} \tag{7.10}$$

as shown in Tab. 7.1 for a precision of 5 bit and the demonstrated 4×4 matrix.

However, the optical losses can significantly be reduced especially by reducing the coupling losses to and from the chip. A coupling loss of below 1.5 dB per coupler is possible [113] and even less if integrating lasers and detectors directly on-chip gaining almost two orders of magnitude. Minimizing also the loss of the waveguide crossings and the directional couplers a future projection of a 64×64 sized matrix (corresponding to a loss of −36 dB) results in 0.025 pJ/MAC with five-bit precision and approximately 1.6 pJ/MAC for eight-bit precision. Although the projected energy metrics are highly competitive to the other approaches, it should be noted that the estimation only accounts for the optical power in the system and electronic control circuits are neglected.

Chapter 8.

8 Conclusions and Outlook

In this work, the potential of the phase-change photonics platform for all-optical signal processing was explored and especially applied to new and unconventional approaches to computing. Based on a single phase-change cell comprised of a waveguide with an evanescently coupled phase-change material as an active element to control the amplitude of light travelling down the waveguide, different routes towards scalable and CMOS compatible architectures for computing have been developed. Three different approaches to tackle the challenges of the modern information age have been formulated and experimentally implemented, all of them based on the principle of in-memory computing circumventing the von Neumann bottleneck present in conventional computers.

In the first part of the thesis a basic arithmetic unit capable of addition, subtraction, multiplication and division was investigated in remembrance of one of the oldest tools for arithmetic – an abacus. The all-optical version of the abacus is based on counting optical pulses in a phase-change material and enables operation directly in base ten or other arbitrary bases suitable for the given task. By operating with optical pulses of a picosecond length, very fast processing speeds and low power operation was first demonstrated in a single PCM-cell and then expanded to multiple cells in a waveguide-crossing array capable of automatically storing the carry-over to higher place values in different PCM-cells. Using a two-pulse switching scheme all elements in the crossing array could be individually addressed paving the way to an all-optical random-access memory with computing capabilities. Throughout this work, the arithmetic unit was operated manually but in future research the abacus could further be optimized to a fully functional standalone arithmetic unit as a building block for larger processors. Especially detecting if the crystalline level of a PCM-cell was reached and a carryover has to be performed would need to be automatized in a more advanced photonic circuit.

The second and third part of this thesis were focussed on solutions for new processors in the area of artificial intelligence and neural networks. In the second part an all-optical neural network was developed and experimentally validated on a simple pattern recognition task. The network architecture was designed to be scalable to many layers by optically separating the

individual layers and allowing to add additional power in every layer. Using ring resonators for multiplexing and demultiplexing optical signals in a distributor and collector structure, optical crosstalk between neurons is avoided. One layer of the network with four neurons consisting of more than 120 ring resonators and 60 synapses based on the basic phase-change cell was experimentally implemented, carrying out a full matrix multiplication and a subsequent non-linear activation function in single time step. The neurons of the proposed photonic design were trained in a supervised as well as in an unsupervised way allowing for high flexibility in the training process. Employing simulations, the potential of the architecture for larger neural networks was investigated and a simple application of language detection in a two-layer network demonstrated. In future research the all-optical neural network could be expanded to larger scales for example by moving to the industry compatible silicon-on-insulator platform and implementing also the light sources and modulators on the photonic chip.

The last experimental part of the thesis was based on developing and experimentally demonstrating a hardware accelerator for matrix multiplications, which are the computationally most expensive task in many machine-learning based systems. The principle of photonic matrix multiplication relies on a waveguide array without any resonant structures allowing to access a wide range of the optical spectrum for processing. By employing phase-change materials as non-volatile matrix elements, the photonic tensor core is capable of very high data throughput in the range of TMAC/s with a compute density way above state-of-the-art electronic or photonic counterparts. Exploiting wavelength division multiplexing and an optical frequency comb, a high degree of parallelization can be achieved. Further developing the matrix could include the use of many incoherent sources to generate a single vector on a single wavelength making use of a broad range of the optical spectrum and enabling a degree of parallelization not possible in electronic circuits. Combining the passive matrix elements with optical modulators and detectors with a bandwidth up to 100 GHz and integrating the light sources on the same chip holds promise to scaling to the PetaMAC/s and ExaMAC/s regime on a single chip in the future, outperforming any known hardware.

In this work, first steps towards scalable all-optical computation on the phase-change photonics platform especially for neuromorphic processors have been developed. Although all-optical processing has the potential to overcome the limitations of electronic circuits in terms of speed and energy efficiency, electro-optic implementations of hardware accelerators combining photonic circuits with conventional CMOS control as a step in between are a valuable goal. Advances in fabrication of photonic circuits will lead to more reliable systems exploiting the full potential of multiplexing on the photonic platform and removing the need to thermally tune individual elements of the circuits. Similar to multilayer integration of electronic circuits, multilayer photonic circuits are gaining more attention and can help solving the disadvantage in device footprint of photonic circuits in the future. Together with the very high modulation bandwidth accessible in photonics up to 100 GHz, co-integration of photonic hardware accelerators with conventional

electronics has the potential to revolutionize data processing in applications in the field of artificial intelligence. In combination with non-volatile phase-change materials that offer low power operation, photonic phase-change processors can also be envisioned to be applicable for edge computing directly on mobile devices.

Going one step further towards all-optical computation, electro-optic conversions between data transfer and computation could be removed leading to faster and more efficient processing, especially because a lot of the input data for cognitive tasks as for example camera data in autonomous driving or microscope images in medical diagnostic are originally present in the optical domain before they are converted to electrical signals using photosensors.

Appendix A.

A Appendix

A.1. Fabrication process

To fabricate the photonic devices throughout this thesis a three-step electron-beam lithography process is used consisting of the marker deposition, definition of photonic structures and the deposition of the phase-change material. The samples are fabricated from a silicon nitride wafer with a layerstack of 325 nm to 343 nm silicon nitride on 3.3 µm silicon oxide and 525 µm silicon.

Step one: alignment markers

1. Cleaning of the chip:
   a) Sonication in acetone for 5 min at high power and rinsing with acetone and isopropanol. Subsequently the sample is dry-blown with a nitrogen gun.
   b) Pre-baking at 200 °C for 5 min to remove adsorbates.
2. Spincoating PMMA 950k 4.5:
   a) Spincoating at 4000 rpm with an acceleration of 1000 rpm/s for 90 sec. The spin speed is adjusted in order to achieve a film-thickness of 250 nm.
   b) Baking at 180 °C for 2 min.
3. Electron-beam exposure with a dose of 1400 µC/cm^2 to define windows for the marker deposition.
4. Development in a mixture of MIBK and isopropanol (1 : 3) for 2 min. Isopropanol is used to stop the development process. After development the sample is dry-blown with nitrogen.
5. 7 nm of chromium and 70 nm of gold are deposited via physical vapour deposition. For the lift-off the sample is sonicated in acetone at lowest power for 5 min, then rinsed with acetone and isopropanol and dry-blown with nitrogen.

Step two: photonics

1. Cleaning of the chip:
   a) Sonication in acetone for 5 min at high power and rinsing with acetone and isopropanol. Subsequently the sample is dry-blown with a nitrogen gun.
   b) Pre-baking at 200 °C for 5 min to remove adsorbates.
2. Spincoating TI Prime (only neeeded if the main resist is ma-N)
   a) Spincoating at 3000 rpm with an acceleration of 1000 rpm/s for 23 sec. No residual drops of TI Prime should be visible after spincoating.
   b) Baking at 120 °C for 2 min with subsequent cool-down on a metal plate for a few seconds. The main resist should be spincoated immediately.

3i. Spincoating ma-N 2403
   a) Spincoating in a two-step process. First step: 400 rpm with an acceleration of 1000 rpm/s for 4 sec. Second step: 3400 rpm with an acceleration of 1000 rpm/s for 60 sec.
   b) Post-bake at 90 °C for 2 min. The film-thickness should be 340 nm.
   c) Electron-beam exposure with proximity effect correction and a dose of 480 µC/cm^2 to define the photonic structures.
   d) Development in MF-319 for 75 sec. Water is used to stop the development process. After development the sample is dry-blown with nitrogen.
   e) Baking at 100 °C for 2 min.

3ii. Spincoating AR-N 7520.12
   a) Spincoating in a two-step process. First step: 100 rpm with an acceleration of 1000 rpm/s for 5 sec. Second step: 2000 rpm with an acceleration of 1000 rpm/s for 60 sec.
   b) Baking at 85 °C for 1 min. The film-thickness should be 300 nm.
   c) Electron-beam exposure with proximity effect correction and a dose of 1400 µC/cm^2 to define the photonic structures.
   d) Development in MF-319 for 75 sec. Water is used to stop the development process. After development the sample is dry-blown with nitrogen.
   e) Bake at 85 °C for 1 min.

4. Reactive ion etching (Oxford 80 RIE): the silicon nitride is completely etched.
   a) Chamber cleaning.
   b) Silicon nitride etching: 55 sccm CHF_3, 5 sccm O_2, 55 mTorr, 125 W, $T = 20$ °C.

b) Resist removal: 50 sccm O_2, 50 mTorr, 100 W, $T = 20\,°C$ for 10 min.

Step three: PCM deposition

1. Cleaning of the chip:
   a) Rinsing the chip with acetone and isopropanol without sonication. Subsequently the sample is dry-blown with nitrogen.
   b) Pre-baking at 200 °C for 5 min to remove adsorbates.
2. Spincoating PMMA 950k 4.5
   a) Spincoating at 4000 rpm with an acceleration of 1000 rpm/s for 90 sec. The spin speed is adjusted in order to achieve a film-thickness of 250 nm.
   b) Baking at 180 °C for 2 min.
3. Electron-beam exposure with a dose of 1400 µC/cm^2 to define windows for the PCM deposition.
4. Development in a mixture of MIBK and isopropanol (1 : 3) for 2 min. Isopropanol is used to stop the development process. After development the sample is dry-blown with nitrogen.
4. PCM deposition: 10 nm of the phase-change material (GST or AIST) and 10 nm indium tin oxide are deposited using radio frequency sputtering with an argon plasma with 30 W (5 mtorr working pressure, 15 sccm).
5. Lift-off in acetone: sonicate for 5 min at lowest power. Rinse with acetone and isopropanol.

A.2. Setup for two-pulse switching

For the experiments demonstrating two-pulse switching on a PCM-crossing (see Sec. 5.4.1) the setup shown in Fig. A.1 is used. The setup is similar to the basic pump-probe setup explained in Sec. 3.1 with the difference that the pump light is generated from a mode-locked laser with a pulse length of one picosecond and repetition rate of 40 MHz. From the generated pulse-train individual pulses can be selected using a custom-made pulse picker. The pump pulse is subsequently split in two optical paths. In one of the paths the optical length can be varied using a variable optical delay line. This way it is possible to control the delay of the two pulses that arrive at the PCM on the intersection.

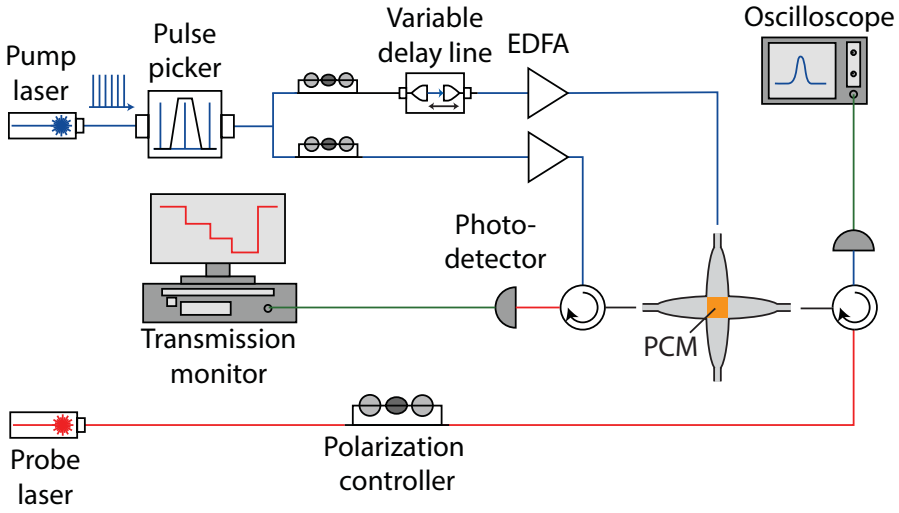

Figure A.1.: Setup for two-pulse switching The setup again consists of a pump- and a probe-path. The probe-path is used to monitor the state of the PCM with a simple transmission measurement. The pump-light is split in two parts. The optical path-length of one of the arms can be varied in order to adjust the delay between the two pulses arriving at the PCM on the crossing.

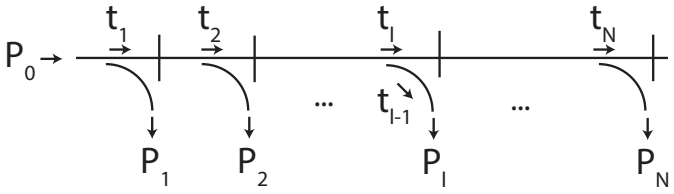

Figure A.2.: Derivation of the splitting ratios of the horizontal directional couplers

A.3. Derivation of the splitting ratios of the directional couplers in the photonic matrix

The splitting ratios of the horizontal couplers (see Fig. A.2) in the photonic matrix can be derived as follows. For a given input power P_0, through-port transmission of the directional couplers t_i and the loss factors l_{DC} for the directional couplers and l_{C} for the waveguide crossings the powers P_1 and P_2 arriving at the first two outputs can be calculated as

$$P_1 = P_0 \cdot (1 - t_1) \cdot l_{\mathrm{DC}}$$

and

$$P_2 = P_0 \cdot t_1 \cdot l_{\mathrm{C}} \cdot l_{\mathrm{DC}}^2 \cdot (1 - t_2).$$

As every output power is directly related to the previous directional couplers the term

$$P_l = P_0 \cdot l_{\mathrm{C}}^{l-1} \cdot l_{\mathrm{DC}}^{l} \cdot (1 - t_l) \cdot \prod_{i=1}^{l-1} t_i \tag{A.1}$$

can be derived for the lth output. As the aim of designing the directional couplers is to achieve equal splitting between all outputs, the condition

$$P_l = P_{l-1} \tag{A.2}$$

must be fulfilled. Plugging Equation (A.1) into Equation (A.2) leads to the expression

$$P_0 \cdot l_{\mathrm{C}}^{l-1} \cdot l_{\mathrm{DC}}^{l} \cdot (1 - t_l) \cdot \prod_{i=1}^{l-1} t_i = P_0 \cdot l_{\mathrm{C}}^{l-2} \cdot l_{\mathrm{DC}}^{l-1} \cdot (1 - t_{l-1}) \cdot \prod_{j=1}^{l-2} t_j$$

that can be simplified to

$$l_{\mathrm{c}} l_{\mathrm{DC}} (1 - t_l) t_{l-1} = 1 - t_{l-1}.$$

Solving for t_{l-1} leads to the result

$$t_{l-1} = \frac{1}{l_c l_{DC} \left(1 - t_l\right) + 1}$$

from which all transmission factors of the directional couplers can be derived given that the last transmission factor t_N is equal to zero because no light should get lost. A similar calculation reveals the transmission factors for the vertical couplers ((7.6)).

Bibliography

[1] R. Chandrasekar. "Elementary? Question answering, IBM's Watson, and the Jeopardy! challenge." *Resonance* 19 (3), 222–241, **2014**. DOI http://dx.doi.org/10.1007/s12045-014-0029-7.

[2] D. Ferrucci. "Build Watson: An Overview of DeepQA for The Jeopardy! Challenge." *2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT)* 4503, **2010**.

[3] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, *et al.*. "Watson: Beyond Jeopardy!" *Artificial Intelligence* 199-200, 93–105, **2013**. DOI http://dx.doi.org/10.1016/j.artint.2012.06.009.

[4] D. Ferrucci. "Introduction to "This is Watson"." *IBM Journal of Research and Development* 56 (3), 1–15, **2012**.

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, *et al.*. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (7587), 484–489, **2016**. DOI http://dx.doi.org/10.1038/nature16961.

[6] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, *et al.*. "Squeezing Deep Learning into Mobile and Emdedded Devices." *IEEE Pervasive Computing* 16 (3), 82–88, **2017**.

[7] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, *et al.*. "Artificial neural networks in medical diagnosis." *Journal of Applied Biomedicine* 11 (2), 47–58, **2013**. DOI http://dx.doi.org/10.2478/v10136-012-0031-x.

[8] R. A. Nawrocki, R. M. Voyles, and S. E. Shaheen. "A Mini Review of Neuromorphic Architectures and Implementations." *IEEE Transactions on Electron Devices* 63 (10), 3819–3829, **2016**. DOI http://dx.doi.org/10.1109/TED.2016.2598413.

[9] H. Wu, P. Yao, B. Gao, and H. Qian. "Multiplication on the edge." *Nature Electronics* 1 (1), 8–9, **2018**. DOI http://dx.doi.org/10.1038/s41928-017-0011-y.

[10] G. Batra, Z. Jacobson, S. Madhav, A. Queirolo, *et al.*. "Artificial-intelligence hardware: New opportunities for semiconductor companies." *Tech. Rep. December*, McKinsey&Company, **2018**. DOI https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies.

[11] M. M. Waldrop. "More than moore." *Nature* 530 (145), 145–147, **2016**. DOI http://dx.doi.org/10.3169/itej.70.324.

[12] G. E. Moore. "Cramming more components onto integrated circuits." *Electronics* 38, 114–117, **1965**. DOI http://dx.doi.org/10.1109/JPROC.1998.658762.

[13] G. E. Moore. "Progress In Digital Integrated Electronics." *IEDM Tech. Digest* , **1975**. DOI http://dx.doi.org/10.1109/N-SSC.2006.4804410.

[14] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, *et al.*. "Design of Ion-Implanted MOSFET's With Very Small Physical Dimensions." *IEEE Journal of Solid-State Circuits* , **1974**. DOI http://dx.doi.org/10.1109/JSSC.1974.1050511.

[15] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, *et al.*. "Principles of Neuromorphic Photonics." In *Encyclopedia of Complexity and Systems Science*, **2018**. DOI http://dx.doi.org/10.1007/978-3-642-27737-5_702-1.

[16] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, *et al.*. "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations." *Proceedings of the IEEE* , **2014**. DOI http://dx.doi.org/10.1109/JPROC.2014.2313565.

[17] C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, *et al.*. "Single-chip microprocessor that communicates directly using light." *Nature* 528 (7583), 534–538, **2015**. DOI http://dx.doi.org/10.1038/nature16454.

[18] R. Won and M. Paniccia. "Integrating silicon photonics." *Nature Photonics* 4, 498, **2010**. DOI http://dx.doi.org/10.1038/nphoton.2010.189.

[19] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, *et al.*. "Deep learning with coherent nanophotonic circuits." *Nature Photonics* 11 (June), 441–446, **2017**. DOI http://arxiv.org/abs/1610.02365.

[20] M. Wuttig, H. Bhaskaran, and T. Taubner. "Phase-change materials for non-volatile photonic applications.", **2017**. DOI http://dx.doi.org/10.1038/nphoton.2017.126.

[21] G. W. Burr, M. J. Brightsky, A. Sebastian, H.-Y. Cheng, *et al.*. "Recent Progress in Phase-Change Memory Technology." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6 (2), 146–162, **2016**. DOI http://dx.doi.org/10.1109/JETCAS.2016.2547718.

[22] S. Raoux, F. Xiong, M. Wuttig, and E. Pop. "Phase change materials and phase change memory." *MRS Bulletin* 39 (08), 703–710, **2014**. DOI http://dx.doi.org/10.1557/mrs.2014.139.

[23] M. Stegmaier. *On-Chip All-Optical Processing with Phase-Change Photonics*. Ph.d thesis, Westfälische Wilhelms-Universität Münster, **2016**.

[24] C. Rios, M. Stegmaier, Z. Cheng, N. Youngblood, *et al.*. "Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics." *Optical Materials Express* 8 (9), 2455, **2018**.

[25] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, *et al.*. "Integrated all-photonic non-volatile multi-level memory." *Nature Photonics* 9 (11), 725–732, **2015**. DOI http://dx.doi.org/10.1038/nphoton.2015.182.

[26] T. Ben-Nun and T. Hoefler. "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis." *ACM Computing Surveys* 52 (4), **2019**. DOI http://dx.doi.org/10.1145/3320060.

[27] F. L. Traversa and M. Di Ventra. "Universal Memcomputing Machines." *IEEE Transactions on Neural Networks and Learning Systems* 26 (11), 2702–2715, **2015**. DOI http://dx.doi.org/10.1109/TNNLS.2015.2391182.

[28] M. Le Gallo, A. Sebastian, R. Mathis, M. Manica, *et al.*. "Mixed-precision in-memory computing." *Nature Electronics* 1 (4), 246–253, **2018**. DOI http://dx.doi.org/10.1038/s41928-018-0054-8.

[29] S. Furber. "Bio-inspired massively-parallel computation." In *Advances in Parallel Computing*, vol. 27, 3–10, **2016**. ISBN 9781614996200. DOI http://dx.doi.org/10.3233/978-1-61499-621-7-3.

[30] S. Schmitt, J. Klahn, G. Bellec, A. Grubl, *et al.*. "Neuromorphic hardware in the loop: Training a deep spiking network on the BrainScaleS wafer-scale system." In *Proceedings of the International Joint Conference on Neural Networks*, 2227–2234, **2017**. ISBN 9781509061815.

[31] Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, *et al.*. "High-performance photonic reservoir computer based on a coherently driven passive cavity." *Optica* 2 (5), 438, **2015**. DOI http://dx.doi.org/10.1364/OPTICA.2.000438.

[32] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer. "Parallel photonic information processing at gigabyte per second data rates using transient states." *Nature Communications* 4, 1364–1367, **2013**. DOI http://dx.doi.org/10.1038/ncomms2368.

[33] T. Ferreira De Lima, B. J. Shastri, A. N. Tait, M. A. Nahmias, *et al.*. "Progress in neuromorphic photonics." *Nanophotonics* 6 (3), 577–599, **2017**. DOI http://dx.doi.org/10.1515/nanoph-2016-0139.

[34] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, *et al.*. "Experimental demonstration of reservoir computing on a silicon photonics chip." *Nature Communications* 5, 1–6, **2014**. DOI http://dx.doi.org/10.1038/ncomms4541.

[35] J. von Neumann. "John von Neumann Collected Works:." *Journal of the American Statistical Association* , **1964**. DOI http://dx.doi.org/10.2307/2283131.

[36] J. Backus. "Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs." *Communications of the ACM* 21 (8), 613–641, **1978**. DOI http://dx.doi.org/10.1145/359576.359579.

[37] A. Francillon and C. Castelluccia. "Code injection attacks on harvard-architecture devices." In *Proceedings of the ACM Conference on Computer and Communications Security*, **2008**. ISBN 9781595938107. DOI http://dx.doi.org/10.1145/1455770.1455775.

[38] S. Mehrgardt and M. Winterer. "Harvard architecture microprocessor with arithmetic operations and control tasks for data transfer handled simultaneously.", **1989**.

[39] Atmel. "8-bit Atmel Microcontroller with 128kBytes In-System Programmable Flash.", **2011**. DOI http://www.atmel.com/images/doc2467.pdf.

[40] M. Asghari and A. V. Krishnamoorthy. "Silicon Photonics: Energy-efficient communication." *Nature Photonics* 5 (5), 268–270, **2011**. DOI http://dx.doi.org/10.1038/nphoton.2011.68.

[41] A. Varma, B. Bowhill, J. Crop, C. Gough, *et al.*. "Power management in the Intel Xeon E5 v3." *Proceedings of the International Symposium on Low Power Electronics and Design* 2015-Septe, 371–376, **2015**. DOI http://dx.doi.org/10.1109/ISLPED.2015.7273542.

[42] E. Mora, J. Vieron, and P. Larbier. "Benefits of INTEL ® XEON ® scalable processors for video compression." DOI https://pdfs.semanticscholar.org/c451/a68402ac800572781ba53ac7a5d4e1188a69.pdf.

[43] K. Sato, C. Young, and D. Patterson. "An in-depth look at Google's first Tensor Processing Unit (TPU).", **2017**. DOI https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.

[44] A. Linn. "The moonshot that succeeded: How Bing and Azure are using an AI supercomputer in the cloud.", **2016**. DOI https://blogs.microsoft.com/ai/project{_}brainwave{_}catapult{_}moonshot/.

[45] N. P. Jouppi, C. Young, N. Patil, D. Patterson, *et al.*. "In-Datacenter Performance Analysis of a Tensor Processing Unit." *Proceedings of ISCA '17* , **2017**. DOI http://dx.doi.org/10.1145/3079856.3080212.

[46] P. W. Shor. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring." *Proceedings 35th Annual Symposium on Foundations of Computer Science* 124–134, **1994**. DOI http://dx.doi.org/10.1016/0024-3205(96)00287-1.

[47] F. Arute, K. Arya, R. Babbush, D. Bacon, *et al.*. "Quantum supremacy using a programmable superconducting processor." *Nature* 574, 505–510, **2019**. DOI http://dx.doi.org/10.1038/s41586-019-1666-5.

[48] C. D. Wright, P. Hosseini, and J. A. V. Diosdado. "Beyond von-Neumann Computing with Nanoscale Phase-Change Memory Devices." *Advanced Functional Materials* 23 (18), 2248–2254, **2013**. DOI http://dx.doi.org/10.1002/adfm.201202383.

[49] H.-S. P. Wong and S. Salahuddin. "Memory leads the way to better computing." *Nature Nanotechnology* 10 (March), 191–194, **2015**. DOI http://dx.doi.org/10.1038/nnano.2015.29.

[50] M. Di Ventra, Y. V. Pershin, and L. O. Chua. "Circuit Elements With Memory: Memristors, Memcapacitors, and Meminductors." *October* 97 (10), 1715–1716, **2009**.

[51] C. D. Wright, L. Wang, M. M. Aziz, J. a. V. Diosdado, *et al.*. "Phase-change processors, memristors and memflectors." *Physica Status Solidi (B)* 249 (10), 1978–1984, **2012**. DOI http://dx.doi.org/10.1002/pssb.201200378.

[52] Y. V. Pershin and M. Di Ventra. "Solving mazes with memristors: A massively parallel approach." *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 84 (4), 1–6, **2011**. DOI http://dx.doi.org/10.1103/PhysRevE.84.046703.

[53] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, *et al.*. "'Memristive' switches enable 'stateful' logic operations via material implication." *Nature* 464 (7290), 873–876, **2010**. DOI http://dx.doi.org/10.1038/nature08940.

[54] Y. V. Pershin and M. Di Ventra. "Neuromorphic, digital, and quantum computation with memory circuit elements." In *Proceedings of the IEEE*, **2012**. DOI http://dx.doi.org/10.1109/JPROC.2011.2166369.

[55] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, *et al.*. "Stochastic phase-change neurons." *Nature Nanotechnology* 11, 693–699, **2016**. DOI http://dx.doi.org/10.1038/nnano.2016.70.

[56] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H. P. Wong. "Materials for Brain-Inspired Computing." *Nano Letters* (12), 2179–2186, **2012**.

[57] H. Völker-Feldmann. *Endokrine Regulation der Androgen-induzierbaren 3[alpha]-Hydroxysteroid-Dehydrogenase in der Mikrosomenfraktion der Rattenniere: Einfluss der Applikation von Antiandrogenen auf die Androgenwirkung in Abhängigkeit von von Geschlecht und gonadalem Status*. Ph.D. thesis, Universität Gesamthochschule Essen, **1981**.

[58] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana. "The SpiNNaker project." *Proceedings of the IEEE* , **2014**. DOI http://dx.doi.org/10.1109/JPROC.2014.2304638.

[59] P. A. Merolla, J. V. Arthur, R. Alvarez-icaza, A. S. Cassidy, *et al.*. "A million spiking-neuron integrated circuit with a scalable communication network and interface." *Science* 345 (6197), 668–673, **2014**.

[60] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, *et al.*. "A wafer-scale neuromorphic hardware system for large-scale neural modeling." In *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, **2010**. ISBN 9781424453085. DOI http://dx.doi.org/10.1109/ISCAS.2010.5536970.

[61] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli. "Photonic Neural Networks: A Survey." *IEEE Access* 7, 175827–175841, **2019**. DOI http://dx.doi.org/10.1109/ACCESS.2019.2957245.

[62] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, *et al.*. "Recent advances in physical reservoir computing: A review.", **2019**. DOI http://dx.doi.org/10.1016/j.neunet.2019.03.005.

[63] G. Van Der Sande, D. Brunner, and M. C. Soriano. "Advances in photonic reservoir computing.", **2017**. DOI http://dx.doi.org/10.1515/nanoph-2016-0132.

[64] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, *et al.*. "Neuromorphic Silicon Photonic Networks using silicon photonic weight banks." *Scientific Reports* 7 (June), 7430, **2017**. DOI http://dx.doi.org/10.1038/s41598-017-07754-z.

[65] D. Querlioz, O. Bichler, A. F. Vincent, and C. Gamrat. "Bioinspired Programming of Memory Devices for Implementing an Inference Engine." *Proceedings of the IEEE* 103 (8), 1398–1416, **2015**. DOI http://dx.doi.org/10.1109/JPROC.2015.2437616.

[66] R. Dunne and N. Campbell. "On the pairing of the Softmax activation and cross-entropy penalty functions and the derivation of the Softmax activation function." *Proc. 8th Aust. Conf. on the Neural Networks* 1–5, **1997**. DOI http://dx.doi.org/10.1.1.49.6403.

[67] P. Ramachandran, B. Zoph, and Q. V. Le. "Swish: a self-gated activation function." *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings* , **2018**.

[68] P. Sibi, S. Allwyn Jones, and P. Siddarth. "Analysis of different activation functions using back propagation neural networks." *Journal of Theoretical and Applied Information Technology* 47 (3), 1344–1348, **2013**.

[69] J. Feng and S. Lu. "Performance Analysis of Various Activation Functions in Artificial Neural Networks." *Journal of Physics: Conference Series* 1237 (2), 111–122, **2019**. DOI http://dx.doi.org/10.1088/1742-6596/1237/2/022030.

[70] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, *et al.*. "Digital selection and analogue amplification coexist in a cortex- inspired silicon circuit." *Nature* 405 (6789), 947–951, **2000**. DOI http://dx.doi.org/10.1038/35016072.

[71] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 770–778, **2016**. DOI http://dx.doi.org/10.1109/CVPR.2016.90.

[72] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11), 2278–2323, **1998**. DOI http://dx.doi.org/10.1109/5.726791.

[73] G. W. Burr, R. M. Shelby, S. Sidler, C. Di Nolfo, *et al.*. "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element." *IEEE Transactions on Electron Devices* 62 (11), 3498–3507, **2015**. DOI http://dx.doi.org/10.1109/TED.2015.2439635.

[74] D. Whitley, T. Starkweather, and C. Bogart. "Genetic algorithms and neural networks: optimizing connections and connectivity." *Parallel Computing* 14 (3), 347–361, **1990**. DOI http://dx.doi.org/10.1016/0167-8191(90)90086-O.

[75] D. Amodei and D. Hernandez. "AI and compute.", **2018**.

[76] D. Hebb. *The organization of behaviour New York*. Wiley & Sons, New York, **1949**.

[77] S. R. Elliott. "Chalcogenide Phase-Change Materials: Past and Future." *International Journal of Applied Glass Science* 6 (1), 15–18, **2015**. DOI http://dx.doi.org/10.1111/ijag.12107.

[78] R. Soref, L. Fellow, and I. Paper. "The Past, Present, and Future of Silicon Photonics." *Selected Topics in Quantum Electronics, IEEE Journal of* 12 (6), 1678–1687, **2006**. DOI http://dx.doi.org/10.1109/jstqe.2006.883151.

[79] G. T. Reed. *Silicon Photonics: The State of the Art*. Wiley, **2008**. ISBN 9780470994535. DOI http://dx.doi.org/10.1002/9780470994535.

[80] M. J. Deen and P. K. Basu. *Silicon Photonics: Fundamentals and Devices*. Wiley, **2012**. ISBN 9781119945161. DOI http://dx.doi.org/10.1002/9781119945161.

[81] A. Alduino and M. Paniccia. "Interconnects: Wiring electronics with light." *Nature Photonics* 1 (3), 153–155, **2007**. DOI http://dx.doi.org/10.1038/nphoton.2007.17.

[82] D. A. Miller. "Optical interconnects to electronic chips." *Applied Optics* , **2010**. DOI http://dx.doi.org/10.1364/AO.49.000F59.

[83] L.-W. Luo, N. Ophir, C. P. Chen, L. H. Gabrielli, *et al.*. "WDM-compatible mode-division multiplexing on a silicon chip." *Nature Communications* 5 (1), 3069, **2014**. DOI http://dx.doi.org/10.1038/ncomms4069.

[84] R. Nagarajan, C. H. Joyner, R. P. Schneider, J. S. Bostak, *et al.*. "Large-scale photonic integrated circuits." *IEEE Journal on Selected Topics in Quantum Electronics* , **2005**. DOI http://dx.doi.org/10.1109/JSTQE.2004.841721.

[85] B. Jalali and S. Fathpour. "Silicon photonics.", **2006**. DOI http://dx.doi.org/10.1109/JLT.2006.885782.

[86] M. Luo and M. Wuttig. "The Dependence of Crystal Structure of Te-Based Phase-Change Materials on the Number of Valence Electrons." *Advanced Materials* 16 (5), 439–443, **2004**. DOI http://dx.doi.org/10.1002/adma.200306077.

[87] M. Born, E. Wolf, and E. Hecht. " Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light ." *Physics Today* , **2000**. DOI http://dx.doi.org/10.1063/1.1325200.

[88] D. K. Gramotnev and S. I. Bozhevolnyi. "Plasmonics beyond the diffraction limit." *Nature Photonics* 4 (2), 83–91, **2010**. DOI http://dx.doi.org/10.1038/nphoton.2009.282.

[89] H. Raether, G. Hohler, and E. A. Niekisch. "Surface Plasmons on Smooth and Rough Surfaces and on Gratings.", **1988**. DOI http://dx.doi.org/10.1007/BFb0048317.

[90] W. D. Sacher, J. C. Mikkelsen, Y. Huang, J. C. Mak, *et al.*. "Monolithically Integrated Multilayer Silicon Nitride-on-Silicon Waveguide Platforms for 3-D Photonic Circuits and Devices." *Proceedings of the IEEE* 106 (12), 2232–2245, **2018**. DOI http://dx.doi.org/10.1109/JPROC.2018.2860994.

[91] K. Shang, S. Pathak, B. Guan, G. Liu, *et al.*. "Low-loss compact multilayer silicon nitride platform for 3D photonic integrated circuits." *Optics Express* 23 (16), 21334, **2015**. DOI http://dx.doi.org/10.1364/oe.23.021334.

[92] W. D. Sacher, J. C. Mikkelsen, P. Dumais, J. Jiang, *et al.*. "Tri-layer silicon nitride-on-silicon photonic platform for ultra-low-loss crossings and interlayer transitions." *Optics Express* 25 (25), 30862, **2017**. DOI http://dx.doi.org/10.1364/oe.25.030862.

[93] B. E. A. Saleh and M. C. Teich. "Fundamentals of Photonics , 2nd Edition.", **2007**.

[94] M. Wuttig and N. Yamada. "Phase-change materials for rewriteable data storage." *Nature Materials* 6 (11), 824–832, **2007**. DOI http://dx.doi.org/10.1038/nmat2077.

[95] S. R. Ovshinsky. "Reversible electrical switching phenomena in disordered structures." *Physical Review Letters* 21 (20), 1450–1453, **1968**. DOI http://dx.doi.org/10.1103/PhysRevLett.21.1450.

[96] W. Welnic and M. Wuttig. "Phasenwechsel-Materialien als universale Speichermedien." *Physik in unserer Zeit* 40 (4), 189–195, **2009**. DOI http://dx.doi.org/10.1002/piuz.200801205.

[97] A. V. Kolobov, P. Fons, A. I. Frenkel, A. L. Ankudinov, *et al.*. "Understanding the phase-change mechanism of rewritable optical media." *Nature Materials* 3 (10), 703–708, **2004**. DOI http://dx.doi.org/10.1038/nmat1215.

[98] D. Lencer, M. Salinga, and M. Wuttig. "Design rules for phase-change materials in data storage applications." *Advanced Materials* 23 (18), 2030–2058, **2011**. DOI http://dx.doi.org/10.1002/adma.201004255.

[99] A. Athmanathan, M. Stanisavljevic, N. Papandreou, H. Pozidis, *et al.*. "Multilevel-Cell Phase-Change Memory: A Viable Technology." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6 (1), 87–100, **2016**. DOI http://dx.doi.org/10.1109/JETCAS.2016.2528598.

[100] D. Loke, T. H. Lee, W. J. Wang, L. P. Shi, *et al.*. "Breaking the Speed Limits of Phase-Change Memory." *Science* 336 (6088), 1566–1569, **2012**. DOI http://dx.doi.org/10.1126/science.1221561.

[101] M. A. Kats, D. Sharma, J. Lin, P. Genevet, *et al.*. "Ultra-thin perfect absorber employing a tunable phase change material." *Applied Physics Letters* , **2012**. DOI http://dx.doi.org/10.1063/1.4767646.

[102] M. A. Kats, R. Blanchard, P. Genevet, Z. Yang, *et al.*. "Thermal tuning of mid-infrared plasmonic antenna arrays using a phase change material." *Optics Letters* , **2013**. DOI http://dx.doi.org/10.1364/ol.38.000368.

[103] P. Guo, A. M. Sarangan, and I. Agha. "A review of germanium-antimony-telluride phase change materials for non-volatile memories and optical modulators.", **2019**. DOI http://dx.doi.org/10.3390/app9030530.

[104] N. Yamada, E. Ohno, N. Akahira, K. Nishiuchi, *et al.*. "High speed overwritable phase change optical disk material." *Japanese Journal of Applied Physics* 26, 61–66, **1987**. DOI http://dx.doi.org/10.7567/JJAPS.26S4.61.

[105] N. Yamada, E. Ohno, K. Nishiuchi, N. Akahira, *et al.*. "Rapid-phase transitions of GeTe-Sb2Te3 pseudobinary amorphous thin films for an optical disk memory." *Journal of Applied Physics* 69 (5), 2849, **1991**. DOI http://dx.doi.org/10.1063/1.348620.

[106] Y. C. Chen, C. T. Rettner, S. Raoux, G. W. Burr, *et al.*. "Ultra-thin phase-change bridge memory device using GeSb." *Technical Digest - International Electron Devices Meeting, IEDM* 10–13, **2006**. DOI http://dx.doi.org/10.1109/IEDM.2006.346910.

[107] M. Salinga, B. Kersting, I. Ronneberger, V. P. Jonnalagadda, *et al.*. "Monatomic phase change memory.", **2018**. DOI http://dx.doi.org/10.1038/s41563-018-0110-9.

[108] R. E. Simpson, P. Fons, A. V. Kolobov, T. Fukaya, *et al.*. "Interfacial phase-change memory." *Nature Nanotechnology* 6, 501–505, **2011**. DOI http://dx.doi.org/10.1038/nnano.2011.96.

[109] G. Lucovsky and R. M. White. "Effects of Resonance Bonding on the Properties of Crystalline and Amorphous Semiconductors." *Physics Review B* 8 (2), 660–667, **1973**. DOI http://dx.doi.org/10.1103/PhysRevB.8.660.

[110] F. Van Laere, T. Claes, J. Schrauwen, S. Scheerlinck, *et al.*. "Compact focusing grating couplers for silicon-on-insulator integrated circuits." *IEEE Photonics Technology Letters* 19 (23), 1919–1921, **2007**. DOI http://dx.doi.org/10.1109/LPT.2007.908762.

[111] D. Taillaert, W. Bogaerts, P. Bienstman, T. F. Krauss, *et al.*. "An Out-of-Plane Grating Coupler for Efficient Butt-Coupling Between Compact Planar Waveguides and Single-Mode Fibers." *IEEE Journal of Quantum Electronics* 38 (7), 949–955, **2002**.

[112] C. R. Doerr, L. Chen, Y. K. Chen, and L. L. Buhl. "Wide bandwidth silicon nitride grating coupler." *IEEE Photonics Technology Letters* 22 (19), 1461–1463, **2010**. DOI http://dx.doi.org/10.1109/LPT.2010.2062497.

[113] H. Gehring, A. Eich, C. Schuck, and W. H. P. Pernice. "Broadband out-of-plane coupling at visible wavelengths." *Optics Letters* 44 (20), 5089, **2019**. DOI http://dx.doi.org/10.1364/ol.44.005089.

[114] H. Gehring, M. Blaicher, W. Hartmann, P. Varytis, *et al.*. "Low-loss fiber-to-chip couplers with ultrawide optical bandwidth." *APL Photonics* 4 (1), 0–7, **2019**. DOI http://dx.doi.org/10.1063/1.5064401.

[115] M. Stegmaier, C. Ríos, H. Bhaskaran, and W. H. P. Pernice. "Thermo-optical Effect in Phase-Change Nanophotonics." *ACS Photonics* 3 (5), 828–835, **2016**. DOI http://dx.doi.org/10.1021/acsphotonics.6b00032.

[116] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, *et al.*. "Integrated all-photonic non-volatile multi-level memory." *Nature Photonics* 9 (11), 725–732, **2015**. DOI http://dx.doi.org/10.1038/nphoton.2015.182.

[117] S. Raoux, J. L. Jordan-Sweet, and A. J. Kellock. "Crystallization properties of ultrathin phase change films." *Journal of Applied Physics* 103 (11), 1–7, **2008**. DOI http://dx.doi.org/10.1063/1.2938076.

[118] M. Salinga, E. Carria, A. Kaldenbach, M. Bornhöfft, *et al.*. "Measurement of crystal growth velocity in a melt-quenched phase-change material." *Nature communications* 4, 2371, **2013**. DOI http://dx.doi.org/10.1038/ncomms3371.

[119] J. Feldmann, N. Youngblood, X. Li, C. D. Wright, *et al.*. "Integrated 256 Cell Photonic Phase-Change Memory with 512-Bit Capacity." *IEEE Journal of Selected Topics in Quantum Electronics* 26 (2), **2020**. DOI http://dx.doi.org/10.1109/JSTQE.2019.2956871.

[120] D. Ielmini and H. S. Wong. "In-memory computing with resistive switching devices." *Nature Electronics* 1 (6), 333–343, **2018**. DOI http://dx.doi.org/10.1038/s41928-018-0092-2.

[121] C. D. Wright, H. Bhaskaran, and W. H. Pernice. "Integrated phase-change photonic devices and systems." *MRS Bulletin* 44 (09), 721–727, **2019**. DOI http://dx.doi.org/10.1557/mrs.2019.203.

[122] X. Li, N. Youngblood, C. Ríos, Z. Cheng, *et al.*. "Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell." *Optica* 6 (1), 1, **2019**. DOI http://dx.doi.org/10.1364/optica.6.000001.

[123] J. Von Keitz, J. Feldmann, N. Gruhler, C. Ríos, *et al.*. "Reconfigurable Nanophotonic Cavities with Nonvolatile Response." *ACS Photonics* 5 (11), 4644–4649, **2018**. DOI http://dx.doi.org/10.1021/acsphotonics.8b01127.

[124] M. Stegmaier, C. Ríos, H. Bhaskaran, C. D. Wright, *et al.*. "Nonvolatile All-Optical 1 × 2 Switch for Chipscale Photonic Networks." *Advanced Optical Materials* 5 (1), 1600346, **2017**. DOI http://dx.doi.org/10.1002/adom.201600346.

[125] I. V. Karpov, M. Mitra, D. Kau, G. Spadini, *et al.*. "Fundamental drift of parameters in chalcogenide phase change memory." *Journal of Applied Physics* , **2007**. DOI http://dx.doi.org/10.1063/1.2825650.

[126] W. W. Koelmans, A. Sebastian, V. P. Jonnalagadda, D. Krebs, *et al.*. "Projected phase-change memory devices." *Nature communications* 6 (May), **2015**. DOI http://dx.doi.org/10.1038/ncomms9181.

[127] D. Ielmini, A. L. Lacaita, and D. Mantegazza. "Recovery and drift dynamics of resistance and threshold voltages in phase-change memories." *IEEE Transactions on Electron Devices* , **2007**. DOI http://dx.doi.org/10.1109/TED.2006.888752.

[128] N. Papandreou, H. Pozidis, T. Mittelholzer, G. F. Close, *et al.*. "Drift-tolerant multilevel phase-change memory." *2011 3rd IEEE International Memory Workshop, IMW 2011* , **2011**. DOI http://dx.doi.org/10.1109/IMW.2011.5873231.

[129] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, *et al.*. "Programming algorithms for multilevel phase-change memory." In *Proceedings - IEEE International Symposium on Circuits and Systems*, **2011**. ISBN 9781424494736. DOI http://dx.doi.org/10.1109/ISCAS.2011.5937569.

[130] A. Cabrini, S. Braga, A. Manetto, and G. Torelli. "Voltage-driven multilevel programming in phase change memories." In *Proceedings of the 2009 IEEE International Workshop on*

Memory Technology, Design, and Testing, MTDT 2009, **2009**. ISBN 9780769537979. DOI http://dx.doi.org/10.1109/MTDT.2009.11.

[131] J. Feldmann, M. Stegmaier, N. Gruhler, C. Riós, *et al.*. "Calculating with light using a chip-scale all-optical abacus." *Nature Communications* 8, **2017**. DOI http://dx.doi.org/10.1038/s41467-017-01506-3.

[132] M. J. Lee, C. B. Lee, D. Lee, S. R. Lee, *et al.*. "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5-x/TaO2-x bilayer structures." *Nature Materials* 10 (8), 625–630, **2011**. DOI http://dx.doi.org/10.1038/nmat3070.

[133] I. S. Kim, S. L. Cho, D. H. Im, E. H. Cho, *et al.*. "High performance PRAM cell scalable to sub-20nm technology with below 4F2 cell size, extendable to DRAM applications." *Digest of Technical Papers - Symposium on VLSI Technology* 203–204, **2010**.

[134] H. Gehring, M. Blaicher, W. Hartmann, and W. H. P. Pernice. "Python based open source design framework for integrated nanophotonic and superconducting circuitry with 2D-3D-hybrid integration." *OSA Continuum* 2 (11), 3091–3101, **2019**. DOI http://www.osapublishing.org/osac/abstract.cfm?URI=osac-2-11-3091.

[135] T. Ito and S. Okazaki. "Pushing the limits of lithography.", **2000**. DOI http://dx.doi.org/10.1038/35023233.

[136] H. J. Levinson. *Principles of Lithography*. SPIE, **2010**. DOI http://dx.doi.org/10.1117/3.865363.

[137] N. Gruhler, C. Benz, H. Jang, J.-H. Ahn, *et al.*. "High-quality Si_3N_4 circuits as a platform for graphene-based nanophotonic devices." *Optics Express* 21 (25), 31678, **2013**. DOI http://dx.doi.org/10.1364/oe.21.031678.

[138] C. D. Wright, Y. Liu, K. I. Kohary, M. M. Aziz, *et al.*. "Arithmetic and biologically-inspired computing using phase-change materials." *Advanced Materials* 23 (30), 3408–3413, **2011**. DOI http://dx.doi.org/10.1002/adma.201101060.

[139] M. Davis and G. Ifrah. "The Universal History of Computing: From the Abacus to the Quantum Computer." *The American Mathematical Monthly* 109 (6), 581, **2002**. DOI http://dx.doi.org/10.2307/2695463.

[140] Y. V. Pershin, L. K. Castelano, F. Hartmann, V. Lopez-Richard, *et al.*. "A Memristive Pascaline." *IEEE Transactions on Circuits and Systems II: Express Briefs* 63 (6), 558–562, **2016**. DOI http://dx.doi.org/10.1109/TCSII.2016.2530378.

[141] T. Fukazawa, T. Hirano, F. Ohno, and T. Baba. "Low Loss Intersection of Si Photonic Wire Waveguides." *Japanese Journal of Applied Physics* 43 (2), 646–647, **2004**. DOI http://dx.doi.org/10.1143/JJAP.43.646.

[142] A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, *et al.*. "Meep: A flexible free-software package for electromagnetic simulations by the FDTD method." *Computer Physics Communications* 181 (3), 687–702, **2010**. DOI http://dx.doi.org/10.1016/j.cpc.2009.11.008.

[143] Y. Ma, Y. Zhang, S. Yang, A. Novack, *et al.*. "Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect." *Optics Express* 21 (24), 29374, **2013**. DOI http://dx.doi.org/10.1364/oe.21.029374.

[144] Y. Liu, J. M. Shainline, X. Zeng, and M. a. Popović. "Ultra-low-loss CMOS-compatible waveguide crossing arrays based on multimode Bloch waves and imaginary coupling." *Optics letters* 39 (2), 335–8, **2014**. DOI http://dx.doi.org/10.1364/IPRSN.2013.IM1A.4.

[145] N. Youngblood, C. Ríos, E. Gemo, J. Feldmann, *et al.*. "Tunable Volatility of Ge2Sb2Te5 in Integrated Photonics." *Advanced Functional Materials* 29, **2019**. DOI http://dx.doi.org/10.1002/adfm.201807571.

[146] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, *et al.*. "All-optical spiking neurosynaptic networks with self-learning capabilities." *Nature* 569, 208–214, **2019**. DOI http://dx.doi.org/10.1038/s41586-019-1157-8.

[147] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer. "Efficient Processing of Deep Neural Networks: A Tutorial and Survey.", **2017**. DOI http://dx.doi.org/10.1109/JPROC.2017.2761740.

[148] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal. "Broadcast and Weight: An Integrated Network For Scalable Phtonic Spike Processing." *Journal of Lightwave Technology* 32 (21), 3427–3439, **2014**. DOI http://arxiv.org/abs/1407.2917.

[149] D. G. Rabus. *Integrated Ring Resonators*. Springer-Verlag Berlin Heidelberg, **2007**. DOI http://dx.doi.org/10.1007/978-3-540-68788-7.

[150] S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, *et al.*. "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses." *Frontiers in Neuroscience* 10 (MAR), 1–12, **2016**. DOI http://dx.doi.org/10.3389/fnins.2016.00056.

[151] K. A. Buchanan and J. R. Mellor. "The activity requirements for spike timing-dependent plasticity in the hippocampus." *Frontiers in Synaptic Neuroscience* 2 (JUN), 1–5, **2010**. DOI http://dx.doi.org/10.3389/fnsyn.2010.00011.

[152] G.-q. Bi and M.-m. Poo. "Synaptic Modification by correlated activity: Hebb ' s Postulate Revisited." *Annual review of neuroscience* 24, 139–166, **2001**. DOI http://dx.doi.org/10.1146/annurev.neuro.24.1.139.

[153] D. Goldhahn, T. Eckart, and U. Quasthoff. "Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages." In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, **2012**. ISBN 9782951740877.

[154] T. Feldmann and F. Zindler. ""Aus äußeren Gründen nicht verfügbar"? – German Anthologies of Flemish Literature Published During WWII." *Journal of Dutch Literature* 9 (2), 35–59, **2018**.

[155] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, *et al.*. "Going deeper with convolutions." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June, 1–9, **2015**. DOI http://dx.doi.org/10.1109/CVPR.2015.7298594.

[156] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* 1–14, **2015**.

[157] "ImageNet." DOI http://image-net.org/.

[158] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, *et al.*. "ImageNet: A large-scale hierarchical image database." In *ImageNet*, **2009**. DOI http://dx.doi.org/10.1109/cvprw.2009.5206848.

[159] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* , **2017**. DOI http://dx.doi.org/10.1145/3065386.

[160] C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, *et al.*. "In-memory computing on a photonic platform." *Science Advances* 5 (2), **2019**. DOI http://dx.doi.org/10.1126/sciadv.aau5759.

[161] Z. Lu, H. Yun, Y. Wang, Z. Chen, *et al.*. "Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control." *Optics Express* 23 (3), 3795, **2015**. DOI http://dx.doi.org/10.1364/oe.23.003795.

[162] T. Herr, M. L. Gorodetsky, and T. J. Kippenberg. "Dissipative Kerr Solitons in Optical Microresonators." *Nonlinear Optical Cavity Dynamics: From Microresonators to Fiber Lasers* 8083, 129–162, **2015**. DOI http://dx.doi.org/10.1002/9783527686476.ch6.

[163] E. Temprana, E. Myslivets, B. P. Kuo, L. Liu, *et al.*. "Overcoming Kerr-induced capacity limit in optical fiber transmission." *Science* , **2015**. DOI http://dx.doi.org/10.1126/science.aab1781.

[164] N. Picqué and T. W. Hänsch. "Frequency comb spectroscopy.", **2019**. DOI http://dx.doi.org/10.1038/s41566-018-0347-5.

[165] S. Randel, A. Kordts, W. Freude, M. Karpov, *et al.*. "Ultrafast optical ranging using microresonator soliton frequency combs." *Science* 359 (6378), 887–891, **2018**. DOI http://dx.doi.org/10.1126/science.aao3924.

[166] A. L. Gaeta, M. Lipson, and T. J. Kippenberg. "Photonic-chip-based frequency combs.", **2019**. DOI http://dx.doi.org/10.1038/s41566-019-0358-x.

[167] T. J. Kippenberg, R. Holzwarth, and S. A. Diddams. "Microresonator-based optical frequency combs." *Science* 332 (6029), 555–559, **2011**. DOI http://dx.doi.org/10.1126/science.1193968.

[168] P. Del'Haye, A. Schliesser, O. Arcizet, T. Wilken, *et al.*. "Optical frequency comb generation from a monolithic microresonator." *Nature* , **2007**. DOI http://dx.doi.org/10.1038/nature06401.

[169] T. Herr, V. Brasch, J. D. Jost, C. Y. Wang, *et al.*. "Temporal solitons in optical microresonators." *Nature Photonics* , **2014**. DOI http://dx.doi.org/10.1038/nphoton.2013.343.

[170] M. H. P. Pfeiffer, A. Kordts, V. Brasch, M. Zervas, *et al.*. "Photonic Damascene Process for Integrated High-Q Microresonator Based Nonlinear Photonics." *Optica* 3 (1), 1–6, **2016**. DOI http://dx.doi.org/10.1364/OPTICA.3.000020.

[171] J. Zheng, Z. Fang, C. Wu, S. Zhu, *et al.*. "Nonvolatile electrically reconfigurable integrated photonic switch." *arXiv* 1–30, **2019**.

[172] N. Farmakidis, N. Youngblood, X. Li, J. Tan, *et al.*. "Plasmonic nanogap enhanced phase change devices with dual electrical-optical functionality." *Science Advances* 5, 1–8, **2019**. DOI http://arxiv.org/abs/1811.07651.

[173] P. J. Grother and K. K. Hanaoka. "NIST Special Database 19 - Handprinted Forms and Characters Database." *Technical Report on Special Database 19* 1–30, **2016**. DOI http://dx.doi.org/10.18434/T4H01C.

[174] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, *et al.*. "Photonic Multiply-Accumulate Operations for Neural Networks." *IEEE Journal of Selected Topics in Quantum Electronics* , **2019**. DOI http://dx.doi.org/10.1109/jstqe.2019.2941485.

[175] L. Lu, J. Wu, T. Wang, and Y. Su. "Compact all-optical differential-equation solver based on silicon microring resonator." *Frontiers of Optoelectronics* , **2012**. DOI http://dx.doi.org/10.1007/s12200-012-0186-9.

[176] H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, *et al.*. "Neuromorphic Photonic Integrated Circuits." *IEEE Journal of Selected Topics in Quantum Electronics* 24 (6), 1–16, **2018**. DOI http://dx.doi.org/10.1109/JSTQE.2018.2840448.

[177] "Nvidia Titan." DOI https://www.nvidia.com/en-us/titan/titan-v/.

[178] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, *et al.*. "Parallel convolution processing using an integrated photonic tensor core." , **2020**. DOI http://arxiv.org/abs/2002.00281.

[179] K. Okamoto, K. Moriwaki, and S. Suzuki. "Fabrication of 64 × 64 arrayed-waveguide grating multiplexer on silicon." *Electronics Letters* 31 (3), 184–186, **1995**. DOI http://dx.doi.org/10.1049/el:19950133.

[180] J. Wang, S. Chen, and D. Dai. "Silicon hybrid demultiplexer with 64 channels for wavelength/mode-division multiplexed on-chip optical interconnects." *Optics Letters* 39 (24), 6993, **2014**. DOI http://dx.doi.org/10.1364/ol.39.006993.

[181] P. O. Weigel, J. Zhao, K. Fang, H. Al-Rubaye, *et al.*. "Bonded thin film lithium niobate modulator on a silicon photonics platform exceeding 100 GHz 3-dB electrical modulation bandwidth." *Optics Express* 26 (18), 23728, **2018**. DOI http://dx.doi.org/10.1364/oe.26.023728.

[182] L. Vivien, A. Polzer, D. Marris-Morini, J. Osmond, *et al.*. "Zero-bias 40Gbit/s germanium waveguide photodetector on silicon." *Optics Express* , **2012**. DOI http://dx.doi.org/10.1364/oe.20.001096.

Zusammenfassung in deutscher Sprache

Einleitung und Motivation

2011 schlug das von IBM entwickelte Computerprogramm *Watson* in der Quizshow *Jeopardy!* die beiden Rekordgewinner in drei Runden mit einem Entstand von $77.147 gegenüber $24.000 und $21.600. *Watson* war damit eine der ersten Demonstrationen von künstlicher Intelligenz (KI), die natürliche Sprache auf dem Niveau von Menschen analysieren, in sekundenschnelle Zusammenhänge zwischen Themengebieten und Doppeldeutigkeiten verstehen und die korrekten Antworten in riesigen Informationsmengen ausfindig machen konnte.

Spätestens mit dem viel beachteten Triumph von *Watson* gilt die künstliche Intelligenz als eine der Schlüsseltechnologien des 21. Jahrhunderts. Seit 2011 kamen weitere beeindruckende Erfolge hinzu, so zum Beispiel der Gewinn des Programms *AlphaGo* von Google DeepMind gegen den damaligen weltbesten Spieler im Brettspiel *Go* Lee Sedol. *AlphaGo* gewann 2016 mit 4:1 in einem Spiel, das, anders als etwa Schach, aufgrund seiner Komplexität mit heutigen Rechnern nicht vollständig vorausberechnet werden kann. *AlphaGo* setzte daher auf maschinelles Lernen und neuronale Netze, die sich in Spielen gegen sich selbst oder mit menschlichen Experten trainierten. Noch deutlicher zeigt sich die Entwicklung künstlicher intelligenter Computerprogramme daran, dass der Nachfolger von *AlphaGo*, *AlphaGo Zero*, seinen Vorgänger in einem Wettkampf über hundert Runden mit 100:0 besiegte.

Auch abseits medienwirksamer Experimente ist die künstliche Intelligenz in unserem alltäglichen Leben angekommen. Smartphones erkennen Gesichter und Sprache, Autos lernen selbstständig zu fahren und Internetsuchmaschinen unterscheiden zwischen relevanten und irrelevanten Informationen oder klassifizieren Bilder innerhalb von Millisekunden. In der medizinischen Diagnostik eröffnet die künstliche Intelligenz neue Möglichkeiten, beispielsweise in der Erkennung von Hautkrebs oder Herzrythmusstörungen. All diese Aufgaben verbindet, dass sie die Verarbeitung von riesigen Datenmengen erfordern, welche konventionelle Computer an ihre Grenzen bringt. Kognitive Aufgaben wie Sprach- und Gesichtserkennung auf modernen Smartphones werden immernoch häufig in *Cloud*-basierten Systemen analysiert. Bis 2025 ist zu erwarten, dass sich die Datenmengen, die durch die verschiedenen Anwendungen künstlicher Intelligenz verarbeitet werden, auf mehr als 800 Exabytes verzehnfachen. Datenmengen, die die Kapazitäten jedes heutigen Rechners überschreiten.

Obwohl die Miniaturisierung von elektronischen Schaltkreisen lange mit den steigenden Anforderungen mithalten konnte, verlangsamt sich das Moore'sche Gesetz, das die Verdopplung der

Anzahl der Transistoren auf einem Mikroprozessor in Abständen von zwei Jahren vorhersagt, stetig. Durch die Verkleinerung der elektronischen Bauelemente steigen die Tunnelströme und damit der Energiebedarf drastisch an. Um dieser Herausforderung gerecht zu werden, braucht es vollständig neue Ansätze für die Datenverarbeitung. Herkömmliche Computer sind zwar gut in genauen arithmetischen Berechnungen, bei kognitiven Aufgaben wie Sprach- und Mustererkennung werden sie jedoch durch menschliche Gehirne um mehrere Größenordnungen in Geschwindigkeit und Energieaufnahme geschlagen. Die Simulation eines Mausgehirns mit 2.5 Millionen Neuronen auf einem Computer dauert etwa 9000 mal länger und verbraucht mehr als die 40 000-fache Energie als das Gehirn einer realen Maus. Der Grund für den massiven Vorteil biologischer Gehirne ist ein vollständig anderer Aufbau. Konventionelle Computer basieren auf der von Neumann Architektur und verarbeiten Daten seriell, Befehl für Befehl. Zudem sind Prozessoreinheit und Daten- und Programmspeicher physisch voneinander getrennt, was einen ständigen Datentransfer zwischen diesen Einheiten nötig macht. Biologische Gehirne verarbeiten Daten parallel ohne Trennung von Speicher und Prozessor. Im Beispiel des Wettbewerbs zwischen *AlphaGo* und Lee Sedol gewinnt der Mensch deutlich im Bereich Leistungsaufnahme mit 77 000 W für das Computerprogramm gegenüber 20 W für das menschliche Gehirn. Von der Natur inspirierte Prozessoren, die die Funktion von Gehirnen nachahmen, stellen somit einen vielversprechenden Ansatz für neuartige Computersysteme dar. Zwar haben sich Standardprozessoren (CPUs) und Grafikprozessoren (GPUs) als Hardwarebeschleuniger für maschinelles Lernen in der letzten Zeit stark verbessert, allerdings sind sie nach wie vor durch die von Neumann Architektur limitiert. Neuere Forschung ist daher dazu übergegangen, spezielle (überwiegend elektronische) Hardware für das Berechnen kognitiver Aufgaben zu entwickeln.

In jüngster Zeit rücken in der KI-Forschung auch optische Ansätze in den Vordergrund. Während optische Glasfasernetze aufgrund ihrer hohen Bandbreite und geringen Verluste schon länger die Datenübertragung auf Langstrecken dominieren, wird die optische Signalverarbeitung zunehmend auch für die Kommunikation zwischen verschiedenen Computerchips und sogar innerhalb einzelner Chips genutzt. Die Berechnungen selbst werden aber zumeist elektrisch durchgeführt, was eine verlustbehaftete Umwandlung optischer Signale in elektrische bedingt. Rein optische Systeme bieten daher die Chance auf eine schnelle und effiziente Datenverarbeitung, die die Limitierungen herkömmlicher Elektronik überwinden kann.

Beispiele für rein optische Architekturen im Bereich der künstlichen Intelligenz wurden schon gezeigt. Allerdings weisen viele optische Schaltkreise durch den Einsatz von thermischen Heizern zum Justieren der einzelnen optischen Elemente eine hohe Leistungsaufnahme auf. Ein Weg diesen Nachteil zu umgehen bietet sich mit sogenannten Phasenwechselmaterialien (PWMs), die z.B. in der optischen Datenspeicherung mit DVDs oder Blu-Ray RE zur Anwendung kommen. PWMs ändern ihre optischen und elektrischen Eigenschaften um mehrere Größenordnung abhängig von ihrem Materialzustand, können reversibel zwischen dem amorphen und kristallinen Zustand umgeschaltet werden und behalten diesen ohne weitere Energiezufuhr bei Raumtempe-

ratur für mehrere Jahre bei.

In dieser Arbeit werden daher photonische integrierte Schaltkreise in Kombination mit Phasenwechselmaterialien für neuartige Prozessorarchitekturen insbesondere im Bereich der künstlichen Intelligenz entwickelt und untersucht. Aufbauend auf früheren Forschungsarbeiten, die Phasenwechselmaterialien mit einzelnen optischen Wellenleitern zur Modulation von Licht kombinieren, werden größere Systeme von optischen Schaltkreisen für unkonventionelle Prozessoren implementiert.

Durchführung und Ergebnisse

Optischer Abacus

Im ersten Schritt wird eine optische arithmetische Einheit zur Addition, Subtraktion, Multiplkation und Division vorgestellt, die das Potential des Rechnens direkt im Datenspeicher aufzeigt. Diese Einheit besteht aus einem Wellenleiter mit darauf aufgebrachtem PWM, an das das Licht, welches durch den Wellenleiter geleitet wird, evaneszent koppelt. Durch die unterschiedliche Absorption von Licht im kristallinen und amorphen Zustand des PWMs, kann die Amplitude des Lichts moduliert werden. Die Funktionsweise des optischen Bauteils basiert auf der eines Abacus. Ähnlich dem Verschieben der Kugeln auf den verschiedenen Stäben eines Abacus, werden mit Hilfe von optischen Pulsen unterschiedliche Kristallisationsgrade im PWM erreicht. Beginnend im amorphen Zustand des PWMs, können optische Pulse praktisch akkumuliert werden. Jeder optische Puls von gleicher Energie induziert einen Kristallisationsschritt im PWM und ändert damit die optische Transmission durch den Wellenleiter. Dadurch wird ein einfacher Zähler ermöglicht, der durch Kombination mehrerer Zellen zu einem optischen Abacus erweitert werden kann.

Auf diese Art können alle Standardrechenarten vollständig optisch und nichtflüchtig durchgeführt werden. Da die Schaltzeiten von Phasenwechselmaterialien unter einer Nanosekunde liegen können, sind Operationen im GHz-Bereich möglich. In dieser Arbeit werden zunnächst Beispiele für die Grundrechenarten experimentell mit einzelnen PWM-Elementen demonstriert und anschließend eine Wellenleiterkreuzungsstruktur entwickelt, die es ermöglicht, viele Zellen miteinander zu verbinden. Mittels einer zwei-Puls-Schalttechnik können alle PWM-Elemente, die sich auf den Wellenleiterkreuzungen befinden, individuell angesteuert werden. Der optische Abacus ist ein Beispiel für das Durchführen von Rechnungen im Speicher, da der Zustand des PWMs ohne äußere Energiezufuhr erhalten bleibt, was den Datentransfer zwischen Speicher und Prozessor überflüssig macht. Ein weiterer Vorteil gegenüber digitalen elektronischen Prozessoren, die binär arbeiten, ist die Möglichkeit beliebig viele unterschiedliche Kristallisations- und damit Transmissionslevel in ein PWM zu programmieren. Die arithmetischen Aufgaben in dieser Arbeit wurden direkt in Basis zehn durchgeführt, was zu einer höheren Rechendichte führt.

Optische neuronale Netze

Im zweiten Schritt wird ein voll-optisches, photonisches neuronales Netz entwickelt und an einer einfachen Mustererkennungsaufgabe getestet. Einzelne Neuronen nehmen Eingangsdaten über mehrere Synapsen (Gewichte) auf, gewichten diese und addieren sie auf. Die gewichtete Summe (Aktivierungsenergie) wird mit einer Schwellwertfunktion abgeglichen, die darüber entscheidet, ob das Neuron ein Ausgangssignal aussendet oder nicht. Diese Einzelbausteine werden nun zu großen Netzwerken, die aus einer Schichtstruktur aus Eingangsschicht, die die Schnittstelle zur realen Welt darstellt, einigen verborgenen Schichten, die die eigentliche Berechnung durchführen und einer Ausgabeschicht, die die Ergebnisse ausgibt, bestehen, verknüpft. Die so entstandene mathematische Struktur eignet sich sehr gut zur Bewältigung von kognitiven Aufgaben, wurde in den letzten Jahren immer weiter verbessert und erreicht heutzutage in der Bilderkennung die Genauigkeit von menschlichen Gehirnen.

Künstliche neuronale Netze werden üblicherweise auf herkömmlichen Computern simuliert, sind also durch die von Neumann Architektur beschränkt, die allerdings der hochparallelen Datenverabeitung in neuronalen Netzen entgegensteht. Für die Berechnung neuronaler Netze optimierte, direkte Hardwareimplementierungen können große Vorteile in Geschwindigkeit und Energieeffizienz bringen.

In dieser Arbeit wird zunächst ein einzelnes optisches Neuron gezeigt, dass die wesentlichen Funktionen Gewichtung, Addition und Aktivierung in einer photonischen Struktur abbildet. Das optische Neuron ist in der Lage nach einem Trainingsprozess einfache Pixelmuster bestehend aus vier Pixeln zu unterscheiden. Der Trainingsprozess kann sowohl überwacht als auch unüberwacht erfolgen, was ein großes Spektrum an Anwendungsmöglichkeiten zulässt. Beim überwachten Lernen sind Trainingsdatensätze aus Eingangsmustern und korrektem Ergebnis vorhanden, die genutzt werden, um die Gewichte des Neurons so zu optimieren, dass sie die gewünschte Funktion aufweisen. Sind solche Datensätze nicht vorhanden und Muster müssen aus einem unbekannten Datenstrom extrahiert werden, kann das selbstständige, unüberwachte Lernen eingesetzt werden. Hierbei adaptiert das Neuron seine Gewichte mittels einer festen Lernregel mit Hilfe einer Rückkopplung vom Ausgang zu den Synapsen selbst. Anschließend wird eine skalierbare photonische Architektur entwickelt, die es ermöglicht viele Neuronen in mehreren Schichten miteinenader zu verbinden. Um die Eingangssignale gleichmäßig auf alle Neuronen zu verteilen, werden eine Verteiler- und eine Sammlerstruktur implementiert, die auf dem Wellenlängenmultiplexverfahren mit Ringresonatoren basieren. Eine Schicht eines solchen photonischen neuronalen Netzwerks bestehend aus vier Neuronen mit sechzig Synapsen wird in dieser Arbeit zur Mustererkennung von kleinen Bildern eingesetzt.

Die vorgestellte Architektur ermöglicht eine ganze Schicht des Netzwerks in einem Zeitschritt zu bearbeiten. Dieser Schritt besteht im wesentlichen aus einer Matrixmultiplikation und der Anwendung der Aktivierungsfunktion. Da keine Wellenleiterkreuzungen in der Signalleitung eingesetzt

werden und die Wellenleiter der verschiedenen Schichten des Netzwerks optisch von einander isoliert sind, lassen sich tiefe Netzwerke realisieren, ohne Limitierungen durch Propagationsverluste und Signalvermischungen. In abschließenden Simulationen von größeren neuronalen Netzwerken, wird das Potential der Strukturen weiter beleuchtet und eine einfache Sprachenerkennung demonstriert.

Photonischer Hardwarebeschleuniger für Matrixmultiplikationen

Im letzten Teil der Arbeit wird ein photonischer Hardwarebeschleuniger für Matrixmultiplikationen, die eine Hauptlast in der Berechnung künstlicher neuronaler Netze ausmachen, beschrieben und experimentell demonstriert. Dieser beruht auf einer photonischen Matrix aus einem Wellenleiterkrezungsarray, in dem alle Phasenwechselmaterialien über Direktionalkoppler separat voneinander in einer einfachen optischen Transmissionsmessung ausgelesen werden können. Anhand von Faltungen zwischen Eingangsbildern und Filtermatrizen wird die Funktion der photonischen Matrixmultiplikation nachgewiesen. Diese Faltungen entsprechen der Hauptoperation in faltenden neuronalen Netzen, die die höchsten Klassifizierungsgenauigkeiten in Bilderekennungsaufgaben erreichen. Im Experiment wird eine Kantenerkennung in Bildern durchgeführt.

Da die photonische Matrixmultiplikation einer passiven optischen Transmissionsmessung entspricht, ist sie in der Geschwindigkeit nur durch die Bandbreite der Modulatoren und Detektoren begrenzt, die bis zu 100 GHz betragen kann. Ein weiterer Vorteil der Photonik gegenüber der Elektronik ist die Möglichkeit der Parallelisierung durch das Multiplexen. Indem mehrere Vektoren auf unterschiedliche Wellenlängen moduliert werden und gleichzeitig durch die photonische Matrix geschickt werden, lassen sich mehrere Vektoren gleichzeitig mit derselben Matrix multiplizieren. Im Versuch wird das Multiplexen von vier Vektoren in einer 4×4 Matrix demonstriert und in einem weiteren Experiment die Modulation der Vektoren mit bis zu 14 GHz gezeigt. Die photonische Matrix erreicht damit Rechengeschwindigkeiten von 4 TOP/s (Terra Operationen pro Sekunde) und kann potentiell bis in den Bereich von POP/s (PetaOP/s) und EOP/s (ExaOP/s) auf einem einzelnen Chip skaliert werden. Geschwindigkeiten und Rechendichten (Berechnungen pro Chipfläche), die von aktuellen elektronischen Hardwarebeschleunigern für Matrixmulitiplikationen unerreicht sind.

Ausblick

In dieser Arbeit werden drei unterschiedliche Ansätze für die optische Datenprozessierung verfolgt. Neben einer arithmetischen Einheit, die im Speicher und direkt in Basis zehn rechnet, werden ein voll-optisches künstliches Neuronales Netz und ein Hardwarebeschleuniger für Matrixmultiplikationen, der sich mit herkömmlicher Elektronik integrieren lässt, vorgestellt.

Während die Arithmetik im Speicher mit Phasenwechselmaterialien einen attraktiven Ansatz für die Umgehung des von Neumann Flaschenhalses darstellt, die Rechengenauigkeit und Ge-

schwindigkeit von herkömmlichen Computern aber bisher unerreicht bleibt, zeigt sich das volle Potential von photonischen Prozessoren in speziellen Hardwareimplementierungen für Anwendungen in der künstlichen Intelligenz. Durch den hohen Parallelisierungsgrad in photonischen Strukturen, der durch Multiplexverfahren gegeben ist, kombiniert mit der hohen Modulationsbandbreite, werden sich in Zukunft elektro-optische und auch voll-optische Prozessoren realisieren lassen, die die Datenraten von elektronischen Prozessoren um Größenordnungen in Geschwindigkeit und Energieeffizienz übertreffen. Mit der weiteren Optimierung der Fabrikation integrierter photonischer Schaltkreise in professionellen Fabrikationsstätten, kompatibel mit herkömmlicher Elektronik, werden in Zukunft immer genauere photonische Elemente gefertigt werden können, die den hohen spektralen Bandbreiten und den Anforderungen an die Wellenlängenmultiplexer gerecht werden. Durch die Integration von Phasenwechselmaterialien erhält man zudem ein aktives Bauelement, mit dem das in Wellenleitern propagierende Licht gezielt beeinflusst werden kann. Da PWMs ihren Zustand bei Raumtemperatur ohne weitere Energiezufuhr beibehalten, ermöglichen sie eine energieeffizente Möglichkeit zur Realisierung photonischer Prozessoren.

List of Publications

Publications related to this thesis

- **J. Feldmann**, M. Stegmaier, N. Gruhler, C. Ríos, H. Bhaskaran, C.D. Wright, W.H.P. Pernice, "Calculating with light using a chip-scale all-optical abacus", *Nat. Commun.*, vol. 8, 1-8 (2017).

- **J. Feldmann**, N. Youngblood, C.D. Wright, H. Bhaskaran, W.H.P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities", *Nature*, vol. 569, 208-214 (2019).

- **J. Feldmann**, W.H.P. Pernice, "Lichtschnelles Nervennetz", *Physik in unserer Zeit*, vol. 6,50, 282-288 (2019).

- **J. Feldmann**, N. Youngblood, X.Li, C.D. Wright, H. Bhaskaran, W.H.P. Pernice, "Integrated 256 cell photonic phase-change memory with 512-bit capacity", *IEEE J. Sel. Topics Quantum Electro*, vol. 26, 2, 208-214 (2020) (invited).

- **J. Feldmann**, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Le Gallo, X. Fu, A. Lukashchuk, A. Raja, J.u Liu, C.D. Wright, A. Sebastian, T. Kippenberg, W.H.P. Pernice, H. Bhaskaran, "Parallel convolution processing using an integrated photonic tensor core", in revision.

- **J. Feldmann**, W.H.P. Pernice, "Photonik mit Phasenwechselmaterialien für Anwendungen in der Künstlichen Intelligenz", in preparation at *Physik Journal*.

- **J. Feldmann**, W.H.P. Pernice, "Hybrid Phase-Change Nanophotonic Circuits", contribution to "21st Century Nanoscience – A Handbook", *Taylor & Francis Books, Inc.*, in press.

Further publications

- J. von Keitz, **J. Feldmann**, C. Ríos, N. Gruhler, C. Rios, C.D. Wright, H. Bhaskaran, W.H.P. Pernice, "Reconfigurable Nanophotonic Cavities with Nonvolatile Response", *ACS Photonics*, vol. 5, 4644-4649 (2018).

- N.Youngblood, C. Ríos, E. Gemo, **J. Feldmann**, Z. Cheng, A. Baldycheva, W.H.P. Pernice, C.D. Wright, H. Bhaskaran, "Tunable Volatility of $Ge_2Sb_2Te_5$ in Integrated Photonics", *Adv. Funct. Mater.*, vol. 29, 1807571 (2019).

Conference contributions

- **J. Feldmann**, M. Stegmaier, N. Gruhler, C. Ríos, H. Bhaskaran, C.D. Wright, W.H.P. Pernice, "All-optical signal processing using phase-change nanophotonics", *19th International Conference on Transparent Optical Networks (ICTON)*, 2017.
- **J. Feldmann**, M. Stegmaier, N. Gruhler, C. Ríos, H. Bhaskaran, C.D. Wright, W.H.P. Pernice, "Calculating with light – an all-optical abacus using phase-change materials", *Frühjahrstagung der DPG in Erlangen (SAMOP)*, 2018.
- **J. Feldmann**, N. Youngblood, C.D. Wright, H. Bhaskaran, W.H.P. Pernice, "All-optical neural networks with phase-change photonics", *EOS Topical Meeting (OμS'19)*, 2019.
- **J. Feldmann**, N. Youngblood, C.D. Wright, H. Bhaskaran, W.H.P. Pernice, "All-optical neural networks with phase-change photonics", *E-MRS Fall Meeting*, 2019.

Curriculum Vitae

Personal Data

Name	Feldmann, Johannes
Date of Birth	17th of February 1990
Place of Birth	Coesfeld, Germany
Nationality	German

Education

12/2015 - present	**Ph.D. in Physics**, *University of Münster*, Germany Supervisor: Prof. Dr. Wolfram Pernice, Institute of Physics Thesis: "Photonic non-von Neumann Processors"
10/2013 - 09/2015	**M.Sc. in Physics**, *University of Münster*, Germany Thesis: "Sputter deposition and characterization of lithium titante thin-film electrodes" Supervisor: Prof. Dr. Gerhard Wilde, Dr. Frank Berkemeier
10/2010 - 09/2013	**B.Sc. in Physics**, *University of Münster*, Germany Thesis: "Electronic conductivity of sputter-deposited Lithium Iron Phosphate thin-films" Supervisor: Prof. Dr. Guido Schmitz, Dr. Drank Berkemeier
2000 - 2009	**Abitur and Secondary School**, *Gymnasium Nepomucenum*, Coesfeld, Germany

Civilian Service

08/2009 - 04/2010	**Hospital logistics** *St. Vincenz-Hospital*, Coesfeld, Germany

Work Experience

05/2010 - 07/2010	**Hospital logistics**, *St. Vincenz-Hospital*, Coesfeld, Germany Continuation of civilian service
09/2013 - 11/2013	**Student Assistant**, *University of Münster*, Münster, Germany Programming and automatizing a measurement setup in the group of Dr. Frank Berkemeier
10/2014 - 02/2015	**Student Assistant**, *University of Münster*, Münster, Germany Tutor for "Physics for chemists"
04/2015 - 07/2015	**Student Assistant**, *University of Münster*, Münster, Germany Tutor for the lab course in "material physics"

Acknowledgments

At this point of my thesis i would like to thank all the people who helped and supported me during the past years. First of all, i am immensely grateful to my advisor Prof. Dr. Wolfram Pernice who gave me the opportunity to work on the exciting topic of optical processors and the freedom to develop and explore my own ideas. His door was always open and i am thankful for the many helpful and inspiring discussions about countless experimental challenges and his guidance and infinite optimism throughout my whole Ph.D-project. I also want to thank Prof. Dr. Gerhard Wilde for co-supervising my thesis.

Special thanks go to Dr. Matthias Stegmaier, Dr. Nico Gruhler and Dr. Anna Ovvyan, who gave me a warm welcome in Karlsruhe where i started my thesis two months before we moved the labs to Münster back in 2015. Matthias supervised my project in the first months and i am very grateful for his support and learned a lot from his immense expertise in photonics and phase-change materials in theoretical questions as well as in the experiments. Together with Carlos Ríos Ocampo he laid the experimental foundations my thesis is based on. His comments, suggestions and challenging questions on all aspects of the setup and phase-change materials were invaluable for the success of this work. I also thank Nico for all the fabrication work he has done for me and the whole group in the times when the cleanroom facilities in Münster were not yet established. Besides many fruitful discussions about photonics and fabrication, we shared an office in Münster and had many interesting and cheerful chats besides work about life and especially football. Similarly, i would like to thank Anna for sharing her nanofabrication expertise and the fruitful discussions.

I am grateful to Dr. Simone Ferrari and Nicolai Walter, who initially organized the labs in Münster and spent a lot of their time setting up all the machines needed for the fabrication. Together with Wladick Hartmann and Fabian Beutel we made up the cleanroom management team in the transition period before Dr. Banafsheh Abasahl took over the management of the Münster Nanofabrication Facility. I would like to thank Wladick, Fabian, Nicolai and Simone for all the time they spent on taking care of countless issues in and around the cleanrooms. Besides work and many scientific discussions they were always supportive and willing to help, no matter how much work had piled up for themselves already. Special thanks to Fabian for his invaluable support with the e-beam system and helping me hold submission deadlines!

I would like to thank Corinna Kaspar for the many (often frustrating) hours we spent in the lab and for discussing work and non-work related topics and especially all the effort she put into

repairing the always broken RIE-System enabling the fabrication of some of the most useful chips of my thesis when the machine was finally running for a day!

I am thankful to Helge Gehring for printing his couplers for one of my last experiments and appreciate very much all the effort he put into building, maintaining and advancing the python libraries for designing chips and controlling our setups in the labs. Many thanks to Maik Stappers, who took care of most of the installation and preparation works for the etching machines in the SoN cleanroom and especially the maintenance of the ICP in the past two years.

I would like to thank Frank Brückerhoff-Plückelmann, Corinna Kaspar and Fabian Beutel for proofreading this thesis. Special thanks go to Frank for taking over my projects and help with the fabrication of new chips!

I would like to express my gratitude to our collaborators in the group of Prof. Harish Bhaskaran at the university of Oxford for their support in sputtering the phase-change materials on my samples. Especially, I would like to mention Nathan Youngblood (assistant professor at the University of Pittsburgh) and Xuan Li, who i both worked together with closely and who joined me for some experiments in our lab in Münster. We had many fruitful discussions and shared ideas regarding phase-change photonic devices and new potential projects. Many thanks also to Maxim Karpov from the group of Prof. Tobias Kippenberg at EPFL in Lausanne, who provided the frequency comb for my final measurements and spent a week in Münster setting it up in the lab.

Thank you very much to all the group members of the AG Pernice and AG Schuck for the great atmosphere and working environment and of course the innumerable kicker matches! Many thanks also to everyone lending parts of his setup to me whenever i needed all the lasers available in CeNTech or any other piece of equipment.

Last but not least i would like to thank my family. My parents for always being supportive on the whole way from school to the university until writing of this thesis and for never stopping to believe in me, although i tried to convince them they shouldn't hundreds of times. Many thanks to my sister Theresia and my brother Heinrich for supporting and encouraging me not only during the time of my thesis. Finally, i would like to thank my wife Mari for giving me strength during many frustrating and stressfull days in the past four years, for spending weekends, nights (literally) and days in the lab with me, being patient, for supporting and encouraging me, all the great moments (and Sunday breakfasts!), being there whenever i need you and your unconditional support and willingness to listen to all the experimental issues throughout this work. I love you to bits!