Supplementary Material: Single-Shot Matrix-Matrix Photonic Processor based on Spatial-Spectral Hypermultiplexed Dispersion

Chao Luan[1*], Ronald Davis III[1], Zaijun Chen[2], Dirk Englund[1], and Ryan Hamerly[1,3]

[1] *Research Laboratory of Electronics, MIT, Cambridge, MA, 02139, USA*

[2] *Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA and*

[3] *NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, CA 94085, USA*

CONTENTS

I. DETAILED COMPARISON OF EXISTING OPTICAL NEURAL NETWORK ARCHITECTURES

Table I compares six existing ONN schemes in terms of their hardware (number of components), gross throughput [MAC/step], and normalized throughput (i.e. optical parallelism) [MAC/step/HW component]. The latter figure is most important since it is directly proportional to throughput density (the scaling factor being the component size). The PNP [1] and weight-bank [2] schemes both involve relatively small components (MZIs and rings), but require N^2/ them to do a single-step matrix-vector product. Therefore, the normalized parallelism is 1/MZI for the PNP and 1/ring for the weight-bank scheme. WDM is problematic in both cases: the PNP, which is based on MZIs, is highly sensitive to wavelength if directional couplers are used, leading in practice to a sub-nanometer window of operation (MMIs would improve this but suffer higher losses and back scattering); likewise, the weight bank only operates at designated wavelengths. In any case, wavelength multiplexing the PNP or weight-bank system would add N WDMs with a parallelism factor of N/WDM; this eats up a lot of chip area for passives alone. The Homodyne-ONN [3] was specifically optimized to maximize performance per unit area. In matrix-matrix form, it consists of $O(N)$ modulators, free-space fan-out optics, and $O(N^2)$ detectors and performs a matrix-matrix product in N steps. Only a single wavelength is used, so the modulators can be compact rings; even so, the parallelism factor is N/ring, a factor of N larger than the weight bank. The detectors only have a parallelism factor of unity; however, detectors can be made very small and tightly packed, i.e. a camera screen. The biggest challenge with this scheme is its use of coherence and cylindrical free-space optics, which make it uniquely sensitive to vibrations and aberrations. (Note that we could achieve N/det if we really wanted to by using WDM to compute the matrix-matrix product in a single step, but this requires $O(N^2)$ modulators and becomes a phase stabilization nightmare). NetCast [4] also scores reasonably well on the metric if only the client is considered, but this analysis is only valid in the edge computing case, which is its target application. Also, note that the modulator must be a broadband MZM, which takes up much more space than the HD-ONN modulators. Finally, the so-called optical tensor core [5] combines a PCM crossbar array with WDM to achieve very high compute density, at least within the crossbar. This allows it to perform a matrix-matrix product in a single time step, achieving a parallelism of N/PCM. Like the WDM variants of the other schemes, it also requires $O(N^2 + 2N)$ WDMs, which takes up a lot of area if on-chip. There is also a very fundamental scaling problem associated with the beam combiners used in the crossbar circuit. Since the combiners are not wavelength-selective, there is a net $O(N)$ optical power loss from the source to the detector; this waste of light will lead to unacceptably large optical powers that ruin the energy efficiency of the design. Moreover, the 95% of input power scattered as stray light may interfere with the actual signal. These problems did not arise in the $4 * 4$ demonstration in the paper, but they will inevitably make scaling this scheme difficult. As this comparison illustrates, most leading ONN designs do not harness the full power of photonics. The PNP, WB are natively single-wavelength and adding wavelength multiplexing introduces some fundamental challenges; as a result the parallelism of the components themselves is only $O(1)$. On the other extreme, the PCM-OTC has $O(N)$ parallelism, but is hampered by $O(N)$ loss in the beam combiners and the need for $O(N^2 + 2N)$ on-chip WDMs. The only design that can achieve high parallelism without fundamental scaling roadblocks is the HD-ONN; however, practical issues of coherence and cylindrical optic aberrations may ultimately limit its performance. This introduces our high parallelism dispersive beam routing optical neural network, a new ONN based on the chromatic dispersion of free-space diffraction gratings. Grating neural network could (1) achieve high component parallelism like the PCM-OTC and HD-WDM-ONN, while (2) not relying on optical interference

Comparison of ONN architectures.

Concept*	Hardware	MACs/step	MACs/step/HW	Caveats
PNP [1]	N^2 MZI	N^2 [1-step MV]	1/MZI	MZI errors, BW limits.
WB [2]	N^2 ring	N^2 [1-step MV]	1/ring	Ring stabilization.
HD-ONN [3, 6]	N ring, N^2 det	N^2 [N-step MM]	N/ring, 1/det	Coherence, cyl. optics.
NetCast [4]	1 MZI, 1 WDM	N [N-step MV]	N/MZI, N/WDM	Weight server.
PCM-OTC [5]	N^2 PCM, N^2+2N WDM†	N^3 [1-step MM]	N/PCM, N/WDM§	Combiner loss, WDMs.
PNP+WDM	N^2 MZI, N WDM	N^3 [1-step MM]	N/MZI, N^2/WDM	MZI BW, WDMs.
WB+WDM	N^2 ring, N WDM	N^3 [1-step MM]	N/ring, N^2/WDM	Ring variations, WDMs.
HD+WDM	N^2 ring, N^2 det	N^3 [1-step MM]	N/ring, N/det	Coherence nightmare.
This work	$2N^2$ MZI, 1 grating	N^3 [1-step MM]	N/MZI, N^3/grating	Dispersive grating parallelism, crosstalk.

Table I: Comparison of existing ONN architectures, variants of these architectures, and this work. The PNP, WB, HD-ONN, NetCast architectures have been experimentally demonstrated but exhibit low parallelism. Proposed enhanced variants, including PNP+WDM, WB+WDM, and HD+WDM, theoretically support high-parallelism operations but face practical limitations in large-scale feasibility and have not been experimentally demonstrated. The PCM-OTC architecture has experimentally demonstrated with high parallelism; however, it encounters serious scaling challenges (wavelength count, WDM count, and crossbar fan-in) for large-scale implementation. In contrast, our proposed architecture achieves high parallelism without being restricted by such scaling limitations. *PNP: Programmable nanophotonic processor. WB: Weight bank. HD-ONN: Homodyne-ONN. PCM-OTC: Phase-change memory optical tensor core. †The system needs two types of WDM, the first type WDM has N^2 channels, PCM-OTC architecture needs N of them, the second type WDM has N channels, and PCM-OTC architecture needs N^2+N of them. §Only consider the N channel WDMs.

or cylindrical imaging like the HD-WDM-ONN or suffering from the fan-in losses of the PCM-OTC crossbar. The grating neural network realizing single-shot hardware-efficient large-scale matrix-matrix multiplication in a relatively robust optical setup.

II. PERFORM MULTIPLICATION AND ACCUMULATION USING OPTICS

A. Subsection 1: E/O modulator characterization

Commercial $LiNbO_3$ intensity modulators are employed to encode the electrical signal into the optical carrier, we characterize the modulator bandwidth over a wide wavelength range, the measured bandwidth is about 20 GHz and shows high consistency over different wavelengths. As shown in Fig. 1.

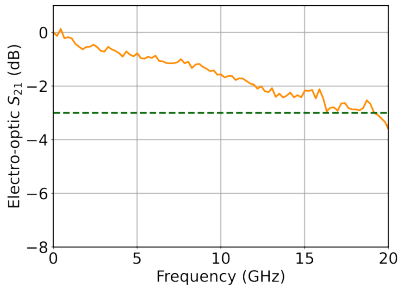

Figure 1: Bandwidth of the modulator. The figure illustrates the frequency response of the modulator, demonstrating E/O response up to 20 GHz.

B. Subsection 2: Static and Dynamic modulator transfer function

The output electric field of a non-chirped single-drive Mach-Zehnder Modulator (MZM) could be written as:

$$E_{\text{out}}(t) = \gamma\, E_{\text{in}}(t) \sin\left(\frac{\pi}{2V_\pi}\big(V(t) + V_{\text{bias}}\big)\right) \tag{1}$$

The corresponding optical output power is:

$$P_{\text{out}}(t) = \gamma^2 P_{\text{in}} \frac{1 - \cos\left(\frac{\pi}{V_\pi}\big(V(t) + V_{\text{bias}}\big)\right)}{2} \tag{2}$$

where:

- $E_{\text{out}}(t)$: Output electric field.
- γ: Optical amplitude scaling factor.
- P_{in}: Input optical power ($P_{\text{in}} = |E_{\text{in}}(t)|^2$).
- V_{bias}: DC bias voltage.
- $P_{\text{out}}(t)$: Optical output power.
- $E_{\text{in}}(t)$: Input electric field.
- $V(t)$: Applied voltage signal.
- V_π: Voltage required for a π-phase shift.

We modeling the modulator power transfer function using two methods, the first method is performing D.C. voltage sweep over a wide voltage range and measuring the output power using a photodetector and an oscillascope, another method is setting the D.C. bias voltage at the power transfer function quadrature point and then apply random AWG RF voltage within the AWG V_{pp} to the modulator and measure the output power, after the measurement, we fit the power transfer function using eq. (2), the fitted power transfer function obtained using these two methods matched well, as shown in Fig. 2.

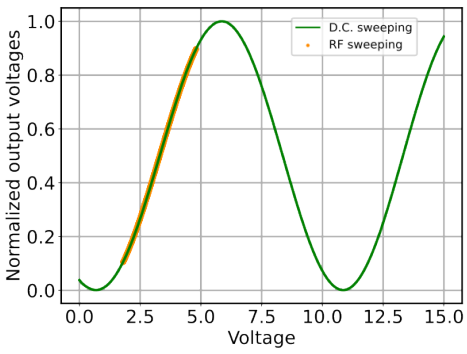

Figure 2: Power transfer function of the modulator. The green line illustrates the D.C. response of the modulator output power over a wide voltage range where the orange dot represents the response when sending random high-speed data from the AWG to the modulator and measuring with photodetector and oscilloscope, these two responses match well.

C. Subsection 3: Perform analog value multiplication using E/O modulators

We use the optical hardware to perform the analog value multiplication, in order to do this, we need to map the desired analog values to the E/O modulator output powers, and extract the corresponding modulator input voltages by solving each modulator's nonlinear transfer function individually, because all these 16 modulators used in the experiment have different transfer function fitting parameters, as illustrated in Fig. 3.

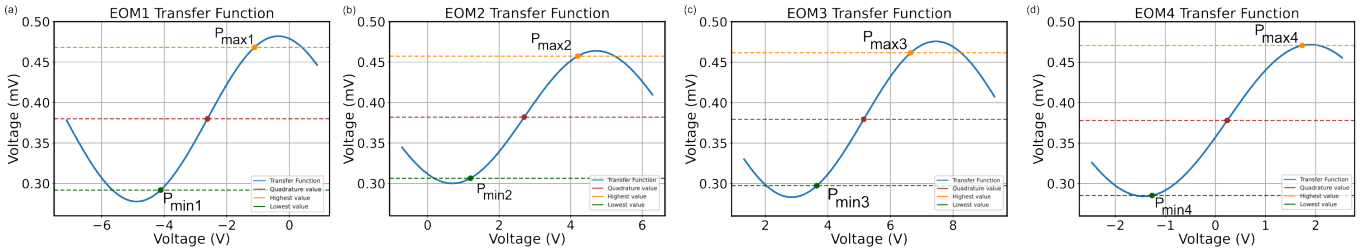

Figure 3: Power transfer function of different modulators. All these measured E/O modulator transfer functions are different and have different calibration parameters.

Here, we take an example of using four modulators to show how this process works. Fig. 3 shows the transfer functions of the four modulators from one weight modulator group. The highest and lowest output powers of these four modulators are P_{max1}, P_{max2}, P_{max3}, P_{max4}, and P_{min1}, P_{min2}, P_{min3}, P_{min4}, respectively. We need to make sure for all these modulators, the same floating numbers correspond to the same modulator output power values, so we select the highest value from P_{min1}, P_{min2}, P_{min3}, P_{min4}, which is P_{min2} to map the lowest floating number P_{min}, and lowest value from P_{max1}, P_{max2}, P_{max3}, P_{max4}, which is P_{max2} to map the maximum floating point value P_{max}. Once we mapped the minimum and maximum analog value numbers to P_{min} and P_{max}, any analog value numbers between [P_{min},P_{max}] could be mapped to the output power according to eq. (3)

$$P_{\text{out}} = (Analog - Analog_{\text{min}}) * \frac{P_{\text{max}} - P_{\text{min}}}{Analog_{\text{max}} - Analog_{\text{min}}} + P_{\text{min}} \tag{3}$$

where:

- P_{max}: Maximum photodetector output power.
- P_{min}: Minimum photodetector output power.
- $Analog_{\text{max}}$: Maximum analog value power.
- $Analog_{\text{min}}$: Minimum analog value power.
- $Analog$: Desired analog value.
- P_{out}: Desired analog value mapped output Power.

Now we have mapped the analog values to the photodetector output voltages. To extract the corresponding AWG input voltages, we calibrate the different E/O modulator transfer functions, which are described in eq. (2). From eq. (2), the extracted input voltage is

$$V(t) = \frac{V_\pi}{\pi} \arccos\left(1 - 2\frac{P_{\text{out}}(t)}{\gamma^2 P_{\text{in}}}\right) - V_{\text{bias}}. \tag{4}$$

Now we can map the desired analog value into the E/O modulator input voltage, all these E/O modulators have different transfer functions and output analog values so the input voltage extraction for each modulator needs to be performed individually. After the encoding, we can perform analog value multiplication and accumulation operations

by sending different AWG voltages into the modulators. In a DNN, the output of the analog value multiplication will serve as the input for the next layer, so we also need to perform the decoding from the photodetector output power to the analog value. For our architecture, in a single time step, 4 MACs are performed through wavelength and space multiplexing. So we use the fact that

$$0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 = 0 \tag{5}$$

$$1 * 1 + 1 * 1 + 1 * 1 + 1 * 1 = 4 \tag{6}$$

According to eq. (5) and eq. (6), we set all the E/O modulator output power to the largest analog value and the photodetector output power P_{maxpd}, which is the maximum case, corresponds to the analog value $4 * Analog_{\mathrm{max}}$; when setting all the E/O modulator output power to smallest analog value, the photodetector has the lowest output power P_{minpd} which corresponds to analog value $4 * Analog_{\mathrm{min}}$. This principle could also be used to calibrate the analog time integration, for the time integrator that integrates over l time steps, the maximum photodetector output power corresponds to the analog value $4 * l * Analog_{\mathrm{max}}$ and the minimum output power corresponds to the analog value $4 * l * Analog_{\mathrm{min}}$. Any output powers between the maximum output power and the minimum output power could be mapped to their corresponding analog value through

$$Analog_{\mathrm{out}} = (P_{\mathrm{out}} - P_{\mathrm{outmin}}) * \frac{Analog_{\mathrm{max}} - Analog_{\mathrm{min}}}{P_{\mathrm{outmax}} - P_{\mathrm{outmin}}} + Analog_{\mathrm{min}} \tag{7}$$

Using these above-mentioned encoding and decoding methods, we perform the Analog value multiplication using the optical hardware, and the optical computing experimental results match well with the theory value, showing high computing accuracy over 8 bits.

D. Subsection 4: Scaling to high sampling rate

We investigate the optical architecture computing bit precision under different data encoding speeds. The AWG we used supports multiple channel implementation, the highest sampling rate is $1\,\mathrm{GSa\,s^{-1}}$, and the highest analog bandwidth is 400 MHz. In order to keep a high bit precision, we down-sample the AWG encoding speed using pulse amplitude modulation. Fig. 4 shows AWG patterns with different sampling rates in the time and frequency domain, different sampling rate AWG signals have different frequency domain waveforms and components, we set the sampling rate to $50\,\mathrm{MSa\,s^{-1}}$, which is limited by the bandwidth (65 MHz) of the ADC used in the experiment. Further increase the sampling rate will lose the high-frequency components of the AWG pattern and cause reduced computing accuracy. Fig. 5 shows the pulse train of the calculated, measured signal traces and the ground truth at $50\,\mathrm{MSa\,s^{-1}}$ sampling rate. The experimental trace and theory trace match well, we extract the central value from the measured pattern at each value level and compare with the ground truth, the calculated error standard deviation is 0.003, which corresponds to over 8 bits computing precision. Fig. 6 shows the error at different sampling rates. Based on the trade-off of the bandwidth and accuracy, the sampling rate is kept at $50\,\mathrm{MSa\,s^{-1}}$.

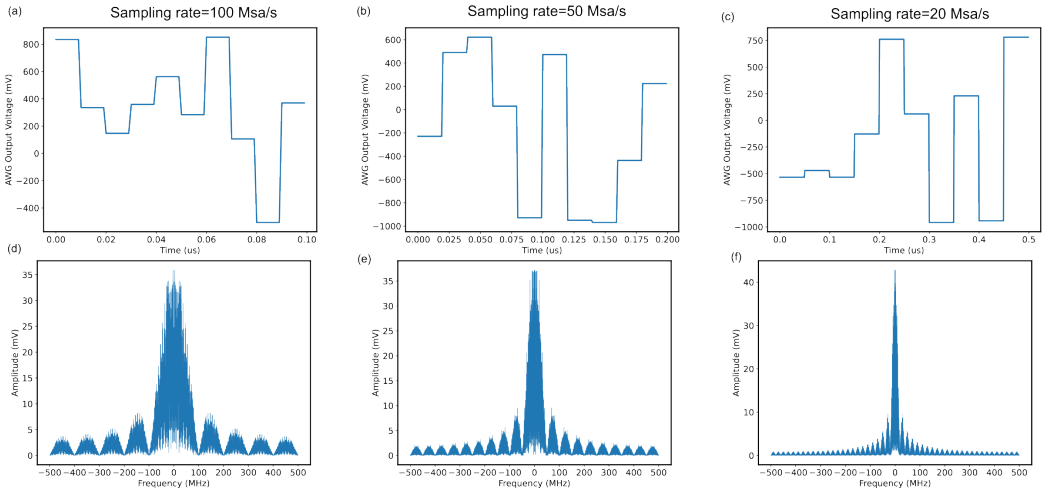

Figure 4: Perform pulse amplitude modulation of the AWG pattern, the top figure shows the pattern after the pulse amplitude modulation and the bottom figure shows the frequency information of these signals, high sampling rate modulation includes more high-frequency component data.

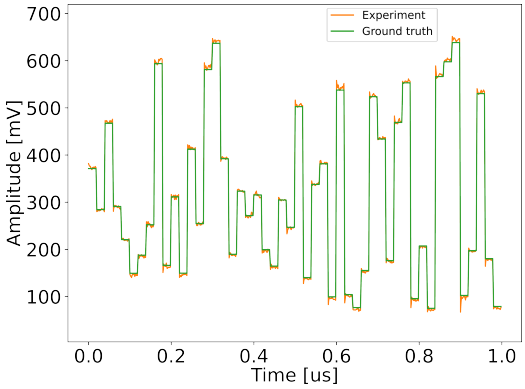

Figure 5: Pulse train of the calculated and measured signal at $50\,\mathrm{MSa\,s^{-1}}$.

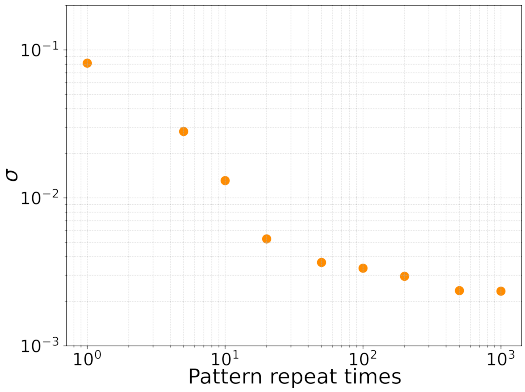

Figure 6: Experiment-theory error standard-deviation at different data sampling rates.

E. Subsection 5: Stabilizing quadrature point drift

$LiNbO_3$ modulator is known to be susceptible to quadrature-point drift, where the voltage corresponding to a fixed output power varies over time, which is attributed to fluctuations in charges accumulated in the phase shifters, this quadrature-point drift happens on the time scale of several seconds. We use a photodetector to provide feedback to D.C. voltage to stabilize the modulator quadrature-point output power. Fig. 7 shows the E/O modulator outputs when the D.C. bias stabilization is employed, the D.C. voltage is gradually decreased, which is used to compensate for the quadrature-point shift and the output power is very stable, thanks to the stable modulator output power and constant transfer function, the modulator calibration just needs to be performed once and the system shows high bit precision.

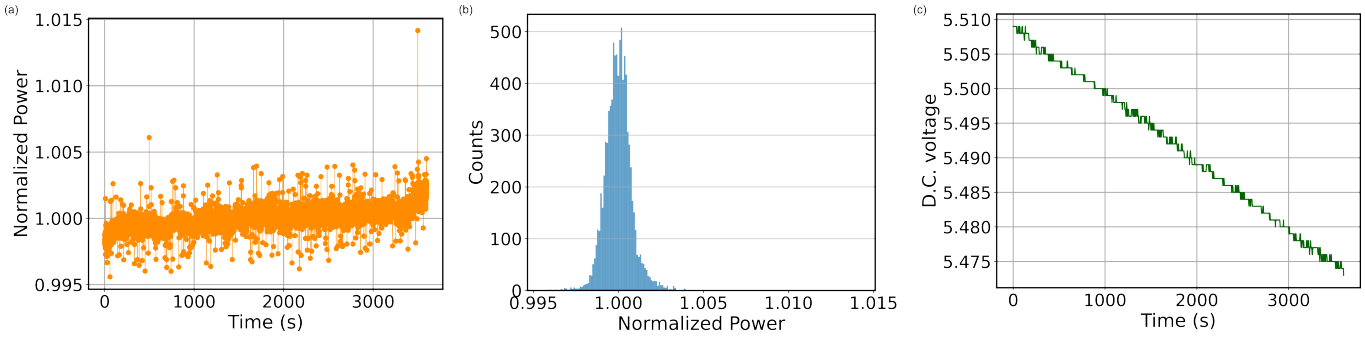

Figure 7: E/O modulator quadrature-point drift measurements. The output power is very stable when the D.C. bias voltage stabilization system is employed.

III. FREE SPACE GRATING BEAM ROUTING SYSTEM AS 3D PARALLEL WAVELENGTH MUX AND DEMUX

The $4 * 4$ grating beam routing system works as parallel, 64-channel input, and 64-channel output Mux and DeMux. We characterize the system performance when the grating beam routing system works as a Mux and DeMux, respectively.

A. Subsection 1: Grating beam routing system parallelism characterization

We characterize the hardware parallelism when sending multiple wavelengths from different channels. To make the results more evident, we let the light incident from different rows of the fiber array so the output distribution can easily be separated. Fig. 8 (a) shows the output image from the receiver fiber array when we send four different wavelengths into each channel, the wavelengths in each different channel are offsetetd with one WDM spacing. The output wavelengths are equally distributed among different channels, adjacent wavelength lights incident from adjacent spatial channels will emitted into the same column of the receiver fiber array. We further investigate the broadband operation of the grating architecture where seven different wavelengths are injected to the grating through the same channel and the output lights are equally distributed to seven different positions, these spots have different intensity because the input light intensity for different wavelengths are different. The grating architecture also has low crosstalk, we measure the crosstalk between different channels, and the measured result is about -50 dB, as indicated in Fig. 9

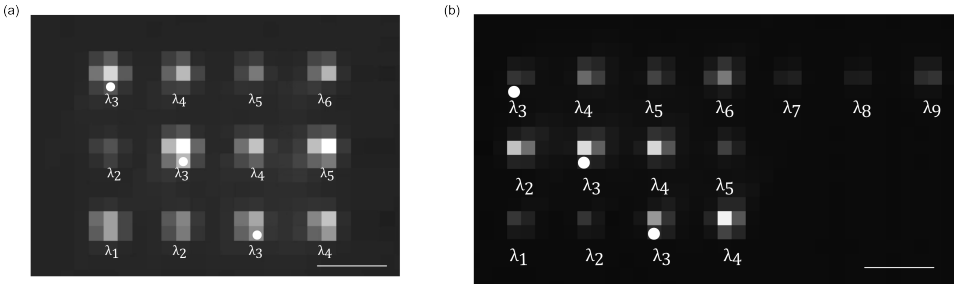

Figure 8: Parallelism measurement of the grating beam routing architecture, the white dot represents the light incident spatial channel and the wavelengths indicate the frequency of the output light that diffracted to this spatial channel.

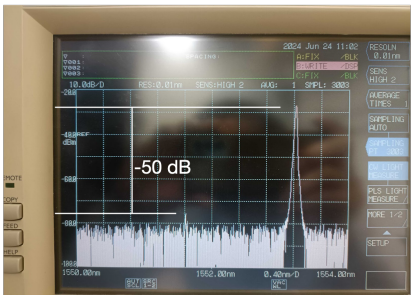

Figure 9: Crosstalk measurement of the grating beam routing system, the crosstalk between different channels are -50 dB.

Having demonstrated the high parallelism and low crosstalk advantages of the grating beam routing system, we characterize the performance of the grating when working in the optical computing architecture as a wavelength Mux and DeMux.

B. Subsection 2: Grating beam routing system works as Mux

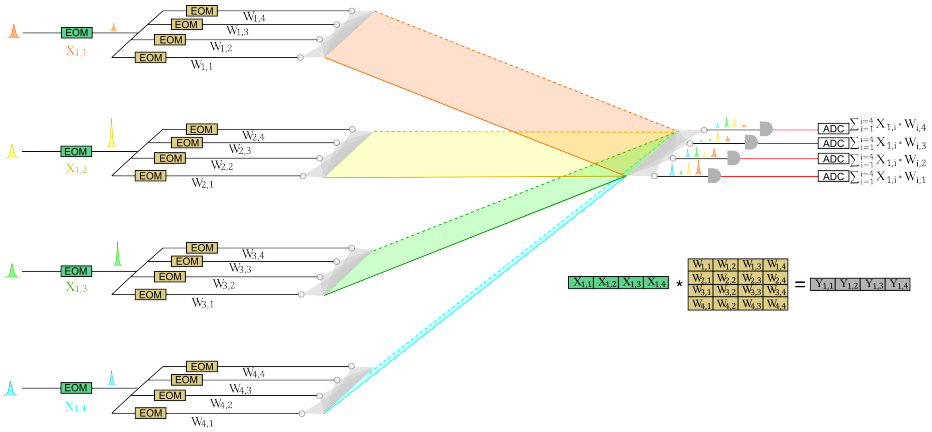

Figure 10: Architecture setup when the grating beam routing system works as Mux.

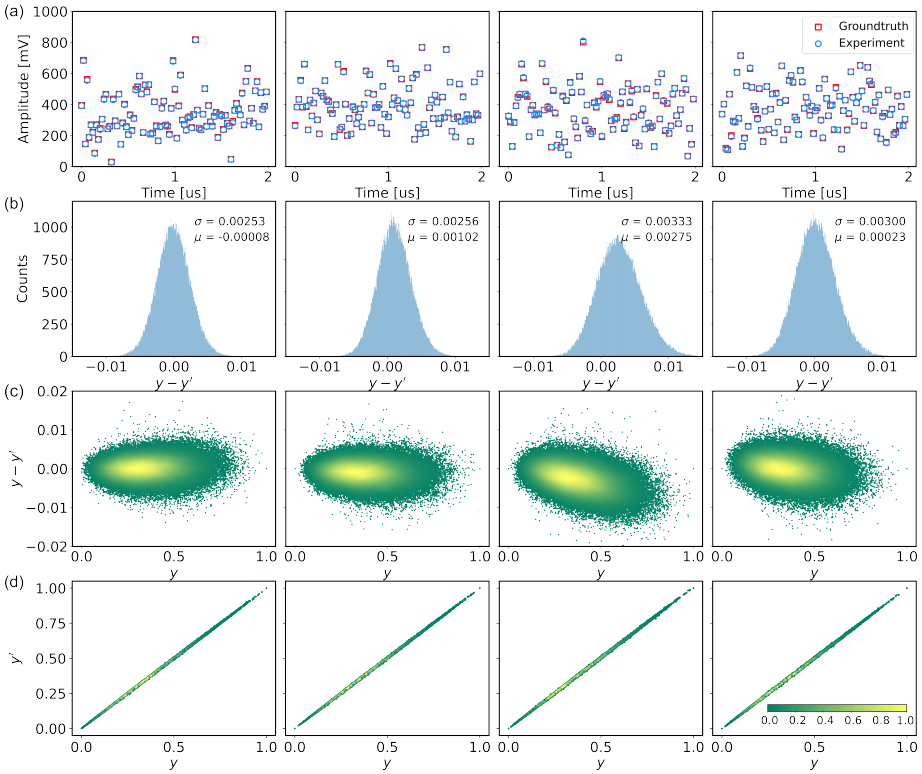

Figure 11: Experimental results of the grating beam routing Mux.

Fig. 10 shows the setup of the system when employing the grating beam routing system as wavelength Mux. Four data modulators are modulated by four different wavelengths, each modulated signal is then spatially fan-out to four weight modulators, the outputs are then guided to the grating beam routing system through fiber array, the grating beam routing system works as wavelength/spatial Mux, different wavelengths are multiplexed to the same channel and the outputs are being recorded by amplified photodetectors. Fig. 11 shows the measured waveform and MAC error distribution. The modulation signal sampling rate is 50 $\mathrm{MSa\,s^{-1}}$. The system shows low error standard deviation and high computing accuracy across multiple channels simultaneously.

C. Subsection 3: Grating beam routing system works as DeMux

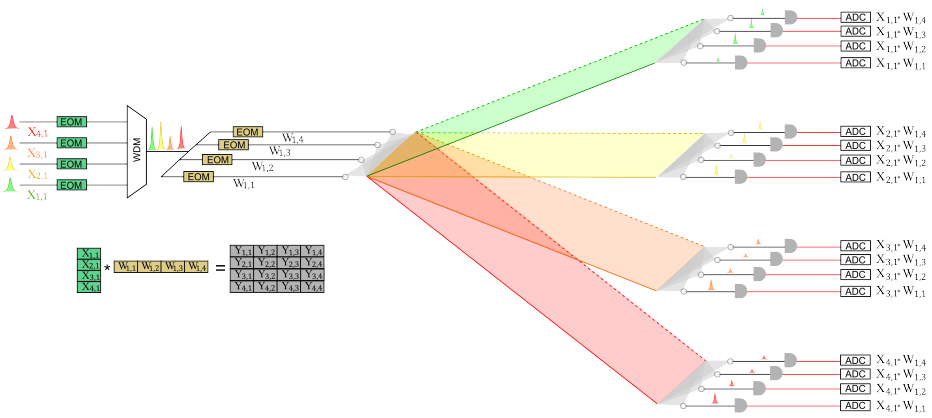

Figure 12: Architecture setup when the grating beam routing system works as DeMux.

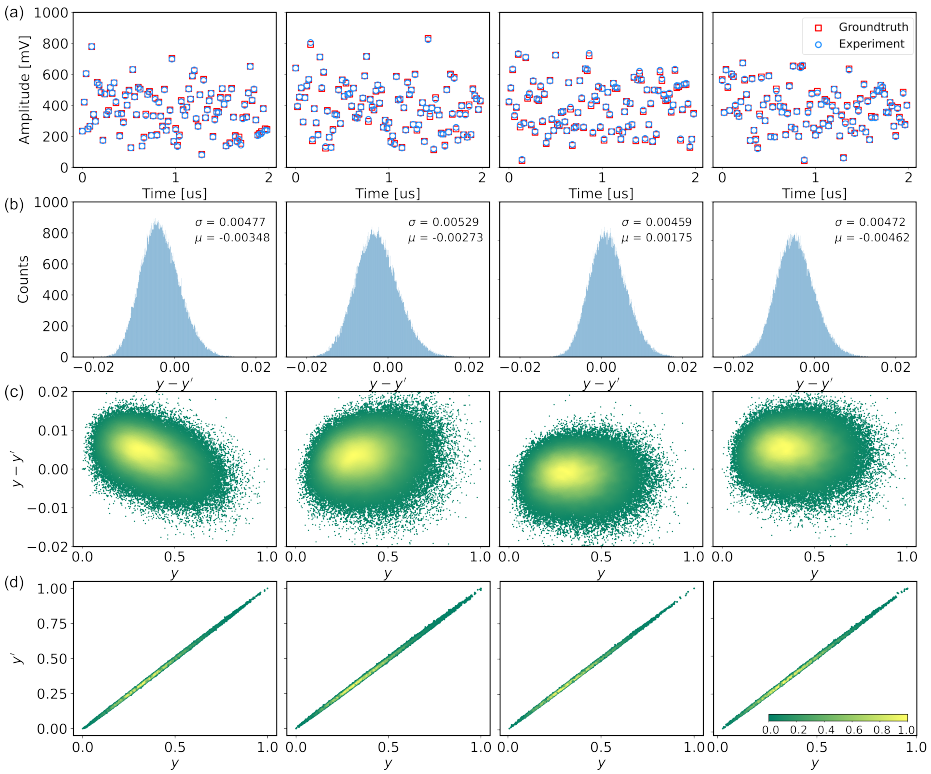

Figure 13: Experimental results of the grating beam routing DeMux.

The grating beam routing system also works as wavelength DeMux. Fig. 12 shows the setup of the system when a commercial WDM multiplex four different wavelengths from different channels into one channel and then spatially fan-out to four different weight modulators, the modulated signals are then demultiplexed by the grating, the system structure is similar to our previous work [4] but provides 3D data parallelism advantage. Fig. 13 shows the measured waveform and multiplication error distribution. The system shows low error standard deviation and high computing accuracy, the results are comparable with work [4], which uses commercial WDM for DeMux and has the system parallelism of $O(N)$, while our grating architecture provides $O(N^3)$ parallelism with only one device.

D. Subsection 4: Minimum wavelength spacing

The minimum wavelength spacing is determined by the crosstalk of these two wavelengths beams when these two beams are focused by the lens in the surface of the receiver fiber array. Fig. 14 shows the concept of how the minimum wavelength spacing is calculated. Broadband light emitted from the input fiber array becomes collimated after the input achromatic doublet, the beam diameter, which is proportional to the focal lens, is about 1.9 cm. Through the grating diffraction, different wavelength components will be scattered at different angles, with the angle difference $\Delta\theta$ determined by the grating angular dispersion, and the distance between these two diffracted beams are determined by both the grating angular dispersion and the distance l between the grating and the receiver achromatic doublet. After propagate the distance l, these two collimated beams are focused in the front surface of the receiver fiber array by the receiver achromatic doublets. The minimum wavelength spacing without channel/wavelength crosstalk is shown in the Fig. 14. In this figure, the shorter wavelength is focused in the center of the receiver fiber array, where another wavelength beam is just focused outside the receiver fiber array core, in this case, the receiver fiber array will only collect the lights from the shorter wavelength. The minimum distance between the focused spots is determined by the receiver fiber array core diameter, which is 50 µm. The numerical aperture (NA) of the multi-mode fiber is 0.22. This corresponds to a critical angle of 12.7°, which is larger than the half solid angle 7.2° of the focused beam. The beam waist is 8 µm, so the focused beam could be totally collected. We also need to consider the evanescent field of the light, which could otherwise cause crosstalk. The evanescent field penetration depth could be calculated through

$$d_p = \frac{\lambda}{2\pi n_1 \sqrt{\sin^2\theta - \left(\frac{n_2}{n_1}\right)^2}} \tag{8}$$

From Eq. (8), the penetration depth is calculated to be 300 nm. Therefore, the minimum distance g shown in the figure is 33 µm, and the diffraction angle difference $\Delta\theta$ is given by:

$$\Delta\theta = \frac{33\,\mu\text{m}}{l}.$$

And the corresponding wavelength difference is

$$\Delta\lambda = \frac{1}{2}\lambda_0 \Delta\theta \cot(\theta_0), \tag{9}$$

where:

- θ_0 is directly back scattered by the grating.

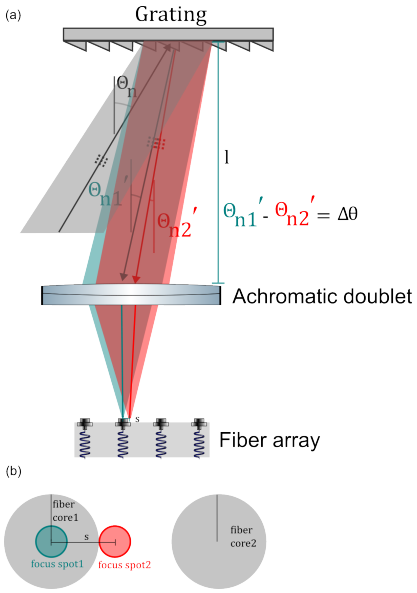

Figure 14: Concept of how to calculate the minimum wavelength spacing of the grating beam routing architecture.

Consider the grating-receiver achromatic doublet distance l of 30 cm, the wavelength difference is

$$\frac{\lambda d}{2l} = 0.09\,\text{nm}.$$

The theoretical resolution of the grating is $\frac{\lambda}{N} = 0.057\,\text{nm}$, which meets the low crosstalk requirement of the system. We need to noticed that the calculated wavelength spacing is the theoretical value, in the real case, we also need to consider the scattering in the fiber surface, consider the availability of the commercial WDM that we need to use to multiplex different lasers, the wavelength spacing of 0.4 nm (50 GHz) is a reasonable value.

E. Subsection 5: Maximum wavelength range we can use

Several factors will influence the maximum wavelength range we can use, like the total bandwidth of the laser source, the bandwidth of the commercial WDM. Here, we consider the grating beam routing architecture itself, and the main limitation for the working wavelength range is the chromatic aberration of the lens, different wavelength lights will have different focal lens, making different spot size of the focused beam in the receiver fiber array front surface.

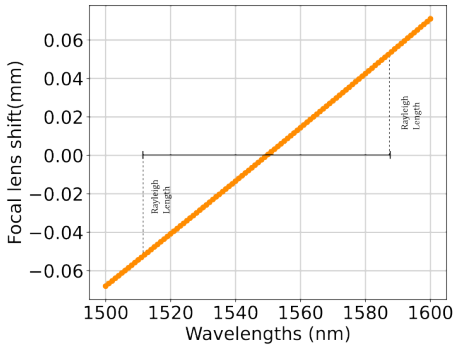

Figure 15: Focal lens shift of different wavelength lights.

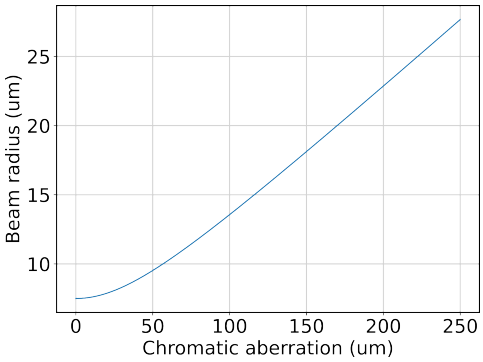

Figure 16: Spot size change caused by the chromatic aberration.

Figure 15 shows the focal lens shift of the achromatic doublets. When the wavelengths change from 1530 nm to 1580 nm, the focal lens shift is only 70 µm. The Rayleigh length of the focused beam is 50 µm, so the focal spot size of different wavelength lights is within the Rayleigh length range and could all be collected by the receiver fiber array. From this point of consideration, the working wavelength range is at least 50 nm. The large numerical aperture (NA) and core diameter of the multimode fiber also provide a large tolerance for the broadband light's chromatic aberration. Figure 16 shows the spot size change versus chromatic aberration. The spot size changes from 7.5 µm to 25 µm, with a small divergence angle of 7°. So the working wavelength range of the grating beam routing architecture is over 50 nm. The input light wavelength spacing could reach 50 GHz meanwhile keeping low crosstalk between adjacent fiber array channels, and the maximum allowed working wavelength range IS over 50 nm, this corresponds to a wavelength channel number $2N-1$ of 125 and fiber array channel number N over 60, which makes an off-axis distance of 4.5 mm. Achromatic doublet lenses have a much reduced sensitivity to centration on the beam axis when compared to spherical singlets and aspheric lenses, the calculated lateral and transverse aberrations is 32 µm, within the tolerance of the Rayleigh range, so the maximum working wavelength range could reach 50 nm.

F. Subsection 6: Compare with commercial Wavelength Mux and DeMux

The current wavelength spacing of our grating beam routing architecture is 200 GHz, which is limited by the available commercial WDM we have that used to multiplex different laser wavelengths. As we discussed in the previous section, the theoretical wavelength spacing of the grating architecture is below 50 GHz, which is comparable with dense WDM, the working wavelength range is over 50 nm, and the spatial channel could reach 120. We characterize the commercial WDM that we used to multiplex the different wavelength lasers, the wavelength spacing for the commercial WDM is 200 GHz, with the insertion loss of 0.7 dB for the first DeMux channel, and gradually increased to 1.6 dB for the fourth channel. The insertion loss will be further increased if more WDM channels are being used. For our grating beam router, the loss is between 0.7 dB to 1.5 dB, which is comparable with the commercial WDM. In addition, our grating beam router supports multichannel, three-dimensional input and output, which is not possible with just one commercial WDM. Let take the $4 * 4$ beam router as an example, in order to achieve the same output throughput, we need 16 commercial WDMs to DeMux the broadband wavelength into 64 different ports, and another 16 commercial WDMs to Mux the different wavelength lights from different channels into 16 Mux ports, as shown in Fig. 17.

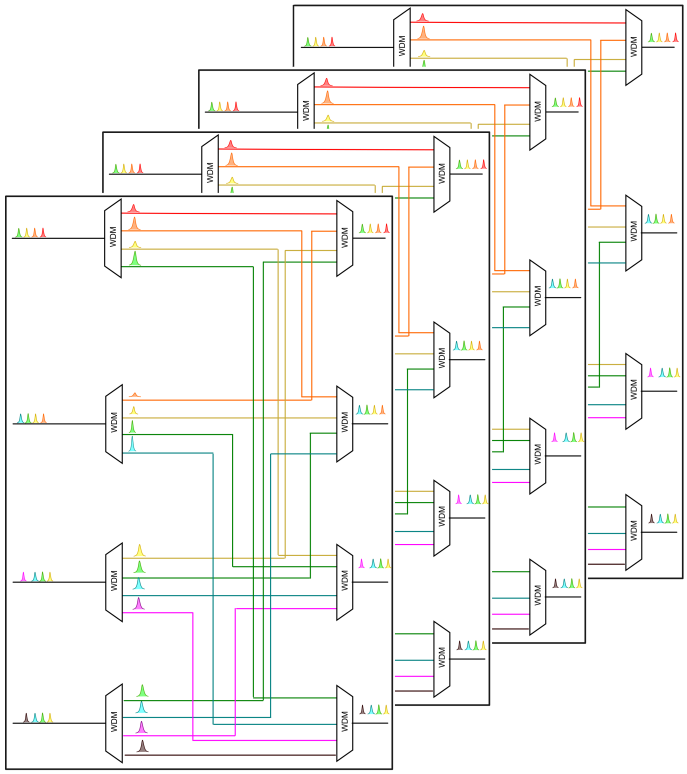

Figure 17: Schematic setup of the grating beam routing system using commercial WDM.

Our grating beam routing system is capable for further scaling with dense wavelength spacing. In order to achieve the same throughput of the grating beam routing system with wavelength spacing of 50 GHz and channel numbers of 120, 240 commercial WDMs are needed.

G. Subsection 7: Chip scale architecture

The fiber array pitch could be further reduced, as illustrated in Fig. 18. The modulators, WDMs, beam splitters, circulators and grating couplers are fabricated in the chip while the collimation lens and the blazed grating are in free space. The chip grating couplers are used to de-multiplex different wavelength lights into differents of the collimation lens and the free space grating according to their wavelength while the free-space grating works as Mux and DeMux by through the wavelength-spatial coupling. In this case, the chip grating pitch could be small.

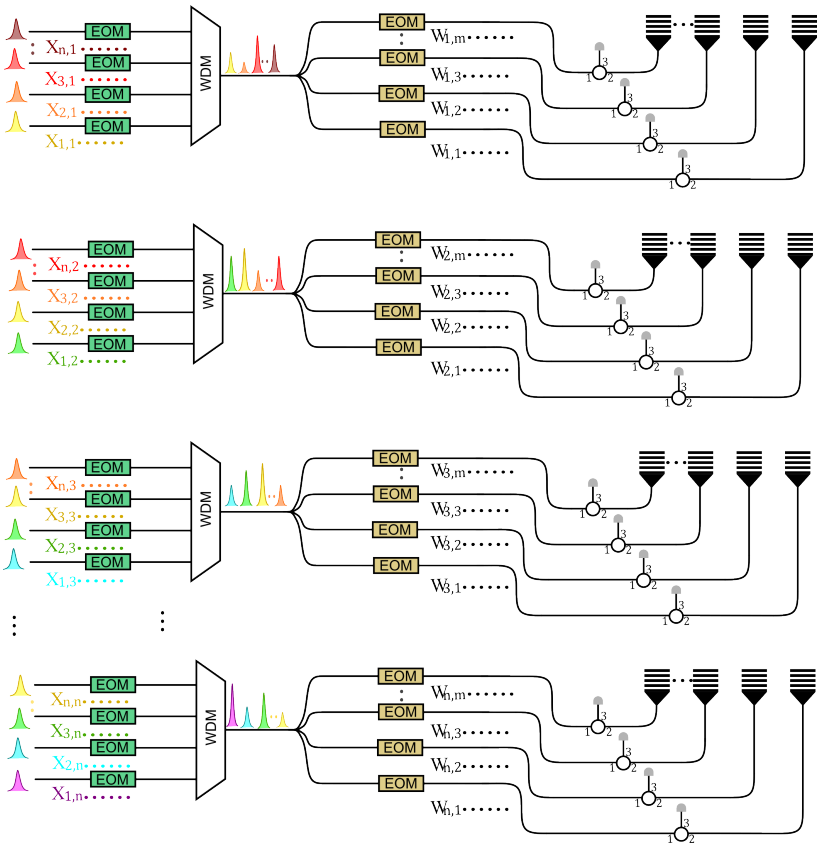

Figure 18: Concept of using chip scale modulators and fiber arrays.

The grating coupler angle does not change things as long as the emitted light is still captured by the lens. To show this, consider Fig. 19. The grating coupler emits a cone of light at an angle $\theta_{\rm gr}$ with a divergence $\theta_{\rm div} = 2\text{NA}$. Everything works in the small-NA limit, so this leads to a spot of size $w = L\theta_{\rm div}$ on the lens. The lens converts this diverging spot to a collimated beam of size w that propagates as a plane wave since the distance to the grating is much less than the Rayleigh length. The grating deflects the angle by an amount $|\theta_{n'} - \theta_n|$, which changes the focal point on the chip.

There is a small amount of beam travel $s = \frac{L'}{|\theta_{n'} - \theta_n|}$ due to the angular change, which leads to a beam angle deflection $\Delta\theta_{\rm gr} = (s/w)\theta_{\rm div}$ when focused to the PIC. As long as $s \ll w$, the beam angle is deflected by much less than the grating coupler divergence, so the light couples near-perfectly into the grating coupler. We find that (for a 45° grating tilt):

$$\frac{s}{w} = \frac{L'|\theta_{n'} - \theta_n|}{w} \leq \frac{d}{w}|\theta_{n'} - \theta_n| \leq \frac{d}{w} N\Delta\theta = \frac{2d}{w}\frac{\Delta\lambda_{\rm BW}}{\lambda} \tag{10}$$

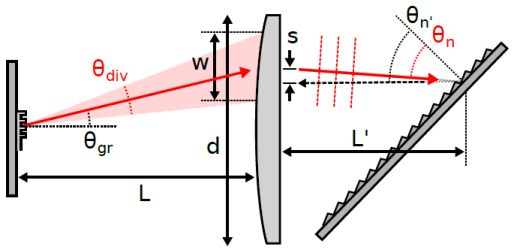

Figure 19: Effect of grating coupler angle and beam travel on the position of the return beam.

Equation (10) shows the relationship between the beam deflection, grating properties, and wavelength. As long as (1) the optical bandwidth is small enough ($\Delta\lambda_{\mathrm{BW}}/\lambda \ll 1$) and (2) the lens is not much bigger than the beam divergence ($d/w = O(1)$), the ratio in Eq. (10) is small and the system does not depend on the grating coupler angle θ_{gr} at all: light emitted at any angle will always return to the target grating coupler (with the same angle). Condition (1) is satisfied for telecom optics ($\Delta\lambda_{\mathrm{BW}}/\lambda = 0.02$ for 30 nm bandwidth). Condition (2) is related to the grating coupler angular dispersion, since the lens must be large enough to collect the light from all colors. This sets a lower bound on d/w:

$$\frac{d}{w} \geq \frac{(\theta_{\mathrm{div}} + 2\Delta\theta_{\mathrm{disp}})L}{\theta_{\mathrm{div}}L} = \frac{\theta_{\mathrm{div}} + n_g\Delta\lambda_{\mathrm{BW}}/\lambda}{\theta_{\mathrm{div}}} \tag{11}$$

(where we have used $d\theta \approx -n_g d\lambda/\lambda$ valid for a near-vertical grating coupler). Assuming some reasonable $n_g \lesssim 4$ and setting $\Delta\lambda = 30$ nm and $\theta_{\mathrm{div}} = 2\mathrm{NA} = 0.28$, the ratio in Eq. (11) is around 1.25, easily satisfying condition (2). While this analysis suggests that grating angles are not particularly important, there are hardware solutions that may lead to more convenient vertical coupling. Micro-lens arrays or 3D-printed optics can verticalize a beam [7], at least for a target wavelength. Some SiPh processes also incorporate broadband TIR mirrors for vertical coupling [8].

IV. FREE SPACE GRATING BEAM ROUTING FOR PARALLEL CONVOLUTION OPERATION

A. Subsection 1: Positive and Negative kernels

For the image convolution, the image matrix data is positive and the kernel matrix contains both positive and negative numbers, we process the positive and negative kernel numbers in different time rounds. We consider the $2*2$ kernel K, where a and b are positive numbers, c and d are negative numbers

$$\mathbf{K} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The $28*28$ data matrix X is

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix}$$

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} * \mathbf{K} \\ &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix} \\ &= \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} * \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ c & d \end{bmatrix} * \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix} \\ &= \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} * \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ -c & -d \end{bmatrix} * \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,28} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{28,1} & x_{28,2} & \cdots & x_{28,28} \end{bmatrix}. \end{aligned}$$

In this way, the kernels are positive and could be processed by direct detection in our optical hardware. E/O modulators and waveshapers are used to encode the kernels into the optical carrier. The positive and negative kernels could also be processed in one time round by setting the intermediate value, which is

$$\frac{P_{\max} + P_{\min}}{2}$$

as the reference level, in this way, the P_{maxpd} corresponds to floating point value 1 and P_{minpd} corresponds to floating point value -1, any values between -1 and 1 could be extracted using the methods described in (4) and (3).

B. Subsection 2: Electrical configurable E/O modulator for convolution operation

Fig. 20 shows the setup of the convolution operation. Two $2 * 2$ kernels are encoded through eight E/O intensity modulators, and each weight modulator encodes one element of these two kernels, in this way, the output value and input voltage for each modulator are constant values and the convolution operation for a single patch could be finished in a single time step.

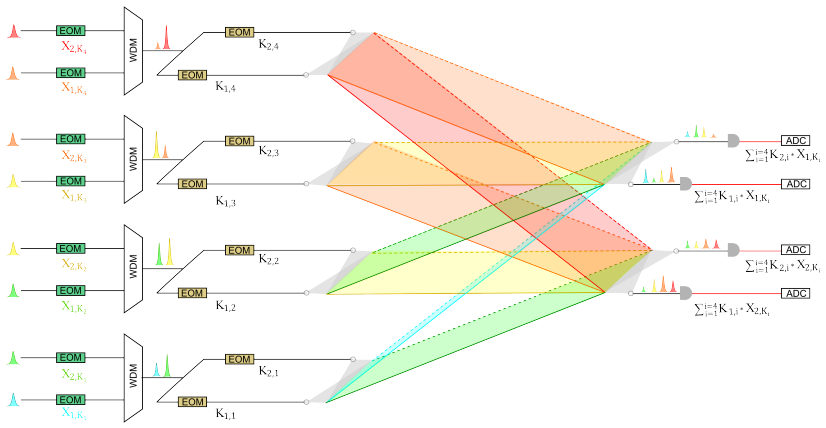

Figure 20: Convolution setup of mapping two $2 * 2$ kernels into the optical hardware, eight kernel modulators and eight data modulators are used to encode the kernel and the data.

The image data matrix needs to be processed before it could be mapped to the data modulators, Fig. 21 shows how to process the image matrix data for convolution.

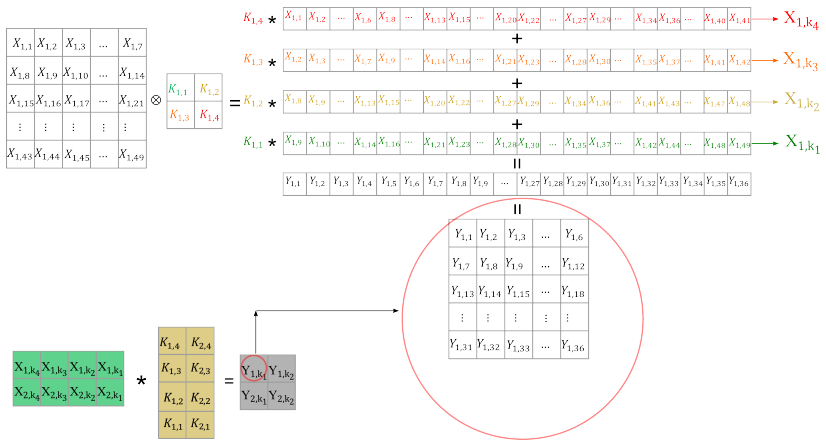

Figure 21: Processing the data matrix for $2 * 2$ kernel convolution, the data matrix is processed and mapped to four data modulators.

Using the analog time integrator, we can map larger kernels into the optical hardware. Fig. 22 and Fig. 23 show the setup and the data processing steps when using the optical hardware to perform a $4 * 4$ kernel convolution, analog time integrator accumulate the multiplication every four-time steps, Fig. 24 shows the convolution outputs of the

image with two $4 * 4$ kernels using the analog time integrator, the optical convolutional results matches well with the digital convolution results, indicating high accuracy of the grating beam routing system.

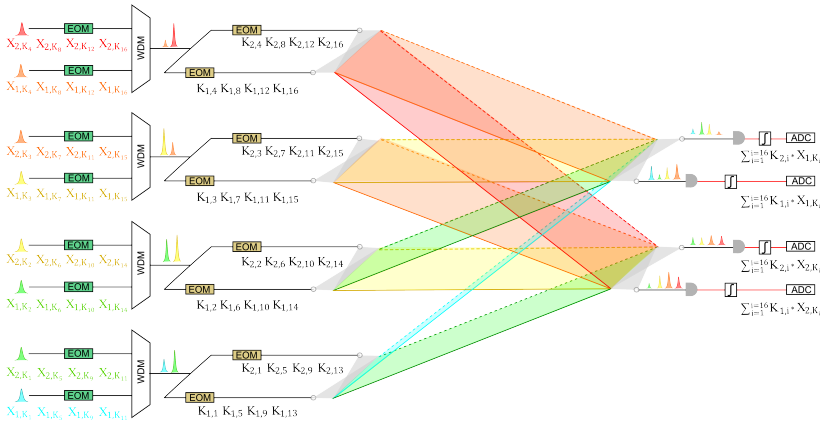

Figure 22: Convolution setup of mapping two $4 * 4$ kernels into the optical hardware, 4 modulators encodes the $4 * 4$ kernel in 4 time steps.

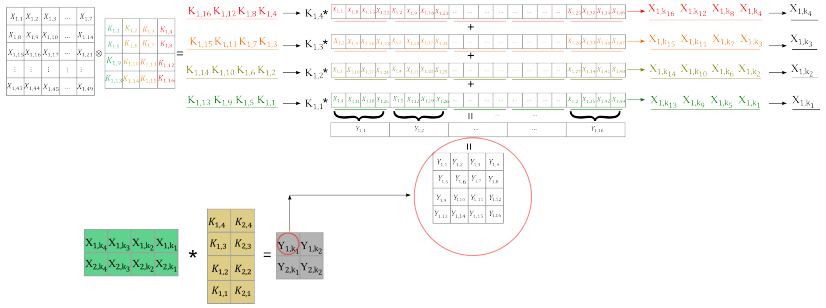

Figure 23: Processing the data matrix for $4 * 4$ kernel convolution, the data matrix is processed and mapped to four data modulators in 4-time steps.

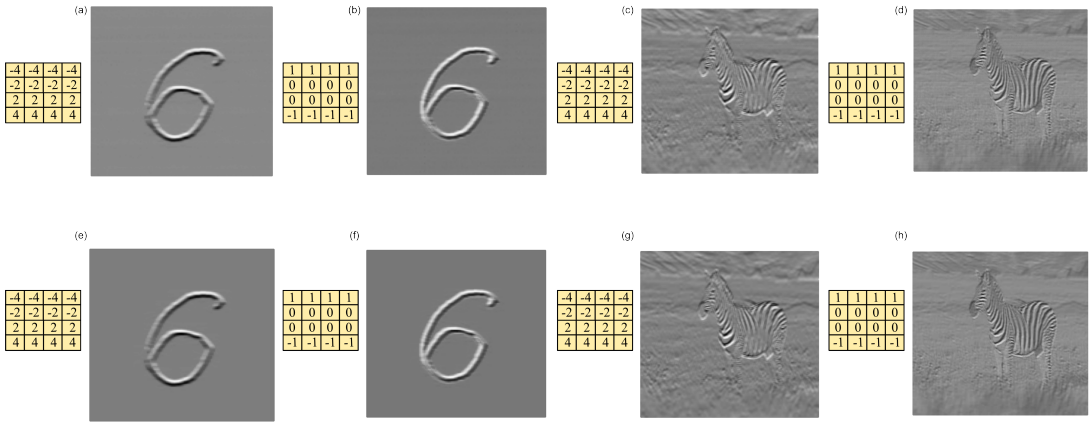

Figure 24: The convolution results for the $4*4$ kernel with image matrix using analog time integrator, the top 4 figures are the convolution images using the optical hardware, where the bottom figures are digital convolution results..

C. Subsection 3: Programmable waveshapers for convolution operation

Programmable waveshapers are also used to perform the convolution operation, compared with the E/O modulator, waveshaper could provide a much larger on-off ratio and extinction ratio of the signal. Fig. 25 shows the setup and results of the convolution operation, two 2∗2 kernels and 4 images are processed simultaneously, the optical convolution results match well with the digital convolution results across multiple wavelengths and spatial channels, which further verifies the high parallelism, low crosstalk and high accuracy of the optical beam routing system. The waveshaper switching speed is low, so we didn't perform the large-size kernel convolution using waveshaper.

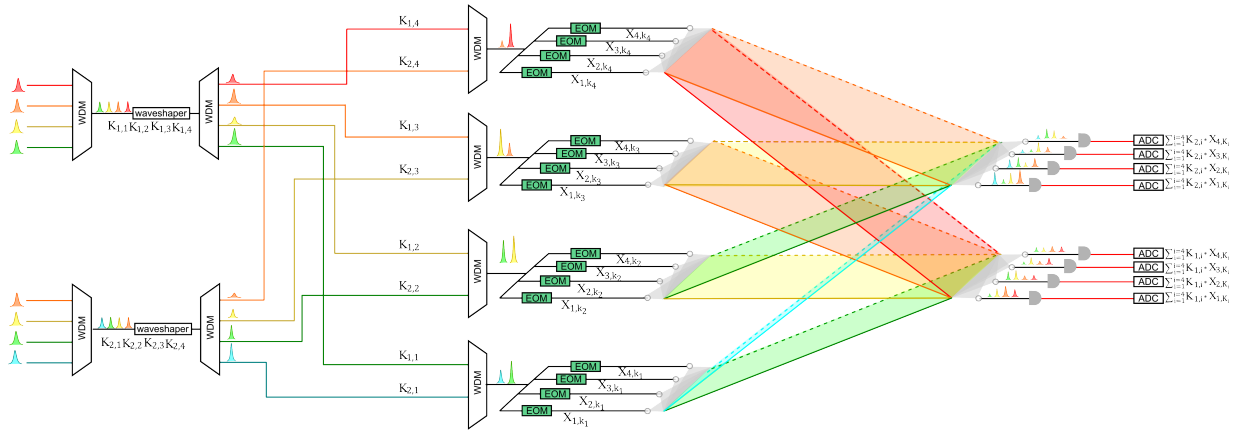

Figure 25: The convolution setup for the $2 * 2$ kernel with image matrix using waveshaper.

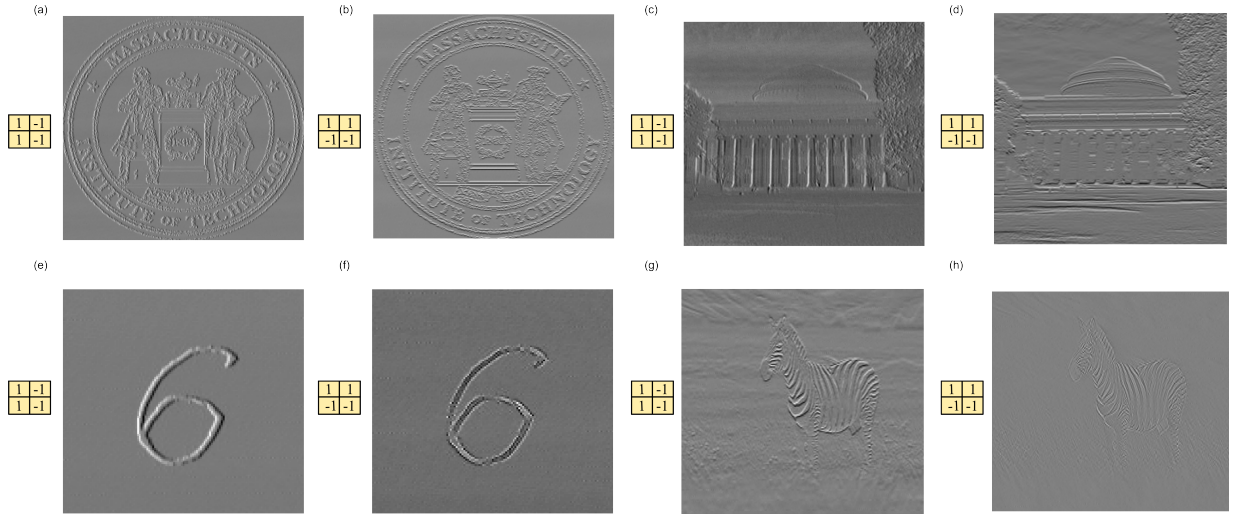

Figure 26: The convolution results for the $2 * 2$ kernel with image matrix using waveshaper.

V. OPTICAL PARALLELISM AND ENERGY EFFICIENCY

The grating beam routing system uses SDM, wavelength Mux, wavelength DeMux, free-space beam routing system, and analog time integrator to achieve low energy, high parallelism operation. The energy consumption of the system is shown in Table II, where N and M are the fiber array channels, T is the analog time integrator integration period.

Table II: Energy consumption of the grating beam routing system based optical hardware

Components	**Energy per Operation**	**MACs per Operation**	**Energy per MAC**	**Number of Components**
Data modulator	$\sim 1\,\mathrm{pJ}$	M	$1/M\,\mathrm{pJ}$	N^2
Weight modulator	$\sim 1\,\mathrm{pJ}$	N	$1/N\,\mathrm{pJ}$	NM
DAC for data modulator	$\sim 1\,\mathrm{pJ}$	M	$1/M\,\mathrm{pJ}$	N^2
DAC for weight modulator	$\sim 1\,\mathrm{pJ}$	N	$1/N\,\mathrm{pJ}$	NM
ADC	$\sim 1\,\mathrm{pJ}$	NT	$1/(NT)\,\mathrm{pJ}$	NM
Photoreceiver	$\sim 1\,\mathrm{fJ}$	N	$1/N\,\mathrm{fJ}$	NM
Analog integrator	$\sim 1\,\mathrm{fJ}$	NT	$1/(NT)\,\mathrm{fJ}$	NM
Nonlinearity	$< 100\,\mathrm{fJ}$	NT	$100/(NT)\,\mathrm{fJ}$	NM
Total	–	–	$(2/M + 2/N + 1/NT)\,\mathrm{pJ}$	–

Table III: Energy consumption of the grating beam routing system with optimized optical DAC

Components	**Energy per Operation**	**MACs per Operation**	**Energy per MAC**	**Number of Components**
ODAC for data encoding	$\sim 40\,\mathrm{fJ}$	M	$40/M\,\mathrm{fJ}$	N^2
ODAC for weight encoding	$\sim 40\,\mathrm{fJ}$	N	$40/N\,\mathrm{fJ}$	NM
ADC	$\sim 1\,\mathrm{pJ}$	NT	$1/(NT)\,\mathrm{pJ}$	NM
Photoreceiver	$\sim 1\,\mathrm{fJ}$	N	$1/N\,\mathrm{fJ}$	NM
Analog integrator	$\sim 1\,\mathrm{fJ}$	NT	$1/(NT)\,\mathrm{fJ}$	NM
Nonlinearity	$< 100\,\mathrm{fJ}$	NT	$100/(NT)\,\mathrm{fJ}$	NM
Total	–	–	$(40/M + 40/N + 1101/NT + 1/N)\,\mathrm{fJ}$	–

For an optical neural network architecture without using the above-mentioned technique, the system performance is shown in Table IV. In this case, the energy consumption is 5 pJ per MAC.

Table IV: Energy consumption of optical hardware without using grating beam routing, optical fanout, and time integration

Components	**Energy per Operation**	**MACs per Operation**	**Energy per MAC**	**Number of Components**
Data modulator	$\sim 1\,\mathrm{pJ}$	1	$1\,\mathrm{pJ}$	N^2M
Weight modulator	$\sim 1\,\mathrm{pJ}$	1	$1\,\mathrm{pJ}$	$N^2 * M$
ADC for data modulator	$\sim 1\,\mathrm{pJ}$	1	$1\,\mathrm{pJ}$	$N^2 * M$
ADC for weight modulator	$\sim 1\,\mathrm{pJ}$	1	$1\,\mathrm{pJ}$	$N^2 * M$
DAC	$\sim 1\,\mathrm{pJ}$	1	$1\,\mathrm{pJ}$	$N^2 * M$
Photoreceiver	$\sim 1\,\mathrm{fJ}$	1	$1\,\mathrm{fJ}$	$N^2 * M$
Analog integrator	$\sim 1\,\mathrm{fJ}$	1	$1\,\mathrm{fJ}$	$N^2 * M$
Nonlinearity	$< 100\,\mathrm{fJ}$	1	$100\,\mathrm{fJ}$	$N^2 * M$
Total	–	–	$5\,\mathrm{pJ}$	–

VI. CNNS AND DNNS FOR IMAGE CLASSIFICATION

A. Subsection 1: Digital training of the network

The CNN and FC NN were trained on a standard digital computer in Python with the PyTorch library on 50,000 training images for the MNIST and Fashion-MNIST datasets. 10,000 images were reserved for validation sets to fine-tune the network hyperparameters and optical setup. Each dataset was normalized by its standerd deviation, and L2 regularization with $L_2 = 0.0001$ and 10% dropout were applied to each layer. Gaussian noise was also added to each activation value at every layer, with a standard deviation of $0.25 * \delta$, where δ is the standard deviation of the activation value across a batch. The nonlinear activation function on the final layer was SoftMax. The Adam optimizer minimized the categorical cross-entropy loss function with a batch size of 100 and learning rate of 0.001. Since we use the noise aware training and the quantization aware training, the actual network is very robust to the noise and suitable for low bit precision tasks. Fig. 27 shows the network testing accuracy for data with different bit precisions.

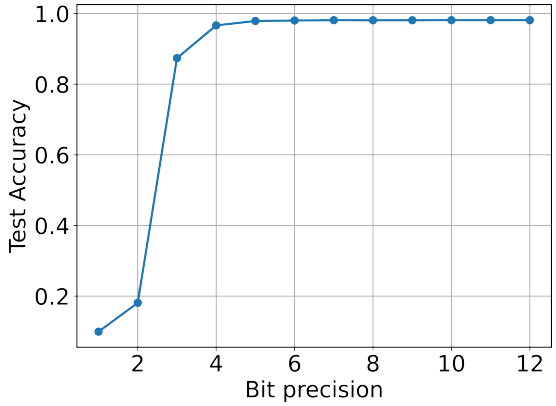

Figure 27: The convolution results for the $2 * 2$ kernel with image matrix using waveshaper.

B. Subsection 2: Implementation of the optical CNNs and DNNs

We perform a benchmark one-layer CNN and a following fully connected DNN inferences using the optical processor. We apply four 2×2 kernels to extract the features of these images, each input image, with 28×28 pixels, is encoded in 729 time-steps to four data modulators . The grating beam routing optical processor performs parallel batch operations, two 2×2 kernels, and two MNIST images are convolved by the grating beam routing processor simultaneously. The outputs from the convolutional layer, which contain 2,916 pixels of 1,000 images, are detected by amplified photodetectors. Series of operations like decoding the detector data from voltages to analog values, ReLU nonlinear activations, and data normalizations are performed digitally. The CNN outputs are fed into a fully connected DNNs, implemented by the same optical system.

The DNNs have one input layer (2,916 neurons), one fully connected hidden layer (100 neurons), and one output layer (10 neurons). These 2,916-pixel values are mapped to four data modulators in 729 time-steps, the first two columns of the weight matrix are mapped to the eight weight modulators, with each weight column encoded by four weight modulators in 729 time-steps, and this process is repeated for all the 100 weight matrix columns. The fully connected

hidden layer output values are recorded by analog time integrators. This is a batch process, and the same operations are performed for two input images simultaneously. Decoding, nonlinear ReLU activation, and normalization are performed digitally and the output results are sent to the system to perform the 100×10 matrix multiplication. Similar batch operations are performed by encoding the weight matrix and data matrix to the modulators in 25 time-steps. The outputs are converted into probabilities by a digital SoftMax operation.

C. Subsection 3: Latency of the image classification Model

We take MNIST images (each with 28×28 pixels) for consideration. For a $N \times M$ channel fiber array, N images with M kernels, each kernel with the dimension of $\sqrt{N} \times \sqrt{N}$ will be processed simultaneously, the accumulation of kernel elements and data pixels multiplication is achieved through WDM. The N input images are mapped to N^2 data modulators, with each image being flattened to $N(28-\sqrt{N}+1)^2$ pixels and encoded by N data modulators in $(28-\sqrt{N}+1)^2$ time steps. The M kernels (each with the dimension of $\sqrt{N} \times \sqrt{N}$) are encoded to $N \times M$ weight modulators, with each modulator has constant input RF voltage and output intensity. In a single clock rate, $N \times\ M$ pixels of $N \times\ M$ new images are generated, $(28-\sqrt{N}+1)^2$ time steps are needed to finish the $(\sqrt{N} \times \sqrt{N})$ convolutional operation of M kernels among N images. For a total MNIST image numbers of P, the time consumption for the convolutional operation is $\frac{P}{N}\left(28-\sqrt{N}+1\right)^2 \frac{1}{C}$.

For the subsequent DNN operation, the analog time integration is employed. Suppose the fully connected hidden layer and output layer neurons are D_1, D_2. The weight matrix size is $M(28-\sqrt{N}+1)^2 \times D_1$ and $D_1 \times D_2$. For the hidden layer, each analog time integrator needs to accumulate $\frac{M\left(28-\sqrt{N}+1\right)^2}{N}$ time steps over N different wavelengths before the readout, and this process needs to perform $\frac{D_1}{M}$ times across M different spatial channels for all the D_1 weight matrix columns. The system supports batch operation, in each time step, N images are processed by the system simultaneously. Based on this calculation, in order to process these P images, the needed time is $\frac{D_1 P\left(28-\sqrt{N}+1\right)^2}{CN^2}$. Similarly, the time to process output layer is $\frac{1}{C}\frac{D_2}{M}\frac{P}{N}\frac{D_1}{N}$N. Other operations, including the ReLU nonlinear activation, SoftMax operation, are processed digitally. So the total time needed to perform the image classification is $\frac{D_1 P\left(28-\sqrt{N}+1\right)^2}{CN^2}+\frac{D_1 P D_2}{CMN^2}+\frac{P\left(28-\sqrt{N}+1\right)^2}{CN}$, while for the system with $O(N)$ throughput, the processing time is N^2M longer. For the current 4×4 fiber array, the time to process 1,000 MNIST images is 96 ms. The latency could be reduced by multiplexing more wavelength and spatial channels through a large channel number fiber array, a 30×30 channel fiber array with $10\,\mathrm{GSas}^{-1}$ would reduce the processing time of 1,000 images to 1140 ns.

VII. SYSTEM PICTURE

Fig. 28 shows the setup figure, where from top to bottom, the data modulators, the weight modulators, and the grating beam routing system are shown.

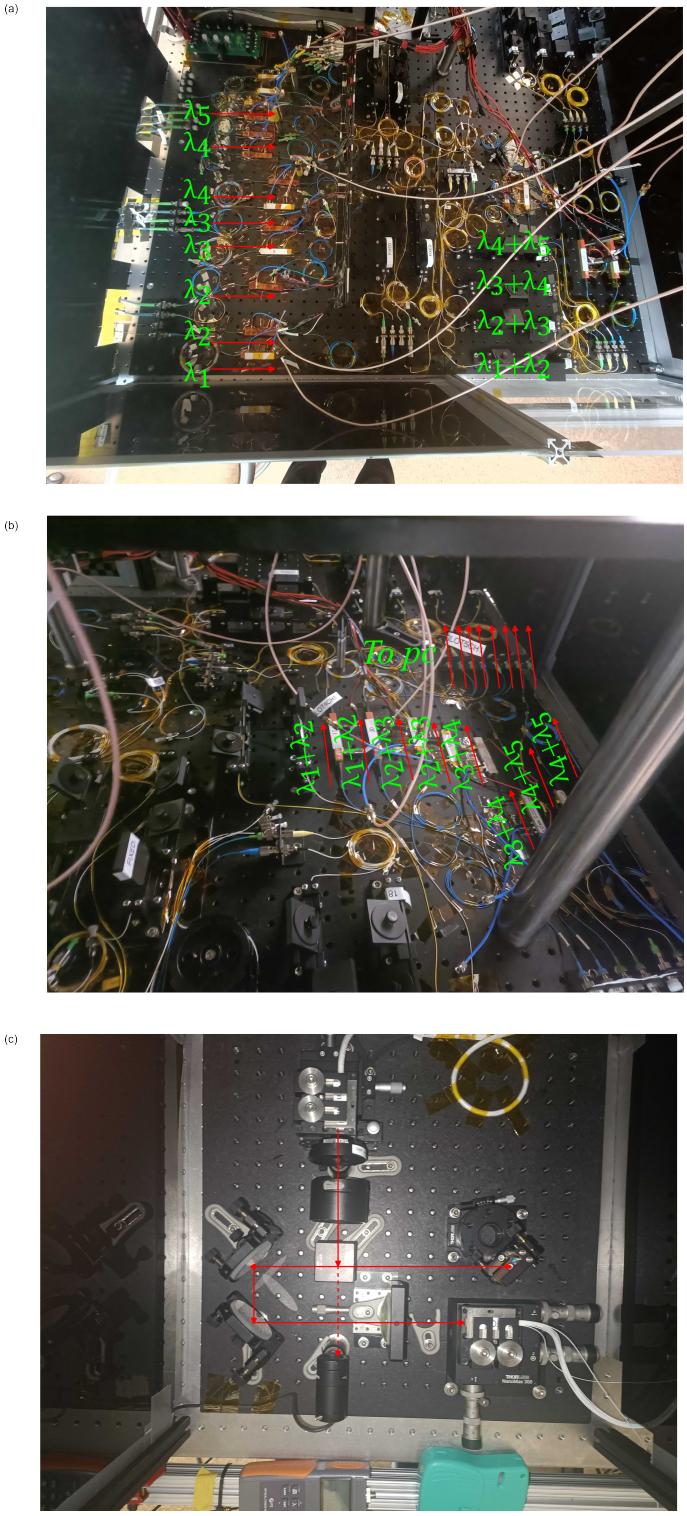

Figure 28: Figure of the builded grating optical neural network setup.

[1] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, *et al.*, Nature Photonics **11**, 441 (2017).

[2] A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Scientific reports **7**, 7430 (2017).

[3] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, Physical Review X **9**, 021032 (2019).

[4] A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein, D. Bunandar, P. B. Dixon, S. A. Hamilton, M. Streshinsky, A. Novack, T. Baehr-Jones, M. Hochberg, M. Ghobadi, R. Hamerly, and D. Englund, Science **378**, 270 (2022).

[5] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, *et al.*, Nature **589**, 52 (2021).

[6] Z. Chen, A. Sludds, R. Davis III, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein, *et al.*, Nature Photonics **17**, 723 (2023).

[7] P.-I. Dietrich, M. Blaicher, I. Reuter, M. Billah, T. Hoose, A. Hofmann, C. Caer, R. Dangel, B. Offrein, U. Troppenz, *et al.*, Nature Photonics **12**, 241 (2018).

[8] T. Aalto, M. Cherchi, M. Harjanne, S. Bhat, P. Heimala, F. Sun, M. Kapulainen, T. Hassinen, and T. Vehmas, IEEE Journal of selected topics in quantum electronics **25**, 1 (2019).