Supplementary Information

1. HITOP Multi-layer implementation

HITOP is an optoelectronic architecture. The input and weights are stored in nearby memories and transmitted to the optical system for matrix operations in the optical domain. The multiplied result is read out with charge accumulation and serialized with nonlinearity activation to the next layer. With a memory unit integrated together, our ONN could be used for multiple layer neural network inference with such working flow:

- Input vectors and weight vectors for the current layer are read out from the digital memory and programmed as time traces.
- Each input and weight time trace, respectively, is sent to the corresponding VCSEL or TFLN MZM (Fig. 1d).
- The output is acquired and accumulated by the photodetectors and time integrator arrays.
- Digital memory stores the spatially distributed signals from the photodetector and time integrator array.
- Reprogram the output by mapping the spatial distributed signal into corresponding time steps. The outputs of the same wavelength at different spatial channels are serialized as the input driving signal to a VCSEL transmitter in the next layer. Here the ReLu nonlinear operation function is encoded using the lasing threshold effect.

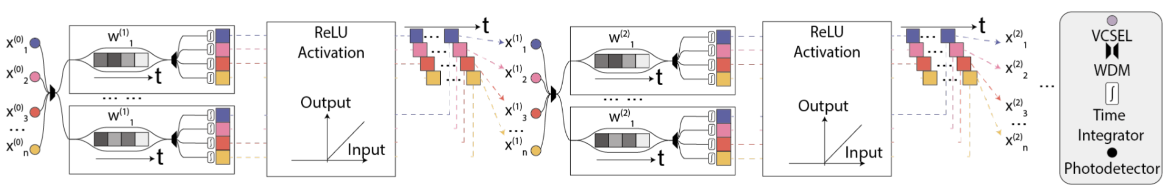


Fig. S1 HITOP forward propagating deep neural network architecture. The network consists of an arbitrary number of layers (2 layers are shown). Each layer computes a matrix-matrix multiplication in wavelength-time-space multiplexing using VCSELs and TFLN MZM (Fig. 1d).

2. Optical power budget of time integrating receivers

The total noise is a combination of additive detector noise, photon shot noise and laser intensity noise. The signal to noise ratio analysis was studied in [73], validated in our previous work for optical computing [35] and adapted here. The total noise spectral density is

$$N_{tot} = \sqrt{[(\eta\alpha)^2 + 2h\nu\eta P + \beta(\eta P)^2]B}$$

where α is the noise equivalent power (NEP) of the detector, η is the detector quantum efficiency, h is the Planck's constant, ν is the laser frequency, P is the laser power, and β is the relative laser intensity noise. The noise spectrum is integrated over the effective Fourier bandwidth of B=1/(2T), where T is the acquisition time. With the root square mean input signal $S = 2\sqrt{2}\eta P$, where a factor of 2 enhancement of the signal account for the differential subtraction using 2 photodetectors in HITOP, the average uncertainty of the signal is determined as

$$\sigma_t = N_{tot}/S = 1/(2\sqrt{T})\sqrt{[(\alpha/P)^2 + 2h\nu/(\eta P) + \beta]}$$

So the total signal to noise ratio is

$$SNR = 1/\sigma_t = 2\sqrt{K/R}[(\alpha/P)^2 + 2h\nu/(\eta P) + \beta]^{-1/2}$$

where T=K/R, K is the integration time steps, and R is clock cycle.

To reach a SNR of 100 (for 6-7 bits of computing precision) at 10 GS/s, a optical power of P=40 μW is required at each PD, with the detector noise of NEP=2 $\mathrm{pW\cdot Hz^{-1/2}}$, and P=120 μW for NEP=10 $\mathrm{pW\cdot Hz^{-1/2}}$, RIN=-135 dBc/ limited by the laser intensity noise at this power level (Fig. S2a). Considering the high speed detectors are relatively noisy (due to limited chip area and quantum efficiency), the NEP=2 $\mathrm{pW\cdot Hz^{-1/2}}$ is practically hard to achieve. So the total power requirement when the system scales up to 1000x1000 channels is P_{tot}=120 W (and P=400 W with ~5 dB loss due to VCSEL-chip coupling [57]), which is challenging for on-chip laser generation and power handling of photonic circuits.

This power requirement is significantly reduced with time integration. Although the computation clockrate is high (e.g. 10 GS/s), the integrating receiver only digitizes the vector products when the accumulation is complete (e.g. 10 MHz with vector size K=1000, or 10 kHz with $K=10^6$). The slow readout detection allows (1) sufficient measurement time to accumulate photon charges, as we previously demonstrated less than 1 photon per MAC operation in Ref. [39]; (2) usage of low-speed photon detectors of high sensitivity. In the following plots, we calculate the SNR evolution with increasing optical power under our experimental conditions of the PD sensitivity (noise equivalent power NEP=2 $\mathrm{pW\cdot Hz^{-1/2}}$ including amplifier) and K=1000. Our system only requires 0.3 μW optical power on each PD, which sums up to 0.3 mW for 1000 channels. Taking into account the loss of VCSEL-to-chip coupling (5 dB demonstrated in [57]) and the TFLN modulator loss (0.1 dB/cm), a total power of about 1 mW per VCSEL is required. This number can be reduced to 0.03 mW with $K=10^6$. Therefore, our VCSEL power of 6~8 mW is sufficient to support the 1000 channels without amplification required in either case.

For the detector NEP=10 pW·Hz$^{-½}$. The power requirement for K=1000 increases to 1.6 µW per PD (1.6 mW for 1000 channels), which is considered the limit of our VCSELs for 1000 channels (Fig. S2d). Again, this requirement is lower by a factor of ~30 (SNR∝T$^{-1/2}$) to 50 nW per PD (50 µW for 1000 channels), when the integration time is K=10^6 (Fig. S2f).

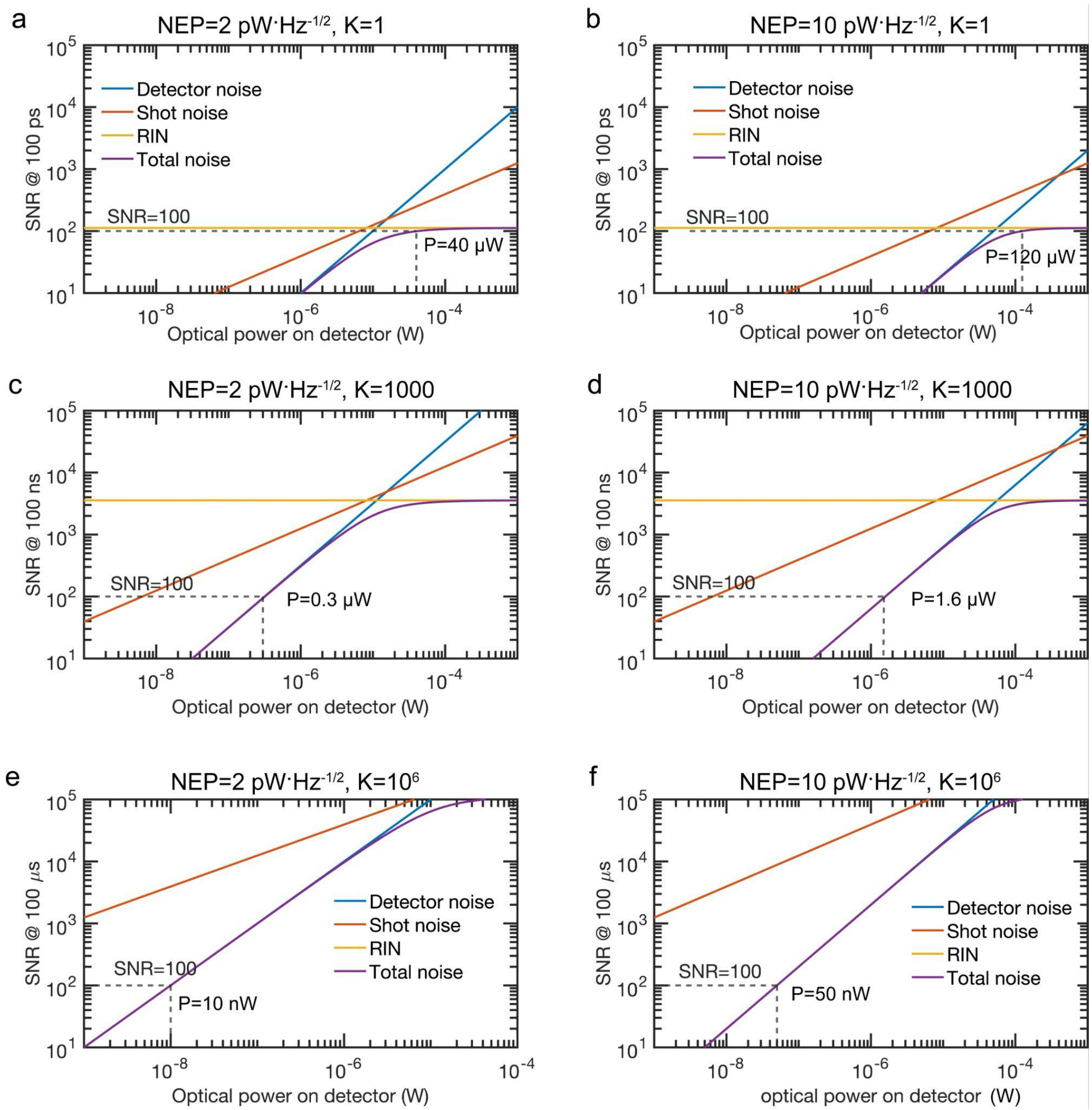


Fig. S2 the evolution of signal-to-noise ratio with input optical power. The power/PD requirement is labeled to reach a SNR=100 at each scenario. The SNR is calculated at the laser frequency of 195 THz, η=0.9, R=10 GS/s and laser RIN (β)=135 dbc·Hz$^{-½}$ a. the detector noise equivalent power NEP=2 pW·Hz$^{-½}$, acquisition time T=K/R=100 ps. b. NEP=10 pW·Hz$^{-½}$, T=K/R=100 ps. c. NEP=2 pW·Hz$^{-½}$, K=1000, T=K/R=100 ns. d. NEP=10 pW·Hz$^{-½}$, K=1000, T=K/R=100 ns. c. NEP=2 pW·Hz$^{-½}$, K=10^6, T=K/R=100 µs. d. NEP=10 pW·Hz$^{-½}$, K=10^6, T=K/R=100 µs.

Table S1 Power budget for different experimental conditions

Power budget	Detector NEP=2 $pW \cdot Hz^{-1/2}$			Detector NEP=10 $pW \cdot Hz^{-1/2}$		
	K=1	K=1000	K=10^6	K=1	K=1000	K=10^6
Power/detector	40 μW	0.3 μW	10 nW	120 μW	1.6 μW	50 nW
Power per laser for 1000 fanout (with 5 dB coupling loss [57])	120 mW	1 mW	30 μW	400 mW	50 mW	150 μW
Total optical power for 1000x1000 channels	120 W	1 W	30 mW	400 W	50 W	150 mW

3. TFLN Device characterization

Our MZMs are designed with sight imbalance path length between two arms, for the ease of calibration, which is not necessary in future full-system development. The free-spectral range is >150 nm, which is negligible in our demonstration with a total bandwidth of <5 nm.

The response of lithium niobate amplitude modulation is a sinusoidal function with increasing driving voltage. When encoding positive and negative weight valuers, the modulator is biased at the quadrature point with driving voltage of (V_{pp}=0.6 V), which is quasi-linear. Without calibration, the encoding accuracy is 6 bits precision due to the slight nonlinearity. The slight nonlinearity at close extinction point can be pre-calibrated using an inverted transfer function to reach a linearity up to >99.6%, which corresponds to 8 bit precision.

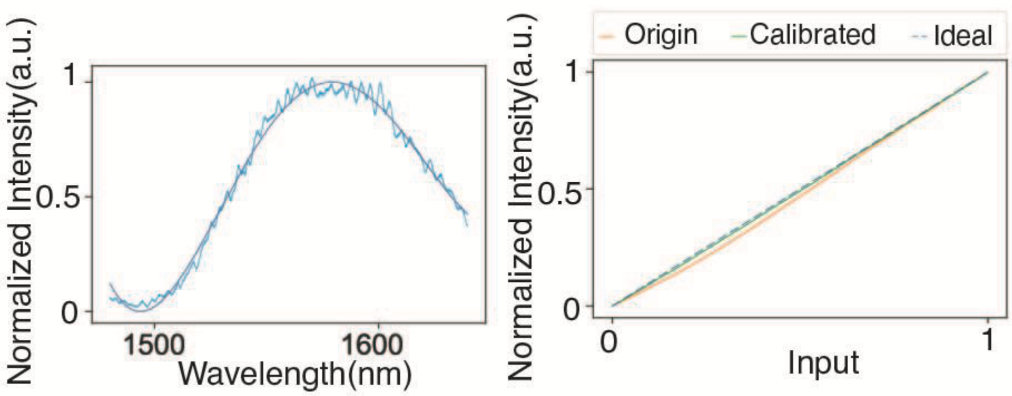


Fig. S3. TFLN MZM characteristics, left: the free spectrum range of a quasi-balance arm MZM, right: Linearity of the transfer function of the MZM, without calibration, we achieved 6 bits of encoding precision. With calibration, the standard deviation of the residue between the calibrated transfer function and the ideal transfer function is less than 0.4%.

4. Data Encoding Accuracy

We investigate the encoding accuracy of our opto-electro transceiver, i.e. VCSEL and TFLN modulator, by applying random data distributed in the range of [0,1]. Due to the linearity of VCSEL response, when driving in the near-threshold region, we achieved 8-bit precession without extra calibration at low clock rates. With the linear calibration of MZM response, we also achieve 8-bit precession encoding accuracy with a single output TFLN MZM. Limited by our photodetectors and the measurement electronics, the encoding accuracy drops to 6 bits when the clock rate goes up to 10 GHz.

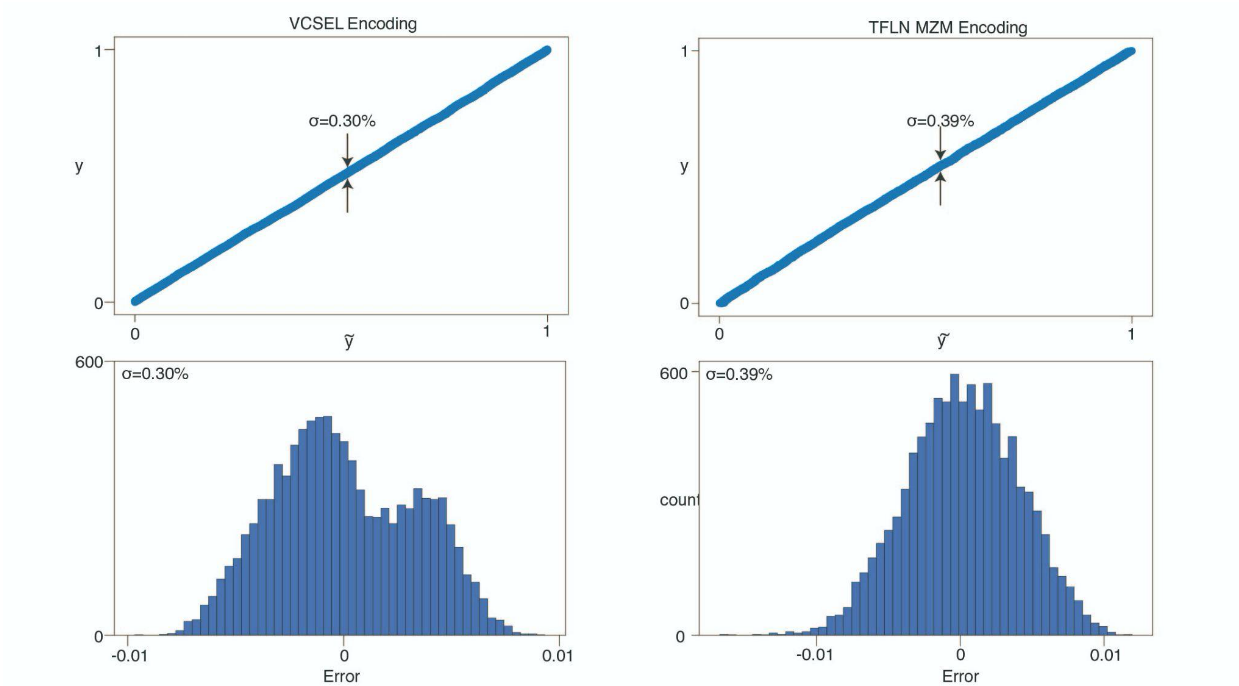


Fig. S4. Input-output voltage response accuracy of the VCSEL encoding and residuals (left) and TFLN MZM encoding (right) showing 8-bit precision.

We verify the VCSEL response accuracy when biasing around or slightly below the lasing threshold (Fig. S5 ~ S7). The VCSELs are biased at V_{dc}=1.2 V and the image data were flattened and encoded with a peak voltage of V_{ac}=0.6 V (same as our experimental conditions). The encoding results match well with the ground truth. For an example image "0", we achieved the encoding errors of 2.26%, 2.3% and 3.11%, respectively, at the data rate of 1 GHz, 5 GHz, and 10 GHz. The decrease in accuracy at 10 GS/s is due mainly to the impedance of drivers and VCSELs, which leads to electronic interference ripples.

The encoding of image "1" and "3" at 10 GS/s exhibit errors of 2.3% and 2.5% respectively (Fig. S6). An average error of 2.78% over the 10 consecutive images (Fig. S7) is observed, corresponding to 5~6 accuracy, which is sufficient for our computing tasks. We didn't experience degradation of accuracy due to the power and spectrum instabilities.

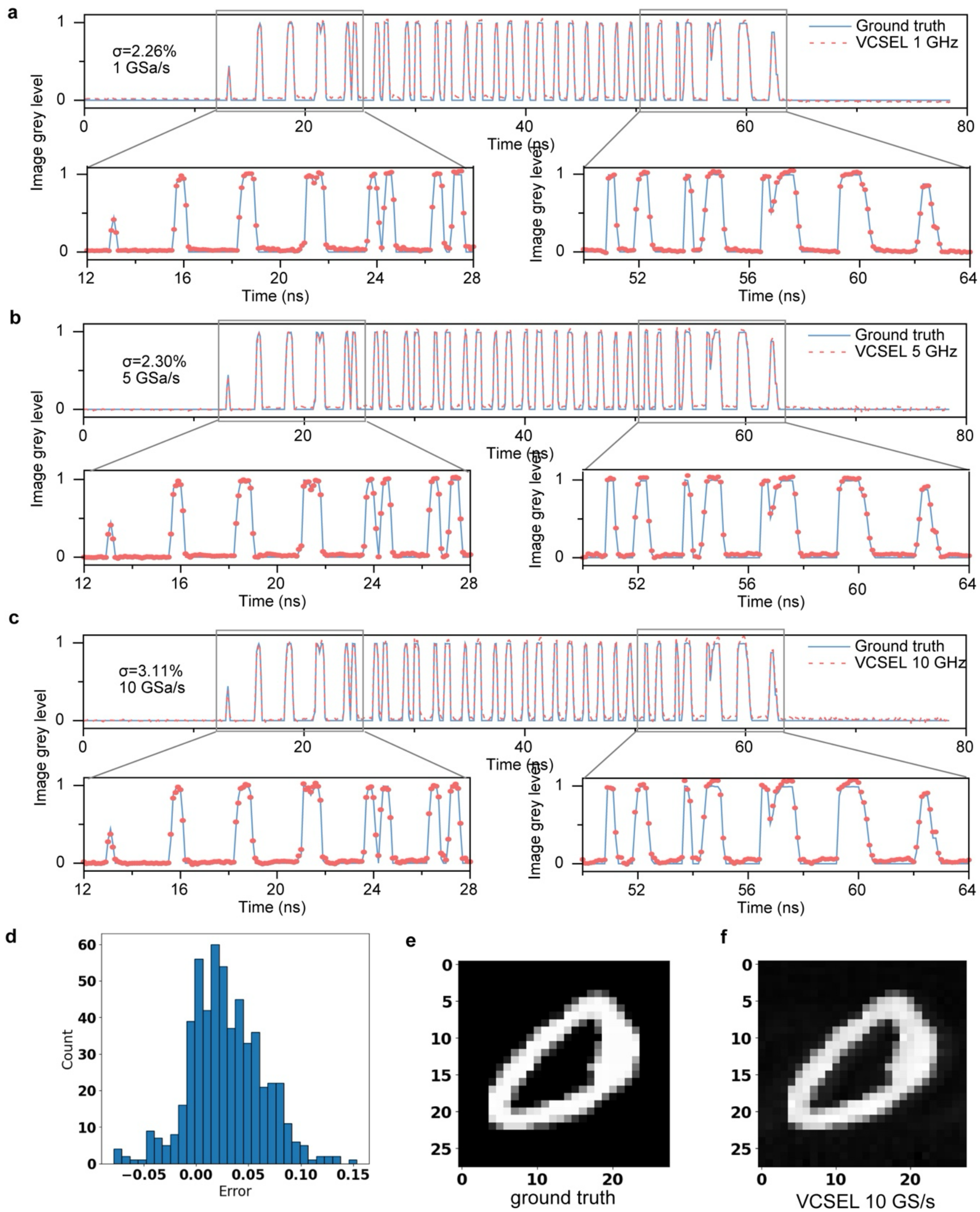


Fig. S5 Verification of image encoding accuracy with VCSELs biasing slightly below lasing threshold for ReLU nonlinearity. **a-c**. The encoding time trace and statistical errors achieved 2.26%, 2.30% and 3.11% at the data rates of 1 GS/s, 5 GS/s and 10 GS/s, respectively. d. error distribution at 10 GS/s. e-f. comparison of VCSEL encoded image with ground truth.

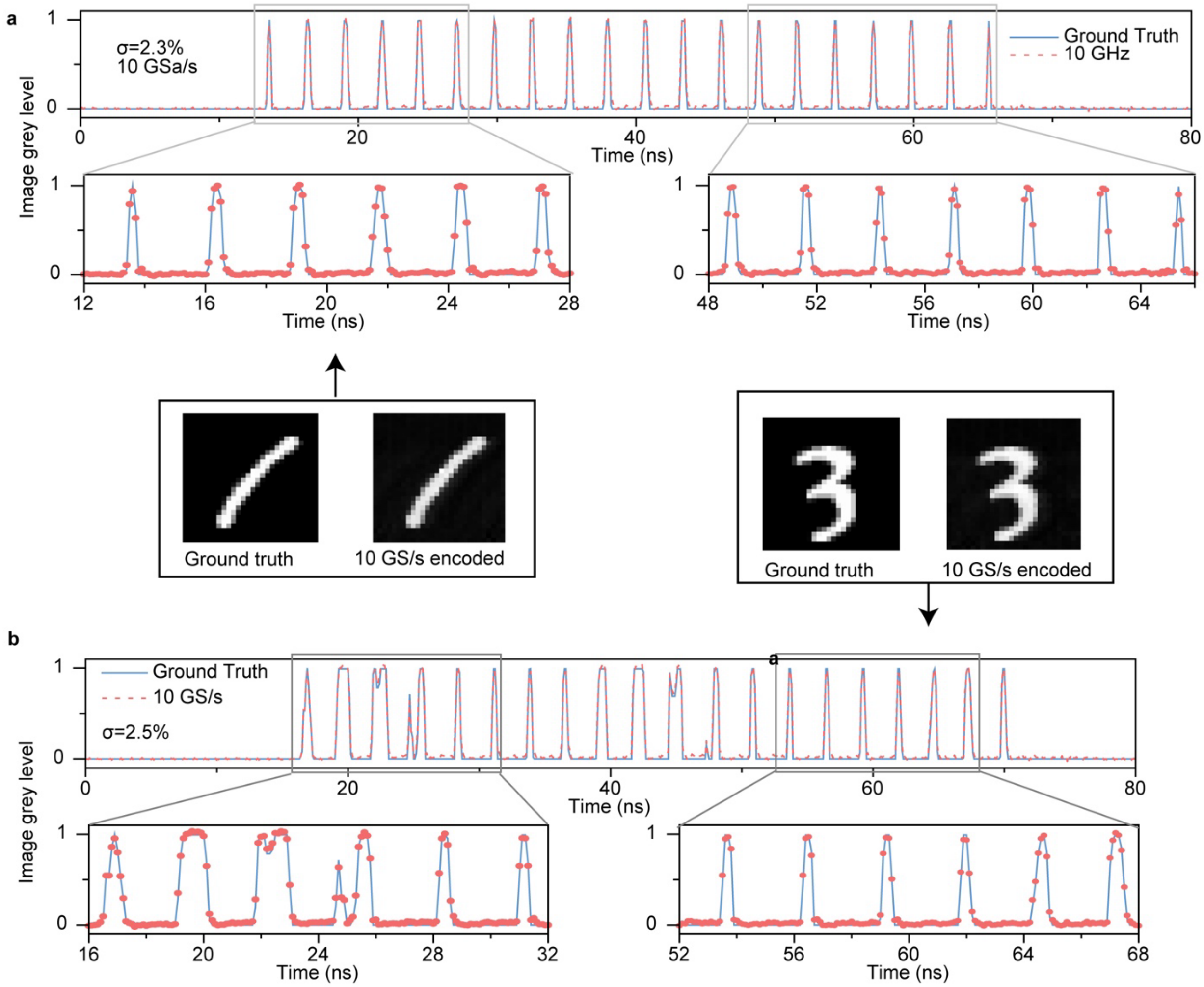


Fig. S6 examples of VCSEL encoding random digit images "1" and "3" using VCSEL at the data rate of 10 GS/s. The result shows 2.3% and and 2.5% errors compared to their ground truth.

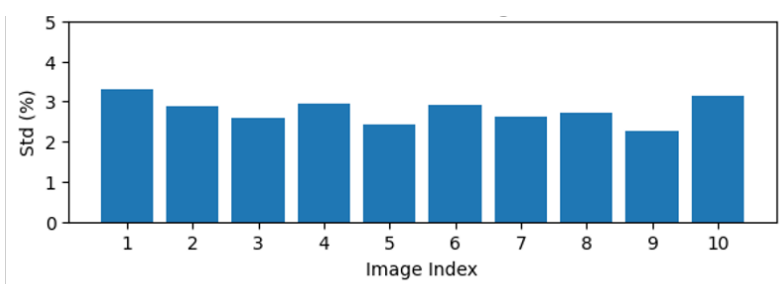


Fig. S7. VCSEL encoding accuracy for consecutively 10 images achieved an average statistical error of 2.78%.

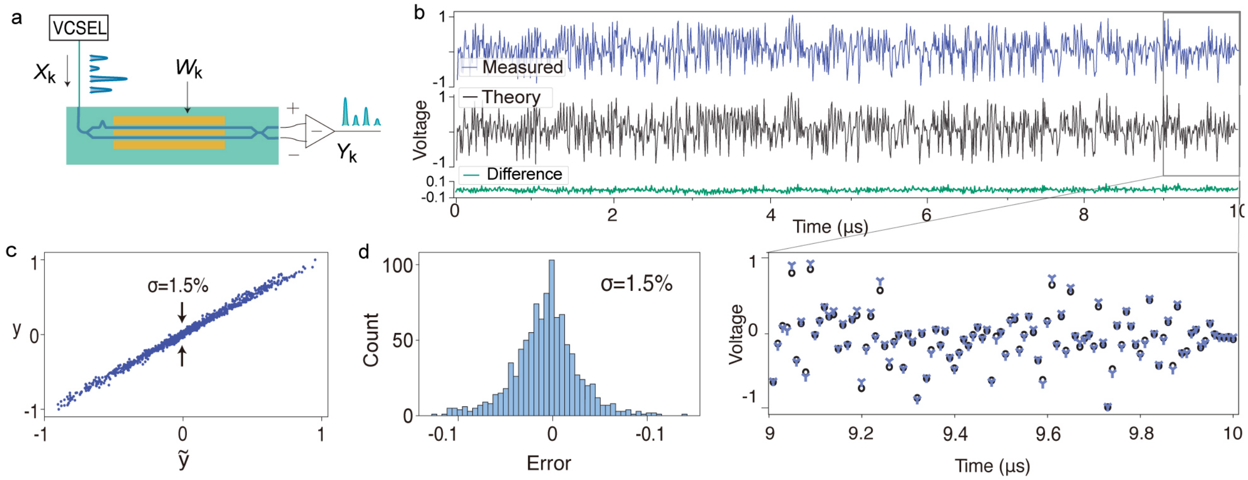


Fig. S8 Experimental verification of HITOP computing accuracy at 100 MS/s. a. A vector computing unit consisting of a VCSEL and a modulator for cascaded modulation. b. Experimental results of multiplication of 2 random-distributed vectors with negative and positive weights. c. correlation of expected values and experimental multiplication results. d. histogram of the multiplication errors between ground truth and experimental results.

5. Neural network models for different tasks

To benchmark the performance of our HITOP system, we trained several neural network models for 3 different datasets. All the models are built using PyTorch. Adam optimizer is deployed for training all these models with different electronic hardwares (CPU, Nvidia GeForce 3070, Nvidia Tesla V100). During the training process, all the parameters are restricted within a range of [-1,1] to match our HITOP system. Additionally noise is applied to enhance the robustness of the model. The hyperparameters and the corresponding digital and experimental accuracy are listed in Table.S2.

Table S2. HITOP experimental demonstrated model structure and accuracy

Dataset	Demonstration	model size	Digital accuracy	HITOP Accuracy	Clockrate	# of Channels
MNIST	Fig. 4d	28x28→10	91.7%	91%	10 GS/s	1 modulator, 1 VCSEL
	Fig. 4e	28x28→100→10	95.9%	95.5%	100 MS/s	7 modulators, 7 VCSELs (simultaneously 2 TFLN modulators and 2 VCSELs)
Fashion MNIST	Fig. 5c	28x28→100→10	94.3%	91.8%		
EMNIST Letters	Fig. 5d	28x28→500→10	93.4%	93.1%		

In addition, we analyzed the required computing precision to implement our AI models. We observed the statistical accuracy of our image classification, by quantizing the output product of each layer to be B=1 bit, 2 bits, 3 bits, ..., 16 bits, respectively, where the signals are normalized and before approximating to the closest level of $1/2^B$. The results are listed in Table S3 and plot in Fig. S4. The fact that our experimental results are close to the maximal digital accuracy indicates that our computing accuracy is higher than 4~5 bits.

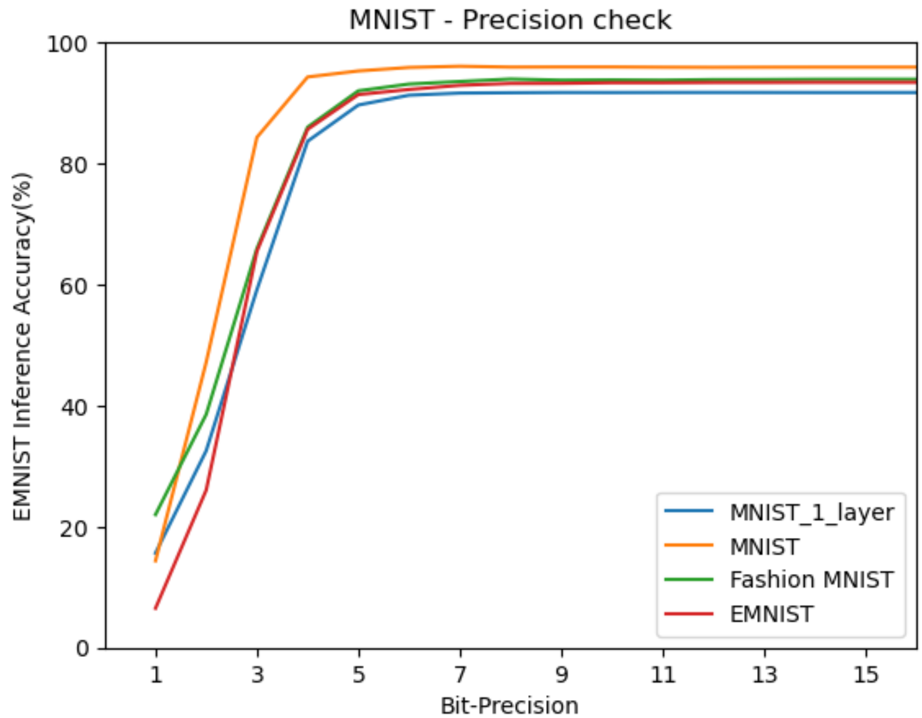


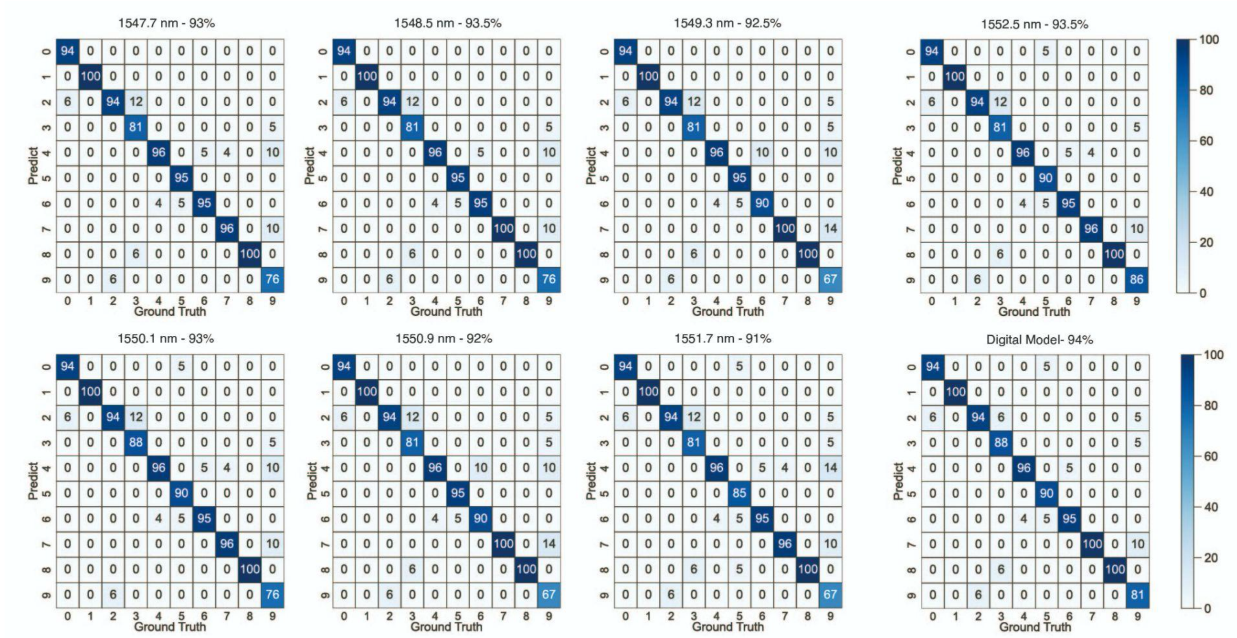
Fig. S9 Model accuracy at different quantization bit precisions. The model accuracy increases with higher computing bit precision and reaches maximum at 4~5 bits of precision.

Table S3. The accuracy of our AI models at different computing precision.

Bit-Precision	MNIST_1_layer	MNIST	FashionMNSIT	EMNIST
1	15.6%	14.4%	22.0%	6.5%
2	32.5%	47.2%	38.6%	26.0%
3	59.1%	84.3%	65.9%	65.5%
4	83.7%	94.3%	86.0%	85.6%
5	89.6%	95.2%	92.0%	91.4%
6	91.2%	95.9%	93.1%	92.2%
7	91.6%	96.0%	93.5%	92.9%
8	91.7%	95.9%	93.8%	93.2%
9	91.7%	95.9%	93.8%	93.3
10-16	91.7%	95.9%	93.8%	93.4%

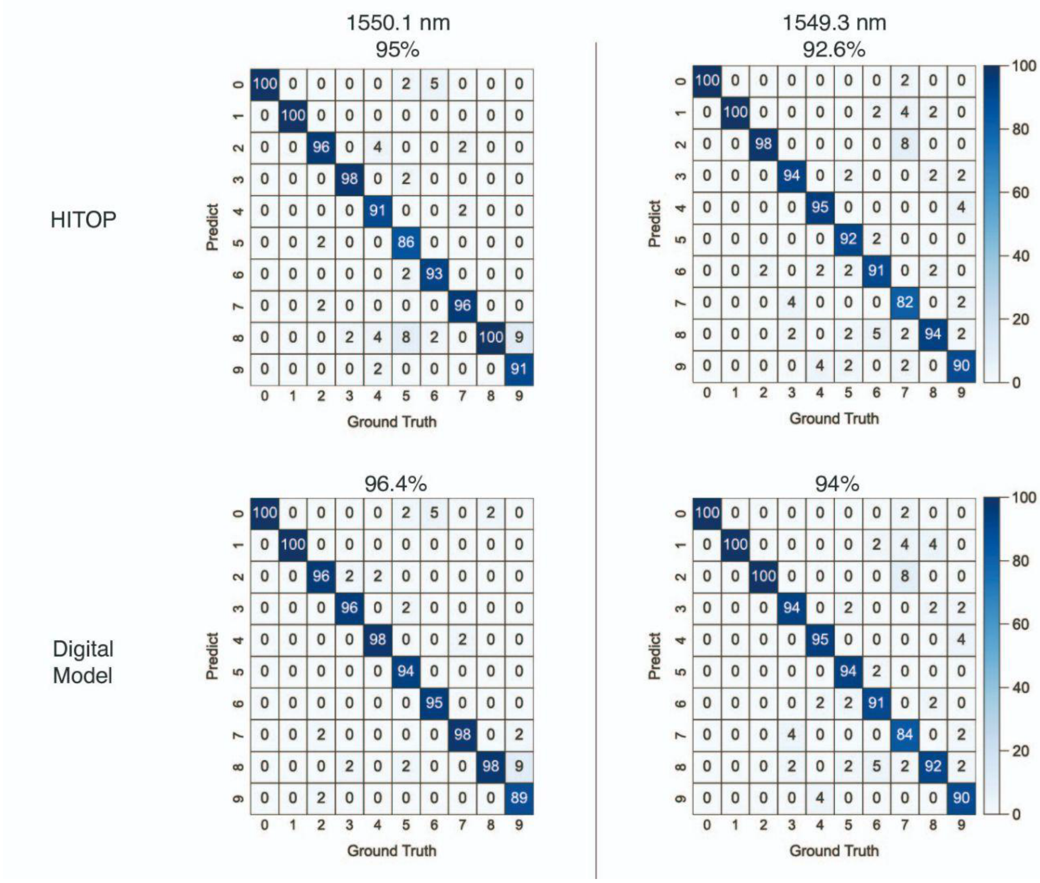
6. MNIST digit classification for multiple channels

To benchmark the optical bandwidth of HITOP. We randomly choose 200 MNIST images from the test set and perform inference on them using HITOP at 7 wavelengths with the 1-layer ANN. The wavelengths are selected to match the corresponding channels of our DWDM. The model has an inference accuracy of 94% while HITOP shows a similar performance at 7 wavelengths from 91% to 93.5%. Which is from 96.8% to 99.5% compared to the electronic inference.



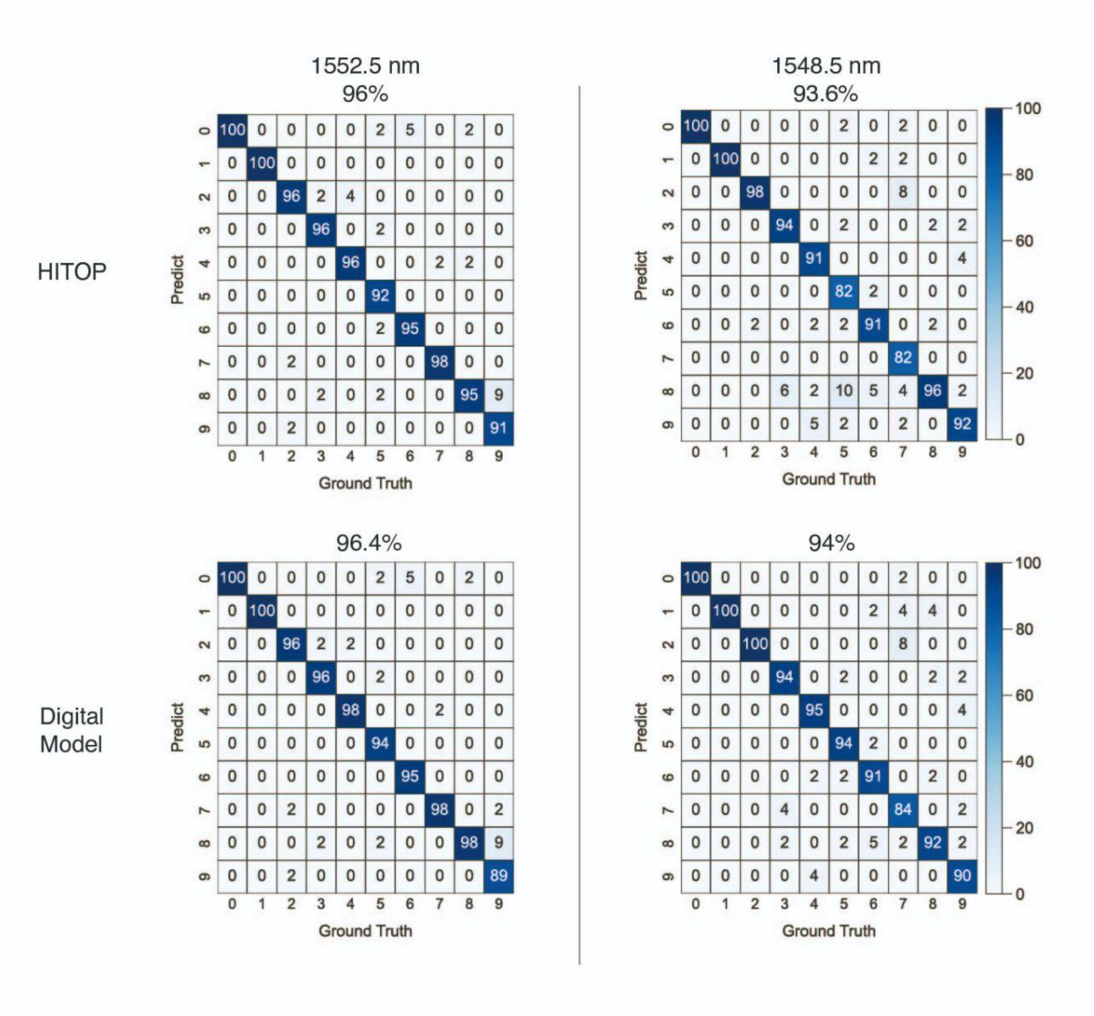
Figs. S10-1 HITOP confusion matrices with 200 MNIST test images at 7 wavelengths.

To verify the parallel computing accuracy of HITOP, we run the MNIST digit inferences with HITOP using 2 wavelength channels simultaneously. 2 sets of 500 MNIST images are randomly picked from the MNIST test dataset. Using the 2-layer DNN model, the electronic processor yields a 96.4% inference accuracy for the first set and 94% for the second set. Firstly, we sent to 2 VCSELs operating at 1549.3 nm and 1550.1 nm, which are the adjacent channels of the DWDM. HITOP shows 95% classification accuracy for the first set and 92.6% for the second set, which is 98.4% and 98.5% compared to the electronic inference.



Figs. S10-2 HITOP confusion matrices with 200 MNIST test images at 2 adjacent wavelengths.

We also perform the same inference using 2 further channels in our system at 1548.5nm and 1552.5nm. HITOP shows similar performance with 96% inference accuracy for the first set and 93.6% for the second set, which is 99.6% and 99.6% compared to the electronic inference.



Figs. S10-3 HITOP confusion matrices with 200 MNIST test images at 2 further wavelengths.

All these benchmark tests verify that HITOP has the parallel computing ability and could have a large data throughput when operating at more wavelengths without a large deviation of the computing accuracy.

6. Wideband modulation of TFLN modulators

Integrated electro-optic modulator based on thin film lithium niobate can support high speed modulation across a wide optical bandwidth. In figure S11, we show the simulated 3-dB EO bandwidth from 1300 nm to 1800 nm based on the extracted microwave and optical properties of the TFLN EO modulator demonstrated in this paper. Due to the fact that the optical group index only shows small variation across the wide optical bandwidth, the TFLN EO modulation can support > 42 GHz EO modulation bandwidth across 1300-1800nm with a small variation of 0.3 GHz.

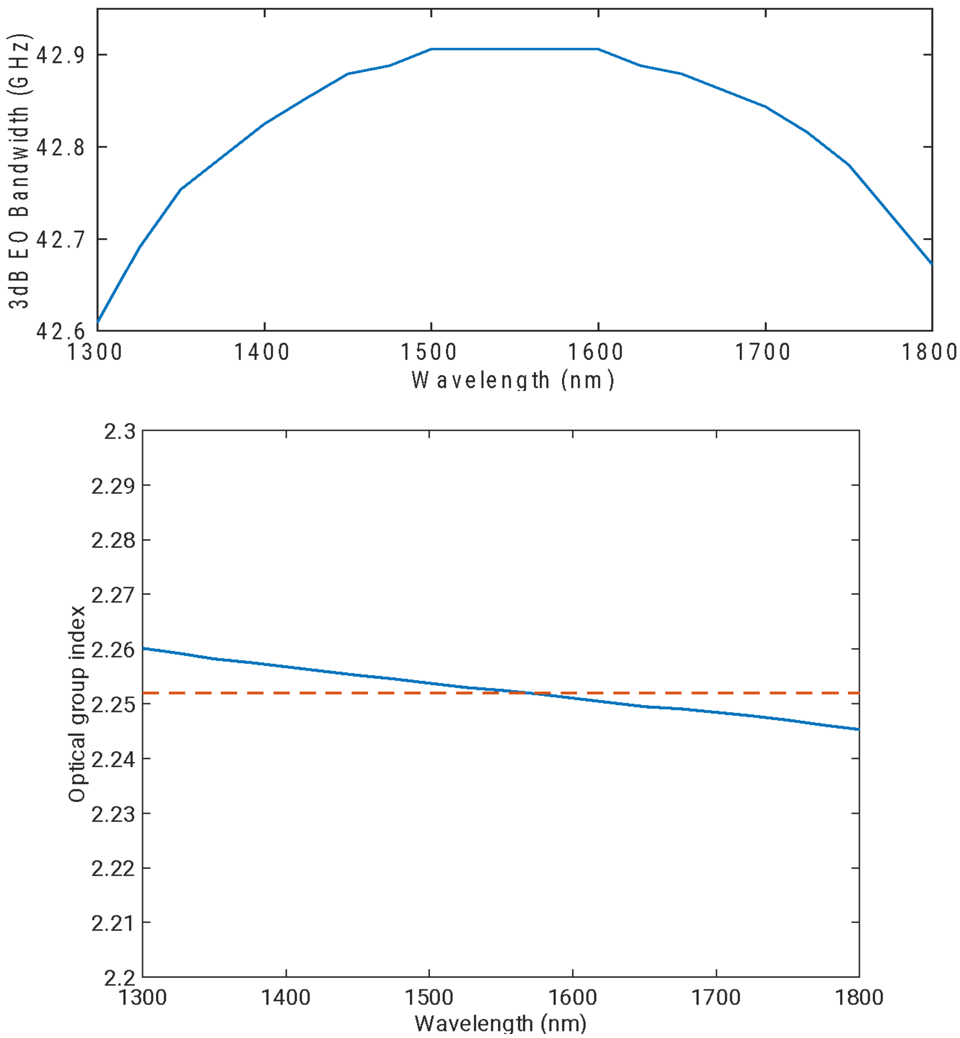


Figure S11 The simulated 3-dB EO bandwidth (top) and optical group index (bottom) as a function of optical wavelength. The EO response simulation is based on the device parameters of the 2-cm-long TFLN modulator device used in this paper, including measured microwave property of the electrodes with 0.7 dB/cm/GHz$^{0.5}$ loss and microwave index of 2.252 (dashed red line). Due to the small variation of the optical group index across the 500-nm optical bandwidth, the 3-dB EO response bandwidth variation is less than 0.3 GHz (from 42.6 GHz at 1300nm to 42.9 GHz at 1550nm).

We further simulation the DC half-wave voltage based on the electrical and optical mode profile [53] at different optical wavelength, shown in Fig S12. Due to the similar EO response curve, both the half-wave voltages at DC and 30 GHz microwave frequency are dominated by the wavevector difference at different optical wavelength. We acknowledge that the difference of the half-wave voltages needs to be pre-calibrated for each VCSEL wavelength.

Finally we provided the measurement results of the same EO modulator at 1.05 μm in Fig S13, which further supports our argument and simulation result that the TFLN modulator can support high speed electro optic modulation across more than 500 nm optical wavelength range. In our paper, 1000 channels each spaced by 50 GHz centered at 1550 nm corresponds to a span from 1375 nm to 1785 nm.

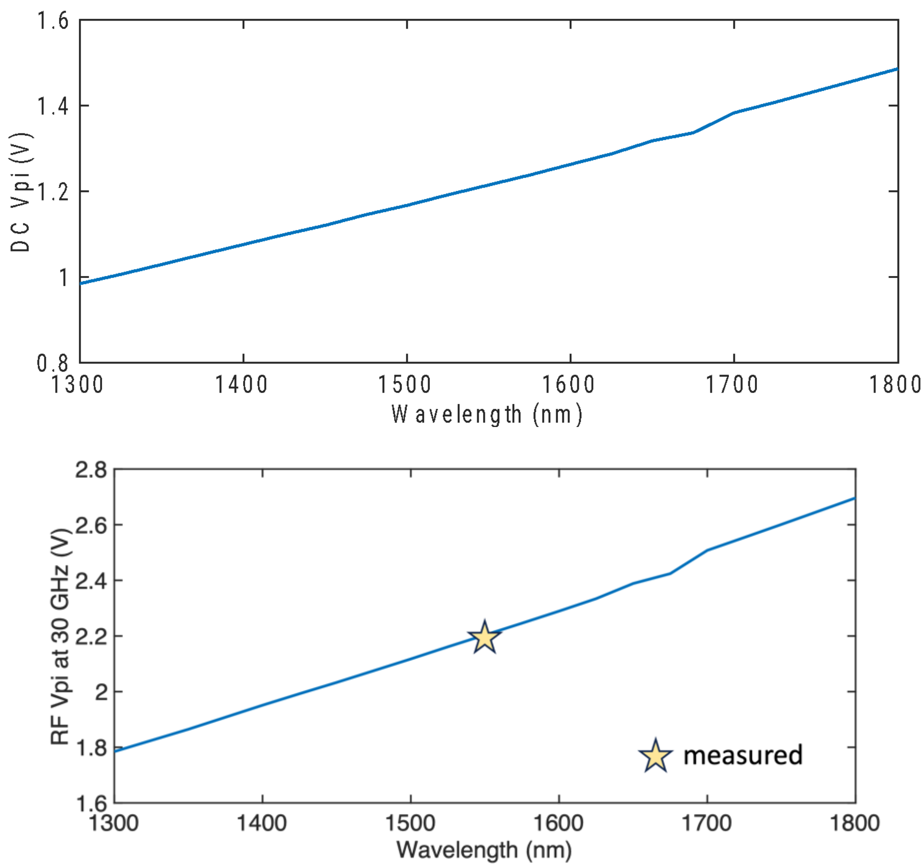


Figure S12 The half-wave voltage at DC (top) and at 30 GHz microwave frequency (bottom) as a function of the optical wavelength.

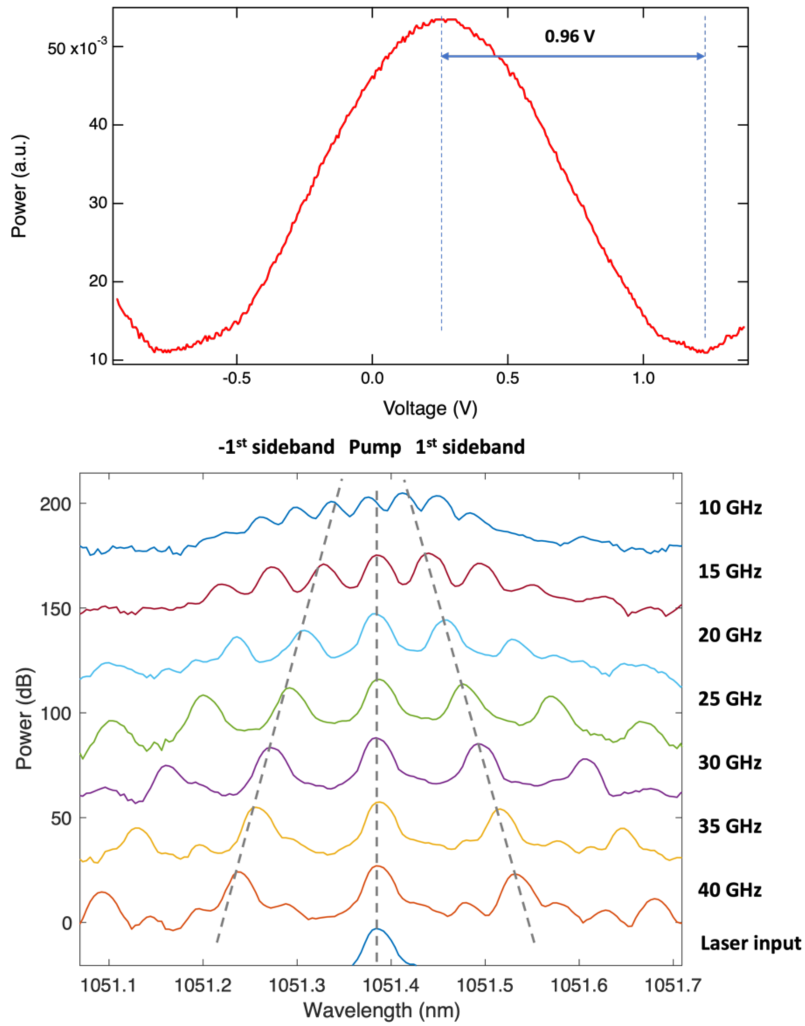


Figure S13 (a) The measured half-wave voltage of 0.96 V at near DC frequency (1MHz) at the optical wavelength of 1.051 μm, as compared to the half-wave voltage of 1.2 V at 1.55 μm. (b) the optical spectrum via driving different microwave frequencies at the optical wavelength near 1 μm. We can extract the half-wave voltage at 30 GHz to be 1.96V, as compared to 2.2 V at 1.55 μm. Both measurement matches well with our simulation results, and it shows the TFLN EO modulator can support a wide range of the optical operating wavelength.