Review and perspective on ultrafast wavelength-size electro-optic modulators
Ke Liu¹,², Chen Ran Ye¹, Sikandar Khan¹, and Volker J. Sorger¹,*

¹ Department of Electrical and Computer Engineering, School of Engineering and Applied Science, George Washington University, Washington, DC 20052, USA; ² The Key Laboratory of Optoelectronics Technology, Ministry of Education, Beijing University of Technology, Beijing, 100124 P.R. China

Received: 11 September 2014, Revised: 1 January 2015, Accepted: 14 January 2015
Published online: 10 February 2015

Key words: Electro-optic modulators, surface plasmon-polariton, light matter interaction, ultrafast.

As electronic device feature sizes scale-down, the power consumed due to onchip communications as compared to computations will increase dramatically; likewise, the available bandwidth per computational operation will continue to decrease. Integrated photonics can offer savings in power and potential increase in bandwidth for onchip networks. Classical diffraction-limited photonics currently utilized in photonic integrated circuits (PIC) is characterized by bulky and inefficient devices compared to their electronic counterparts due to weak light matter interactions (LMI). Performance critical for the PIC is electro-optic modulators (EOM), whose performances depend inherently on enhancing LMI. Current EOMs based on diffraction-limited optical modes often deploy ring resonators and are consequently bulky, photon-lifetime modulation limited, and power inefficient due to large electrical...
Abstract As electronic device feature sizes scale-down, the power consumed due to onchip communications as compared to computations will increase dramatically; likewise, the available bandwidth per computational operation will continue to decrease. Integrated photonics can offer savings in power and potential increase in bandwidth for onchip networks. Classical diffraction-limited photonics currently utilized in photonic integrated circuits (PIC) is characterized by bulky and inefficient devices compared to their electronic counterparts due to weak light–matter interactions (LMI). Performance critical for the PIC is electro-optic modulators (EOM), whose performances depend inherently on enhancing LMIs. Current EOMs based on diffraction-limited optical modes often deploy ring resonators and are consequently bulky, photon-lifetime modulation limited, and power inefficient due to large electrical capacitances and thermal tuning requirements. In contrast, wavelength-scale EOMs are potentially able to surpass fundamental restrictions set by classical (i.e. diffraction-limited) devices via (a) high-index modulating materials, (b) nonresonant field and density-of-states enhancements such as found in metal optics, and (c) synergistic onchip integration schemes. This manuscript discusses challenges, opportunities, and early demonstrations of nanophotonic EOMs attempting to address this LMI challenge, and early benchmarks suggest that nanophotonic building blocks allow for densely integrated high-performance photonic integrated circuits.

Review and perspective on ultrafast wavelength-size electro-optic modulators
Ke Liu¹,², Chen Ran Ye¹, Sikandar Khan¹, and Volker J. Sorger¹,∗

1. Introduction
In the past decade, photonic technologies have become universal in global data communications [1, 2]. Ever-increasing data bandwidth, falling power consumption requirements, and cost margins of onchip optical interconnects [3, 4] have set a photonic roadmap trend for photonic components [5, 6]. Overall, the three main driving forces stand out for making a case of photonic integration, include (i) power consumption, (ii) available bandwidth, and (iii) functionality per area (i.e. packing density). Miller recently discussed roadmap-like device requirements for optical interconnects and argued that the power consumption of logic (i.e. central processing unit (CPU)) input/output (I/O) is a major reason for replacing electrical interconnects with optical ones onchip [7]. In detail, the single device power requirement has to be about 10 fJ/bit or less by the year ~2020, which is approximately two orders of magnitude below the current power levels. Breaking down the power consumption of computing, for instance, within data centers shows that it is not the computational logic and will be the limiting factor moving forward, but the data I/O between the CPU, random-access memory (RAM), graphics processing unit (GPU), cache, and other onchip or on-board parts [4,8,9] (Fig. 1a). It is interesting to mention that the release of the 2005 International Technology Roadmap for Semiconductors (ITRS) has pointed to optical interconnects as a potential solution to address the ever-increasing I/O requirements for about one decade now [10]. As the photonics “frontier” gets closer to the computing core, the number of interconnect links will increase, thus integration density is a critical factor to consider. However, the latter will be challenging with current technology, as microrings are typically used for both modulators and (de)multiplexers (Fig. 1b). Generally, photonic interconnect links consist of a light source, (de)multiplexers if wavelength-division-multiplexing is used, modulators, and photodetectors including their amplifiers (Fig. 1b). Placing the discussion of EOMs into the context of links is important as emerging technologies utilize a) a variety of waveguide types, and b) subdiffraction-limited optical modes, which could lead to impedance mismatches between passive and active devices. However, the promise of using optics versus electronics lies in addressing the arising bottleneck between
computation and communication (Fig. 1a). While there is much research pursued in utilizing silicon-on-insulator (SOI) or III-V-based platforms for photonic-electric hybrid networks, more investigations should be focused on incorporating emerging nanophotonic devices into advanced networks.

This paper concentrates on a single component within a photonic link – the electro-optic modulators (EOMs). Within the aforementioned power and bandwidth context, future EOMs must satisfy a range of challenging requirements, consisting of high speed, low drive power, and compact footprints. Here, we discuss scaling laws, concepts, and examples of EOMs that push this performance boundary for the next-generation devices. An interesting new photonic design scheme is to hybridize not only electronic with optical components, but also utilizing nanophotonic building blocks such as plasmonics and optical subdiffraction-limited components towards designing the high-performance networks (Fig. 1c).

2. Scaling laws

EOMs are inherently dependent on the strength of the interaction between light and matter. This interaction can be fundamentally described by an induced dipole moment or polarization, $P$, which acts upon the functional switching material of the modulators (Fig. 2). The total polarization is the sum of the linear, $P_L$, and nonlinear, $P_{NL}$, polarization, with the latter depending on the higher-order terms of electric field strength. The key for an ultrafast, compact, and power-efficient EOM is to make use of field-enhancement techniques that increase the nonlinear polarization. For instance, narrow gap slot waveguides [13,14] or plasmonic...
propagating modes allow for electric field intensity enhancements of up to ~30 times. Therefore, fundamentally a linear effect may be enhanced by that factor, or a 2nd (3rd) order nonlinear effect by 2–3 (3–4) orders of magnitude.

The design choices for waveguide modal field enhancement [15] range from slot-waveguides [13], plasmonic metal–insulator–metal (MIM) [16], insulator–metal–insulator (IMI) [17], channel plasmon [18], gap waveguides [19], metal grooves [20] to hybrids thereof [21, 22]. To illustrate the potential we show the nonlinear effect enhancement for one of these strong waveguide confinements, namely a hybrid plasmon-polariton (HPP) waveguide mode [23–25] (Fig. 3). This mode offers many advantages relating to photonic integration and active optoelectronics; (a) deep-subdiffraction limited modes (λ^2/400), (b) metal–oxide–semiconductor (MOS) synergy, i.e. the plasmon mode confining metal can simultaneously be utilized as an electrical contact and heat sink, while allowing access to semiconductors for optoelectronic functionality, (c) SOI–platform integration such for low routing losses, and (d) it provides an integrated electrical contact that may also be used as a heat sink. We see that if the optical confinement of a waveguide mode is increased (i.e. thinner gap height), the nonlinear response significantly strengthens (Fig. 3). In contrast, diffraction-limited silicon core does not experience such enhancement (dashed red line in Fig. 3c). We highlighted the three sections in Fig. 3c to guide our discussion; the optical mode in the yellow-shaded region C is mainly concentrated in the silicon core similar to a nonplasmonic, pure SOI waveguide. Thus, the nonlinear effect is relatively weak due to low electric field strengths (vanishing mode hybridization). In the region A the mode hybridization has completed the increase in the optical confinement leading to the signal strength that follows a straight line in the semilog plot. The steep slope of section B can be understood as a gap-height range where the strong optical confinement of the hybrid mode starts to become dominant over the SOI bulk waveguide mode. This is contrasted by flatter slopes in sections A and C, which have different origins; in C the weakly confined SOI bulk mode dominates, whereas for polarization charges of the index jump across the gap become screened for the low gap height limit (section A).

High-speed EOMs need to be power efficient and ultra-compact. Next, we provide a brief discussion highlighting the physical insights of LMI-enhanced EOMs as they relate to the device’ power consumption (i.e. energy-per-bit, \( E_{\text{bit}} \)).

\[
E_{\text{bit}} = \frac{1}{2} CV_b^2 = \frac{1}{2} V_b Q_{\text{charge}} \propto \frac{A_{\text{mode}}}{V_{\text{mode}}} \propto F_p \propto \frac{V_{\text{mode}}}{Q}
\]

\[
F_p \propto \frac{Q}{V_{\text{mode}}}. \quad (2)
\]

**Figure 2** Understanding the fundamentals of light–matter interaction and defining the underlying performance of electro-optic modulators. The nonlinear polarization depends on the higher-order electric field densities inside the device.

**Figure 3** (a) The hybrid waveguide can be seamlessly integrated with a SOI platform. For a passive HPP waveguide the materials such as silicon oxide and aluminum can be deployed with 100% CMOS compatible process. For the functional device of a photonic integrated circuit, the thin oxide layer may be replaced by a material with a strong nonlinear coefficient, \( n_2 \). (b) The electric field strength of a SOI-based hybrid plasmon waveguide mode shows the strong field enhancement inside the thin gap layer coinciding with the nonlinear material. Simulation parameters: width = 500 nm, heights: functional region = 20 nm, silicon = 250 nm, aluminum = 150 nm. Refractive indices: \( n_{\text{Si}} = 3.5–0.001i, n_{\text{Al}} = 2.1, n_{\text{Ge}} = 4.51–15.4i, \lambda = 1550 \text{ nm} \). (c) Nonlinear signal strength vs. HPP gap height. Compared to pure SOI waveguide (red dashed line), the strong optical field can produce a 500-fold enhanced nonlinear signal. This allows the LMI device length to be reduced by the same factor creating a path for (sub)micrometer-size photonic components. The signal strength is computed by the product of \( |E_{\text{max}}|^3 \) and the optical nonlinear coefficients, \( n_2 \text{Si} \) and \( n_2 \text{As}_2\text{Se}_3 = 440 \) and \( 1200 \times 10^{-20} \text{ m}^2/\text{W} \), respectively. Panels 1–4 on the right are the corresponding normalized electric field strengths for the HPP gap heights at \( h = 8, 15, 30, 100 \text{ nm} \), respectively. Notice that the field strength in panel 1 is chosen to be saturated (white spot), and serves as the normalization field for the larger gap heights.
The energy-per-bit of an EOM is customarily quoted as \( \frac{1}{2} CV^2 \), where \( C \) is the device capacitance, and \( V \) is the driving voltage [26]. This expression takes the resistive energy loss of the modulator device itself into account during the charge–discharge cycles, ignores the power consumed by the driver circuit, and hence provides a lower bound for the switching energy [27]. For a charge-based EOM (e.g. plasma dispersion effect, free carriers) the energy-per-bit is proportional to the total accumulated charge, \( Q \), times the applied voltage, \( V \). Comparing the modulation efficiency of a cavity-based EOM (e.g. microrings) with a noncavity device (i.e. linear) the cavity’s finesse \( F \) reduces the power consumption literally \( \text{(Eq. (1))} \) [28]. Furthermore, the smaller the optical mode can be made, the more efficient the modulator becomes. The latter is interesting, as it shows the need to confine the optical mode to tiny physical spaces.

A brief summary of the latter is provided below. Going back to the energy-per-bit (Eq. (1)), the relationship between the 4th and 5th is interesting as it bears the underlying scaling laws or EOMs; term 4 states that introducing a resonant enhancement reduces the length of the modulator by a scaling factor proportional to the cavity’s finesse \( F \) [28], i.e. \( L_{\text{cavity}}/L_{\text{non-cavity}} = \frac{1}{2}F \), where, \( F = \Delta \nu_{\text{FSR}}/\Delta \nu_{\text{FWHM}} \). \( \Delta \nu_{\text{FSR}} \) is the free spectral range \( \text{(FSR)} \), and \( \Delta \nu_{\text{FWHM}} \) is the resonant peak’s linewidth \( \text{(i.e. full width at half-maximum \text{(FWHM)})} \). At the same time the finesse is proportional to the cavities’ \( Q \)-factor \( (Q = \nu_{\text{res}}/\Delta \nu_{\text{FWHM}}, \nu_{\text{res}} \) is the frequency at the resonant peak). Thus, together with the optical mode area, the 4th term is inversely proportional to the optical mode volume divided by the \( Q \)-factor, which is inversely proportional to the Purcell factor, \( F_P = \frac{Q}{V_{\text{mode}}} \) [29], where \( V_{\text{mode}} \) is the effective mode volume in units of cubic half-wavelengths. While the Purcell factor has been used in recent years to discuss laser physics of nanoscale devices [30–35], it was not brought into the context of modulators until now. Another way to interpret the Purcell factor is to term it as “optical concentration factor” where the aim is to enhance the interaction time of the electromagnetic field with the electronic wavefunction \( \text{(i.e.} Q \text{ enhancement)} \), while increasing the optical field intensity.

It is therefore understandable that such research focused on investigating high \( Q \)-factor cavity irrespective of how large the optical mode volume may be. However, increasing the cavity size is not a viable solution as it limits modulation speed, requires a large wafer footprint, and introduces tuning overheads. In contrast, here we have shown that the modulation efficiency can alternatively be reduced by shrinking the optical mode volume, while accepting low cavity qualities. Interestingly, the latter approach is synergistic with increasing the obtainable modulation speed since the EOMs frequency response is often limited by the electrical capacitance \( (i.e. f_{\text{lim}} = 1/RC, R \) is the contact resistance, and \( C \) is the device capacitance).

It is worth pointing out that only in recent years have these scaling laws become significant due to the downscaling of modulators into wavelength-compact sizes [36–43].

<table>
<thead>
<tr>
<th>Active Volume</th>
<th>Modulator Dissipation</th>
<th>Optical Concentration Factor ~ Q/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 ( \mu )m(^3))</td>
<td>5 ( \mu )J</td>
<td>Not required</td>
</tr>
<tr>
<td>(100 nm(^3))</td>
<td>50 aJ</td>
<td>( \sim 10^2 )</td>
</tr>
<tr>
<td>(10 nm(^3))</td>
<td>0.5 aJ*</td>
<td>( &gt; 10^4 )</td>
</tr>
</tbody>
</table>

Applying this fundamental scaling law (Table 1) we find that for a modulator with size of about wavelength \( \text{(i.e.} \) active volume = 1 \( \mu \)m\(^3\)) \) does not require any Purcell enhancement because the mode is not diffraction limited, and no additional losses apply. Scaling the mode volume to a cube of 100 nm allows for tens of attojoule efficient devices; however, the requirement to provide strong optical confinement increases the Purcell factors to about 100. Such \( Q/V_{\text{mode}} \) values, while challenging, are close to the experimental demonstrations using photonic crystal cavities [44] and plasmonic waveguides [33]. For EOMs of quantum-dot dimensions the LMI enhancement requirements appear (at the moment) unfeasibly high. However, demonstration of such a device would be driven by only a few electronic charges.

Concluding this section, with onchip EOMs pushing for sub-fJ/bit efficiencies, LMI enhancements in the form of storing optical energy \( \text{(i.e.} Q \text{ factor)} \) relative to the optical mode volume \( (V_{\text{mode}}) \) are critically needed for continuously growing performance of optoelectronic devices. In terms of operational \( \text{(i.e. modulation)} \) speed of logic and telecom devices, Fig. 4 qualitatively compares the area of nanophotonic optoelectronics and EOMs discussed in this manuscript with other technological options [45]; in brief, pure electronics communication links onchip are fundamentally challenged and hence limited by capacitive charging of electrical wires [46]. Nonlinear photonics, on the other hand, may operate with extremely fast \( \text{(i.e.} \) fs) interaction times, however, due to weak LMIs they are inherently requiring bulky devices and have high power consumption. Subwavelength optics and potentially emerging material choices allow for ultrafast and efficient device designs beyond the current integrated-circuit interconnect delays.

3. Modulation mechanisms and limitations

Light signals can be modulated either in amplitude, phase, polarization, or frequency, and combinations thereof. In practice, most optical modulation is performed by amplitude modulation, since it is difficult for photodetectors to
distinguish a change in frequency or phase unless an interference technique is adopted. In order to construct an effective optical modulator the modification of its refractive index or absorption coefficient of a material system is required in a controllable fashion.

So far, the light modulation mechanisms may be divided into three “effect” categories, thermo-optic, electro-optic (EO), and electroabsorption. The thermo-optic effect results from a temperature-dependent refractive index of a material, such as in tuning the refractive index of silicon via resistive heating [47]. Although thermo-optic modulators (and switches) have several drawbacks such as significant power consumption, slow transition time (usually > 1 μs), the mechanism shows operational devices and subsystems onchip [48, 49]. The electro-optic effect includes Pockels and Kerr, and carrier density effect [50,51] in general. Electroabsorption devices involve the quantum-confined Stark effect (QCSE) [52–55], or the Franz–Keldysh effect (FKE) [56, 57]. The electro-optic effect originates from the interaction between a charge distribution (atom or molecule) and an external applied electric field. The electroabsorption devices such as a heterostructure formed by a small-bandgap material sandwiched between two layers of a larger-bandgap material, the Stark effect may be greatly enhanced by bound excitons. In this paper, the two mechanisms of electro-optic and electroabsorption effects will be reviewed since both are predominately utilized to form an ultrafast compact optical modulator.

Simply applying the aforementioned individual effect (i.e. Pockels, Kerr, and QCSE) is not sufficient to realize a high-performance electro-optic modulator. Figure 5 shows a schematic design flow of building an EOM. Different materials possess various optical properties, especially for those novel emerging materials such as graphene and transparent conductive oxide (TCO). Combining an EO mechanism with a functional switching (i.e. voltage-induced) material alters the refractive index of the material. In addition, many design features are developed to enhance those modulation effects through LMIs, including cavity resonators, surface plasmon-polariton (SPP) or hybrid waveguides, and slot-type waveguides.

An initial analysis towards designing an ultrafast electro-optical modulator is to compare the various modulation mechanisms. Naturally the different mechanisms have various advantages and drawbacks, and they should be considered in different operating environments and applications. In the following subsection, we focus on three dominant modulation mechanisms and their limitations on the realization of high-performance EOMs.

3.1. Pockels and Kerr effects

The Pockels and Kerr effects are two well-developed EO mechanisms, and have been used in actual device designs for many years. However, most devices based on both effects are phase-shifter-based modulators, such as the LiNbO_3-based designs, and InGaAsP/InP-based optical switches [58]. Phase modulation may also be configured for intensity modulators if the two interfering phase-shifted lightwaves create a destructively interfering intensity profile depending on their phase difference. Unfortunately, both effects are either weak (i.e. Δn ~ 10^-8 for the Kerr effect in silicon at an applied field of 10^4 V/cm [59]) or are not compatible with group-IV semiconductor materials such as the SOI platform [60, 61]. The Pockels effect, also known as a linear electro-optic effect, is an anisotropic effect proportional to an applied electric field. With the uniform electrical field it can be expressed by [62]

\[
\Delta n_{\text{pockets}} = \frac{1}{2} r_{ij} n_0 E
\]
Table 2 Calculation of refractive-index change due to Pockels effect and Kerr effect for various materials at 1550 nm wavelength. For the purpose of comparison an electric field $E = \frac{V}{d}$ is applied, including a constant voltage $= 1.0$ V and $= 50$ nm. Here, the Pockels coefficient, $\sim 1.68$ pm/V for InP, $\sim 1.6$ pm/V for GaAs, and $\sim 30$ pm/V for LiNbO$_3$ [62], and the Kerr coefficient, $k$, $\sim 5.8 \times 10^{-15}$ cm$^2$/V$^2$ for InGaAsP/InP [63], $\sim 7.7 \times 10^{-16}$ cm$^2$/V$^2$ for GaAs [63], and $\sim 2.3 \times 10^{-10}$ cm$^2$/V$^2$ for LiNbO$_3$ [64].

<table>
<thead>
<tr>
<th>Material</th>
<th>Silicon</th>
<th>InP</th>
<th>GaAs</th>
<th>LiNbO$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta n_{\text{Pockels}}$</td>
<td>$\sim 0$</td>
<td>$5.3 \times 10^{-4}$</td>
<td>$6.3 \times 10^{-4}$</td>
<td>$3.3 \times 10^{-5}$</td>
</tr>
<tr>
<td>$\Delta n_{\text{Kerr}}$</td>
<td>$10^{-5}$ @ 1.3 $\mu$m</td>
<td>$3.7 \times 10^{-4}$</td>
<td>$5.2 \times 10^{-5}$</td>
<td>$1.0 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

where $r$ is the Pockels coefficient, and $r_{ij}$ is the third-rank electro-optical tensor, represented by a matrix with 6 rows and 3 columns in a contracted notation. The form of the tensor $r_{ij}$ and the relations between the various elements can be derived from symmetry considerations. $n_0$ is the unperturbed refractive index, and $E$ is the applied electric-field strength. Table 2 shows the calculation of refractive-index change for various materials. Conventional unstrained silicon does not exhibit the Pockels effect due to the centrosymmetric crystal structure. However, the Pockels effect has recently been exploited in a strained silicon rib-waveguide by depositing a Si$_3$N$_4$ layer on top of the waveguide for the asymmetrical distortion of the silicon crystal [65]. Although the Pockels effect may be induced in silicon through strain, a larger modulation voltage is required to reach as high as 122 pm/V coefficient in the electro-optic Mach–Zehnder modulator [66].

The Kerr effect is a second-order electric-field effect where a change in the real part of the refractive index, $\Delta n_{\text{Kerr}}$, is proportional to the square of an applied electric field. The index change is expressed as

$$\Delta n_{\text{Kerr}} = \frac{1}{2} kn_0 E^2$$

where $k$ is the Kerr coefficient. The theoretical evaluation of $\Delta n_{\text{Kerr}}$ at 1.31 $\mu$m wavelength in silicon is of the order of $10^{-3}$ at an applied field of $2 \times 10^5$ V/cm (i.e. 20 $\mu$V/µm) (Table 2) and $10^{-4}$ at an applied field of $10^6$ V/cm (i.e. 100 $\mu$V/µm) [59] due to the Kerr effect. However, if the electric field used for the larger index change of $10^{-4}$ level is beyond the breakdown field for lightly doped silicon, making an optical modulator through the Kerr effect in silicon is thus unfeasible. Regarding the modulation speed, both the Pockels and the Kerr effects have a fast response time of the refractive-index change in the order of 100 fs or less [67]. In practice, a modulation speed of 100 GHz has experimentally been demonstrated for electro-optical polymer Mach–Zehnder modulators [68].

3.2. Carrier effect

Borrowing the electrostatics arguments of MOS capacitors, a bias voltage can change the carrier concentration in an accumulation (or inversion) layer of the devices. This carrier change can result in a refractive-index change, and has been deployed in semiconductor modulators for decades. A widely investigated method is the free-carrier plasma dispersion effect, which is intrinsically compatible with complementary metal–oxide–semiconductor (CMOS) semiconductor technology [69, 70]. Here, the concentration variation of free carriers may change the refractive index of the functional switching material (i.e. Si). This index change due to the injection of carriers in silicon can be derived from the Kramers–Kronig relation [62], and is given by

$$\Delta n(\omega) = \frac{c}{\pi} \rho \int_0^\infty \frac{\Delta \alpha(\omega_1)}{\omega_1^2 - \omega^2} d\omega_1,$$  \hspace{1cm} (5)

where $p$ denotes the Cauchy principal value, $\Delta \alpha(\omega) = \alpha(\omega, \Delta N) - \alpha(\omega, 0)$, $\Delta N$ is the change in free-carrier concentration within crystalline silicon, $\omega$ is the angular frequency, and $c$ is the speed of light. Soref and Bennett experimentally estimated the changes in the refractive index from absorption spectra for a wide range of electron and hole densities with a wide range of wavelengths [59]. For a telecom wavelength in the 1.55-$\mu$m region the expressions relating the change in $\Delta_n$ and $\Delta_\alpha$ due to injection or depletion of carriers in silicon are

$$\Delta n = \Delta n_e + \Delta n_h = -[8.8 \times 10^{-22} \Delta N_e + 8.5 \times 10^{-18}(\Delta N_h)^{0.8}]$$  \hspace{1cm} (6)

$$\Delta \alpha = \Delta \alpha_e + \Delta \alpha_h = 8.5 \times 10^{-18} \Delta N_e + 6.0 \times 10^{-18} \Delta N_h$$  \hspace{1cm} (7)

where $\Delta n_e$ and $\Delta n_h$ are the changes in refractive index caused by the changes in free-electron and free-hole carrier concentrations (i.e. $\Delta N_e$ and $\Delta N_h$), respectively. $\Delta \alpha_e$ and $\Delta \alpha_h$ are the absorption coefficient changes, respectively. According to the doping process and physical property of a doped semiconductor material, the free-carrier concentrations are equal to each other, i.e. $\Delta N_e = \Delta N_h = \Delta N$. Using Eqs. (6) and (7) both the refractive-index change and the induced absorption coefficient can be further evaluated as a function of the voltage-dependent change of the carrier concentration, $\Delta N$, respectively. This is the so-called carrier plasma dispersion effect, which can be generated electrically or optically in silicon for the required free-carrier concentration [71, 72]. For the case of electrical generation the carrier density interacts with the propagating light through three different mechanisms including carrier accumulation, carrier injection, and carrier depletion. In most cases p-i-n diode structures are formed around an optical
waveguide electrically controlling the injection of electrons and holes into the path of the propagating light. An overview of each case is given below.

The first electrical mechanism to achieve on-chip high-speed modulation is carrier accumulation. This option based on monolithically integrated silicon was experimentally demonstrated with an operating speed >1 GHz by the Intel group [69], which is one of the Intel breakthroughs that opened a curtain for the roadmap of silicon photonics. The design contains a MOS capacitor on the top of SOI, where light travels in the silicon region confined by the gate (polysilicon) and buried oxides (SiO$_2$) (Fig. 6a). Its operation principle is similar to that of a MOS transistor, i.e. the gate voltage controls the charge density under the gate oxide, and thus the refractive index may be tuned in the thin charge-accumulated silicon layer. The modulation speed of carrier accumulation-based silicon modulators is not limited by the relatively long minority carrier lifetime, but by the resistance and capacitance of the device instead. Therefore, these modulators generally suffer from the high capacitance and the lack of modulation efficiency due to the low number of carriers involved.

The mechanism of carrier injection used for a silicon modulator, just as its name implies, is to inject charge carriers into a SOI waveguide by a forward-biased p-i-n diode structure. The basic p-i-n architecture consists of the optical waveguide embedded into highly p- and n-type doped silicon regions (Fig. 6b). The optical mode is confined into the intrinsic silicon region to avoid high losses due to the heavily doped poly-Si gate, while the mode may overlap with the current of charge carriers injected through the intrinsic region. The device bandwidth was first proposed in the gigahertz regime [50], and was later demonstrated with an operation speed at >10 Gb/s utilizing microring and Mach–Zehnder interferometer (MZI) structures, respectively [73, 74]. However, a drawback of the refractive-index change by the carrier-injection mechanism is its relatively low speed due to long carrier recombination times (e.g. ~0.4 ns to ~1.8 ns at the effective carrier concentration of 2.5×10$^{17}$ cm$^{-3}$ to 2.5×10$^{18}$ cm$^{-3}$ in silicon [75]), rather than the device capacitance as for carrier depletion. Naturally, a trade-off between the intrinsic silicon region width and the higher modulation speed exists for these devices.

A carrier-depletion-based modulator can be implemented by incorporating a p-n junction inside an optical waveguide core, whereas the p-n junction is formed through only lightly doped p- and n-type regions (Fig. 6c). Optical modulation is achieved by reverse biasing the junction and then extraction of the free carriers. Note, the intrinsic dynamic response of the device is only dominated by the time required for carriers sweeping out from the p-n junction. This process typically takes a few picoseconds since carrier drift is ultimately at a saturated velocity under a large electric field [76, 77]. The potential bandwidth of such a device may exceed 50 GHz. However, carrier-depletion modulators are generally less efficient than their carrier-injection counterpart, since this effect involves a lower number of carriers than in the case of the injection regime. Despite this drawback, the carrier-depletion mechanism is the alternative option that manipulates free-carrier densities, avoiding the modulation speed limited by the minority carrier lifetime. In short, the carrier-depletion-based modulators may offer the merits of processing simplicity and higher operation speed but with reduced modulation efficiency. Towards improving this modulation option, much attention has been focused on raising the modulation efficiency, while maintaining other performance metrics (e.g. low optical loss) unchanged [78–82]. For example, the interdigitated p-n junctions (orthogonally oriented to the direction of light propagation) serves as an effective solution to enhance the overlapping between an optical mode and the depleted region, breaking the finite space-charge extension region of modulation (typically ~100 nm) due to carrier depletion. A high modulation efficiency of $V_{PEL} = 0.62$ V cm is obtained for a MZI modulator, while keeping both a high modulation speed up to 40 Gbit/s and an acceptable insertion loss of 2.8 dB [81].

### 3.3. Quantum-confined Stark effect

The quantum-confined Stark effect and the Franz–Keldysh effect are both electric-field-induced changes in optical absorption. However, the QCSE relates to the FKE as being the similar effect but for the limit of a thin (< 20 nm) active region thicknesses.

Before introducing the QCSE, we first discuss the FKE, which is usually observed in conventional bulk semiconductor materials. Unlike the Pockels/Kerr effects, the Franz–Keldysh effect gives rise to both electrorefraction and electroabsorption, although primarily the latter phenomenon. This FKE effect originates from distortion of the energy band of a uniform, bulk semiconductor material upon applying an electric field, leading to the electron and hole wavefunctions above and below the bandgap to extend its energy (Fig. 7a). This process may be explained as a transition from step-function-like distributions to Airy-function-like, causing some electron states to become available slightly below the conduction band, and simultaneously some hole states slightly above the valence band. This effect reduces the apparent electronic bandgap of a semiconductor, allowing optical absorption for longer wavelengths relative to the status without the electric field applied. The refractive-index change induced by the FKE in silicon achieves ~10$^{-4}$ level at a wavelength of 1.07 μm when a field of 20 V/μm is used [59], and the index change is higher than that evaluated for the Kerr effect (Table 2). Since the FKE in pure silicon significantly falls in the wavelength range of 1.3–1.5 μm [84], the bandgap engineering work of Ge-on-Si material needs to be performed for a silicon modulator operating at telecom wavelengths [56, 85], such as enabling substantial optical modulation at 1.55 μm through the inclusion of a small amount of Si (i.e. 0.75%) into Ge [86]. One more example using epitaxial GeSi heterostructures for bringing the silicon bandgap close to the direct bandgap transition at ~0.8 eV is adopted to enhance the FKE effect. A demonstrated device of a
Figure 6 Schematic manipulation of free-carrier concentrations in plasma-dispersion effect based silicon optical modulators [51] by (a) a carrier-accumulation mechanism through a MOS structure; a thin SiO$_2$ barrier serving as the insulating layer isolates the waveguide by two halves forming a capacitor structure, (b) a carrier-injection mechanism through a p-i-n structure; an “intrinsic region” formed in the waveguide separates heavily doped p- and n-regions. Electrical charge carriers (i.e. electrons and holes) can be injected into the “intrinsic” waveguide region as the device is forward biased, (c) a carrier-depletion mechanism through a p-n structure; lightly doped p- and n-type regions forming a p-n diode in the waveguide, and increased width of the depletion area dependence on reverse bias of the device.

Figure 7 [83] (a) Optical absorption change by the Franz–Keldysh effect in a bulk semiconductor with an electric field applied. The electron and hole wavefunctions leaking into the bandgap region allow for the energies of photon absorption below $E_g$. (b) Schematic illustration of the quantum-confined Stark effect without (left) and with (right) an applied electric field.

High-speed GeSi electroabsorption modulator (EAM) shows a 40.7-GHz bandwidth with a 2.8-V reverse bias at 1550 nm wavelength [56].

New phenomena arise if semiconductor heterostructures are fabricated in such low dimensions that quantum effects dominate. The QCSE is very similar to FKE in that it is also an electrical-field-induced change in optical absorption. Considering a single quantum well placed in an electric field perpendicular to the quantum well layer, the so-called QCSE forces a shift of the bandgap towards lower energies causing a decrease of the absorption coefficient for energies above the bandgap (Fig. 7b). The electric field pulls the electrons and holes headed for opposite sides of the layer, and an overall net reduction to the energy of the electron–hole pair is thus produced accompanying a corresponding shift of the exciton absorption. In terms of the optical modulation application, this leads to a sharper rise of the absorption edge than that in the bulk material. Bastard et al. found the electric-field dependence on the lowest transition energy in a quantum well with infinite barrier height [87], i.e. $\Delta E_g = E^2 W_{QW}^3$. The equation shows that the exciton energy shift is proportional to both the square of the applied electric field $E$ and the fourth power of the quantum well width $W_{QW}$.

Many earlier QCSE-based devices (e.g. optical modulator) rely on III-V semiconductors as the constituent materials for the quantum wells [52, 53]. There are some demonstrations for integrating III-V QCSE devices onto silicon substrates containing drive electronics. However, for the on-chip interconnect applications, the integration of III-V into a silicon fabrication line still remains challenging due to bonding and alignment issues [88, 89]. The QCSE in thin Ge quantum-well structures was first reported in 2005 [90], where the QCSE strength and its performances are comparable to that in III-V materials at room temperature around a wavelength of 1.45 $\mu$m. The effect is observed from Ge/Si$_{0.15}$Ge$_{0.85}$ multiple quantum wells (MQW) on relaxed Si$_{0.1}$Ge$_{0.9}$ buffer layers directly grown on a Si substrate, forming a type-I band structure (i.e. thin QW surrounded by barrier materials). This discovery paves the way for the realization of QCSE-based Ge/Si modulators [85, 91, 92] through standard silicon CMOS manufacturing processes. The modulation bandwidth of these modulators was evaluated using a pump-probe measurement (i.e. differential transmission spectroscopy), which predicts that a Ge quantum well may potentially operate up to a speed of 100 GHz when a high electric field intensity is applied [54]. In practice the SiGe electroabsorption modulators utilizing QCSE have so far been able to achieve 23 GHz bandwidth from a 3 $\mu$m x 90 $\mu$m size Ge/SiGe MQW waveguide with a swing voltage of 1 V between 3 and 4 V [55].
4. Benchmarking

4.1. Modulation mechanism benchmark

To summarize the performance of each modulation mechanism, Fig. 8 shows a review of an EOM’s energy-per-bit as a function of modulation speed from the literature. To enhance differentiation, different colors symbolizing the various modulation mechanisms are chosen to categorize devices, and those corresponding literature values represented by stars or triangles loosely give a surrounding area for the particular type of EOM. The performance sweet spot is at the bottom-right corner, in which we find nanophotonic and plasmonic devices. In the opposite corner, thermal-optic modulators switches are found because of the limited thermal response time. Although exhibiting a high energy-per-bit and a low switching speed, the mechanism of the thermo-optic effect is simple and the formed devices are potentially compact [109, 110] compared to those built by carrier-density effects (excluding nanophotonics) due to the relatively larger thermo-optic coefficient in silicon (i.e. $\sim 1.8 \times 10^{-4}$/K). The region circled between the energy-per-bit in the range of $1 \times 10^4$ fJ/bit and the speed of GHz level shows a cluster of effects for traditional diffraction-limited devices involving mechanisms with carrier effects, Pockels/Kerr, and QCSE effects. In particular, the diagonal line in the yellow rectangle region represents the switching energy-per-bit dependence on the modulation speed in cavities with different $Q$ factors [27]. High-$Q$ cavities exhibit sharp resonant peaks in spectra and thus require a smaller frequency shift at low modulation frequencies, while their long cavity photon lifetime results in a lower onset frequency of the switching energy rise. The diagonal yellow line can serve as a performance boundary, and the performance of these classical diffraction-limited device designs cannot reach the boundary owing to a weak modal index change, i.e. $\Delta n_{\text{eff}}(V_{\text{bias}})$. Those early plasmonic-based EOMs may break off from this trend line either by deploying a Si/TiN/Si$_3$N$_4$/GZO/Si multilayer modulator [105] or a hybrid plasmon-polariton design [42, 43]. This indicates that optical, traditionally weak, nonlinear effects such as electro-optic modulation can be boosted by increasing the optical density of states through an enhanced electric-field strength. By means of this field enhancement the device interaction length can be decreased to a fraction of what diffraction-limited designs require. However, we have to point out the successful demonstration of THz bandwidth experiments is still challenging for plasmonic-based EOMs due to either the current limited technological measurement speed (e.g. Agilent) or the imperfect devices caused by tight tolerance of nanofabrication processing. Below, we discuss the figure of merit (FOM) of some emerging EOMs designs, such as plasmonic-based ones.

4.2. Enhancing light–matter interactions via waveguide designs

The key for an EOM is to create a strong interaction of the propagating optical data signal with the actively modulated material inside the device. With the existing weak interaction between electronic systems (i.e. solids) and light at telecom wavelengths, EOMs are challenged by fundamental tradeoffs; for instance the moduation efficiency (i.e. extinction-ratio/voltage) and energy-per-bit metrics form the tradeoffs with the insertion loss, the device length, and the speed of the device. Thus, if the LMI is enhanced, the EOM can be optimized. Two methods are generally followed; (a) deploying optical resonators (i.e. cavities), or (b) engineering the waveguide dispersion towards creating high optical density-of-states.

Regarding method (a), the energy-per-bit function of an EOM can be lowered by orders of magnitude via increasing the cavity quality factor [27, 28]. However, high-$Q$ cavities come with several undesired features such as (i) large device footprints hindering dense integration, (ii) high thermal sensitivity often requiring active spectral tuning, and (iii) tight fabrication tolerances. In particular, the large footprints of high-$Q$ enhanced EOMs limit the operational speed of these devices due to high capacitive delays. This is why in the review we essentially focus on waveguide-integrated designs that may achieve strong LMI by utilizing optical confinement concept in waveguides. Before we discuss the waveguide-based enhancement techniques, it is interesting to note that the fundamental device performance of classical photonic EOMs has not been reached by most devices, as indicated by the dashed black box and trend line in Fig. 8.
In light of the second option, (b), of enhancing the LMI, we provide a brief review of waveguides involving strong optical confinement. In detail, the 2D optical confinement and propagation length are compared, which are both indicators of the modulation efficiency and insertion loss for EOMs, respectively. In particular, the following waveguide geometries are contrasted (Fig. 9): coplanar waveguide, channel plasmon-polariton (CPP) [18], gap plasmon waveguide (GPP) [19], metal–insulator–metal [16], insulator–metal–insulator [17], dielectric-loaded surface plasmon-polariton (DLSPP) [22], and the hybrid plasmon-polariton. Starting the discussion with the MIM design, this type of waveguide does feature a good confinement in one dimension (normal to a metal plane), however, the fundamental mode eventually becomes cut off, while the second dimension is scaled down as well. However, the $\lambda/5$ confinement is achieved with propagation lengths of about 3 $\mu$m at visible frequencies [16]. Its inverse structure, i.e. the IMI, performs reasonably well with a mode area of about 1/7th of the diffraction limit of its wavelength, and its propagation length exceeds 20 $\mu$m [17]. However, in Verhagen’s work the refractive indices of the insulator sections are relatively low (\textasciitilde 1–1.5). The propagation length is reduced to single digit micrometers when high-index semiconductors are brought into close contact with the metal, which limits the usage of these waveguides for optoelectronics engineering. The CPP may theoretically have propagation lengths approaching 100 $\mu$m, and some photonic functionality is successfully demonstrated, but the actual component footprints are approaching that of pure photonic areas. This results in little advantage of such weakly bound plasmonic modes over SOI, which has a much lower optical loss than the CPPs. The CPPs’ optical confinement is also usually far from being below the diffraction limit. In addition, the elaborate focused-ion-beam milling for creating the opening-angle sensitive V-groove CPP waveguide seems to be not a feasible way for a reliable and cost-effective design. DLSPP waveguides are close relatives of HPP waveguides, except for the absence of a low-dielectric layer between the metal and the higher-dielectric material [22]. With the theoretical optical confinement of both being comparable, the additional loss of DLSPPs makes them unfavorable over HPP waveguides [23–25, 33]. The HPP waveguide system exhibits a theoretical optical confinement of 1/100th of the diffraction limit and yet offers tens of micrometer propagation lengths.

Experimentally, Sorger’s group demonstrated 1/50th of the diffraction limit with the propagation lengths exceeding 30 $\mu$m at telecommunication wavelengths [42].

In conclusion, the nonresonant LMI enhancement options exist in the form of choosing a waveguide design with high optical confinement (i.e. subdiffraction-limited). Although the MIM, IMI, and gap plasmon optical modes can provide subdiffraction-limited optical confinement, they suffer from high optical losses in particular when higher-dielectric materials are introduced. The HPP waveguide is a promising option since it allows for the synergistic designs of using metals as strong optical mode confinement, heat sink, and electrical contacts simultaneously, while offering the seamless integration into an SOI platform [39, 41, 42].

### 4.3. EOM performance benchmarks

Factors such as the scalability, modulation depth, and power consumption of EOMs are the key performance figures that need to be considered when looking towards advancements of compact and high-performance EOMs. The two proposed benchmarks for the figure of merit, (a) the modulation-efficiency vs. optical loss [41], and (b) the energy vs. speed $FOM_{device} = f_1 \text{dB} (E/\text{bit})$ [27] are discussed here. In the former case, we compare various material systems for electroabsorption modulators, which indicates the dominant cases of devices based on strong LMI such as plasmonic-based components over traditional (i.e. diffraction-limited) ones.

As for the first benchmark, the modulation efficiency of an EOM can be described by incorporating the obtainable complex modal index change, $\Delta n_{mode}$, voltage bias, $V_{bias}$, per device length, $L$, versus the power penalty (i.e. (insertion loss, IL)$^{-1}$)

$$FOM_{physical} = \frac{\Delta n_{mode} \text{ (or } \Delta \kappa_{mode} \text{) }}{V_{bias} L} \times \frac{1}{IL}. \quad (8)$$

Note that for the strong $\Delta n_{mode}$ (real part), and $\Delta \kappa_{mode}$ (imaginary part), we can design an interferometric Mach–Zehnder modulator and a linear EAM, respectively. This benchmark could be used to compare the devices early in their development cycle. That is, many researchers utilize numerical evaluation (often finite-difference time-domain) tools to gain first insights into a device performance, which
usually rely on eigenmode, and power-transmittance results. The physical FOM can be directly used to compare those emerging devices. Figure 10a shows the FOM (i.e. modulation efficiency) of electroabsorption modulators dependence on the power penalty. Those references for each color symbol can be found in Ref. 41. High-performance EAMs are found in the upper-right corner of the Fig. 10a, having a maximum modulation efficiency and a minimum insertion loss/power penalty, whereas the dashed line may be interpreted as a technology progression of device advancements. A cluster in the lower-left corner can be found in particular for traditional (i.e. diffraction-limited) device architectures, and the region extends to the lower-right corner, representing the device improvements on losses that have been made in recent years. Note, the gray shaded area in Fig. 10a denotes an empirical region of the traditional device architectures, whereas the trend line in the energy/speed comparison indicates a fundamental device limit. The array of traditional EAM devices in the lower-right corner (highlighted by the gray-shaded region) shows the apparent trend that exemplifies the difficulties of designing a device that is able to incorporate strong light-matter interactions and simultaneously reduce the onchip insertion loss. However, plasmonic devices that deploy a MIM waveguide are able to break off from this trend [37]; in particular, the plasmonic HPP and MOS waveguide modes possess the capability of electric-field enhancement and, thus, boost optical nonlinear effects (e.g. electro-optic modulation), ending up in the upper right hand corner in Fig. 10a [39–42]. Taking advantage of this enhancement, the device interaction length is possible to be downscaled to a fraction of what the classical diffraction-limited designs require, and the common concerns about large losses of plasmonic waveguide becomes less significant.

The technological advancements [27] are possible if plasmonics is utilized in the EOM designs exhibiting a strong dependence on the $Q$-factor (Fig. 10b). The switching energy-per-bit as a function of modulation frequency, $f_{3\text{dB}}$, in cavities operating at 1550 nm wavelength reduces strongly with the increasing cavity $Q$-factor. However, there is a significant trade-off for the bandwidth of such high-$Q$ EOM designs, due to the long photon lifetime. The gray region in Fig. 10b, corresponds to the minimum energy-per-bit of classical EOMs reachable, and shows that the current nondiffraction-limited EOMs actually are not at their fundamental limits yet. The minimum energy-per-bit, $E_{\text{bit}}$, of these classical EOMs can be calculated by fitting the limited performance of the curves at their high frequency ends, which is expressed by

$$E_{\text{bit}} = K \cdot f_{3\text{dB}}^2,$$

where $K$ is a constant associated with the functional switching material [27]. However, a hybrid plasmonic nanocavity is beneficial to achieve deep sub-λ confinement in a single-mode cavity. The tradeoff between modulation efficiency and switching speed can be minimized by utilizing the hybrid plasmonic nanocavity due to its smaller capacitance. Note that for the maximized performance shown at the bottom-right corner (Fig. 10b), the bandwidth is extending into the THz range and the power consumption is approaching the attojoule regime, which is about 2–3 orders of magnitude lower compared to those of state-of-the-art devices.

**5. Ultracompact EOM devices demonstrations**

In this section we briefly survey the proposed and demonstrated ultracompact and fast modulator designs. Among
them, plasmonic-based modulators may allow for the incorporation of subwavelength-scaled structures, yet break the size barriers of classical diffraction-limited devices, and thus lead to strong optical-mode confinements. Note, the discussed devices are representative examples and shall not be taken as an exhaustive list.

5.1. Silicon field-effect plasmonic modulator (PlasMOStor)

Dionne et al. demonstrated the first plasmonic waveguide-based EOM termed PlasMOStor (Fig. 11) [37]. This design utilizes an MIM waveguide with a sandwiched functional silicon/oxide core, whose carriers are electrically modulated. In the absence of an applied field at the gate (0 V), carriers are depleted, facilitating light guiding through the silicon core with the losses expected from the plasmonic MIM mode. The device supports two modes: a photonic mode (red arrow in Fig. 11b) and a plasmonic mode (blue arrow in Fig. 11b), which are both generated at the beginning of the device and can interfere either constructively or destructively at the end. The device operates by pushing the photonic mode to cutoff through the silicon free-carrier modulation. A remarkable result is the low voltage bias (0.7 V) for operating the device at $\lambda = 1.55 \mu m$ wavelength. The plasmonic and photonic (core) modes interfere which can be probed by changing the length of the device (Fig. 11c), whereas the transmission is increased by inducing the charge accumulation. The theoretical fits to the experimental data with a range of 2.6 $\mu$m to 7.5 $\mu$m indicate that a higher modulation ratio may be possible at the shorter source–drain separations. A prototype of the PlasMOStor may achieve a simulated modulation ratio approaching 10 dB, while incurring an internal waveguide propagation loss $< 1 \text{ dB}$, additional losses from the mode insertion of $-12.8 \text{ dB}$, and the extraction loss of $-3 \text{ dB}$ through the source and drain slits. The empirically determined switching voltage of 0.7 V and a capacitance of 14 fF produce a switching energy of 6.8 fJ for a 4-$\mu$m$^2$ large PlasMOStor device. Although potentially capable of gigahertz operation, a circuit analysis of the PlasMOStor photodiode (i.e. gate-voltage provider) coupled system for the purpose of all-optical modulation predicts the bandwidth only up to 15 GHz due to the limited bandwidth of an involved photodiode.

5.2. Resonance-based plasmonic modulator

Brongersma’s group proposed a submicrometer MIM-based EOM (Fig. 12) with a 3-dB extinction ratio along with modest power attenuation ($< 3 \text{ dB}$), compact size ($< 1 \mu m$), and low voltage (1 V). The device takes advantage of both the resonance-based configuration and MIM waveguide structure [38]. The metallic slab with an infinitely extended trough slit along with the two interfaces at its ends forms a Fabry–Pérot-type resonator for the propagating MIM mode in the nanogap. Although the MIM mode in the slit is intrinsically lossy, both the energy density inside and the transmittance through the slit simultaneously reach their local maxima, when the plasmonic resonance condition is fulfilled. Therefore, the strong LMI together with a good power throughput is expected from such a device. While considering a plasmonic modulator based on an inplane MIM waveguide (Fig. 12a), the resonator is formed by the section of MIM waveguide with a length $L$ and a pair of rectangular stubs with sizes $d_x \times W_x$. Both the rectangular stubs may serve as SPP mirrors, which can be designed to satisfy the Fabry–Pérot resonant conditions (Figs. 12b and c). The device was proposed to increase the imaginary part of $\Delta n$ that raises the power dissipation in the cavity medium, and consequently the overall strength of Fabry–Pérot resonance is suppressed. Alternatively, the modulation of the real part $\Delta n$ shifts the central wavelength of resonance, as the optical path per transit is changed. While a modest change of $\Delta n = 0.004$ or $\Delta n = 0.008$ in the cavity index is satisfied, the device may achieve 50% modulation depth with an estimated swing voltage of $\sim 1 \text{ V}$ and a capacitance of $\sim 1 \text{ fF}$. The ultrafast operation at the low power dissipation is thus promised.

5.3. Quantum-well waveguide modulator

Miller’s group investigated the first Ge/SiGe quantum-well (QW) waveguide modulator using the quantum-confined Stack effect, and the device is monolithically integrated on a SOI waveguide through direct-butt coupling and a selective epitaxial growth structure (Fig. 13) [103]. The functional modulator section consists of a vertical p-i-n diode with 15 pairs of Ge/Si$_{10.13}$Ge$_{0.85}$ QWs in the intrinsic region on a single-mode TE-supported waveguide. In order to obtain a maximum overlap, the epitaxial layers are designed such that both the silicon core of the SOI waveguide and the QWs of the functional switching section are on the same vertical level (Figs. 13a and b). This may ensure that the optical mode in the SOI waveguide can be efficiently modulated by Ge QWs when passing through the functional section. To measure the surface normal modulators on a fabricated chip, the photocurrent is collected at the various bias voltages at room temperature (Fig. 13c). The functional switching section of the integrated device has a footprint of only 8 $\mu$m$^3$, while a direct modulation of more than 3 dB under a 1-V swing voltage is provided. However, the relatively high insertion loss ($\sim 12 \text{ dB}$) appears to originate from the mode mismatch between the thin silicon waveguide core (310 nm) and the thick Ge/SiGe region ($\sim 1.5 \mu m$).

5.4. Graphene-based optical modulator

A graphene-based broadband optical modulator has been introduced by Zhang’s group (Fig. 14) [111]. The device has several distinctive advantages such as strong
light–graphene interaction owing to a much stronger interband optical transition possessed by a monolayer graphene, broad bandwidth, and high-speed operation, as well as CMOS compatible capability. This 1.2-GHz fast graphene modulator offers a modulation efficiency of 0.1 dB/μm, which is achieved by actively tuning the Fermi level of the monolayer graphene sheet by applying different driven voltages ($V_D$). While at a large negative $V_D$ ($<-1$ V) there are no electrons available for the interband transition and hence the graphene appears to present the transparent
property. At a low driven voltage (–1 V < \( V_D \) < 3.8 V), the Fermi level of graphene is close to the Dirac point and the interband transition occurs when electrons are excited by incoming photons. In addition, at a larger positive \( V_D \) (> 3.8 V), all electron states are filled up and no interband transition are allowed (Fig. 14c). As a result, a graphene-based optical modulator with a small footprint of 25 \( \mu \text{m}^2 \) was demonstrated at 1.2 GHz operation under the ambient conditions for a broad optical bandwidth between 1.35 to 1.5 \( \mu \text{m} \) wavelengths.

5.5. Silicon–organic hybrid (SOH) slot waveguide modulator

Leuthold’s group demonstrated a high-speed strip-loaded slot waveguide modulator based on SOH technology (Fig. 15a) [112]. The slot between the waveguides is filled by organic cladding material (i.e. EO polymer) with a high \( \chi^{(2)} \) nonlinearity (Pockels effect). One of the key features for the device is the strong LMI in the slot design, which enhances the quasi-TE mode due to the lateral index.
discontinuity (Fig. 15b). This in turn creates the intense optical nonlinearity effect of the EO polymer material inside the nanoscale slot, owing to a large and efficient overlap between the electrical and optical mode when a voltage is applied to the Si rails. The measured frequency response of the modulator was found to be about 100 GHz (at 3-dB signal strength) figure of merit of $V_{\pi}L = 11$ V·mm (Fig. 15d). Using a similar SOH concept this group, collaborating with other teams, recently developed an inphase (I) and quadrature-(Q) modulator based on advanced modulation formats [113, 114]. The IQ modulator is constructed by nesting two MZI modulators, and both modulators are made to interfere with a phase shift of $\pi/2$. The highest single-carrier single polarization data rate of 112 Gbit/s on the SOI platform is observed through the generation of an optical 16 quadrature amplitude modulation (QAM) signal at 28 GBd. 16 QAM is an advanced modulation format requiring complete control of amplitude and phase of the modulated signal, which reduces the energy consumption of the modulator to as low as 640 fJ/bit for a driving voltage of $V_{pp} = 5$ V.

5.6. λ-size plasmonic waveguide modulators

Most plasmonic modulators can be controlled by gap plasmonic modes through a charge-accumulation layer. Two different layouts are summarized here. The first proposed layout comprises oxide and semiconductor layers embedded in a metal from both sides. The metal–insulator–metal structures are formed in such a type of waveguide, and light may be manipulated on a subwavelength scale due to the confined light in the gap with a typical size of ~100 nm or less. This design may also be simplified since two metal surfaces can serve as electrodes.

Such a subwavelength plasmonic modulator in a Ag–SiO$_2$–indium tin oxide (ITO)–Ag waveguide structure has been demonstrated [104], where a five percent change in the average carrier density from $9.25 \times 10^{20}$ to $9.7 \times 10^{20}$/cm$^3$ is obtained. The absorption coefficient of the SPP is modulated by applying a voltage between the two silver electrodes. However, the total plasmonic loss in the geometry is up to 18 dB only for a device length of 2 μm. The other functional EO field-effect plasmonic modulator is developed with a revolutionary size of 25 × 30 × 100 nm$^3$ [115]. The structure shows a 5-nm gap including both HfO$_2$ and ITO thin films sandwiched between the two aluminum metal layers. The loss coefficient of 11 dB/μm is predicted for the nanowire–MIM waveguide. Therefore, the aforementioned designs may be extremely compact. However, there is always a trade-off between modulation depth and transmittance; although a high confinement of light is achieved in these MIM configurations, the propagation length is obviously limited due to larger losses associated with both metal layers. Recently, Babicheva et al.
proposed a CMOS-compatible plasmonic modulators with a multilayer structure [105]. By replacing both metal layers by semiconductor dielectric materials, a long-range SPP mode passes through the multilayer region for light modulation. The best performance of 46 dB/μm modulation depth, 0.29 dB/μm propagation losses in the off-state was obtained for the Si/TiN/Si3N4/GZO/Si multilayer modulator.

In order to overcome a limited propagation length and higher loss of a SPP mode, the hybrid-plasmonic integration architecture is emerging by incorporating nanoscale plasmonics into classic photonic elements. As to the second design, a MOS stack is placed on top of a silicon waveguide to form a hybrid-plasmonic modulator [39]. In this configuration, a transparent conducting oxide material is used for the functional switching part of the MOS-capacitor design upon applying a voltage bias, and the carrier density can be tuned at the TCO/oxide interface (i.e. electron-accumulation layer) of the MOS stack. Here, we introduce a concept of epsilon-near-zero (ENZ) material, such as ITO, where its real part of permittivity can be modified electrically by an applied bias [116]. For example, raising the applied bias starting from zero voltage increased the carrier concentration, resulting in a lower real part of the permittivity based on the Drude model. This produces an index range where the real part of the accumulation layer’s permittivity approaches zero, which physically leads to a dramatic change in its optical property. For the two-modulator states (ON/OFF) the material either exhibits a dielectric response or a metallic response to incident light. Lu et al. numerically reported an electroabsorption modulator based on tunable aluminum-doped zinc oxide (AZO) materials and slot waveguides [117]. The AZO layer employed as the functional slot serves as an ENZ material, which can be changed between ENZ (high absorption) and epsilon-far-from-zero (low absorption) by accumulation carriers. The simulation results show a 3.7-dB modulation depth and a 0.82-dB insertion loss for a 250-nm long device, respectively. Vasudev et al. proposed a similar MOS-capacitor-based modulator consisting of a nanowire SOI waveguide coated with layers of 5-nm HfO2 and 10-nm ITO [116]. An electron-accumulation layer exists at the ITO/HfO2 interface, which modifies the ITO local optical permittivity and creates an ENZ region that perturbs the waveguide mode into a highly absorptive OFF state, when a negative bias is applied. The electroabsorption modulation thus occurs compared to the case without an applied bias (ON state). This device allows for a 3-dB modulation depth at the device length of 30 μm in a nonresonant structure.

The fabrication of these waveguide-integrated hybrid-plasmonic modulators is reasonably simple, and the footprint is comparable with the classical silicon waveguide size. The propagation losses originate mainly from one metal interface compared to two in the MIM waveguide design. Figure 16 shows the schematic modulation mechanism of the MOS-based EOM design; an accumulation layer is formed in the MOS capacitor at the ITO/SiO2 interface by applying a voltage bias between the metal and silicon (Figs. 16a and b). The effective index of ITO changes from being a dielectric to a quasi-metallic state when a voltage bias is applied. We have reported a typical plasmonic EOM based on a SOI waveguide deposited with an ITO–SiO2–Au stack (Figs. 16c and 17) [39, 41, 42]. The portion of an optical mode close to being cut-off residing inside the silicon core is desired, and allows a small impedance mismatch between the MOS switching node and the SOI waveguide. This also denotes an optimum modulation case where most of the energy in the mode is pushed into the MOS gap, increasing the overlap factor of the ITO layer and the mode. Such a modulator using ITO (Fig. 17) exhibits a low insertion loss of ~1 dB for the 5-μm long device, a subwavelength compact size of 0.78λ, and the broadband operation of > 500 nm with a nonresonant MOS mode.

Graphene is an emerging monolayer material with the feature of gate-variable optical conductivity (i.e. σc) controlled by chemical potentials, μc. The chemical potential determining the complex conductivity can be used to electrically tune the refractive index of a material. The optical conductivity of graphene can therefore be converted into a
Sub-λ-size ITO and graphene-based electro-optic modulator [42]. (a) EO functional switching materials used in a MOS structure. The insets exhibit the electric field intensity profiles on the cross section of the MOS structure, showing the field intensity mainly concentrated in the SiO₂ gap layer for the ON-state and shifted to the ITO layer for the OFF-state. (b) The functional hybrid-plasmonic modulation node is placed on a SOI waveguide platform. The electrical ground may be connected either to the silicon waveguide or to the ITO layer, with the latter preferred to avoid additional optical losses of highly doped silicon. (c) The real and imaginary parts of the complex refractive index of graphene as a function of chemical potential at a 1550-nm wavelength, and the inset shows the bias voltages (i.e. 3.1–4.8 V) required to achieve the transition from a "dielectric" state to a "metallic" state.

The real and imaginary parts of the complex refractive index of graphene may then be calculated as a function of chemical potential (Fig. 17c). The shaded region \( (\mu_c = [0.49, 0.58] \text{ eV}) \) is an interesting region, where \( \kappa \) dramatically increases, but \( n \) is almost constant. In this region the optical property of graphene changes from a being "dielectric" to "metallic". The bias voltage required to achieve this transition is in the range of 3.1–4.8 V and depends on the actual device electrostatics. This strong extinction coefficient change can be used for electroabsorption modulation using graphene as a functional switching material. Such a device would exhibit an insertion loss of about 2 dB for a 2-μm long graphene-based EOM, while delivering an extinction ratio of about 18 dB [42].

Despite summarizing both experimental and theoretical EOM studies, no clear optimum design conclusion can be drawn at this time since each has its own strengths or weaknesses. However, a common design feature towards high-performance EOMs naturally include strong light–matter interactions and improved integration synergies. In this regard (i) emerging materials with strong index tunability are encouraging, (ii) synergistic electric-optical integration, and (iii) LMI enhancement via either increasing the field strength and/or enhancing the optical density of states.

6. Onchip integration considerations

The wafer footprint and functionality of LMI-enhanced EOMs can be contrasted to pure photonics components. For example, a MZI EOM is schematically depicted in Fig. 18a showing a typical footprint of 100 × 200 μm [55], which far exceeds the footprint of a plasmonic modulation island [39]. As such, the same 100 by 200 μm footprint for the pure SOI modulator could contain a tens-of-times higher functionality than the plasmon/SOI hybrid devices are deployed, saving the precious material and design space (Fig. 18b). These hybrid crossbar NOCs are being investigated by a variety of computer architecture groups such as in [8,9,11,12]. Returning to the EOM device level, for providing a balanced picture, however, there is one potential disadvantage of such photonic/plasmon hybrid links, i.e. the impedance mismatch between the plasmonic functional switching node and the pure SOI waveguide section. In order to quantify this parasitic coupling loss, we experimentally deployed the cut-back method, where the transmission through a SOI-HPP-SOI waveguide section was measured as a function of the HPP section length [39]. Using this method we were able to quantify the SOI-to-HPP coupling loss (≈0.23 dB/coupler) and the intrinsic HPP propagation loss (≈0.1 dB/μm) at the same time. This coupling loss value could be improved if tapered couplers or adiabatic nanofocusing techniques are utilized to match the impedance between the SOI and the HPP section. The latter, however, would increase the overall device footprint, and more work is needed for this concern. The initial results are nevertheless encouraging: Wassel et al. compared the power consumption as a function of signalling distance for the various link technologies (Fig. 18c) [12]. Naturally, the pure (all) plasmonic link requires higher optical power due to the strong optical losses of these plasmon waveguides. While photonic (i.e. classical...
diffraction-limited solutions) are relatively flat, the hybrid case outperforms all-electrical signalling for links exceeding 700 μm. It would be interesting to revisit this benchmark since within the last 2 years a number of ultrasmall and power-efficient plasmonic EOMs have been proposed and demonstrated.

The rapid development of ultrafast EOMs is motivated by the demands for higher optical transceiver bandwidth. Electronic drivers configured for optical modulators are essential building circuits in these transceiver chips. In earlier work, onchip modulator drivers were wire bonded to amplify electrical data towards combining COMS

Figure 18  (a) and (b) Footprint comparison between pure photonic (e.g. SOI) functional devices. Here, a Mach–Zehnder EOM is contrasted to a plasmon/SOI hybrid approach comprised on a low loss SOI portion (yellow) and the HPP mode (gray) as the functional area. For the wafer footprint of the EOM in panel (b) hybrid plasmonic EOMs can have a significant higher functionality offering a leap forward in NOC technology. (c) Break-even-point analysis of optical interconnect links for the various technology nodes after [12].
electronics with silicon photonics demonstrating a 410-Gbps optoelectronic transceiver [119]. To date, various integration methods have been deployed such as wire bonding [120], flip-chip assembly [121, 122], and monolithic integration [123]. In terms of industry perspective, monolithic integration (i.e. photonics and electronics integrated on one chip) holds strong promise for providing low power consumption and cost-effective solutions. In general, the ideal driver offers a maximized voltage swing across the modulator, low power dissipation, and delivery of a high transmission rate signal. However, in practice technological demonstration of a 200-GHz fast optical modulator integrated with electronic drivers onchip is still challenging. One of the main difficulties is to reduce parasitic (i.e. capacitive) effects especially at the speed signal > 50 Gbps. An improper design of electric connections between the driver and the modulator can degrade the performance of photonic parts. The second challenge is to keep the power consumption low for high-speed drivers. For example, the energy-per-bit of a single microdisk modulator may be as small as 3 fJ/bit, however, the monolithic integration of the driver circuit with the modulator consumes 840 fJ/bit [120]. Roadblocks like these require designers to rigorously develop holistic engineering strategies that can optimize both the photonic device and driver performance.

7. Conclusion and outlook

In conclusion, the performance of an EOM is inherently dependent on the capability to increase the weak interaction between matter and telecom photons. Here, we summarize the tradeoffs that exist between the various performance parameters, namely, insertion loss (IL), extinction ratio (ER), bandwidth (BW), energy efficiency (E/bit), device length (L), and integration density (ID) as a function of length and voltage scaling, respectively. Reducing the physical device length inflicts tradeoffs between ER on the one side and IL, BW, E/bit, and ID on the other, whereas lowering the drive voltage results in trading off IL and E/bit for ER, BW, L, and ID (where ER = const. was assumed for a larger L to make up for the reduced voltage).

Electro-optic modulation being an optical nonlinear effect relies on the electric-field strength inside the device. Various design features have been reviewed towards improving these modulation effects through (i) novel materials, e.g. high EO coefficient polymer, graphene, and ITO, (ii) subdiffraction-limited waveguide integration, e.g. Ge/SiGe QW waveguide modulator, and (iii) field enhancement and/or density-of-state enhancing schemes. A general EOM design involves a trade-off between modulation efficiency and optical losses. Benchmarking EOMs shows the possible trend featuring strong LMs and minimal onchip insertion loss simultaneously. Plasmonic technology possesses advantages relating to speed and scalability. Examples of incorporating plasmonics components into silicon photonics architecture have been demonstrated either using plasmonic waveguides and switches towards Tb/s routing for optical interconnects [124], or using high-speed plasmonic phase modulators for a hybrid silicon–plasmonic chip [107]. This indicates that nanophotonics including plasmonics show great promise for the advancement of next-generation chip-scale integrated photonic circuits.

The design of electro-optic modulators has relevance for advancing photonic/electronic photonic integrated circuits, and future network-on-chip design including inter- and intracore architectures. Compared to silicon photonics nanophotonic devices are a premature technology and hence show limited functionalities to date. However, the next-generational technology for photonic integrated circuits past microrings will likely include nanophotonic elements due to their ability to overcome the fundamental scaling limits of photonics. Fundamental analysis for EOMs has shown that they are able to provide sub 1 fJ/bit efficiencies and might deliver 100+ Gbps fast data that speeds a single channel (i.e. wavelength). However, more research is needed on investigating (a) loss mechanisms of photonic-to-plasmonic couplers, (b) material control and incorporation for functional EO modulation, (c) potential heating issues at higher data rates, and (d) integration of driver designs for ultrafast devices. The latter may be done concurrently with the device design, since it is likely not possible to drive EOMs as a normal electrical load, and it might be interesting to introduce transmission-line concepts into plasmonics.

Acknowledgements. We acknowledge support from the Air Force Office of Scientific Research (AFOSR) under the award numbers FA9550-14-1-0215 and FA9550-14-1-0378, and partial support by funds from the National Natural Science Foundation of China under grant 61377059, the Beijing Municipal Natural Science Foundation under grant 4142004, and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

Received: 11 September 2014, Revised: 1 January 2015, Accepted: 14 January 2015

Published online: 10 February 2015

Key words: Electro-optic modulators, surface plasmon-polariton, light–matter interaction, ultrafast.

Ke Liu received the Ph.D. degree in integrated optics from the City University of Hong Kong, Hong Kong, in 2005. He has been a professor of Electronic Information and Control Engineering at the Beijing University of Technology since 2011. Recently, he joined Prof. Sorger’s research group for visiting purposes, at The George Washington University, USA. From 2006 to 2011, he worked at the University of North Carolina at Charlotte and the University of Texas at Dallas, respectively, as a Research Associate. His current research interests include silicon photonics, plasmonic nanolasers and InP-based photonic integrated circuits. He is members of IEEE, OSA and SPIE.
Chen Ran Ye received the B.S. degree in Automation from North China University of Technology, China, in 2010 and the B.S. degree in Electrical Engineering from Southern Polytechnic State University, Marietta, GA, in 2010 and the M.S. degree in Electrical Engineering from University of Florida, Gainesville in 2012. Currently, she is working as a research assistant. Recently, she entered the field of nanophotonics on joining the Nanophotonic Lab lead by Prof. Sorger at The George Washington University. Her research interest includes the investigation of novel electro-optic devices, the design, simulation and fabrication of gate-level computer architecture via nanophotonics.

Sikandar Khan received the B.S. degree in Electrical engineering from the University of Engineering & Technology Peshawar, Pakistan, in 2007 and the M.S. degree in computer engineering from George Washington University, Washington D.C., in 2010 and currently is a Ph.D. student in the Department of Electrical and Computer Engineering at the George Washington University, Washington D.C. From 2007 to 2008, he was a Research Assistant with the Electrical Lab in UET Peshawar. His research interests include electro-optic modulators, switches and photonic integrated circuits. His current research focus is nanophotonic materials and device concepts for the telecom technology that can bring ultrafast switching and high-bandwidth connections for all the devices.

Volker J. Sorger is an assistant professor in the Department of Electrical and Computer Engineering, and the director of the Nanophotonics Labs at The George Washington University. He received his PhD from the University of California Berkeley. His research areas include optoelectronic devices, plasmonics and nanophotonics, including novel materials. He has received multiple awards such as the AFOSR YIP award, MRS Graduate Gold award, and Intel Fellowship to name a few. Lastly, he is the executive chair of the OSA Nanophotonics technical group, editor-in-chief for “Nanophotonics”, CTO of BitGrid LLC, and a member of IEEE, OSA, SPIE, and MRS.

References


© 2015 by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
www.lpr-journal.org


