Ultracompact electro-optic waveguide modulator based on a graphene-covered $\lambda/1000$ plasmonic nanogap

SHINHO KIM,¹ SERGEY G. MENABDE,¹ JOEL D. COX,²,³ TONY LOW,⁴,⁵ AND MIN SEOK JANG¹,⁶

¹School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
²Center for Nano Optics, University of Southern Denmark, Odense DK-5230, Denmark
³Danish Institute for Advanced Study, University of Southern Denmark, Odense DK-5230, Denmark
⁴Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA
⁵tlow@umn.edu
⁶jang.minseok@kaist.ac.kr

Abstract: The extreme field confinement and electro-optic tunability of plasmons in graphene make it an ideal platform for compact waveguide modulators, with device footprints aggressively scaling orders of magnitude below the diffraction limit. The miniaturization of modulators based on graphene plasmon resonances is however inherently constrained by the plasmon wavelength, while their performance is bounded by material loss in graphene. In this report, we propose to overcome these limitations using a graphene-covered $\lambda/1000$ plasmonic nanogap waveguide that concentrates light on length scales more than an order of magnitude smaller than the graphene plasmon wavelength. The modulation mechanism relies on interference between the non-resonant background transmission and the transmission mediated by the gate-tunable nanogap mode, enabling modulation depths over 20 dB. Since the operation of the device does not rely on graphene plasmons, the switching behavior is robust against low graphene carrier mobility even under 1000 cm²/Vs, which is desirable for practical applications.

© 2021 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

The realization of optical devices unconstrained by the diffraction limit has been a central focus of plasmonics, which capitalizes on the ability of plasmons—coherent charge density oscillations in metals—to concentrate light on length scales well below its wavelength in free space. In particular, metallic 2D materials (graphene [1], black phosphorus [2], etc.) naturally support low-dimensional surface plasmon-polaritons with wavelengths as short as $\lambda_0/100$, where $\lambda_0$ denotes the free-space wavelength. Furthermore, large variations in the Fermi level of 2D materials can be electrostatically controlled by gating, which substantially modulates the associated plasmonic response [3]. As a representative example, graphene plasmons (GPs) have been widely employed in a wide range of devices such as active metasurfaces [3,4], molecular sensors [5–7], and photodetectors [8], along with polaritons in other van der Waals crystals [9,10]. More recently, it has been discovered that the polaritonic modes in 2D materials can be atomically confined also in the transverse direction by sandwiching a thin dielectric layer by a 2D material and a metal substrate [11–14].

The tunable optical properties of two-dimensional materials have also been utilized to develop compact electro-optic waveguide modulators [15–17], where the extremely confined nature of 2D polaritonic modes can dramatically reduce device footprints down to the extreme sub-diffraction regime [18,19]. An ultracompact device footprint is desirable not only for high-density on-chip
integration, but also for low device capacitance, which enables high-speed operation with low switching energy [20]. In such devices, the minimum lateral dimension is fundamentally dictated by the polariton wavelength ($\lambda_p$). For example, two recently proposed graphene plasmon based waveguide modulators have lateral sizes of about $\lambda_p \sim \lambda_0/50$ at mid-infrared frequencies [18,19]. Furthermore, the modulation performance of most graphene plasmonic devices strongly depend on the carrier mobility of graphene, which places practical limits on the large-scale realization of these devices using graphene grown by chemical vapor deposition (CVD).

In this paper, we theoretically demonstrate that an efficient transmission modulation of mid-infrared plasmonic waveguide modes can be achieved even with an active device length of $\lambda_0/1000$, which is more than an order of magnitude shorter than the graphene plasmon wavelength. Here, the non-resonant background transmission interferes with transmission through the $\lambda_0/1000$ nanogap mode, where the latter can be electrostatically tuned by altering the carrier density of graphene. The transmission modulation properties can be elaborately tailored by engineering the geometrical parameters of the device, and the resulting modulation depth can exceed 20 dB. In contrast to previously investigated structures [18,19], the operation of this device does not rely on the graphene plasmon resonances. Consequently, its switching behavior is robust against low graphene carrier mobility, which is strongly desirable for practical applications.

2. Device structure and transmission modulation characteristics

The proposed nanogap plasmonic switch consists of a graphene-insulator-metal structure connecting two identical metal-insulator-metal (MIM) waveguides, which serve as the input and output optical channels, as illustrated in Fig. 1(a). The carrier concentration in graphene can be tuned electrostatically by applying a gate bias voltage between the graphene sheet and the underlying metal. The materials of the insulating core and metal claddings of the MIM waveguides are SiO$_2$ and Au, respectively. The target free space wavelength is chosen to be $\lambda_0 = 7.5$ µm, at which the optical absorption in SiO$_2$ [21] and the optical phonon scattering in graphene [22,23] are negligible. The thickness of the core ($d$) is chosen to be far below the diffraction limit, which enables an efficient coupling between the waveguide mode and the nanogap cavity mode. With the deep subwavelength thickness, the MIM waveguides only support the fundamental transverse magnetic mode ($\text{TM}_0$). The thickness of the cladding ($h$) is set to be larger than the skin depth of Au to prevent leakage of electromagnetic energy from the waveguides.

We solve Maxwell’s equations and characterize the waveguide response by using the commercial finite element method (FEM) software COMSOL Multiphysics. The port boundary condition is applied at the input and output terminations of the waveguide in order to calculate the amplitude and phase of the transmission and reflection coefficients for the $\text{TM}_0$ mode. The reflection and transmission coefficients are measured at the output ports, and the propagation loss and phase delay in the waveguide region is compensated as described in Appendix 3. Graphene is modeled as a thin layer with an effective relative permittivity $\varepsilon = 1 + i \sigma / \varepsilon_0 \omega \tau$, where the layer thickness $\tau$ is assumed to be 0.3 nm. The optical conductivity of graphene $\sigma$ is calculated within the random phase approximation in local limit as a function of frequency $\omega$ and Fermi level $E_F$ [24]. Dielectric functions of Au and SiO$_2$ are taken from [21].

Full-wave electromagnetic simulations reveal that the transmittance ($T$) of the MIM waveguide mode across the gap exhibits a strong dependence on the gap width ($g$) and the Fermi energy of graphene ($E_F$) as shown in Fig. 1(b). Here, $d$ and $h$ are chosen to be 50 nm and 30 nm, respectively. Interestingly, we find that the resulting resonant transmission exhibits qualitatively different behavior depending on the gap width. For $g > 40$ nm, where the gap width is comparable to the graphene plasmon wavelength, the transmission across the gap exhibits a highly asymmetric Fano-like profile. The Fano resonance originates from the interference between the resonant transmission (mediated by the graphene plasmon modes) and the non-resonant background transmission, as recently explained by Jang et al. [18]. The $n$-th order resonance condition
Fig. 1. (a) Schematic of the MIM waveguide-integrated plasmonic modulator. Each MIM waveguide supports the TM\textsubscript{0} guided mode whose transmission is modulated by the gate voltage. For the carrier mobility of graphene $\mu = 10,000$ cm\textsuperscript{2}/Vs and a fixed $\lambda_0 = 7.5$ $\mu$m, the optimized geometric parameters are: core thickness $d = 50$ nm, cladding thickness $h = 30$ nm, and gap width $g = 5$ nm. (b) The transmittance $T$ map as functions of Fermi level $E_F$ and the geometric parameter $g$. (c) The distribution of the normalized Re$\{E_x\}$ inside the each nanocavity geometries at minimum transmittance when $E_F = 0.24$ eV (left), and $E_F = 0.26$ (right).

for graphene plasmons bounded by metallic walls, which provide a nearly $\pi$ phase shift upon GP reflection, can be analytically deduced from the phase matching condition, $\phi_{GM} + \beta_G g = (n + 1)\pi$ [25], where $\phi_{GM}$ is the reflection phase shift of graphene plasmons at the graphene-MIM waveguide boundary and $\beta_G$ is the propagation constant of the graphene plasmon mode that emerges for $E_F > 0.08$ eV. The black dashed curves in Fig. 1(b) indicate the dispersions of zeroth, first, and second order graphene plasmon resonance. Note that, for $g > 40$ nm, the transmission modulation occurs at the vicinity of the graphene plasmon resonance as explained earlier.

Conversely, as the gap becomes much narrower than the graphene plasmon wavelength, there appears a new, unexpected transmission modulation regime, where the dependence of $T$ on $E_F$ and $g$ is significantly different from what was described earlier. In this regime, the transmission line shape becomes much broader and more symmetric. More importantly, the transmission modulation no longer occurs at the Fermi level associated with a graphene plasmon resonance at the operation wavelength. We attribute this discrepancy partly to the inherently broad nature of the zeroth-order graphene plasmon resonance. Since the gap width is much smaller than the graphene plasmon wavelength, the propagation phase of graphene plasmon, $\beta_G g$, becomes marginal while the graphene-metal reflection phase shift $\phi_{GM}$ remains nearly $\pi$ [18]; consequently, the phase matching condition, $\phi_{GM} + \beta_G g = \pi$, is roughly satisfied even though $E_F$ deviates significantly from the exact matching condition. In addition, as the gap width is reduced to lengths of a few nm, the nanogap mode, which is confined in the gap between the top metal claddings, begins to play an important role in mode conversion dynamics. The difference between the two regimes is also evident in the electric field distributions near the nanogap region at $E_F$, where transmission modulation occurs. In the wide gap regime ($g = 80$ nm), the electric field is concentrated on the surface of graphene and decays as it gets further away (Fig. 1(c)). In contrast, the electric field in a narrow gap ($g = 5$ nm) is confined in the air gap region and does not show vertical intensity variation inside the gap, implying that this phenomenon is closely related to the nanogap mode rather than the graphene plasmon mode.
3. Operation principles

To unravel the operation mechanism of the nanogap transmission modulation, we decompose the net transmission into two channels: the background transmission \( t_B \) and the transmission through the nanogap mode \( t_{NG} \), as schematically shown at Fig. 2(a). The background transmission \( t_B \) is inherently independent of \( E_F \) and can be estimated from the transmission of the nanogap structure without the graphene sheet (see Appendix 3). In contrast, the transmission through the nanogap mode \( t_{NG} \) varies as a function of \( E_F \). The nanogap mode corresponds to a vertically propagating plasmonic mode squeezed in between the left and right top metal claddings in the transverse direction, as illustrated in Fig. 2(a). Once the input MIM waveguide mode impinges on the nanogap region, a certain fraction of the wave couples to the nanogap mode. The nanogap mode then propagates up and down along the gap and experiences reflection at the top and bottom ends of the gap. Each time the nanogap mode interacts with the graphene-covered bottom end of the gap, some fraction of the wave outcouples to the output MIM waveguide. This process can be analytically described by using a modified Fabry-Perot model for the nanogap transmission,

\[
t_{NG} = \frac{c_{MN} c_{NM} e^{i \beta_{NG} h}}{1 - r_{NM} r_{NF} e^{i 2 \beta_{NF} h}}
\]

where \( c_{ij} \) and \( r_{ij} \) denote the coupling and reflection coefficient with \( i,j = N \) (nanogap), \( M \) (MIM waveguide), or \( F \) (free space), which are individually determined by separate full-field electromagnetic simulations as described in the Appendix 3, and \( \beta_{NG} \) is the propagation constant of the nanogap mode.

For a specific geometry, \( \beta_{NG} \) and \( r_{NF} \) are independent of the graphene properties, while \( c_{MN} \) and \( c_{NM} \) depend weakly only on \( E_F \) (Fig. 2(b)), so that the modulation of \( t_{NG} \) is mainly achieved by the dependence of \( r_{NM} \) on the graphene carrier concentration. We note that the phase factors of \( r_{NF} \) and \( r_{NM} \) are close to zero as the wave impedances of the free space and the dielectric are sufficiently high, and the phase of \( c_{NM} \) and \( c_{MN} \) are close to zero and \( \pi \), respectively, as imposed by the continuity of the electric field at the boundary. As a result, the net phase of \( t_{NG} \) roughly varies around \( \pi \), which is ideal to achieve destructive interference with \( t_B \) and therefore an efficient modulation of waveguide transmission. Figure 2(c) shows the resulting amplitude and phase of \( t_{NG} \) as a function of \( E_F \) for \( g = 5 \text{ nm}, \ d = 50 \text{ nm}, \) and \( h = 30 \text{ nm} \), which exhibits a broad and symmetric Lorentzian-like resonant feature. The amplitude of \( t_{NG} \) has a maximum at the center Fermi level of \( E_F = 0.26 \text{ eV} \), and its phase linearly decreases from \( 1.24 \pi \) to \( 0.67 \pi \) as \( E_F \) varies from 0.1 to 0.5 eV.

Figure 2(d) compares the net transmission amplitude \( |t_{NG} + t_B| \) obtained from full-wave simulations and analytical modeling, and confirms that the analytical model successfully captures the dependence of \( T \) on \( E_F \). At \( E_F = 0.26 \text{ eV} \), the phase difference between \( t_{NG} \) and \( t_B \) becomes \( \pi \) while their amplitudes are similar, which induces the transmission minimum by destructive interference (“off” state, \( T_{off} \)). By defining the charge neutrality point as the “on” state, \( T_{on} \), the on/off transmittance ratio (modulation depth) exceeds 100. The required gate bias difference for the switching operation is calculated to be 11.5 V, a value which can be further reduced by setting the on-state \( E_F \) closer to the off-state \( E_F \). We also note that the active area of the device is unprecedentedly small, since the gap width is more than 1000 times shorter than the free space wavelength, and more than 10 times shorter than the graphene plasmon wavelength. This extreme miniaturization is advantageous not only for high-density integration purposes, but also for high-speed and energy efficient operations owing to the reduced device capacitance.
4. Structural parameter and carrier mobility dependence

The transmission modulation characteristics of the device can be altered by engineering the geometric parameters of the system. As shown in Fig. 1(b), in the narrow gap regime (g < 40 nm), the transmittance dip becomes broader and is shifted to a higher $E_F$ with increasing g. The broadening of the transmittance dip can be attributed to the fact that the nanogap mode becomes less confined and thus more efficiently leaks out to free space (i.e., $|r_{NM}|$ and $|r_{NF}|$ decrease) as the gap widens. Figure 3(a) shows that, despite broadening of the $t_{NG}$, the difference between $|t_{NG}|$ and $|t_B|$ stays similar at the destructive interference condition. Consequently, the minimum transmittance value marginally depends on the gap width. The increasing gap width also affects the phase of $t_{NG}$ as shown in Fig. 3(b) (note that the phase of the background transmission $t_B$ is nearly zero regardless of the gap size). As the gap widens, more carriers in graphene are required to induce destructive interference between $t_{NG}$ and $t_B$ with their phase difference being equal to $\pi$, which explains the transmittance dip shift toward higher $E_F$. The dependence of the coupling and reflection coefficients on the gap width is discussed in Appendix 4.

Figure 4(a) illustrates the dependence of the net transmittance $T(E_F)$ on the cladding thickness $h$ for a given core thickness $d = 50$ nm. We note that the minimum transmittance Fermi level is linearly dependent on $h$ when $d$ and $g$ are fixed to 50 and 5 nm, respectively (Fig. 4(b)). This can be understood by recognizing that the increase of the metal cladding thickness ($\Delta h$)
is equivalent to the addition of $2i\beta_{NG}\Delta h$ phase shift of the propagating nanogap mode in the modified Fabry-Perot model (Eq. (1)), leading to a phase shift of $t_{NG}$ relative to $t_B$. Consequently, the destructive interference between $t_{NG}$ and $t_B$ occurs at a shifted Fermi level as shown in Fig. 4(c). On the other hand, the increasing core thickness $d$ reduces the effective index of the guided mode, resulting in the decrease (increase) of coupling (reflection) coefficients, which makes the transmittance dip narrower, as shown in Fig. 4(d). As $d$ increases, the amplitude of the background transmission $|t_B|$ increases (Fig. 4(e)), leading to an increased $T_{on}$. However, in this case, the $E_F$ at which the minimum transmittance occurs stays nearly the same, since the phase difference between $t_{NG}$ and $t_B$ is preserved (Fig. 4(f)). At the same time, this also causes $T_{off}$ to rise by making the destructive interference between $t_B$ and $t_{NG}$ less complete due to their enlarged amplitude mismatch. Consequently, the modulation depth ($T_{on}/T_{off}$) of the device decreases with increasing $d$ and $h$ as shown in Fig. 5(a).

In graphene-based active optical devices, the carrier mobility of graphene is one of the key factors that limits the device performance. The carrier mobility is inversely proportional to the non-radiative loss of the devices, and is thus desired to be as high as possible. The experimentally reported mobility of graphene reaches up to a few millions of cm$^2$/Vs if the sample is prepared by mechanical exfoliation and measured at a low carrier concentration under cryogenic conditions [26]. However, large-scale optical device applications necessitate CVD-grown graphene with high carrier concentration ($n > 5 \times 10^{12}$ cm$^{-2}$) operating at room temperature, which tends to have a significantly deteriorated carrier mobility ranging from several hundred to a few thousand cm$^2$/Vs [27,28]. To characterize the tolerance on the carrier mobility of our device, we calculate the transmittance for the various graphene carrier mobility $\mu$ spanning 500 to 50,000 cm$^2$/Vs. As shown in Fig. 5(b), the switching behavior of the nanogap device is remarkably stable under carrier mobility variation, and one can see a noticeable degradation when the mobility drops under 1000 cm$^2$/Vs. We attribute this to the operating mechanism of the device, which employs the impedance modulation of the graphene-covered nanogap cavity mode rather than the propagation of graphene plasmons that are highly sensitive to the ohmic loss.
Fig. 4. (a) The waveguide transmittance $T$ as a function of the graphene Fermi level $E_F$ and the cladding thickness $h$ at fixed core thickness $d = 50$ nm. (b) The amplitude of $t_{NG}$ (solid) and $t_B$ (dashed) as a function of $E_F$ at different $h; d = 50$ nm. (c) The phase difference between $t_{NG}$ and $t_B$ as a function of $E_F$ at different $h; d = 50$ nm. (d) The waveguide transmittance $T$ as a function of $E_F$ and core thickness $d$ for when $h = 30$ nm. (e,f) Same as in (b,c) for different $d; h = 30$ nm. Dotted line in (c,f) indicates the destructive interference condition $\phi_{NG} - \phi_B = \pi$. The gap width $g = 5$ nm in all cases.

Fig. 5. (a) The maximum modulation depth as a function of $d$ and $h$. (b) Transmittance as a function of $E_F$ for different carrier mobility $\mu$ of graphene. Geometric parameters are $h = 30$ nm, $d = 50$ nm, and $g = 5$ nm. Inset: transmittance in “on” and “off” states as a function of $\mu$. 
5. Discussion on quantum and non-local effects

Since the length scale of electromagnetic field confinement in the proposed device is comparable to the electron wavelength of its constituent plasmonic materials, a discussion of quantum and non-local effects that may influence the local classical electromagnetic analysis is deemed necessary. Let us first discuss the quantum tunneling and quantization effect. First, the probability of quantum tunneling across a few-nm air gap should be negligible, because the threshold length scale associated with electron tunneling is \(< 1\) nm for a 1 eV work function difference, which is sufficiently smaller than the gap width of our device \([29,30]\). Second, it is well known that the optical properties of nanostructures, including graphene plasmonic nanoresonators, can be significantly altered by electronic energy quantization when the size of the system is comparable to the electron wavelength (typically \(\sim 10\) nm) \([31]\). We note that the metal on graphene leads to doping due to a difference in work function \([32]\). Hence, the graphene underneath can be modeled as doping variations across the graphene nanogap, which has minimal effect on electrical conduction due to Klein tunneling \([33]\).

The nonlocal effect originates from the spatial spreading of electron wave function. It has been theoretically predicted that the plasmonic properties of sub-10 nm graphene nanoresonators deviate from that of the local classical electromagnetic model due to non-local effects \([34]\). We speculate that the non-local effect would also be present in our case, but its influence on the device operation would be smaller compared to the case of a physically patterned graphene plasmonic resonator for the following reasons: First, the operation mechanism of the proposed nanogap device does not rely on the plasmonic properties of graphene, and thus the in-plane spatial field variation in the graphene is almost negligible. Second, due to the conducting boundaries between the graphene and the Au top cladding, the current density distribution in the proposed device is much smoother compared to the graphene plasmonic resonators with abrupt physical terminations. From the perspective of the hydrodynamic model, the reduced gradient of the current density should result in a smaller hydrodynamic pressure for electrons and thus a weaker non-local effect \([35]\). Although numerical methods for quantitatively estimating the non-local effects have been under active development, it is still difficult to rigorously treat our system, which involves multiple materials and non-trivial geometries, using existing computational methods. Therefore, we defer more rigorous numerical investigations of the nanogap non-local effects, which we anticipate will lead to only quantitative corrections, for future studies.

6. Conclusion

In summary, we theoretically investigate an exotic transmission modulation behavior across a sub-10 nm MIM waveguide gap covered by graphene. Here, the interference between the background transmission and the transmission through the tunable nanogap mode enables large on-off transmission modulations of \(T_{\text{on}}/T_{\text{off}} > 100\) with the active device length being \(\lambda_0/1000\), which is more than 10 times shorter than the wavelength of highly confined graphene plasmons. In addition, the device operation is robust under graphene carrier mobility variation, exhibiting a well-preserved switching behavior even at a low mobility of 1000 cm\(^2\)/Vs. The ultra-compact form factor, high modulation efficiency, and robustness makes the proposed nanogap waveguide architecture a promising platform for future integrated-photonic systems.

Appendix 1: transmission mechanisms via the waveguide-integrated nanogap

We schematically illustrate the charge distribution of the plasmonic modes in the nanogap region in Fig. 6; both modes contribute to the waveguide transmission. The tightly confined nanogap mode (Fig. 6(a)) is a product of coupling between the graphene and MIM gap plasmons. On the other hand, the background transmission (Fig. 6(b)) is mediated by the continuum of the modes that exist in the gap region excluding the nanogap mode. The two main transmission channels –
via the nanogap and via the background modes – have different dependencies on the Fermi level and can be separately calculated as described in Appendix 3.

\[
\text{(a) (b)}
\]

Nanogap Mode  
Background Transmission

**Fig. 6.** The charge distribution schematics of (a) the nanogap mode and (b) background transmission in the MIM waveguide-integrated plasmonic cavity.

**Appendix 2: derivation of the modified Fabry-Perot cavity model**

We have developed a semi-analytical model to calculate the transmission via the Fermi level-dependent nanogap mode by summing up the infinite series of the partial transmissions occurring at every round trip of the nanogap plasmon as schematically shown in Fig. 7. The complex partial transmission for each iteration is given as:

\[
\begin{align*}
t_1 &= c_{MN}e^{i\beta_{NG}h}r_{NF}e^{i\beta_{NF}h}c_{NM} = c_{MN}c_{NM}e^{2i\beta_{NG}h}r_{NF} \\
t_2 &= c_{MN}e^{i\beta_{NG}h}r_{NF}e^{i\beta_{NF}h}r_{NM}e^{i\beta_{ch}h}r_{NF}e^{i\beta_{NF}h}c_{NM} = t_1 \left( r_{NM}r_{NF}e^{2i\beta_{ch}h} \right) \\
&\vdots \\
t_N &= t_1 \left( r_{NM}r_{NF}e^{2i\beta_{ch}h} \right)^{N-1}.
\end{align*}
\]

Therefore, the cumulative transmission via the nanogap mode \( t_{NG} \) is

\[
t_{NG} = \sum_{k=1}^{\infty} t_k = \frac{t_1}{1 - r_{NM}r_{NF}e^{2i\beta_{ch}h}} = \frac{c_{MN}c_{NM}e^{2i\beta_{NG}h}r_{NF}}{1 - r_{NM}r_{NF}e^{2i\beta_{ch}h}}.
\]

Here, \( \beta_{NG} \) is the vertical propagation constant of the nanogap plasmon. \( c_{ij} \) and \( r_{ij} \) are the coupling and reflection coefficients, respectively, for different interfaces where \( i, j = N \) (nanogap), M (MIM waveguide), and F (free space), and are obtained from full-wave simulations as described below.

**Fig. 7.** Conceptual illustration of the nanogap plasmon mode propagation and coupling.
Appendix 3: calculation of $c_{ij}$, $r_{ij}$ and $t_B$

The reflection coefficients ($r_{ij}$), coupling coefficients ($c_{ij}$), and the background transmission coefficient ($t_B$) can be obtained by performing electromagnetic simulations of the structures illustrated in Fig. 8(a) and (b). Figure 8(a) is a schematic of a plasmonic waveguide T-junction connecting Au-SiO$_2$-Au (ASA) and Au-Air-Au (AAA) waveguides. The cladding thickness of the AAA waveguide is set to 100 nm, which is sufficiently thicker than the skin depth of Au. Figure 8(b) shows a schematic of a terminated AAA waveguide with an open end.

$$c_{MN} = \frac{E_5 \exp(-in_{ASA}k_0L_5)}{E_1 \exp(-in_{ASA}k_0L_1)}, \quad c_{NM} = \frac{E_6 \exp(-in_{ASA}k_0L_6)}{E_2 \exp(-in_{ASA}k_0L_2)}, \quad r_{NM} = \frac{E_5 \exp(-in_{ASA}k_0L_5)}{E_2 \exp(-in_{ASA}k_0L_2)}, \quad t_B = \frac{E_6 \exp(-in_{ASA}k_0L_6)}{E_1 \exp(-in_{ASA}k_0L_1)}.$$  

Here, $E_i$ is the complex amplitude of the transverse electric field component of TM$_0$ mode measured at the port number $i$; $n$, $k_0$, and $L_i$ are the effective index of the TM$_0$ mode, free space wavevector, and distance between port $i$ and the nanogap waveguide junction. In $t_B$ calculation, the Fermi level of graphene is assumed to be 0.01 eV to exclude graphene plasmon transmission. Coefficient $r_{NF} = \frac{E_8 \exp(-in_{AAA}k_0L_8)}{E_7 \exp(-in_{AAA}k_0L_7)}$ is calculated in the open-ended waveguide structure shown in Fig. 8(b).

Appendix 4: dependence of $c_{ij}$, $r_{ij}$ and $t_B$ on geometrical parameters

The dependence of the coupling, and reflection coefficients on the nanogap width $g$ are presented in Fig. 9. The coupling coefficients $c_{MN}$ and $c_{NM}$ have identical amplitude with $\pi$ phase difference owing to the nature of the E-type T-junction. $r_{NF}$ stays near-unity and exhibits a marginal dependence on $g$ due to the huge mode mismatch between the nanogap mode and the free space radiation. On the other hand, the amplitude of $r_{NM}$ decreases with a considerable amount of phase variation as the nanogap mode becomes less confined with increasing $g$. 

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig8.png}
\caption{(a) T-junction, and (b) the open-ended MIM waveguides. Numbers indicate the incident and reflection ports, and colored arrows show the propagation direction of the waveguide mode.}
\end{figure}
Fig. 9. Amplitude and phase of (a) the coupling coefficient $c_{NM}$ and the reflection coefficients (b) $r_{NF}$ and (c) $r_{NM}$ as functions of the gap width $g$. (d) dependence of the effective index of nanogap mode $n_{NG}$ on $g$. The other device parameters are set as $h = 30$ nm, $d = 50$ nm, $\mu = 10,000$ cm$^2$/Vs, and $E_F = 0.26$ eV, respectively.

Funding. National Research Foundation of Korea (2017R1E1A1A01074323, 2019K1A3A1A0464929, BK21 FOUR Program); Villum Fonden (16498); Syddansk Universitet; Det Frie Forskningsråd (0165-00015B).

Disclosures. The authors declare no conflicts of interest.

References