Plasmon Damping in Electronically Open Systems

Kirill Kapralov and Dmitry Svintsov[*]
Center for Photonics and 2d Materials, Moscow Institute of Physics and Technology, Dolgoprudny 141700, Russia

(Received 6 May 2020; accepted 4 November 2020; published 1 December 2020)

Rapid progress in electrically controlled plasmonics in solids poses a question about possible effects of electronic reservoirs on the properties of plasmons. We find that plasmons in electronically open systems [i.e., in (semi)conductors connected to leads] are prone to an additional damping due to charge carrier penetration into contacts and subsequent thermalization. We develop a theory of such lead-induced damping based on the kinetic equation with microscopic boundary conditions at the interfaces, followed by perturbation theory with respect to transport nonlocality. The lifetime of the plasmon in an electronically open *ballistic* system appears to be finite, of the order of conductor length divided by carrier Fermi velocity. The reflection loss of the plasmon incident on the contact of the semiconductor and perfectly conducting metal also appears to be finite, of the order of Fermi velocity divided by wave phase velocity. Recent experiments on plasmon-assisted photodetection [Nat. Commun. **9**, 5392 (2018)] are discussed in light of the proposed lead-induced damping phenomenon.

DOI: 10.1103/PhysRevLett.125.236801

Plasmons represent collective oscillations of charge carriers and an electromagnetic field. Both instances can freely propagate in space. Free propagation of electromagnetic waves leads to the radiative decay of plasmons which has been studied extensively [1–4]. Charge carriers may also leak away from plasmonic system if it is coupled to electronic reservoirs (contacts). This process would also lead to plasmon damping. Unlike radiative damping, "contact damping" has gained very little attention because most plasmonic systems studied so far were either electronically closed (nanoparticles [1–3]) or extended [4–6]. With the rapid progress in the electrical control of plasmons [7,8] and electrical readout of plasmon-enhanced photocurrent [9,10], this damping pathway becomes urgent.

There are several experimental evidences for the important role of contacts on plasmon damping that have not received due attention. First, interference of launched and reflected plasmons is readily observed at the edges of 2D and 1D semiconductors [5,11,12], but scarcely observed at contacts of semiconductors and metals [7]. Second, photocurrent spectroscopy of plasmon resonance in transistor structures with close leads provides generally larger linewidths [13,14] compared to electromagnetic transmission measurements in grating-gated semiconductors with distant leads [15,16]. In recent measurements of plasmon-enhanced photovoltage in a graphene bilayer transistor [10], the visibility of plasmon resonance was enhanced by $p-n$ junction barrier at the metal-graphene interface. These factors tell us that the behavior of plasmons at the contact of the semiconductor and metal is not simply refection at impedance discontinuity.

The theory of plasmon decay in electronically open systems is still lacking. Its first theoretical evidence appeared in numerical simulations of plasmons in confined 2D systems with fixed electron distributions at contacts [17]. Further evidence appeared in simulations of current-driven plasmon instability [18], though hardly distinguishable from bulk damping. Another approach to the problem lies in finding the dynamic conductance of lead-coupled conductors and analyzing its peaks as a function of

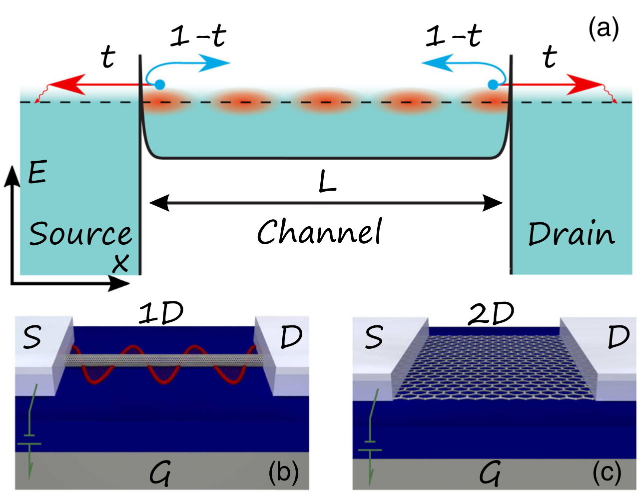

FIG. 1. (a) Electron occupation in a semiconductor channel coupled to metal leads vs energy and coordinate in the presence of plasma wave. Blue filled regions correspond to equilibrium electrons, red—to nonequilibrium ones, excited by the plasmon electric field. A nonequilibrium electron incident on a contact can either transmit and thermalize (with probability t) or be reflected back from a contact (with probability $1-t$). (b) Open plasmonic resonators with 1D and 2D channels (nanotube and graphene are shown as example) coupled to source and drain contacts.

frequency [19]. Such calculation could be performed only for 1D systems with restrictive assumption of the fully screened Coulomb interaction [20]. Remarkably, the latter approach hints that plasmon damping in open systems is tightly linked to the finite conductance of ballistic systems (given by the Landauer [21] and Sharvin [22] formulas in 1D and 2D, respectively).

In this Letter, we present an analytical theory of plasmon decay in 1D and 2D (semi)conductors coupled to leads [23]. A schematic of this process is shown in Fig. 1(a): a nonequilibrium electron participating in plasma oscillation penetrates into a contact and is thermalized therein. We find that such damping appears to be due to the nonlocality of the current-field response, i.e., it vanishes if the Fermi velocity v_0 in the channel tends to zero. Our effect should be distinguished from a plethora of *nondissipative* nonlocal effects in plasmonic systems [24–26]. In the low-temperature limit, the damping is of order tv_0/L, where L is the distance between leads, v_0 is the Fermi velocity, and t is the electron transmission probability at semiconductor-metal interface. An extra contribution to damping oscillating with frequency is found. It appears to be due to synchronization between the carrier transit and plasma oscillation. We find that plasmons incident on the semiconductor-metal contact experience reflection loss due to the above mechanism. It occurs even for ballistic semiconductors and perfectly conducting metals.

Nonlocal conductivity in electronically open system.— The main building block for the evaluation of plasmon losses in open systems is the conductivity kernel $\sigma(x,x')$ linking the current density $j(x)$ and electric field $E(x)$

$$j(x) = \int_0^L \sigma(x,x')E(x')dx'. \tag{1}$$

We find $\sigma(x,x')$ in the d-dimensional semiconductor channel with metal leads located at $x=0$ and $x=L$ [Fig. 1(a) and 1(c)]. The metals are assumed to be perfectly conducting, and strong electron scattering maintains the equilibrium Fermi distributions therein. The latter fact (being a definition of perfect contact [27]) is justified when the carrier collision frequency in metal $\gamma_m \sim 10^{13},\ldots,10^{14}\ \mathrm{s}^{-1}$ [28] exceeds that in the channel and the plasmon frequency ω. Electron distribution in the channel obeys the classical kinetic equation being valid if $\hbar\omega$ is below electron Fermi energy ε_F. Above requirements are fulfilled for terahertz plasmons. In the presence of electric field $E(x)e^{-i\omega t}$, the distribution function $f(x,\mathbf{p}) = f_0(\mathbf{p}) + \delta f(x,\mathbf{p})e^{-i\omega t}$ obeys

$$-i(\omega+i0)\delta f + v_x\frac{\partial \delta f}{\partial x} - eE(x)\frac{\partial f_0}{\partial p_x} = 0, \tag{2}$$

where v_x is the x component of electron velocity, and f_0 is the equilibrium Fermi function. We shall focus on ballistic systems with long momentum relaxation time $\tau\omega \gg 1$ to distinguish the bulk and contact damping.

A nonequilibrium electron incident on the semiconductor-metal junction can either penetrate and thermalize therein with probability t, or undergo specular reflection with probability $r = 1 - t$. These considerations relate the distributions of left- and right-moving electrons at the contacts [29]:

$$\delta f(0,p_x) = r\delta f(0,-p_x),$$
$$\delta f(L,-p_x) = r\delta f(L,p_x). \tag{3}$$

Solving the kinetic equation [Eq. (2)] in the low-temperature limit $k_BT/\varepsilon_F \ll 1$, we obtained the nonlocal conductivity kernel of the form [30]

$$\sigma(x,x') = \sigma_D d\frac{i\omega}{2v_F}\left\langle F\left(\frac{x}{L},\frac{x'}{L},\Omega_\theta\right)\cos\theta\right\rangle_{\cos\theta>0}, \tag{4}$$

$$F(\xi,\eta,\Omega_\theta) = e^{+i(|\xi-\eta|/\Omega_\theta)} - 2r\frac{\cos[\frac{\xi+\eta-1}{\Omega_\theta}] - r\cos[\frac{\xi-\eta}{\Omega_\theta}]e^{+i/\Omega_\theta}}{e^{-i/\Omega_\theta} - r^2e^{+i/\Omega_\theta}}, \tag{5}$$

where $\sigma_D = in_0e^2/\omega m$ is the local Drude conductivity, n_0 is the electron density, $\Omega_\theta = v_F\cos\theta/\omega L$ is the dimensionless transit frequency of electron moving at angle θ, and the angular averaging $\langle\cdots\rangle_{\cos\theta>0}$ is performed over right-moving carriers. The local Drude conductivity $\sigma(x,x') \approx \sigma_D\delta(x-x')$ is restored in the limit of small normalized transit frequency $\Omega_t = v_F/\omega L \ll 1$, i.e., at high frequencies and low Fermi velocities.

Damping rate in electronically open resonator.— Plasmon frequencies ω_n and field distributions $E_n(x)$ in open resonators (n enumerates discrete modes) can be found once material relations Eqs. (1) and (4) are supplemented with Poisson's equation. An exact solution of such a spectral problem for bounded nonlocal conductors looks impossible. Fortunately, evaluation of damping rate due to contacts, $\gamma_{\mathrm{cont}} = -\omega_n''$ is feasible with the aid of energy balance considerations and perturbation theory with respect to nonlocality [30,44]. The former states that the damping rate equals the loss rate of wave energy Q divided by twice the stored energy W, $\gamma_{\mathrm{cont}} = Q/2W$. Both Q and W are functionals of yet unknown field distributions $E_n(x)$. The essence of perturbation theory is to replace the true eigenfunctions $E_n(x)$ and frequencies ω_n with those obtained within the local Drude model of conduction, $E_n^{(0)}(x)$ and $\omega_n^{(0)}$. The damping rates obtained within such approximation will be valid to the leading order in transit frequency $\Omega_t = v_F/\omega L$.

The above premises are sufficient to derive expressions for lost and stored energy, Q and W, and hence, the damping rate due to contacts:

$$\gamma_{\rm cont} = \frac{1}{2}\frac{\int_0^L E_n^{(0)}(x')\sigma'(x,x')E_n^{(0)}(x)dxdx'}{\int_0^L \frac{\sigma_D''}{\omega_n^{(0)}}|E_n^{(0)}(x)|^2dx}, \qquad (6)$$

where the prime and double prime denote real and imaginary parts. The numerator of Eq. (6) represents the power developed by the field over the oscillating particles, this energy dissipating eventually in contacts [31]. The denominator of Eq. (6) is the total stored energy of plasma oscillation. Upon derivation, we have used Brillouin's formula for energy in dispersive media [32] and equality of average kinetic and potential energies.

One may wonder how the information about carrier leakage is encoded in the expression for damping rate Eq. (6) where integration is performed over the channel interior. In fact, it is contained in the conductivity kernel that is sensitive to microscopic boundary conditions. The finite dissipative (real) part of conductivity kernel appears only for permeable contacts ($r < 1$), and disappears for perfectly reflecting ($r = 1$) ones.

The effect of boundaries on bulk current $j(x)$ is visualized in Fig. 2(a), where we show its real and imaginary parts induced by the field of plasmon mode with $n = 2$, $E_2^{(0)}(x) \propto \cos 2\pi x/L$. Both $j'(x)$ and $j''(x)$ possess short fringes of wavelength $2\pi v_0/\omega$, while the generating field has a long wavelength of L. The fringe amplitude tends to zero for perfect reflection, and they are localized near the contacts in the presence of bulk damping. They are nothing but transit-time oscillations of current carried by Fermi surface electrons. Precisely these short fringes lead to nonzero work produced by the field on particles. The emergence of short-wavelength structures for a long-wavelength generating field allows us to interpret contact damping as spatially localized Landau damping.

We now quantify the contact contributions to plasmon damping in several experimentally relevant structures, such as (a) a 1D nanotube field-effect transistor (FET) and (b) a

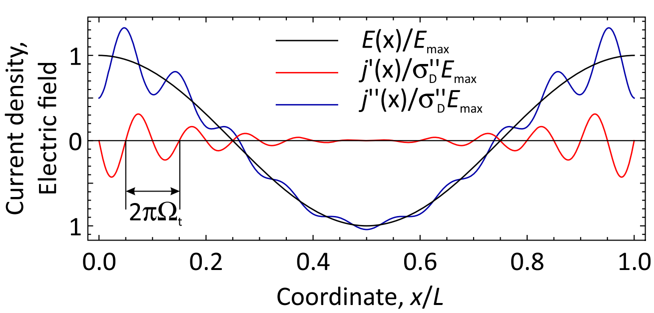

FIG. 2. Spatial distribution of the electric field (black) and induced current (red, blue) in a 1D conductor coupled to perfectly transmitting leads ($r = 0$) for the second plasmon mode ($n = 2$) at transit frequency $\Omega_t = 0.016$. To visualize the localization of the current, we have added a small imaginary part to the frequency in the conductivity kernel [Eq. (4)] mimicking the internal scattering, $\omega \to \omega + i\tau^{-1}$, $L/v_0\tau = 2$

FET with a 2D channel [Figs. 1(b) and 1(c)]. The zero-order field distributions $E_n^{(0)}$ are known here exactly [33] if the vertical extent of contacts much exceeds the plasmon wavelength:

$$E_n^{(0)}(x) = E_{\rm max}\cos(\pi n x/L). \qquad (7)$$

For structures with keen contacts, they can be used as an approximation with $\sim 10\%$ accuracy [34]. Evaluation of integrals in Eq. (6) with fields [Eq. (7)] yields the following estimates of damping $\gamma_{\rm cont}$:

$$\frac{\gamma_{\rm cont}}{\omega_n} = -\alpha(d)t\Omega_t - \beta(d)t^2\Omega_t^{(d+1)/2}\Phi_{\rm osc}(\Omega_t), \qquad (8)$$

$$\Phi_{\rm osc}(\Omega_t) = \sum_{k=1}^{\infty}\frac{r^{k-1}}{k^{\frac{d-1}{2}}}\cos\left[\frac{k}{\Omega_t} + \frac{\pi}{4}(d-1) + \pi n k\right]. \qquad (9)$$

The numerical prefactors α and β depend on channel dimensionality and are given by

$$\alpha(1) = 1; \qquad \alpha(2) = \frac{8}{3\pi}; \qquad \beta(1) = 1; \qquad \beta(2) = \sqrt{\frac{8}{\pi}}.$$

The contribution to damping given by first term of Eq. (8) is the inverse mean escape time of a free electron from the channel. It is linear in Fermi velocity and inverse to channel length, as shown in Fig. 3 with dashed lines. The coefficient α decreases in higher dimensions due to the longer transit time of the electron between the source and drain, *averaged* over the Fermi surface.

The second (oscillatory) term of the Eq. (8) describes possible resonances between the electric field and bouncing electrons [35]. An enhancement of damping occurs if the plasmon field acts in phase with most of the charge carriers, while its reduction occurs when the field and carriers oscillate in counterphase. The resonances become narrower with the reduction of transmittance t due to effective prolongation between particle-field interactions, as seen from the comparison of Figs. 3(a) and 3(b). The narrowing is most pronounced in 1D systems, where carrier transit times are not spread due to different directions of motion. It is possible to show that oscillatory features in $\gamma_{\rm cont}$ will be smeared both at finite temperature and finite internal damping, while the leading Ω_t term will not.

Energy loss of a wave incident at contact.—The contact mechanism of damping would also manifest itself in plasmonic interference phenomena near the metal-semiconductor contacts. Analysis of such interference patterns became an established tool for the determination of spectra and the propagation length of plasmons [5,11].

We now realize that plasmon reflection from semiconductor-metal contact can never be perfect as some fraction of carriers would penetrate into metal and thermalize therein. Attenuated reflection from such contact is a

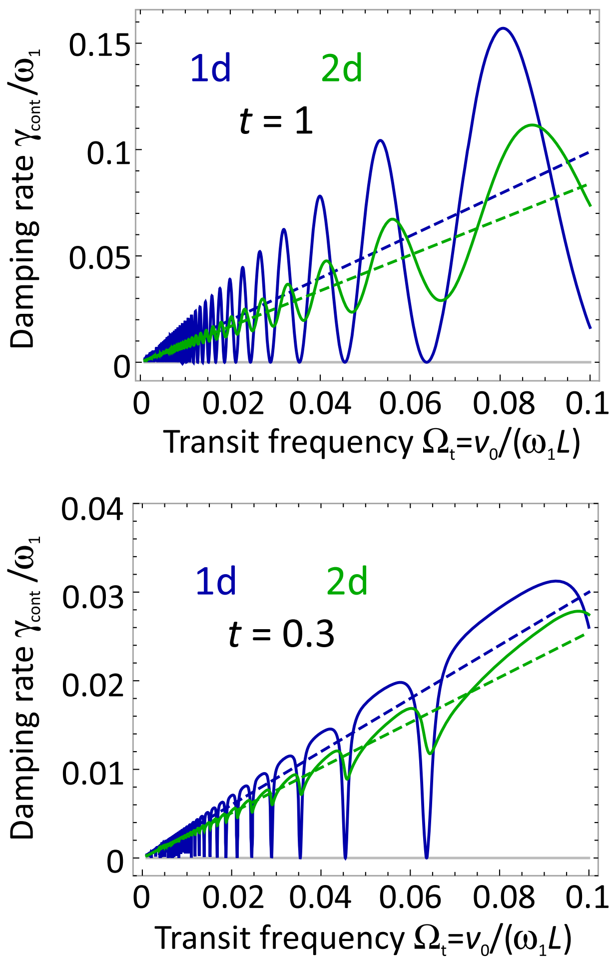

FIG. 3. Plasmon damping rate γ (normalized by eigenfrequency ω_1) in a ballistic semiconductor coupled to leads vs the electron transit frequency. Blue and green lines correspond to 1D and 2D channels, respectively; the upper panel corresponds to perfectly transmitting (Ohmic) contact, lower panel—to contact with transmittance $t = 0.3$. Dashed lines show the nonoscillatory part of damping $\alpha(d)t\Omega_t$.

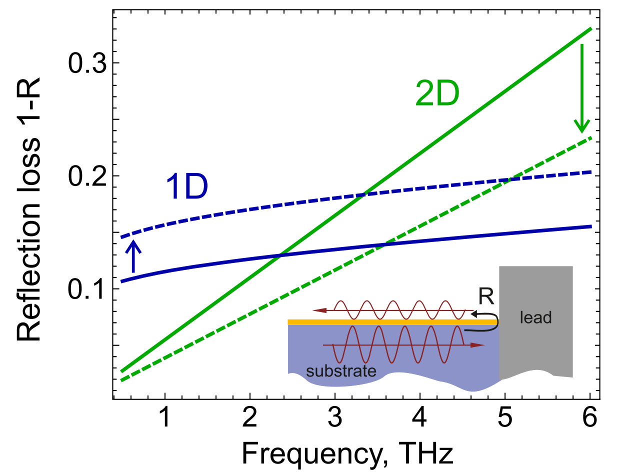

FIG. 4. Plasmon reflection loss upon scattering at semiconductor-metal contact vs wave frequency. Blue and green lines correspond to 1D (semiconductor nanotube [12]) and 2D (GaAs quantum well) plasmons; solid lines correspond to carrier densities $n = 5 \times 10^5\ \mathrm{cm}^{-1}$ in 1D and $5 \times 10^{11}\ \mathrm{cm}^{-2}$ in 2D; dashed lines—to higher densities of $10^6\ \mathrm{cm}^{-1}$ in 1D and $10^{12}\ \mathrm{cm}^{-2}$ in 2D. The radius of the 1D conductor is $a = 5$ nm, effective mass $m^* = 0.1 m_0$, background dielectric constant $\kappa = 4$. Inset shows a schematic of lossy plasmon reflection at the contact.

consequence of nonlocal conductivity; no attenuation would have occurred if the semiconductor was described by a local Drude conductivity.

The method for calculation plasmon reflection loss $A_{\mathrm{pl}} = 1 - R_{\mathrm{pl}}$ at the semiconductor-metal contact (located at $x = 0$) is similar to the evaluation of damping in a confined structure. Namely, it equals the Joule losses (expressed through nonlocal conductivity) divided by energy flux in an incoming wave S_{inc}:

$$A_{\mathrm{pl}} = \frac{1}{S_{\mathrm{inc}}} \int_{-\infty}^{0} E^{(0)}(x') \sigma'(x, x') E^{(0)}(x) dx dx'. \quad (10)$$

The last necessary element for evaluation of losses is the conductivity kernel in a semi-infinite semiconductor channel $\sigma(x, x')$. It is related to the nonlocal conductivity of the extended system $\sigma_\infty(x - x')$ via $\sigma(x, x') = \sigma_\infty(x - x') - r\sigma_\infty(x + x')$ [29]. It can be obtained from conductivity in the finite-length channel Eq. (4) by taking the limit $L \to \infty$ and keeping in mind small wave damping. As before, Eq. (10) is perturbative with respect to nonlocality, for this reason it is evaluated on field profiles obtained within the local model of conduction $E^{(0)}(x) = E_{\max} \cos qx$. Direct integration in Eq. (10) leads us to

$$A_{\mathrm{pl}} = \xi(d) t \frac{v_F}{\omega / q}, \quad (11)$$

where $\xi(1) = 1$, $\xi(2) = 16/3$. The result is remarkably simple: reflection loss of a plasmon incident on metal contact is the ratio of carrier Fermi and plasmon phase velocities, timed by the transmission coefficient for individual carrier t.

When the dispersion of 2D and 1D plasmons is known ($\omega_{2D} \propto q^{1/2}$ and $\omega_{1D} \propto q \ln^{1/2}(qa)^{-1}$, where a is the transverse size of quasi-1D conductor [36]), we can obtain the frequency dependence of reflection losses shown in Fig. 4. 1D plasmons exhibit enhanced loss at low frequencies due to the relative smallness of the phase velocity. Changes in refection losses with increasing carrier density n are governed by the interplay of the Fermi velocity enhancement $v_0 \propto n^{1/d}$ and "stiffening" of the plasmon dispersion $n^{1/2}$. In 1D systems, the Fermi velocity enhancement is dominant, and losses raise at higher density. In 2D systems (both in those with parabolic bands and graphene), the Fermi velocity changes slightly, and reflection losses become negligible at high density.

Discussion and possible experimental manifestations.—Lead-induced damping can make a contribution to the net plasmon damping in any electrically controlled plasmon resonator (e.g., in experiments with gate tuning of plasmons). The most pronounced effect is expected in plasmon-enhanced photodetectors, where a semiconductor channel acts as a plasmonic resonator and photocurrent generator [13,14,37]. The plasmon lifetime in such detectors based on bilayer graphene was inferred in Ref. [10] from the width of gate-tunable photovoltage oscillations. The extracted lifetime $\gamma^{-1} \sim 0.3, \ldots, 1$ ps was well below the double transport relaxation time (~4 ps) predicted by Boltzmann kinetic theory in an extended system. Moreover, the lifetime decreased at larger gate voltage (corresponding to higher Fermi velocities). The latter trend is in agreement with Eq. (8) for contact damping [38].

The magnitude of lead-induced damping under conditions of Ref. [10] is estimated using Eq. (8) as $8v_F/3\pi L \approx 1.7 \times 10^{11}\ \mathrm{s}^{-1}$ (taking $v_F = 10^6$ m/s and $L = 5\ \mu$m). It is of the same order as "bulk damping" $(2\tau_p)^{-1} = 2.5 \times 10^{11}\ \mathrm{s}^{-1}$. Yet, the net damping rate is even above $\gamma_{\mathrm{cont}} + (2\tau_p)^{-1}$, which signalizes on extra plasmon decay mechanisms (radiative decay is a likely candidate). Remarkably, measurements of plasmon-resonant detection in submicron III-V transistors ($L = 150$ nm) [13] also reported a plasmon lifetime ~200 fs smaller than the expected 800 fs from mobility measurements. The magnitude of the measured lifetime is close to $L/v_F \sim 150$ fs expected for the contact mechanism.

It looks like the contact mechanism of plasmon damping sets a limit for the downscaling of plasmonic nanosystems. A straightforward way for the reduction of such damping is to induce weakly transparent barriers for electrons near the contacts. One may therefore state that plasmonic structures should benefit from high *contact resistance* and *long channels*, which is contrary to requirements for conventional high-frequency transistors. Another possible way to reduce contact damping is to shift the carrier transport into the hydrodynamic regime. In this regime, the path of a single carrier to the contact would be prolonged due to frequent electron-electron collisions [39], and so will the plasmon lifetime. On the other hand, damping of plasmons in a short ballistic channel should neutralize the ringing response to pulses of the gate voltage. The latter was considered as a limiting factor to the response time in ultrascaled high-mobility transistors [40].

A slightly modified form of our damping mechanism would be present in open plasmonic resonators without metal contacts. An example of such a resonator comprises a *bounded* gate in close proximity to an *extended* 2D system, the edges of the gate acting as limiters for the plasmon field [41–43]. Because of the transport nonlocality (emerging ultimately from finite Fermi velocity), carriers do not follow the on-site field and can escape the resonator. The probability of carrier return from the *extended* channel to the finite gated region tends to zero, which acts as effective thermalization.

To conclude, we have shown that coupling of a semiconductor system to metal leads induces extra plasmon damping. The damping appears to be due to the electron penetration into the leads and subsequent thermalization. The contribution of this mechanism to the damping rate in a semiconductor of length L is roughly v_F/L; the contribution to the plasmon reflection loss at the semiconductor-metal interface is roughly the ratio of v_F and the wave phase velocity.

The authors thank Aleksandr S. Petrov for the rigorous derivation of Eq. (6) using the perturbation theory for 2D plasmons. The work was supported by the Grant No. 16-19-10557 of the Russian Science Foundation.

*svintcov.da@mipt.ru

[1] J. Crowell and R. H. Ritchie, Phys. Rev. **172**, 436 (1968).

[2] T. Kokkinakis and K. Alexopoulos, Phys. Rev. Lett. **28**, 1632 (1972).

[3] T. V. Teperik, V. V. Popov, and F. J. García de Abajo, Phys. Rev. B **69**, 155402 (2004).

[4] I. V. Kukushkin, J. H. Smet, S. A. Mikhailov, D. V. Kulakovskii, K. von Klitzing, and W. Wegscheider, Phys. Rev. Lett. **90**, 156801 (2003).

[5] A. Woessner, M. B. Lundeberg, Y. Gao, A. Principi, P. Alonso-González, M. Carrega, K. Watanabe, T. Taniguchi, G. Vignale, M. Polini, J. Hone, R. Hillenbrand, and F. H. L. Koppens, Nat. Mater. **14**, 421 (2015).

[6] G. Ni, A. McLeod, Z. Sun, L. Wang, L. Xiong, K. Post, S. Sunku, B.-Y. Jiang, J. Hone, C. R. Dean *et al.*, Nature (London) **557**, 530 (2018).

[7] A. Woessner, Y. Gao, I. Torre, M. B. Lundeberg, C. Tan, K. Watanabe, T. Taniguchi, R. Hillenbrand, J. Hone, M. Polini *et al.*, Nat. Photonics **11**, 421 (2017).

[8] D. Li and M. I. Stockman, Phys. Rev. Lett. **110**, 106803 (2013).

[9] M. B. Lundeberg, Y. Gao, A. Woessner, C. Tan, P. Alonso-González, K. Watanabe, T. Taniguchi, J. Hone, R. Hillenbrand, and F. H. Koppens, Nat. Mater. **16**, 204 (2017).

[10] D. A. Bandurin, D. Svintsov, I. Gayduchenko, S. G. Xu, A. Principi, M. Moskotin, I. Tretyakov, D. Yagodkin, S. Zhukov, T. Taniguchi *et al.*, Nat. Commun. **9**, 1 (2018).

[11] Z. Shi, X. Hong, H. A. Bechtel, B. Zeng, M. C. Martin, K. Watanabe, T. Taniguchi, Y.-R. Shen, and F. Wang, Nat. Photonics **9**, 515 (2015).

[12] S. Wang, S. Zhao, Z. Shi, F. Wu, Z. Zhao, L. Jiang, K. Watanabe, T. Taniguchi, A. Zettl, C. Zhou *et al.*, Nat. Mater. **19**, 986 (2020).

[13] W. Knap, Y. Deng, S. Rumyantsev, and M. S. Shur, Appl. Phys. Lett. **81**, 4637 (2002).

[14] J. D. Chudow, D. F. Santavicca, and D. E. Prober, Nano Lett. **16**, 4909 (2016).

[15] A. V. Muravjov, D. B. Veksler, V. V. Popov, O. V. Polischuk, N. Pala, X. Hu, R. Gaska, H. Saxena, R. E. Peale, and M. S. Shur, Appl. Phys. Lett. **96**, 042105 (2010).

[16] D. Heitmann, Surf. Sci. **170**, 332 (1986).

[17] A. Satou, V. Ryzhii, V. Mitin, and N. Vagidov, Phys. Status Solidi (b) **246**, 2146 (2009).

[18] C. B. Mendl and A. Lucas, Appl. Phys. Lett. **112**, 124101 (2018).

[19] Y. M. Blanter, F. W. J. Hekking, and M. Büttiker, Phys. Rev. Lett. **81**, 1925 (1998).

[20] The boundary conditions on contacts cannot be posed in a consistent way under the assumption of the fully screened Coulomb interaction. The local relation between potential φ and electron density n, $C\varphi = en$ leads to a fixed electron density at the grounded contacts. This differs generally from microscopic boundary conditions derived from the kinetic equation.

[21] R. Landauer, Philos. Mag. **21**, 863 (1970).

[22] Y. V. Sharvin, Sov. Phys. JETP **48**, 655 (1965), http://www.jetp.ac.ru/cgi-bin/dn/e_021_03_0655.pdf.

[23] The studied damping channel is also relevant for bulk (3D) plasmons, but its theoretical description is more complicated. The reason lies in the degeneracy of the 3D plasmon spectrum when spatial dispersion of conductivity is neglected.

[24] P. Gonçalves, T. Christensen, N. M. Peres, A.-P. Jauho, I. Epstein, F. H. Koppens, M. Soljačić, and N. A. Mortensen, arXiv:2008.07613.

[25] E. J. C. Dias, D. A. Iranzo, P. A. D. Gonçalves, Y. Hajati, Y. V. Bludov, A.-P. Jauho, N. A. Mortensen, F. H. L. Koppens, and N. M. R. Peres, Phys. Rev. B **97**, 245405 (2018).

[26] C. Ciracì, R. T. Hill, J. J. Mock, Y. Urzhumov, A. I. Fernández-Domínguez, S. A. Maier, J. B. Pendry, A. Chilkoti, and D. R. Smith, Science **337**, 1072 (2012).

[27] S. Datta, *Quantum Transport: Atom to Transistor* (Cambridge University Press, Cambridge, England, 2005), see Chap. "8.4 What constitutes a contact (reservoir)?".

[28] D. I. Yakubovsky, A. V. Arsenin, Y. V. Stebunov, D. Y. Fedyanin, and V. S. Volkov, Opt. Express **25**, 25574 (2017).

[29] G. Reuter and E. Sondheimer, Proc. R. Soc. A **195**, 336 (1948).

[30] See Supplemental Material available at http://link.aps.org/supplemental/10.1103/PhysRevLett.125.236801 for (I) detailed solution of kinetic equation in bounded channel with partially transparent contacts and (II) derivation of plasmon loss rate from perturbation theory for 2D plasmons.

[31] The direct flux of electron kinetic energy into contact $S = \sum_{\mathbf{p}} \delta^2 f v_x m v^2/2$ [where $\delta^2 f$ is distribution function being second order in the field] is proportional to the third power of the carrier velocity, and is smaller than the Joule-type loss proportional to the first power of v_0.

[32] L. D. Landau, L. Pitaevskii, and E. Lifshitz, *Electrodynamics of Continuous Media*, Vol. 8 (Elsevier, New York, 2013), see Chap. 80 "The field energy in dispersive media".

[33] D. Svintsov, Phys. Rev. Applied **10**, 024037 (2018).

[34] V. Ryzhii, A. Satou, I. Khmyrova, A. Chaplik, and M. S. Shur, J. Appl. Phys. **96**, 7625 (2004).

[35] A. P. Dmitriev and M. S. Shur, Appl. Phys. Lett. **89**, 142102 (2006).

[36] S. Das Sarma and W.-y. Lai, Phys. Rev. B **32**, 1401 (1985).

[37] M. Dyakonov and M. Shur, IEEE Trans. Electron Devices **43**, 380 (1996).

[38] Other possible damping pathways not related to transport relaxation, such as interband absorption and electron viscosity, would be more pronounced at low carrier densities and short wavelengths. Both correspond to low gate voltages, which contrasts to experimental observations. Damping due to dielectric losses in surrounding media [5] is irrelevant at THz frequencies below the frequency of the optical phonon.

[39] R. Gurzhi and S. Shevchenko, Sov. Phys. JETP **27**, 863 (1968), http://www.jetp.ac.ru/cgi-bin/dn/e_027_05_0863.pdf.

[40] S. Rudin, G. Rupper, and M. Shur, J. Appl. Phys. **117**, 174502 (2015).

[41] A. A. Zabolotnykh and V. A. Volkov, Phys. Rev. B **99**, 165304 (2019).

[42] I. Epstein, D. Alcaraz, Z. Huang, V.-V. Pusapati, J.-P. Hugonin, A. Kumar, X. M. Deputy, T. Khodkov, T. G. Rappoport, J.-Y. Hong, N. M. R. Peres, J. Kong, D. R. Smith, and F. H. L. Koppens, Science **368**, 1219 (2020).

[43] V. M. Muravev, P. A. Gusikhin, A. M. Zarezin, I. V. Andreev, S. I. Gubarev, and I. V. Kukushkin, Phys. Rev. B **99**, 241406(R) (2019).

[44] A. S. Petrov and D. Svintsov, Phys. Rev. B **99**, 195437 (2019).