When the RC Model Stops Predicting Your Rise Time

The two response-time limits that no amount of load reduction removes

Jezdan Gomez
Opto Diode

Abstract.
Note 03 of this series gave the circuit response time of a photodiode as t_r = 2.2RC and showed that the capacitance falls with reverse bias. Read as the whole story it invites two moves, reducing the load resistance and raising the bias, and both stop working. Response time is limited by three mechanisms, not one: the circuit time constant, the transit of carriers across the depletion region, and the diffusion of carriers generated behind it. Only the first contains the load. For a 100 mm^2 silicon photodiode on a 300 μm active layer, read at 800 nm with the external load already taken to zero, the circuit term at zero bias is 2.2 ns while the measured transition is 23.3 ns, a factor of 10.6. Widening the depletion region to lower the capacitance lengthens the transit, so the two limits move in opposite directions with bias and their sum has a minimum: for this device 1.20 ns at 32.9 V, beyond which more bias makes the detector slower. The diffusion term is set by wavelength, because the absorption depth decides how much light lands outside the field. At fixed bias it takes the response from 2.21 ns at 800 nm to 2710 ns at 1000 nm.

1. A limit the load cannot reach

The first-order speed model for a photodiode is the one note 03 of this series develops: the diode is a current source charging its own junction capacitance through the resistance in the loop, the 10 to 90% rise time is 2.2RC, and the capacitance falls as the square root of reverse bias. The model is quantitative and it reproduces published rise-time specifications for large-area devices. Its two levers are obvious. Lower the resistance and lower the capacitance, and the response gets faster without limit.

Neither lever runs out gracefully. An engineer who moves a large-area detector from a 50 Ω load into a transimpedance front end has replaced the load with a virtual ground, so the only resistance left in the loop is the diode's own series resistance. The circuit time constant collapses by more than an order of magnitude and the measured response does not follow it down. Raising the reverse bias then produces a further improvement that is smaller than the capacitance ratio promises, stops, and eventually reverses.

The model is not wrong. It is incomplete. Photodiode response speed is set by three mechanisms acting together: drift of carriers across the depletion region, diffusion of carriers generated outside it, and the capacitance of the depletion region working against the resistance in the circuit [1]. The *RC* term is the only one of the three that contains anything the circuit designer owns. The other two are properties of the silicon, and they are what the response falls back on once the circuit term is out of the way.

2. Three limits, combined

Each mechanism sets a characteristic time, and the three combine as the rise times of cascaded stages do, in quadrature [7]:

$$t_r \simeq \sqrt{t_{\mathrm{RC}}^2 + t_{\mathrm{tr}}^2 + \left(f_{\mathrm{diff}} t_{\mathrm{diff}}\right)^2} \tag{1}$$

where f_{diff} is the fraction of the photocurrent that arrives by diffusion rather than by drift. That weight is necessary: the diffusion time is enormous compared with the other two, and without it every photodiode would be predicted to be microseconds slow at every wavelength. Section 5 derives it. Figure 1 shows where each term comes from.

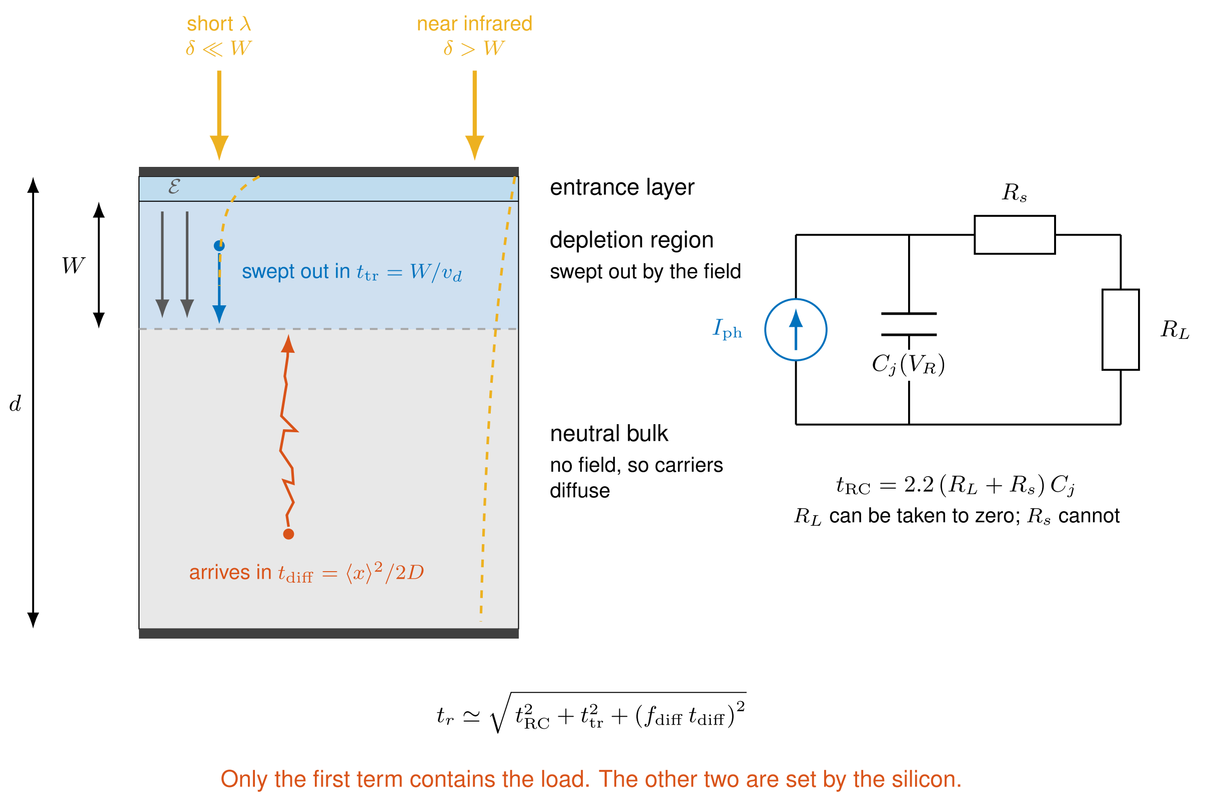

Figure 1. Where the three limits originate. Carriers generated inside the depletion region are swept out by the field and arrive within a transit time t_{tr}. Carriers generated behind the depletion edge see no field and return by diffusion, which is slow and depends on the square of the distance. The circuit sees only the sum of the two arrivals, and adds its own *RC* constant on top. The absorption depth δ decides how the incident light is divided between the first two populations, which is why the answer depends on wavelength. The load R_{L} appears in one term out of three.

What Eq. (1) is and is not. The three-mechanism decomposition is the standard one [1]. The quadrature combination is the cascaded rise-time rule [7] applied to it, not a solution of the transport problem, and the two internal terms are sometimes added directly instead [3]. Quadrature is the conservative choice of the two, and nothing in this note turns on it: the terms differ by more than an order of magnitude wherever the conclusion matters, so whichever term is largest governs under either rule.

3. The circuit term, and its floor

Note 03 gives the circuit contribution directly, and this note takes it unchanged:

$$t_{\mathrm{RC}} = 2.2\left(R_L + R_s\right)C_j\left(V_R\right) \tag{2}$$

with $C_{\mathrm{j}}(V_{\mathrm{R}})$ the bias-dependent junction capacitance of note 03, following the one-sided abrupt-junction law with an effective built-in potential of 0.87 V fitted there to two published endpoints, 1.0 nF at 0 V and 260 pF at 12 V for a 100 mm² device. Note 03 shows this term reproducing a published 30 ns rise-time specification into 50 Ω to within 5%, which is the regime it governs and the reason it is trusted.

The load R_{L} is the term an engineer controls, and it can be taken to zero. The series resistance of the undepleted bulk and the contacts cannot [3, 6]. Take the worked device of this note, defined in Section 6, at 12 V bias and 800 nm, with a series resistance of 1 Ω. Reducing the load from 50 Ω to zero takes the measured response from 29.2 ns to 2.21 ns, a factor of 13.2 where Eq. (2) alone promises an unbounded improvement. The intermediate steps show where the return goes: 10 Ω gives 6.64 ns, 5 Ω gives 4.04 ns, and 1 Ω gives 2.42 ns. The last of those steps halves t_{RC} and buys 8.8% of response time. Deleting the series resistance as well, which no design can do, would reach 2.13 ns.

4. Carrier transit across the depletion region

A carrier generated inside the depletion region is separated by the field and drifts to the edge. The time it takes is the width divided by the drift velocity [1, 3, 5]. The width is fixed by the same electrostatics that fix the capacitance, so it can be read straight off Eq. (1) of note 03:

$$W\left(V_R\right) = \frac{\varepsilon_0 \varepsilon_{\mathrm{Si}} A}{C_j\left(V_R\right)} \tag{3}$$

For the 100 mm² device this puts the depletion edge at 10.4 μm at zero bias, 39.8 μm at 12 V, and 112 μm at 100 V. This is the trade at the centre of the note. Every volt of reverse bias that lowers the capacitance does so by pushing that boundary further from the junction, and the carriers have correspondingly further to travel [1].

The drift velocity is not constant. It rises in proportion to the field at low field and saturates near 1×10^7 cm/s in silicon at high field [2, 4]. The interpolation between the two regimes is

$$v_d = \frac{\mu\mathcal{E}}{\left[1 + \left(\mu\mathcal{E} / v_s\right)^2\right]^{1/2}}, \qquad \mathcal{E} = \frac{V_{\mathrm{bi}} + V_R}{W} \tag{4}$$

$$t_{\mathrm{tr}} = \frac{W}{v_d} \tag{5}$$

Using the mean field across the depletion region and the electron mobility of silicon, the device at 12 V sits at 3.23 kV/cm, which returns a drift velocity of 4.00×10^6 cm/s, 40.0% of the saturation value. The transit time is therefore 0.997 ns rather than the 0.398 ns that W/v_{s} would give, a factor of 2.50. Assuming saturated velocity in a large partially depleted detector understates the transit term by that much; saturation is a condition of high-field devices, not a property of silicon.

The bias dependence of Eq. (5) is the part worth holding on to. In the mobility-limited regime W grows as the square root of bias and the mean field grows with it, so the two cancel and the transit time is nearly constant. As the velocity begins to saturate the cancellation fails and the transit time starts to grow. Across 0 to 100 V it rises by a factor of 1.57, from 0.920 to 1.44 ns, while the capacitance term falls by a factor of more than three over the same range. One limit is falling and the other is rising, which is what puts a floor under their sum.

5. The diffusion tail

Light that is not absorbed inside the depletion region is absorbed behind it, in field-free silicon. Those carriers are not lost: they wander until they reach the depletion edge, at which point the field collects them normally. What changes is when they arrive. Diffusion carries a carrier a distance x in a time [3]

$$t_{\mathrm{diff}} = \frac{x^2}{2D} \tag{6}$$

with D the minority-carrier diffusion coefficient, 12.4 cm^2/s for holes in silicon at 300 K [4]. The square is what makes this term behave unlike the other two. A carrier born 1 μm outside the field arrives in 0.40 ns; one born 100 μm outside arrives in 4.0 μs. Two orders of magnitude of distance cost four orders of magnitude of time, which is a scaling neither of the other two terms has. The result on an oscilloscope is not a delayed edge. It is a fast edge carrying the drift population followed by a long tail carrying the diffusing one, on both the rise and the fall [3].

Two quantities decide how much this matters, and both are geometric. The first is how far the diffusing carriers have to come. For an active layer of thickness d and light of absorption depth δ incident on the front, the mean depth of generation beyond the depletion edge is

$$\left\langle x \right\rangle = \delta - \frac{d - W}{e^{(d-W)/\delta} - 1} \tag{7}$$

which tends to δ when the remaining silicon is much thicker than the absorption depth, and to $(d - W)/2$ when the light passes almost straight through. The second is how much of the charge takes that route at all:

$$f_{\text{diff}} = \frac{e^{-W\,/\,\delta} - e^{-d\,/\,\delta}}{1 - e^{-d\,/\,\delta}} \tag{8}$$

Both depend on δ, and note 02 of this series tabulates δ for silicon across more than four orders of magnitude. Three of its rows carry this note: 0.1 µm at 400 nm, 11.8 µm at 800 nm, and 156 µm at 1000 nm, the last of these because silicon is an indirect-gap material and absorbs weakly near its band edge [6, 8]. The consequence for speed is immediate. At 12 V the depletion edge of the worked device sits at 39.8 µm, so at 600 nm only 8.0×10^{-8} of the charge arrives by diffusion and the tail does not exist. At 800 nm the fraction is 3.4%, at 900 nm it is 29.9%, and at 1000 nm it is 73.6%, at which point most of the signal is arriving the slow way.

The front side is not the problem. A symmetric term exists on the entrance side, where carriers generated in the front layer must diffuse inward to reach the field. It is negligible here for the reason note 02 makes central: the distances are a fraction of a micrometre. Eq. (6) with an entrance layer 0.3 µm thick and the electron diffusion coefficient gives 12.9 ps. The front-side structure governs ultraviolet *responsivity*, which is note 02's subject; it does not govern speed.

6. The budget against bias

The worked device is the 100 mm² silicon photodiode of note 03: 1.0 nF junction capacitance at zero bias falling to 260 pF at 12 V, on an active layer 300 µm thick, with a series resistance of 1 Ω and the external load taken to zero. It is read at 800 nm, where the absorption depth is 11.8 µm and comparable to the depletion width, so all three terms are live. Table 1 carries it through Eqs. (2) to (8) and Figure 2 draws the same numbers.

Table 1. Response-time budget of the worked 100 mm² device against reverse bias at 800 nm, with the external load at zero and R_S = 1 Ω. Capacitances at 0 and 12 V are published typical values; the other rows follow note 03's square-root law and are an extrapolation of it beyond 12 V. All times in ns.

Bias (V)	C_j (pF)	W (µm)	t_{RC}	t_{tr}	$f_{diff}t_{diff}$	t_r, Eq. (1)
0	1000	10.4	2.2	0.920	23.1	23.3
2	551	18.8	1.21	0.933	11.3	11.4
5	385	26.9	0.847	0.953	5.67	5.81
12	260	39.8	0.572	0.997	1.89	2.21
30	168	61.7	0.369	1.10	0.294	1.20
50	131	79.2	0.288	1.21	0.066	1.24
100	92.9	112	0.204	1.44	0.004	1.46

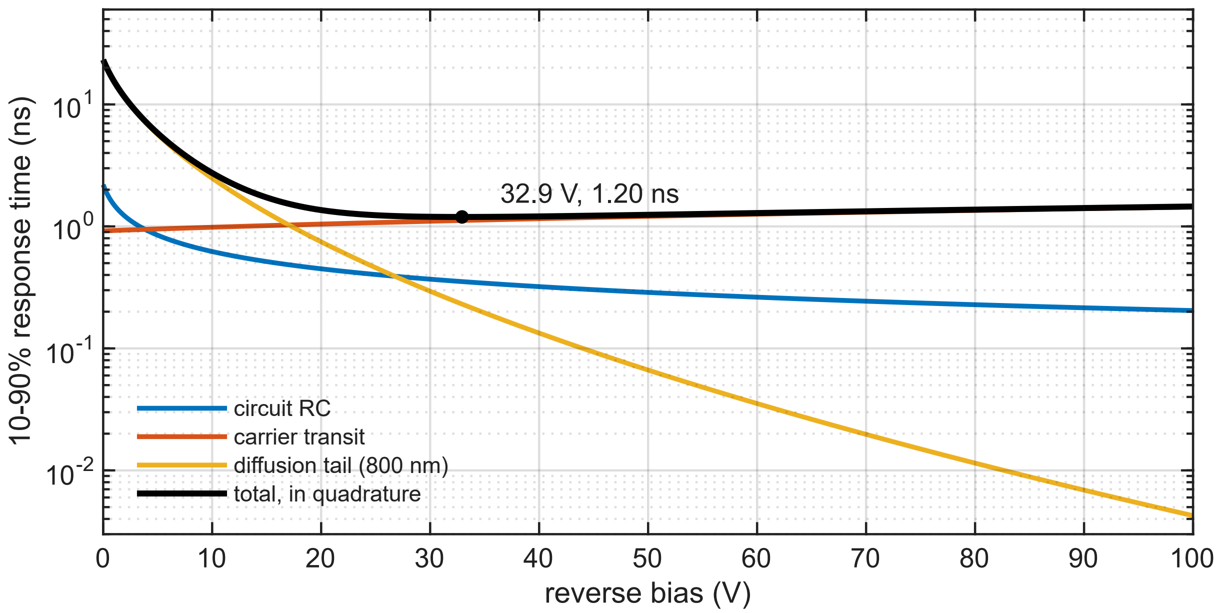

Figure 2. The three terms of Eq. (1) against reverse bias for the worked device at 800 nm, with the load already at zero. The circuit term falls with the capacitance and the transit term rises with the depletion width, so the total does not fall monotonically: it reaches 1.20 ns at 32.9 V and then turns. Below about 10 V the total belongs to the diffusion tail rather than to either of the other two, which is why reducing the load there changes almost nothing.

Three readings of the table are worth stating separately. First, at zero bias the circuit term is 2.2 ns and the response is 23.3 ns. An *RC* calculation is wrong by a factor of 10.6, and every part of the discrepancy is the diffusion tail. Expressed as bandwidth through $f_{3\mathrm{dB}} = 0.35/t_\mathrm{r}$ [7], the circuit promises 159 MHz and the detector delivers 15.0 MHz.

Second, the improvement available from bias is real and large, 19.5× from zero bias to the optimum, but it is not the capacitance ratio doing the work. Most of it is the depletion region growing out to meet the light and taking charge out of the diffusion population. Third, the optimum exists. Past 32.9 V the transit term governs alone, and by 100 V the response has degraded 22% from its best value. The bias that minimises capacitance is not the bias that minimises response time.

7. The budget against wavelength

Holding the bias at 12 V and sweeping the wavelength isolates the diffusion term, because the other two do not depend on it at all. The circuit term stays at 0.572 ns and the transit term at 0.997 ns, a combined 1.15 ns that Table 2 reproduces at every wavelength below 700 nm to three figures.

Table 2. Response time against wavelength at 12 V for the same device, where the depletion edge sits at 39.8 μm. Absorption depths are note 02's tabulated values interpolated log-linearly in α. Below 700 nm every entry is within rounding of the 1.15 ns floor set by the circuit and transit terms. Times in ns.

Wavelength (nm)	δ (μm)	δ/W	$f_{\mathrm{diff}}t_{\mathrm{diff}}$	t_r, Eq. (1)
700	5.36	0.134	0.0068	1.15
800	11.8	0.295	1.89	2.21
850	19.7	0.495	20.7	20.8
900	33	0.828	130	130
1000	156	3.92	2710	2710

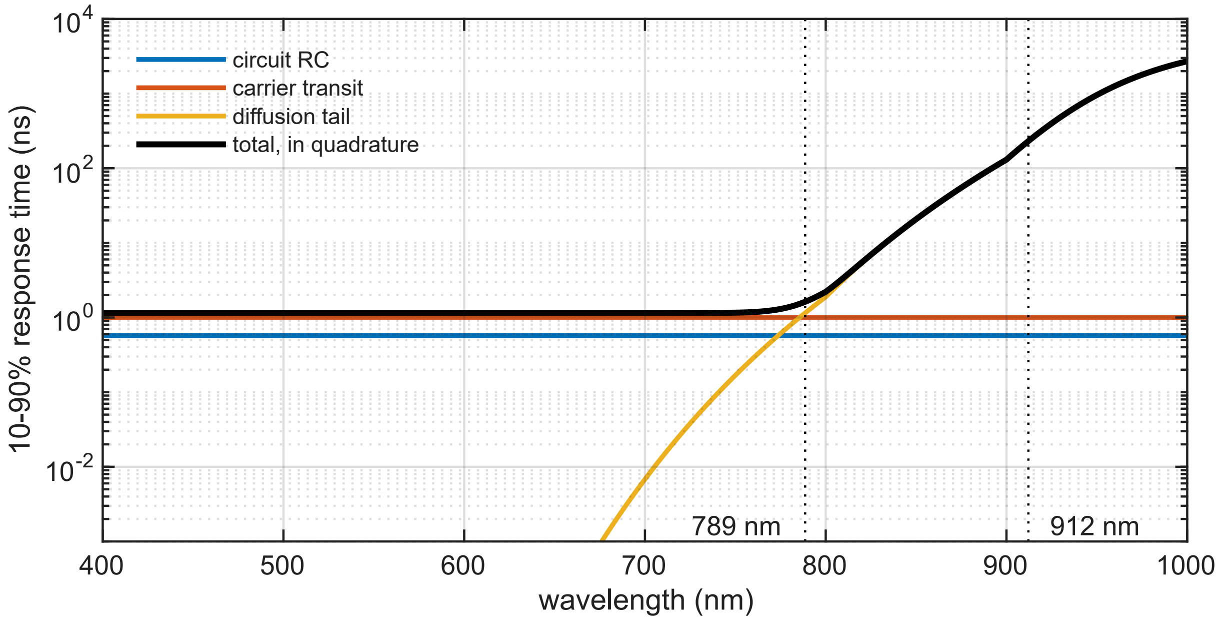

Figure 3. Response time against wavelength at 12 V. Two of the three terms are horizontal lines: neither the circuit nor the transit knows what colour the light is. The diffusion term climbs off the bottom of the plot through the red and near infrared and overtakes the other two at 789 nm, before the absorption depth has reached the depletion width at 912 nm. The same detector in the same circuit is 2360× slower at 1000 nm than at 600 nm.

The order of the two marked wavelengths is the useful detail. The intuitive threshold is where the absorption depth reaches the depletion width, 912 nm here, on the reasoning that the light starts landing outside the field at that point. The response-time crossover arrives earlier, at 789 nm, where the absorption depth is only 10.8 μm and a quarter of the depletion width. Because Eq. (6) is quadratic in distance, a small minority of the charge arriving late is enough to govern the transition: at 800 nm the 3.4% of carriers that diffuse take 55.8 ns to do it, and that product already exceeds the 1.15 ns the rest of the device is capable of.

At 1000 nm the same arithmetic runs to its conclusion. The mean generation depth beyond the depletion edge is 95.5 μm, those carriers need 3680 ns to return, and 73.6% of the signal is in that population. A detector specified at 30 ns and used at 1064 nm is not out of specification when it produces a microsecond tail. It is being asked a question its specification never answered.

Key result. Reducing the load resistance removes one term out of three. For the worked device at 800 nm it takes the response from 29.2 to 2.21 ns and no further, and at zero bias it leaves a response 10.6× slower than t_r = 2.2*RC* predicts. What remains is silicon: the transit of carriers across a depletion region that bias makes wider, and the diffusion of carriers from behind it that wavelength decides the size of.

8. What to do with it

The design rules that follow are not the ones the *RC* model alone suggests.

- Compute all three terms before selecting a load resistance. If the diffusion term already exceeds the circuit term, as it does for this device below about 10 V at 800 nm, a transimpedance front end buys almost nothing in speed over a modest resistor, and the reasons to choose one are the noise and linearity arguments of note 05 instead.
- Bias to the minimum of Eq. (1), not to the minimum of the capacitance. The capacitance keeps falling to whatever bias the part is rated for, and Eq. (1) does not: for this device the optimum is 32.9 V, and running it at 100 V instead costs 22%.
- Ask what wavelength the response time was specified at. A published rise time is a measurement at one wavelength, and Table 2 spans 2360× across the band of a single silicon detector.
- Where a near-infrared measurement must be fast, the lever is a fully depleted active layer rather than more bias on a partially depleted one. If the field reaches the back contact there is no field-free region left to diffuse across, and the diffusion term disappears rather than shrinking. Note 03's full-depletion knee is the same boundary seen from the capacitance side.
- Reducing the active area still works, and it works on the term that responds to nothing else. Area enters the capacitance directly, so it moves t_{RC} without touching either of the other two, which is the point note 03 makes and this note does not weaken.

Verdict. Note 03 gives the circuit time constant and the bias dependence of the capacitance that sets it, and both are correct. They stop being the answer once the load is small, once the bias is large, or once the wavelength is long. Past those points the response belongs to the silicon: to the transit across a depletion region that bias makes wider, and to the diffusion tail from the region behind it. Budget all three, take the minimum of the sum rather than the minimum of any one term, and state the wavelength alongside every rise time quoted or demanded.

Notes

Symbols. t_r 10 to 90% response time; t_{RC}, t_{tr}, t_{diff} its circuit, transit and diffusion terms; f_{diff} the fraction of the photocurrent arriving by diffusion; C_j junction capacitance; A active area; W depletion width; d active-layer thickness; R_L, R_s load and series resistance; V_R reverse bias; V_{bi} built-in potential; $\mathscr{E}$ mean electric field in the depletion region; μ low-field mobility; v_d, v_s drift and saturation velocity; D minority-carrier diffusion coefficient; $\delta = 1/\alpha$ absorption depth; $\langle x \rangle$ mean depth of generation beyond the depletion edge; ε_0 vacuum permittivity; $\varepsilon_{Si} \approx 11.7$ relative permittivity of silicon; f_{3dB} the bandwidth $0.35/t_r$.

Basis. The detector is the 100 mm² silicon photodiode of note 03: 1.0 nF published typical junction capacitance at zero bias and 260 pF at 12 V, following the one-sided abrupt-junction law with an effective V_{bi} = 0.87 V fitted there. Table 1's rows above 12 V extrapolate that law; the device remains partially depleted throughout, since W reaches only 112 μm of the 300 μm layer at 100 V, but a real product may reach its full-depletion knee first and its capacitance–voltage curve shows where. Everything else is a stated input rather than a specification of any product: an active-layer thickness of 300 μm, R_s = 1 Ω (note 03 quotes 0.5 to 5 Ω as the published range), ε_{Si} = 11.7, μ = 1350 cm²/V·s and D = 12.4 cm²/s at 300 K [4], and $v_s = 1 \times 10^7$ cm/s [2]. Absorption coefficients are note 02's tabulated values [8], with the 900 nm penetration depth of 33 μm quoted in [6] added as a sixth anchor and log-linear interpolation in α between anchors. Equation (4) uses the mean field $(V_{bi} + V_R)/W$ rather than the triangular peak field of an abrupt junction, which understates the velocity near the junction and overstates it near the depletion edge, and the exponent 2 of [2] which that reference assigns to electrons. The quadrature combination of Eq. (1) is the cascaded rise-time rule [7]; the underlying three-mechanism statement is [1]. Weighting the diffusion term by f_{diff} is a first-order treatment of a two-component waveform as one rise time and is not exact. Minority-carrier lifetime is assumed long compared with t_{diff}; where it is not, the tail is truncated and the missing charge appears as a responsivity deficit instead of a slow arrival [1, 9].

References

1. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ISBN 978-0-471-14323-9, p. 672 ("the response speed is limited by a combination of three factors: (1) drift time in the depletion region, (2) diffusion of carriers, and (3) capacitance of the depletion region", and the depletion-width compromise between transit time and $R_L C$), pp. 664–666 (the same trade stated as capacitance against transit time), p. 677 ($t_r = W_D/v_s$) and p. 679 ("the diffusion process is a slow one"; charge generated more than a diffusion length outside the depletion region is lost entirely).
2. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007, ISBN 978-0-471-14323-9, p. 36, Eq. (81): the empirical drift-velocity law $v_d = \mu_0\mathscr{E}/[1 + (\mu_0\mathscr{E}/v_s)^{C_2}]^{1/C_2}$ with $C_2 \approx 2$ for electrons; p. 37 ($v_s \approx 1 \times 10^7$ cm/s in silicon); pp. 83 and 85 (one-sided abrupt-junction depletion width and the depletion capacitance ε_s/W_D per unit area).
3. S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, 2nd ed., Pearson, 2013, ISBN 978-0-273-77417-4, p. 397, Eq. (5.5.3) ($t_{drift} = W/v_d$, and that widening W raises quantum efficiency while slowing the response); pp. 399–400, Example 5.5.2 ($t_{diff} = \ell^2/2D$, worked for a 1 μm diffusion distance); p. 423 (the series resistance of the neutral regions); and p. 424 (the fast drift edge followed by the slow diffusion region on both the rise and the fall, and the 10 to 90% definition of rise time).

4. D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, 3rd ed., McGraw-Hill, 2003, ISBN 0-07-232107-5, p. 168 ("the drift velocity of electrons in silicon … saturates at approximately 10^7 cm/s"), p. 176 (the Einstein relation $D/\mu = kT/q$), and p. 177, Table 5.2 (silicon at 300 K: μ_n = 1350 cm²/V·s, D_n = 35 cm²/s, μ_p = 480 cm²/V·s, D_p = 12.4 cm²/s).

5. A. Rogalski, *Infrared and Terahertz Detectors*, 3rd ed., CRC Press, 2019, ISBN 978-1-138-19800-5, p. 262, Eq. (12.135): $\tau_\mathrm{t} = w_\mathrm{dep}/v_\mathrm{s}$, with v_s "about 10^7 cm/s"; p. 263 (the depletion width of a p–i–n as a trade between response speed and quantum efficiency).

6. S. M. Sze and M. K. Lee, *Semiconductor Devices: Physics and Technology*, 3rd ed., Wiley, 2012, ISBN 978-0470-53794-7, p. 285 (penetration depth $\delta = 1/\alpha$, and the slow rise of α with decreasing wavelength in an indirect-gap material), p. 326 (the neutral regions as resistive extensions of the electrodes), and p. 328 ("the penetration depth is about 33 µm … at the wavelength 900 nm … most incident photons are absorbed outside the depletion region where there is no field to separate the EHPs").

7. N. Massa, "Fiber optic telecommunication," in *Fundamentals of Photonics*, Module 1.8, University of Connecticut / CORD, 2000, p. 310 (the 10 to 90% definition of rise time), p. 332, Eqs. (8-22) and (8-23) (BW = $0.35/t_\mathrm{r}$, and the total rise time of a chain as the square root of the sum of the squares of its parts), and p. 333 (the photodiode as one of the terms in that sum).

8. M. A. Green, "Self-consistent optical parameters of intrinsic silicon at 300 K including temperature coefficients," *Sol. Energy Mater. Sol. Cells* **92**, 1305–1310 (2008), doi:10.1016/j.solmat.2008.06.009; the absorption coefficients tabulated in note 02 of this series.

9. D. Birtalan and W. Nunley, eds., *Optoelectronics: Infrared-Visible-Ultraviolet Devices and Applications*, 2nd ed., CRC Press, 2009, ISBN 978-1-4200-6780-4, p. 61 (a pair generated more than a diffusion length from the junction recombines with no contribution to photocurrent) and p. 66 (a wider depletion layer absorbs more photons but lengthens the transit time and reduces bandwidth).