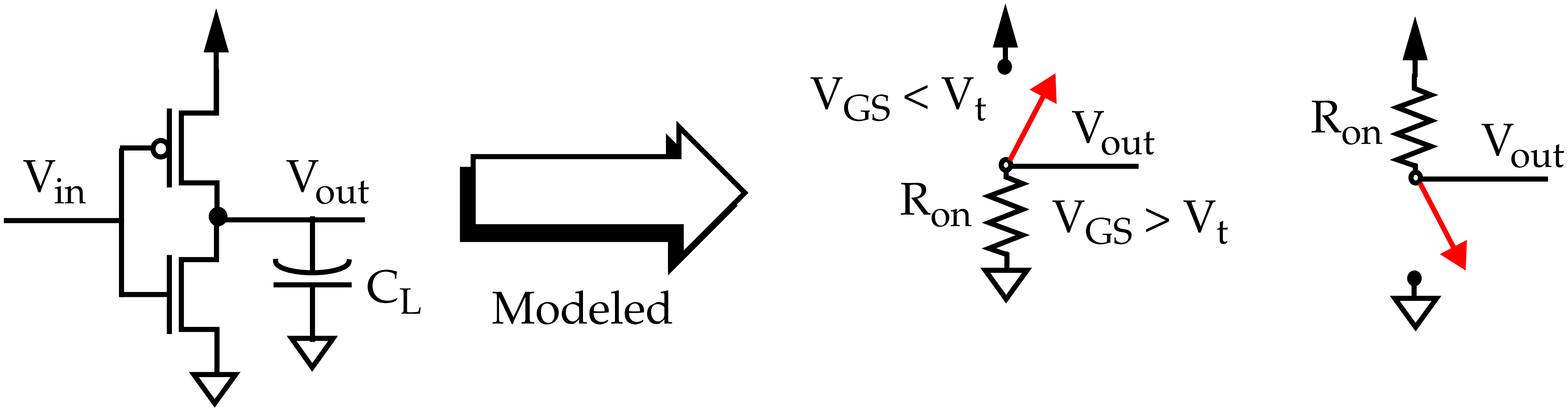
The Inverter

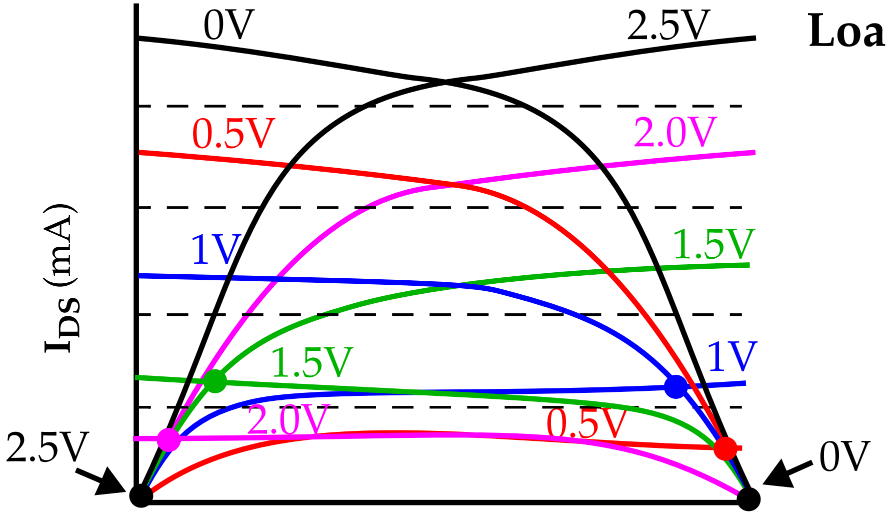
The electrical behavior of complex circuits (adders, multipliers) can be almost completely derived by extrapolating the results obtained for inverters!



Observations:

- **Fully restored** (V_{DD} and GND) output levels results in high noise margins.
- **Ratioless**: Logic levels are not dependent on the relative device sizes.
- **Low output impedance** in steady state (kΩ connection to either V_{DD} or GND), increases robustness to noise.
- **High input impedance**: fanout is theoretically unlimited for static operation, transient response is impacted however.
- **Low static power dissipation**: No path between power and ground.

The Inverter DC current characteristics

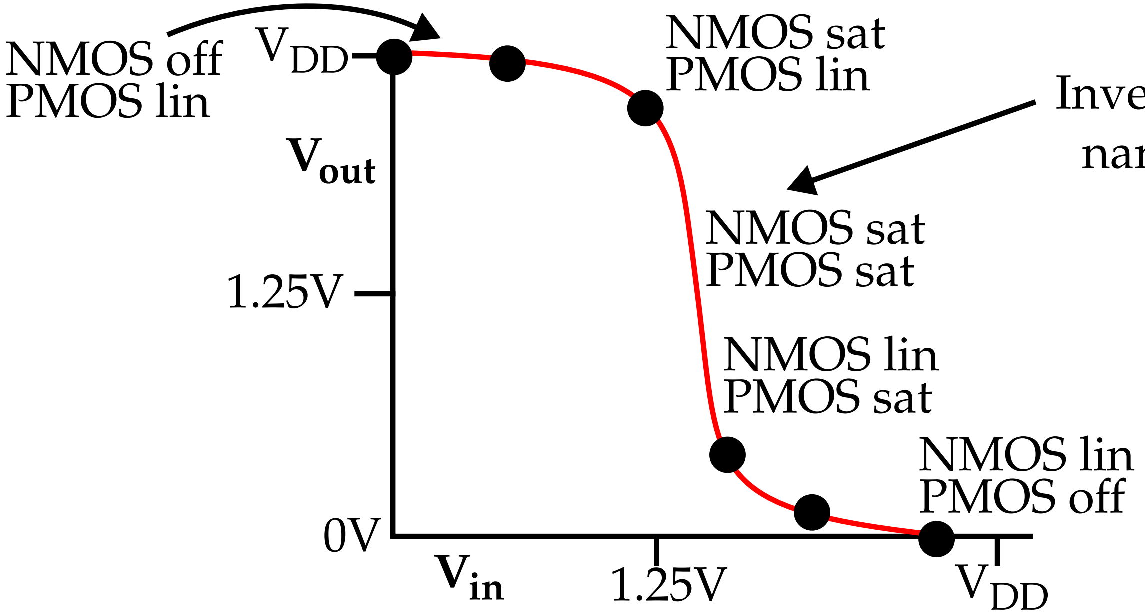


Load-line plot

Here, PMOS curves have been mirrored around x and shifted.

The current of the NMOS and PMOS device MUST be equal.

All points are located at either the high or low output levels.

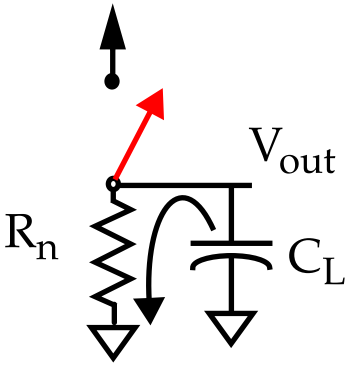


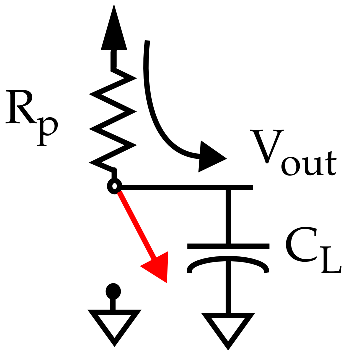
Inverter exhibits a very narrow transition zone.

Inverter Models

It is possible to approximate the transient response to an RC model.

The response is dominated by the output capacitance of the gate, C_L.





Load capacitance, C_L, is due to *diffusion*, *routing* and *downstream* gates.

The propagation delay assuming an instantaneous input transition is R_pC_L.

This indicates a fast gate is built by keeping either or both of R_p and C_L small.

R_p is reduced by increasing the W/L ratio.

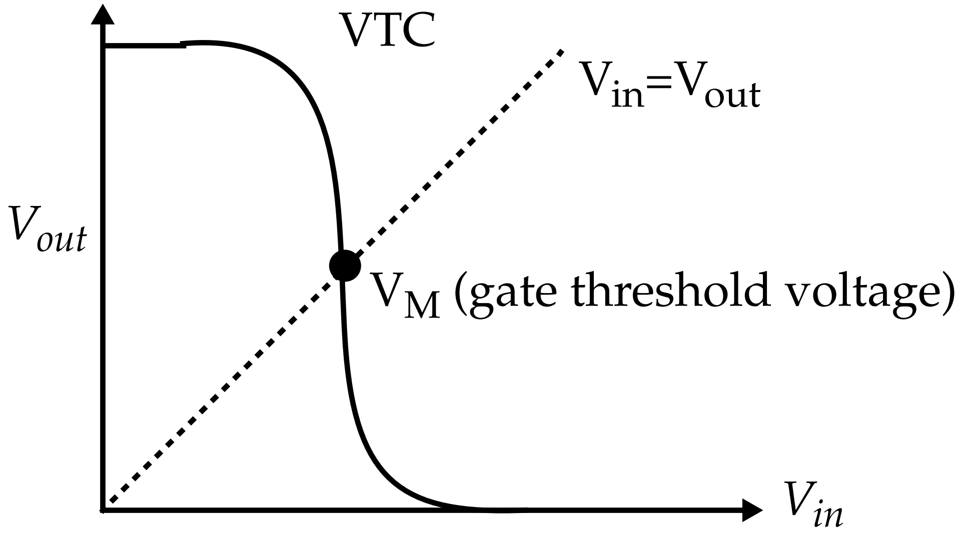
Bear in mind that, in reality, $R_{n/p}$ is a nonlinear function of the voltage across the transistor.

Switching Threshold

Previously, we defined V_M as the **inverter threshold voltage** but did not derive an analytical expression for it.

The same is true for V_{IH} and V_{IL}, and consequently the noise margins (see text for this analysis).

V_M is defined as the intersection of the line $V_{in} = V_{out}$ and the inverter VTC.



In this region, both the NMOS and PMOS transistors are in saturation since $V_{DS} = V_{GS}$.

Switching Threshold

Therefore, the value of V_M can be obtained by equating the NMOS and PMOS currents (assuming devices are velocity saturated).

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right)^2$$

Solving for V_M:

$$V_M = \frac{r\left(V_{Tn} - \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1 + r}$$

with $r = \dfrac{k_p V_{DSATp}}{k_n V_{DSATn}} = \dfrac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n}$

Further simplified:

$$V_M \cong \frac{r V_{DD}}{1 + r}$$

V_M is set by the ratio *r*, and *r* compares the relative driving strengths of the PMOS and NMOS transistors.

It is desirable to have *r* = 1, i.e., V_M situated in the **middle** of the available voltage swing ($V_{DD}/2$) to provide comparable low and high noise margins.

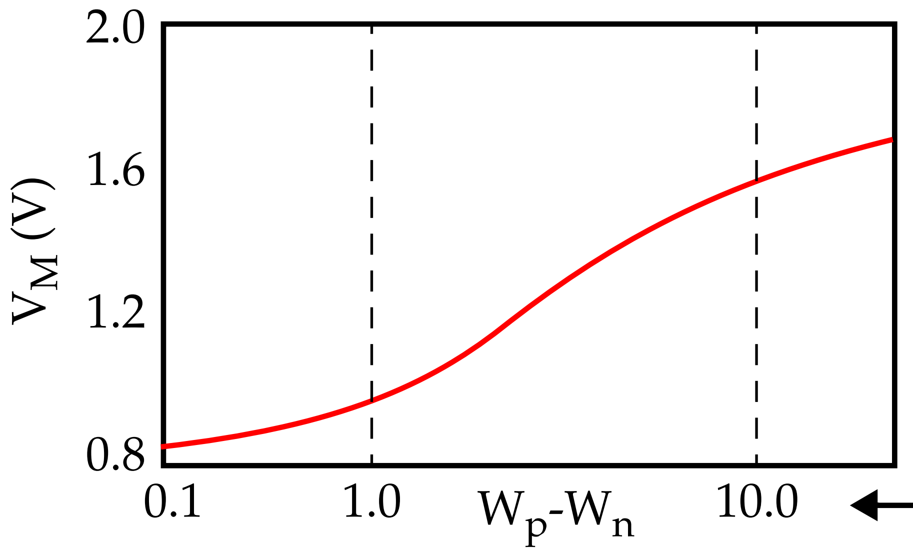
Switching Threshold

The required ratio can be determined for any value of V_M using:

$$\frac{(W/L)_p}{(W/L)_n} = \frac{V_{DSATn} k'_n \left(V_M - V_{Tn} - \frac{V_{DSATn}}{2} \right)}{V_{DSATp} k'_p \left(V_{DD} - V_M + V_{Tp} + \frac{V_{DSATp}}{2} \right)}$$

Using a generic 0.25 µm CMOS process, this means making the PMOS **3.5** times wider than the NMOS.

V_M plotted as a function of the PMOS-to-NMOS ratio.



x scale gives the log of the ratio of PMOS and NMOS widths as the difference.

Switching Threshold

Observations from plot:

- V_M is relatively **insensitive** to variations in device ratio.

  Small variations in the ratio (3.0 -> 2.5) do not disturb the VTC much.

  Industry sets the ratio of PMOS width to NMOS width to values smaller than that needed for an exact symmetry.

  For example, setting the ratio to 3, 2.5 and 2 yields switching thresholds of 1.22 V, 1.18 V and 1.13 V, respectively.

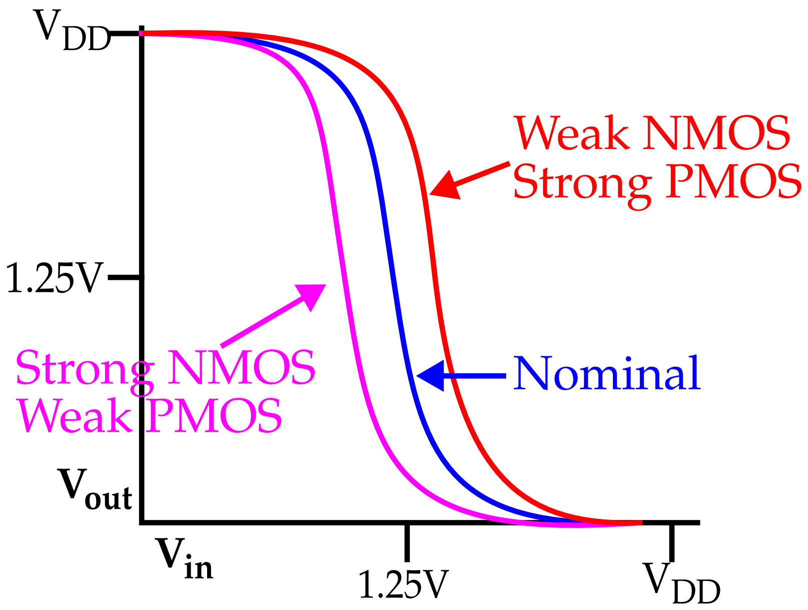
- Increasing the width of the PMOS or the NMOS moves V_M toward V_{DD} or GND, respectively.

  This feature may be desirable in some applications, e.g., when the input signal is noisy (see text).

  Bear in mind that when the ratio of V_{DD} to V_T is relatively *small*, e.g. 2.5/ 0.4 = 6), moving V_M *a lot* is difficult and requires *very large* differences in the width ratios.

Inverter Threshold

Robustness under process variations:



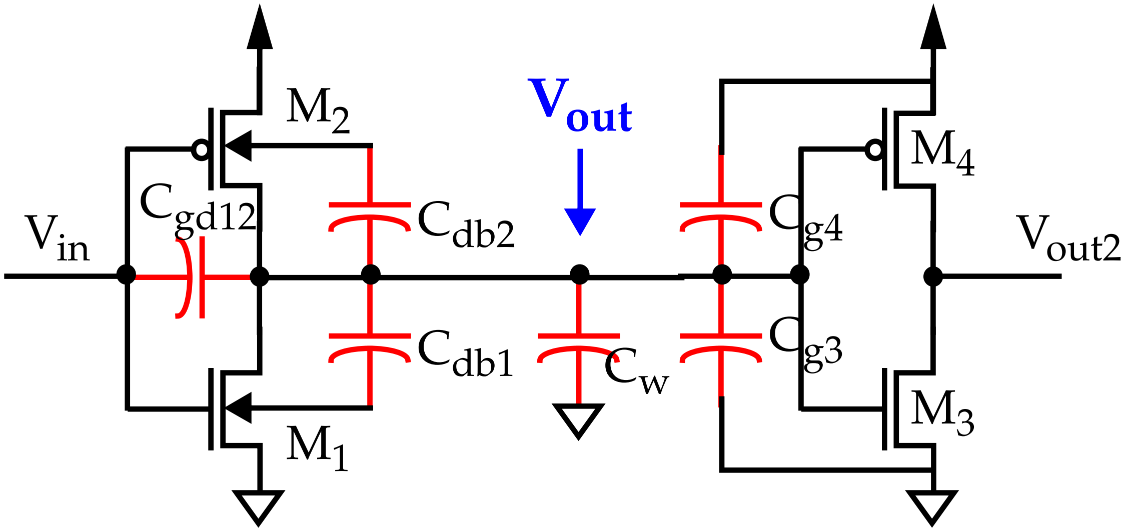
Process variations will cause only small shifts in the transfer curve.

The functionality of the gate is **not** effected however, and this feature has contributed in a big way to the popularity of the static CMOS gate.

Dynamic Behavior

Propagation delay is determined by the time it takes to charge/discharge the load cap, C_L, so it's worth looking closely at C_L before developing a delay model.

Simple propagation delay models **lumps** all capacitances into C_L.



All cap influencing transient response of node V_{out}.

In this analysis, assume V_{in} is driven by an ideal voltage source with fixed rise/fall times.

Dynamic Behavior

Gate-drain capacitance:

- C_{gd12}: Capacitance between the gate and drain of the first inverter.

  M_1 and M_2 are either in **cut-off** or in **saturation** during the first half (up to 50% point) of the output transient.

  It is reasonable to assume that only M1 & M2 *overlap capacitances* contribute.

  Remember, gate cap is either completely between gate/bulk (cut-off) or gate/src (sat).

  In the lumped model, we need to replace the C_{gd12} with a capacitor to GND.

  The value of this capacitor is given as $C_{gd} = 2{*}C_{GD0}{*}W$ where C_{GD0} is overlap capacitance per unit width.

  Note it is doubled due to the **Miller effect**.

Dynamic Behavior

Diffusion capacitances:

- C_{db1} and C_{db2}: Capacitances due to the reversed biased *pn*-junction.

  These caps are quite nonlinear (voltage dependent).

  We linearized these caps over the voltage range of interest:

$$C_{eq} = K_{eq} C_{j0}$$

with C_{j0} the junction cap/unit area under zero bias conditions.

The bottom plate and sidewall zero bias values can be obtained from the SPICE model CJ and CJSW parameters.

K_{eq} was derived in an earlier lecture.

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)}[(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}]$$

Example

Consider a 0.25 µm 2.5 V technology and the previous inverter chain.

Assume ϕ_0 is 0.9 V for both NMOS and PMOS and $m = 0.5$.

Let's compute C_{db1} for the NMOS transistor.

Propagation delay is computed between the 50% points.

This is the time-instance when V_{out} reaches 1.25 V.

For the high-to-low (**H-to-L**) transition, we linearize over {2.5 V, 1.25 V} and for the low-to-high transition over {0, 1.25 V}.

H-to-L: V_{out} is initially 2.5 V: V_{high} = -2.5V. At 50%, V_{low} = -1.25V. K_{eq}= **0.57**.

L-to-H: V_{out} is initially 0 V: V_{low} = 0. At 50%, V_{high} = -1.25 V. K_{eq} = **0.79**.

Sidewall capacitance can be computed in a similar way (see text).

Also, similar, but reversed, values are obtained for PMOS device.

This linearized simplification has only minor effects on logic delays.

Dynamic Behavior

Wire capacitance:

- C_w: The capacitance is dependent on the length and width of the interconnecting wires and is growing in importance.

Gate capacitance of fan-out:

- C_{g3} and C_{g4}: Includes both *overlap* and *gate* capacitance of each transistor:

$$\begin{aligned} C_{fan-out} &= C_{gate}(NMOS) + C_{gate}(PMOS) \\ &= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox}) \end{aligned}$$

But what about the *Miller effect*?

We can safely ignore it here by assuming the driven gate's output does **not** change until **after** the 50% point of the input is reached.

We also assume, with about a 10% over-estimation error, that the *channel cap* of the driven gate remains constant over this interval.

Dynamic Behavior

Text gives the capacitance calculated from the **layout** of a two-inverter chain.

Results are given as follows:

- Overlap capacitance:

  NMOS: 0.31 fF/μm

  PMOS: 0.27 fF/μm

- Bottom junction capacitance:

  NMOS: 2.0 fF/μm^2

  PMOS: 1.9 fF/μm^2

- Sidewall junction capacitance:

  NMOS: 0.28 fF/μm

  PMOS: 0.22 fF/μm

- Gate capacitance:

  NMOS = PMOS: 6.0 fF/μm^2

- Wire capacitance:

  C_{wire}: 0.12 fF

Total load for **H-to-L**: **6.1 fF**, for **L-to-H**: **6.0 fF**

In text, this cap is almost evenly split between *intrinsic* and *extrinsic* srcs.

Propagation Delay: First-Order Analysis

One way to compute delay is to integrate the capacitor (dis)charge current:

$$t_p = \int_{v_1}^{v_2} \frac{C_L(v)}{i(v)} \mathrm{dv}$$

But both C_L(v) and i(v) are **nonlinear** functions of v.

Instead, we can use a simple switch model given earlier to derive an approximation.

Here, both the "on" resistance and load capacitance are replaced by a constant elements, assigned average values over the region of interest.

Although we didn't cover it in class, the average "on" resistance is given by:

$$R_{eq} = \frac{1}{V_{DD}/2} \int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)} \mathrm{dV} \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{7}{9}\lambda V_{DD}\right)$$

with $I_{DSAT} = k'\frac{W}{L}\left((V_{DD} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)$ (see text)

Propagation Delay: First-Order Analysis

The linearized load capacitance is derived as shown previously.

Propagation delay is then computed using a first-order linear RC network model:

$$t_{pHL} = \ln(2) R_{eqn} C_L = 0.69 R_{eqn} C_L$$
$$t_{pLH} = \ln(2) R_{eqp} C_L = 0.69 R_{eqp} C_L$$

Assuming that the equivalent load cap is approximately the same for either transition.

The propagation delay is the average of the two:

$$t_p = \frac{t_{pHL} + t_{pLH}}{2} = 0.69 C_L \left(\frac{R_{eqn} + R_{eqp}}{2} \right)$$

This indicates to make rise and fall times identical, it is necessary to make the "on" resistance of the NMOS and PMOS equal.

See text for a good example.

Propagation Delay: First-Order Analysis

Minimizing propagation delay amounts to:

- Reducing C_L.

  Which is composed of self-loading (diffusion) (*intrinsic*), routing and fan-out (*extrinsic*) capacitance.

  Careful layout can reduce diffusion and interconnect caps.

- Increase W/L ratio of the transistors.

  Warning: doing so **increases** the self-loading and therefore C_L!

  Once intrinsic (self-loading) cap starts to dominate the extrinsic load cap (wires + fan-out), increasing the width doesn't help delay.

- Increase V_{DD}.

  The delay of a gate can be modulated by modifying the supply voltage.

  This allows the designer to trade off energy dissipation for performance.

  However, rising above a certain level yields on a minor improvement.

  Also, reliability concerns (oxide breakdown, hot-electron effects) set firm upper bounds.