Dark (and Bright) Secrets of RF Design

Prof. Tom Lee

Stanford University

IBIC, 17 Sept. 2014

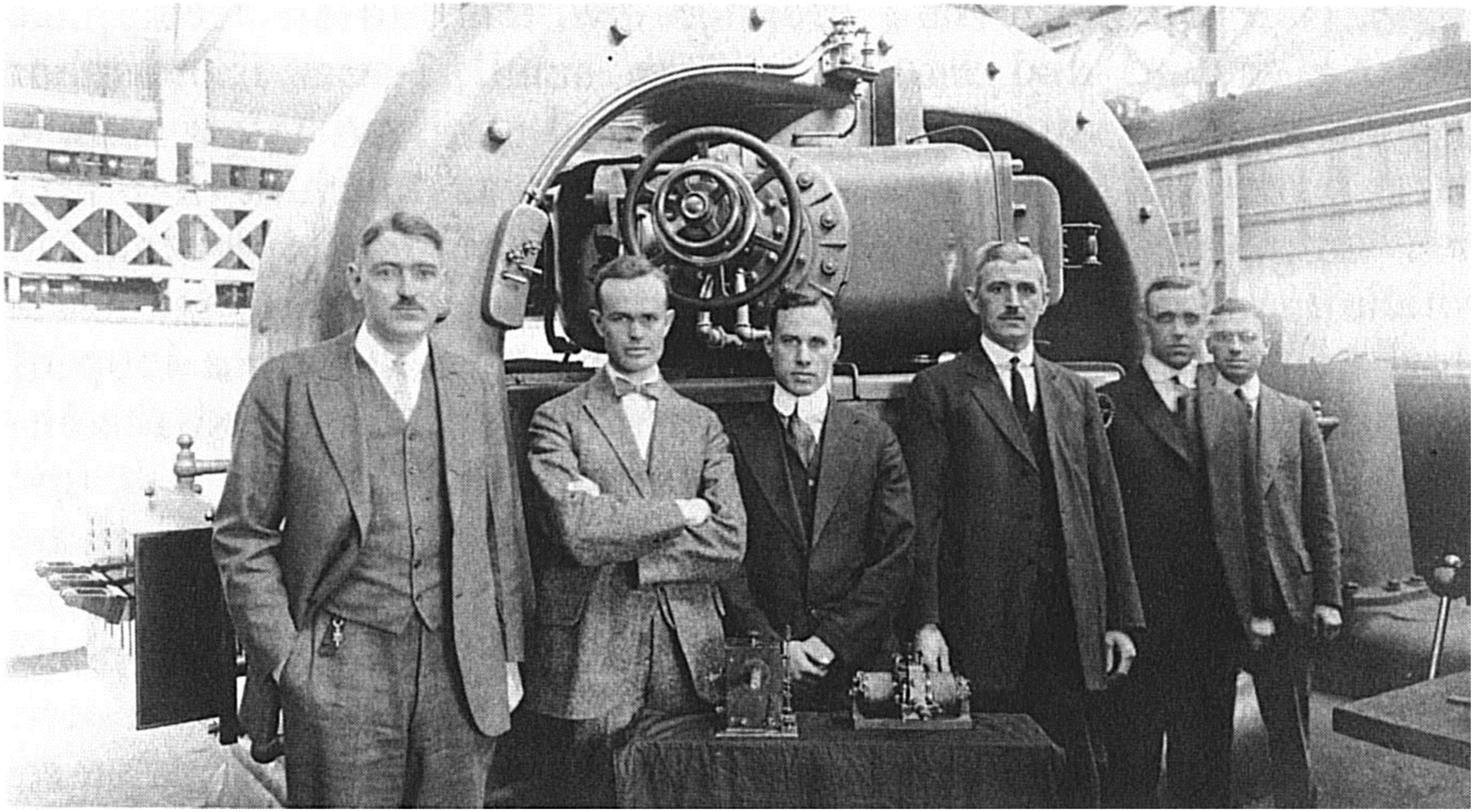
International Beam Instrumentation Conference

Monterey, California, USA September 14–18, 2014

Why RF design is hard

- Can’t ignore parasitics:
  - 100fF is 320Ω@5GHz; 1.6Ω@1THz
- Can’t squander device power gain.
- Can’t tolerate much noise or nonlinearity.
  - 1μV amplitude = 10fW@50Ω
- Can’t expect accurate models, but you still have to make it work.

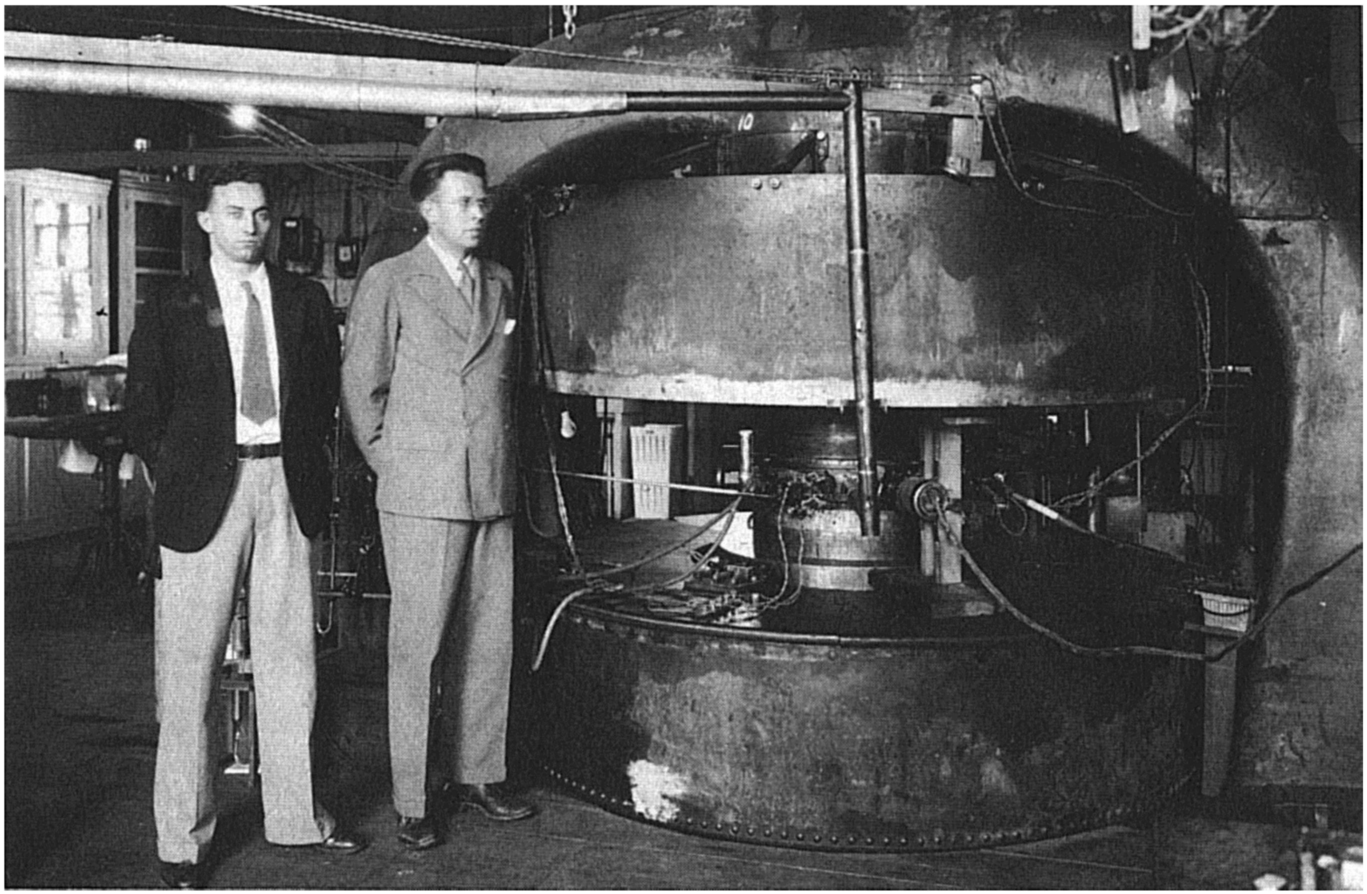
RF and the accelerator story: 1917



Leonard Fuller (left) standing with Federal coworkers.
On table: Original Poulsen model on left, first Federal arc on right.)

From early wireless to the cyclotron

- Ernest O. Lawrence (right) poses with M. Stanley Livingston, 1933. Note Federal’s magnet.



Traditional RF design flow

H7N9

- Put on wizard hat.

- Obtain chicken (don't ask).
- Design first-pass circuit.
- Utter magical Latin incantations ("*semper ubi sub ubi*...*omnia pizza in octo partes divisa est...e pluribus nihil*").
- Test circuit. Weep.
- Adjust chicken. Iterate.

Dark secrets: A partial list

- FETs: Textbooks lie
- The two-port noise model: Why care?
  - Optimum noise figure vs. maximum gain
- Impedance matching: When, and why?
- Linearity and time invariance: Huh?
- Mixers: Deceptions, myths and confusion
- Gain-bandwidth limits aren't
- Strange impedance behaviors (SIBs)

FETs: What Your Textbooks May Not Have Told You

The standard story

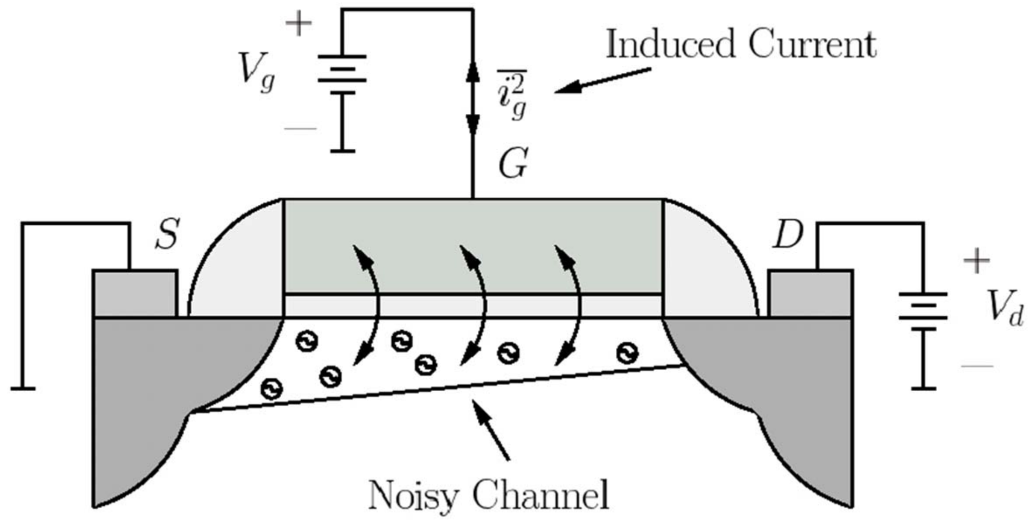
- "Gate-source impedance is that of a capacitor."
- A capacitor is lossless. A capacitive gate-source impedance therefore implies *infinite* power gain at all frequencies.

Dark secret: Gate impedance has a real part

- Gate-source impedance *cannot be* purely capacitive.
  - True even if gate material is superconductive.
- Phase shift associated with finite carrier transit speed means gate field does nonzero work on channel charge.
- Therefore, power gain is not infinite.
- *There is also noise associated with the dissipation.*

Noisy channel charge

- Fluctuations couple capacitively to both top and bottom gates.
  - Induces noisy gate currents.
  - Bottom-gate term is ignored by most models and textbooks.



[Shaeffer]

- Gate noise current. $\overline{i_g^2} = 4kTB\delta\frac{(\omega C_{gs})^2}{5g_{d0}}$
- Real component of Y_g. $\mathrm{Re}\left[Y_g\right] = \frac{(\omega C_{gs})^2}{5g_{d0}}$

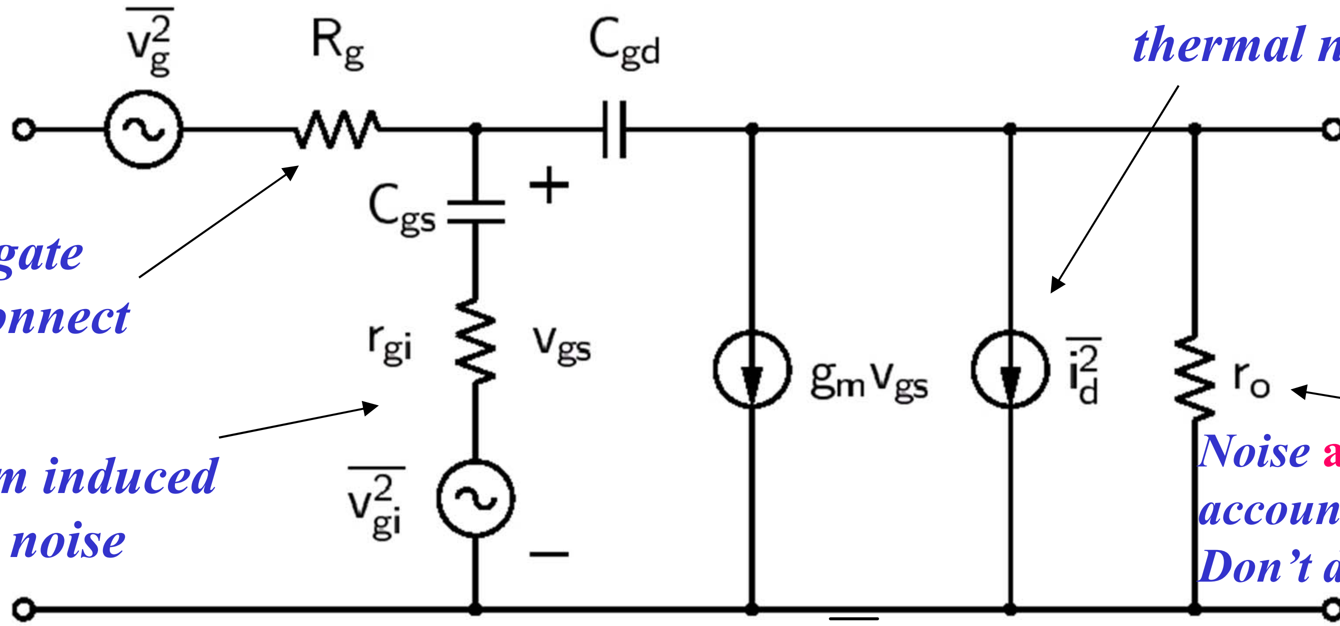
Sources of noise in MOSFETs

- (Thermally-agitated) channel charge.
  - Produces current noise in both drain *and* gate.
- Interconnect resistance.
  - Series gate resistance R_g is very important.
- Substrate resistance.
  - Substrate thermal noise modulates back gate, augments drain current noise in some frequency range.

(All) FETs and gate noise: Basic model

From channel thermal noise

From gate interconnect



From induced gate noise

Noise ***already*** *accounted for – Don't double count!*

$$\overline{i_d^2} = 4kT\gamma g_{d0}$$

$$\overline{v_{gi}^2} = 4kTB\delta r_{gi}$$

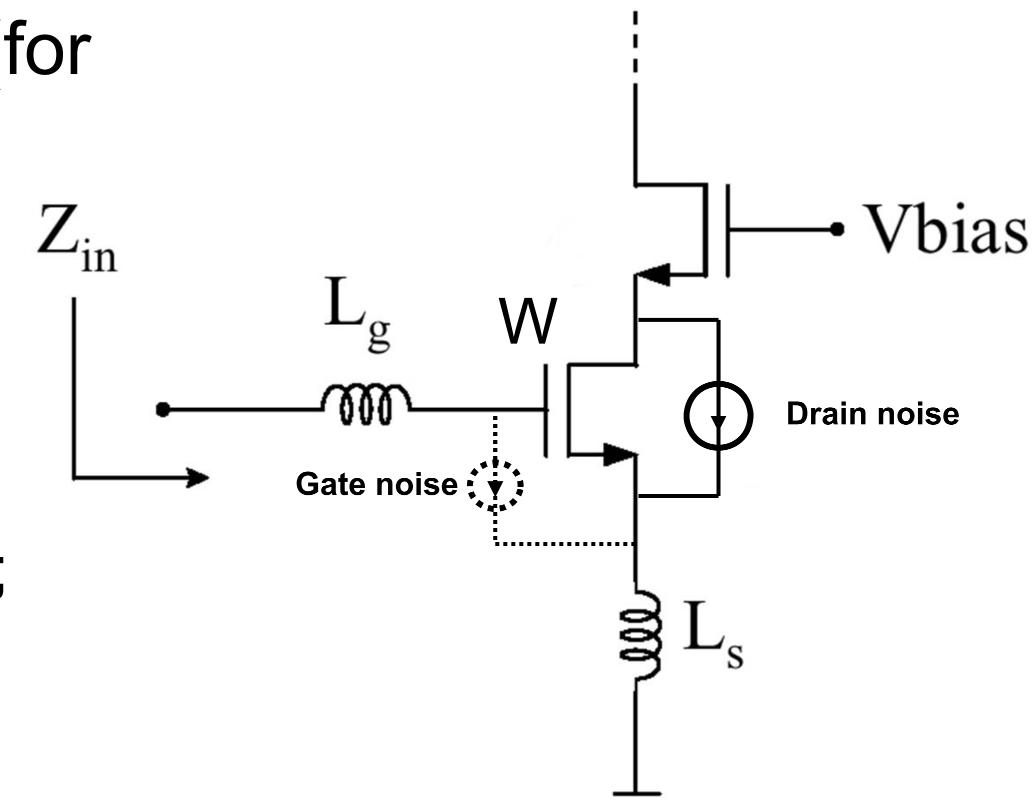
$$r_{gi} = \frac{1}{5g_{d0}}$$

(Note the placement of v_{gs}.)

Important: Common error is to define Vgs as across Cgs alone.

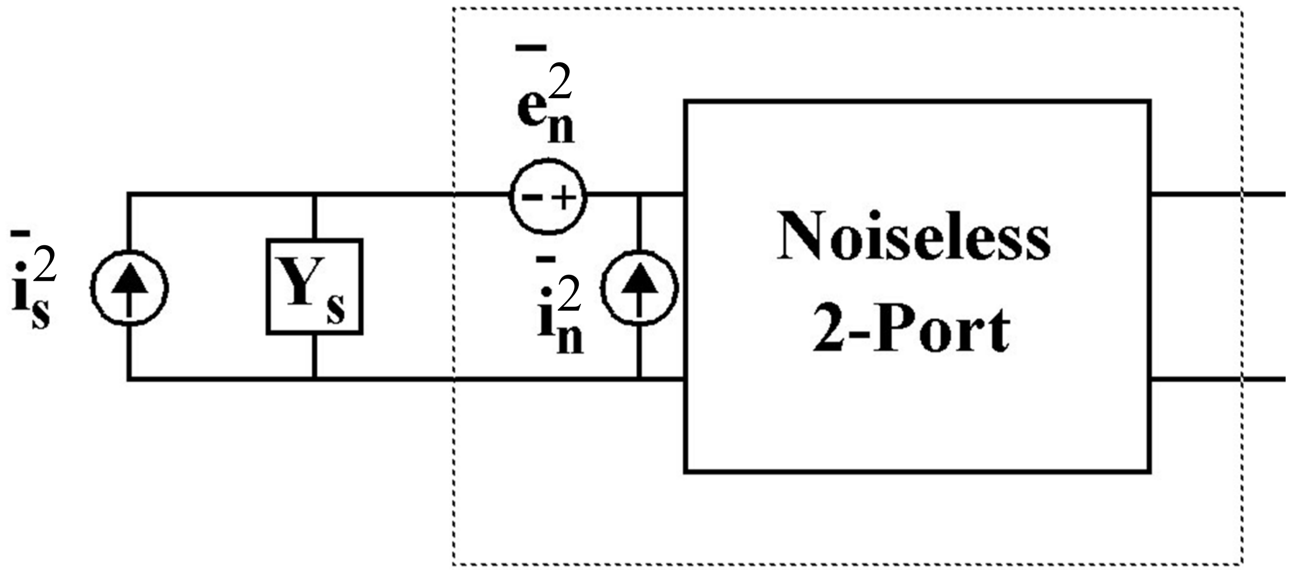
Gedanken experiment: Gate noise is real

- Let W → 0 while maintaining resonance and current density (for fixed f_T).
  - Gain stays fixed.
  - I_D → 0.
- If you ignore gate noise:
  - Output noise → zero; absurd prediction of zero power and noiseless gain.



The Two-Port Noise Model

Two-port noise model



$$F = \frac{\overline{i_s^2} + \overline{|i_n + Y_s e_n|^2}}{\overline{i_s^2}} = 1 + \frac{\overline{i_u^2} + |Y_c + Y_s|^2 \overline{e_n^2}}{\overline{i_s^2}}$$

- The *IRE* chose not to define *F* directly in terms of equivalent input noise sources. Instead:

Two-port noise model

Let $$R_n \equiv \frac{\overline{e_n^2}}{4kT\Delta f},$$

$$G_u \equiv \frac{\overline{i_u^2}}{4kT\Delta f},$$

and

$$G_s \equiv \frac{\overline{i_s^2}}{4kT\Delta f}.$$

Conditions for minimum noise figure

$$B_s = -B_c = B_{opt}$$

$$G_s = \sqrt{\frac{G_u}{R_n} + G_c^2} = G_{opt}$$

$$F_{min} = 1 + 2R_n[G_{opt} + G_c] = 1 + 2R_n\left[\sqrt{\frac{G_u}{R_n} + G_c^2} + G_c\right].$$

$$F = F_{min} + \frac{R_n}{G_s}\left[(G_s - G_{opt})^2 + (B_s - B_{opt})^2\right]$$

No secret: Murphy wants to hurt you

- Minimum NF and maximum power gain occur for the same source Z *only if three miracles occur simultaneously*:
  - G_c = 0 (noise current has no component in phase with noise voltage); *and*
  - $G_u = G_n$ (conductance representing uncorrelated current noise equals the fictitious conductance that produces noise voltage); *and*
  - $B_c = B_{in}$ (correlation susceptance happens to equal the actual input susceptance)

To match, or not to match

Why match?

- Conjugate match maximizes power transfer.

- Terminating a *T*-line in its characteristic impedance makes the input impedance length-independent.
  - Also minimizes peak voltage and current along line.

- Choosing a standard impedance value (e.g., 50Ω) facilitates assembly, fixturing and instrumentation; it addresses *macroscopic* concerns.

Why *not* match?

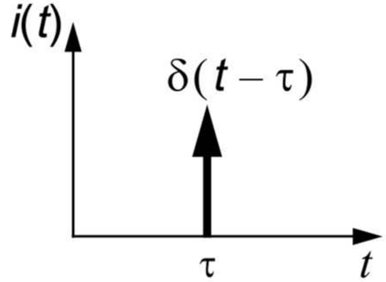
- Amps generally exhibit best NF when mismatched.
- Some amps are more stable when mismatched.
- Many matching networks may provoke instability at frequencies away from design center.
  - Parametric pumping often causes problems in PAs.
- If power gain is abundant, can afford mismatch, resulting in a simpler or smaller circuit.
  - A 741 op-amp uses no impedance matching of any kind.

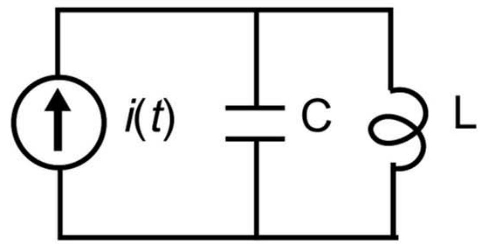
Dark secret: Circuits are neither linear nor time-invariant

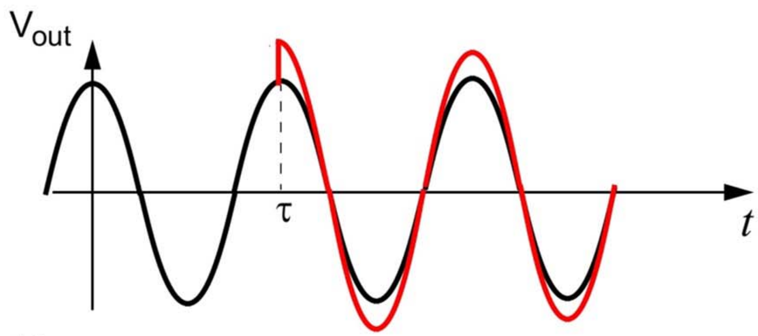
LTI, LTV and all that

- A system is linear if superposition holds.
- A system is TI if an input timing shift only causes an equal output timing shift.
  - Shapes stay constant.
- If a system is LTI, it can only scale and phase-shift Fourier components.
  - Output and input frequencies are the same.
- If a system is LTV, input and output frequencies can be different, ***despite being linear***.
- If a system is nonlinear, input and output frequencies will generally differ.

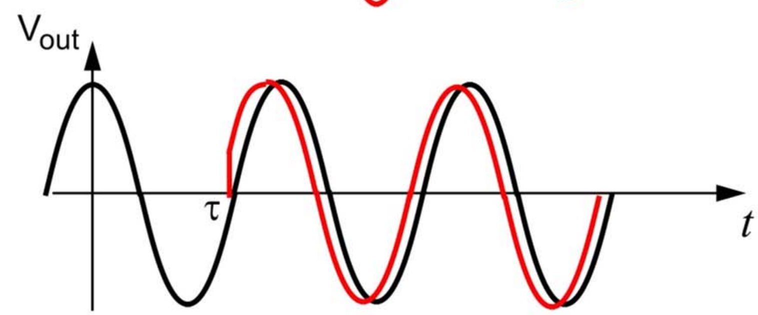
Are oscillators LTI?







Impulse injected at peak

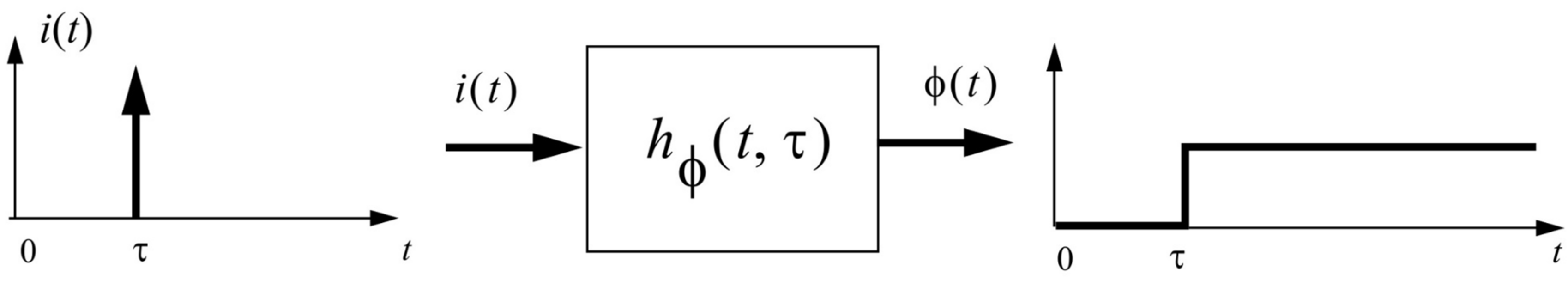


Impulse injected at zero crossing

Even for an ideal LC, the phase response is time-variant.

Phase impulse response

The phase impulse response of an oscillator is a step:



The unit impulse response is:

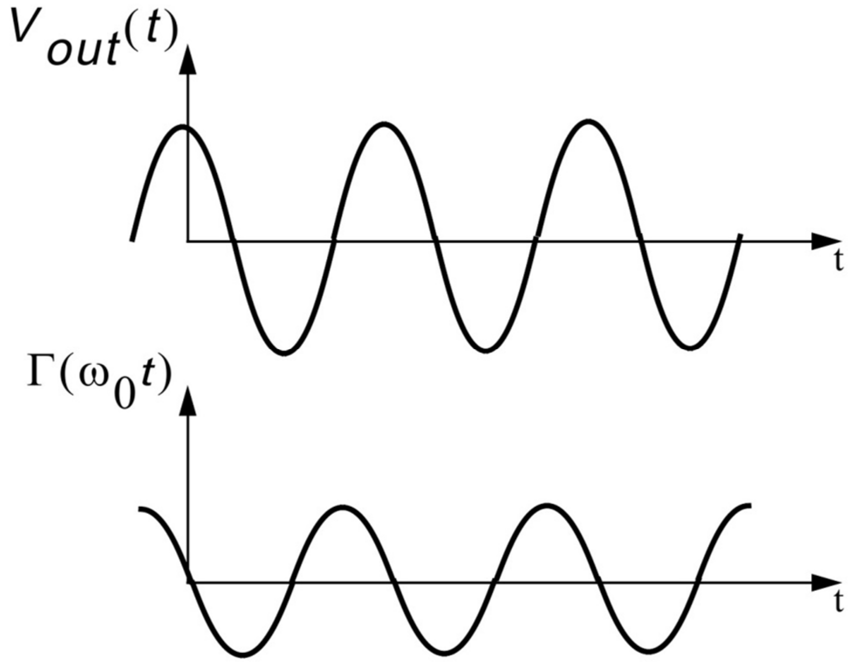
$$h_\phi(t,\tau) = \frac{\Gamma(\omega_0 t)}{q_{\max}} u(t-\tau)$$

$\Gamma(x)$ is a dimensionless function, periodic in 2π, describing how much phase change results from impulse at

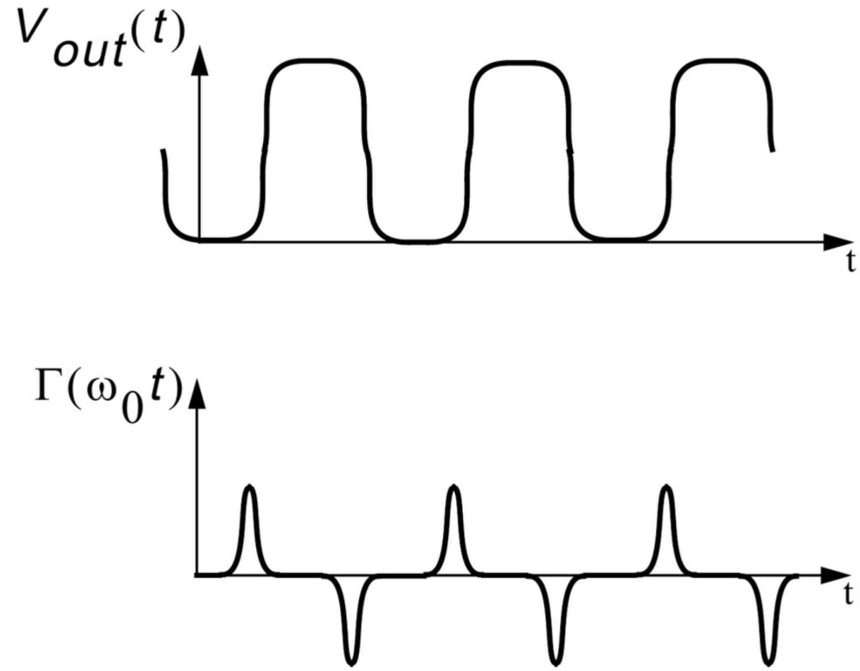
$$t = T\frac{x}{2\pi}$$

Impulse sensitivity function (ISF)

LC oscillator

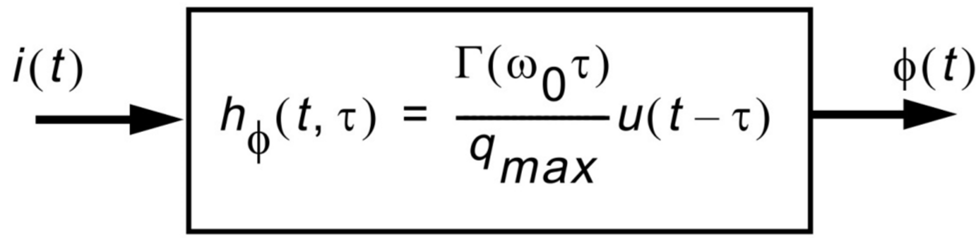


Ring oscillator



The ISF quantifies the sensitivity to perturbations at all instants.

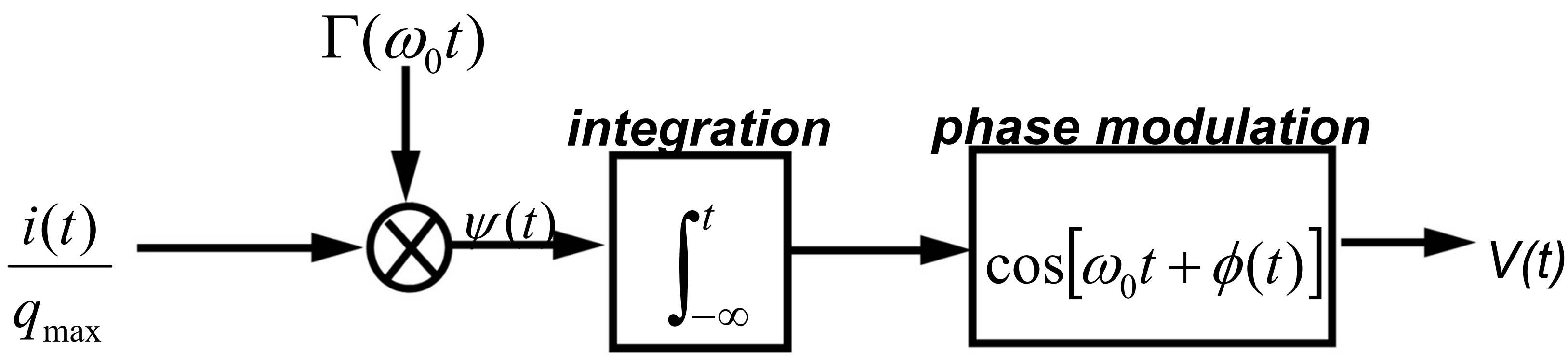
Phase response to arbitrary inputs



Use the superposition integral to compute phase response:

$$\phi(t) = \int_{-\infty}^{\infty} h_\phi(t,\tau) i(\tau) d\tau = \frac{1}{q_{\max}} \int_{-\infty}^{t} \Gamma(\omega_0 t) i(\tau) d\tau$$

Block diagram representation of superposition integral is:



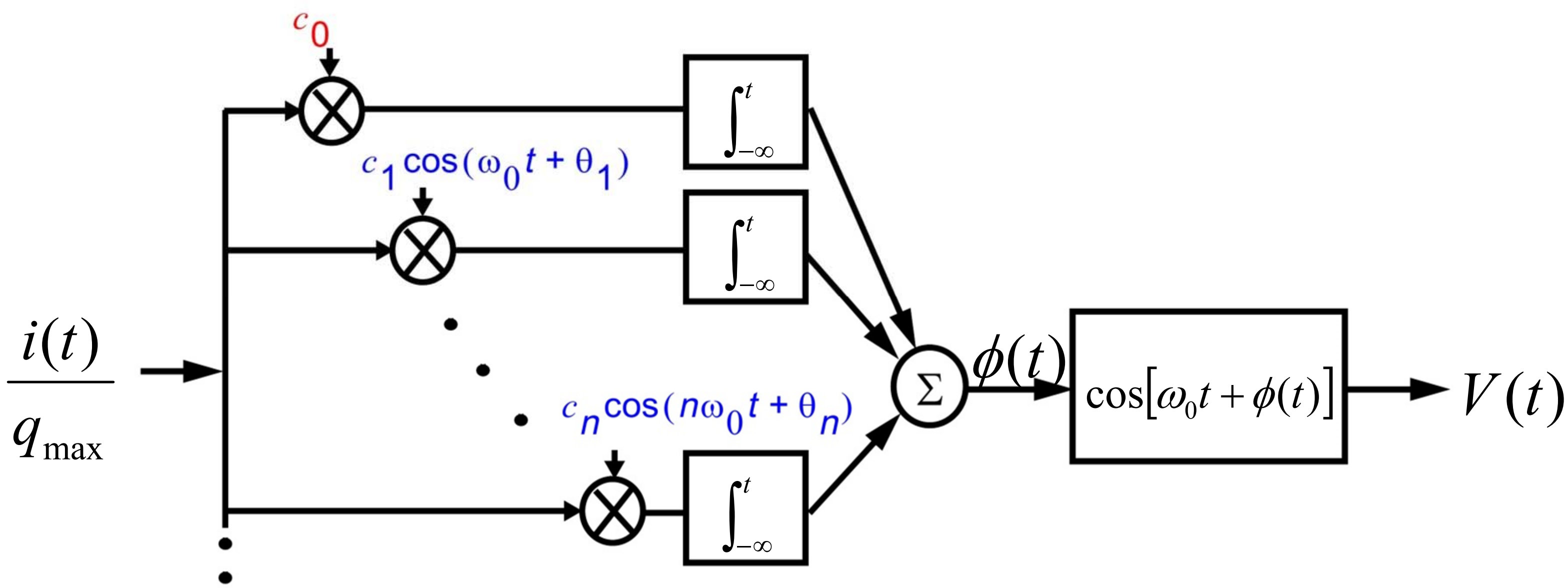
ISF in greater detail

ISF is periodic, expressible as a Fourier series:

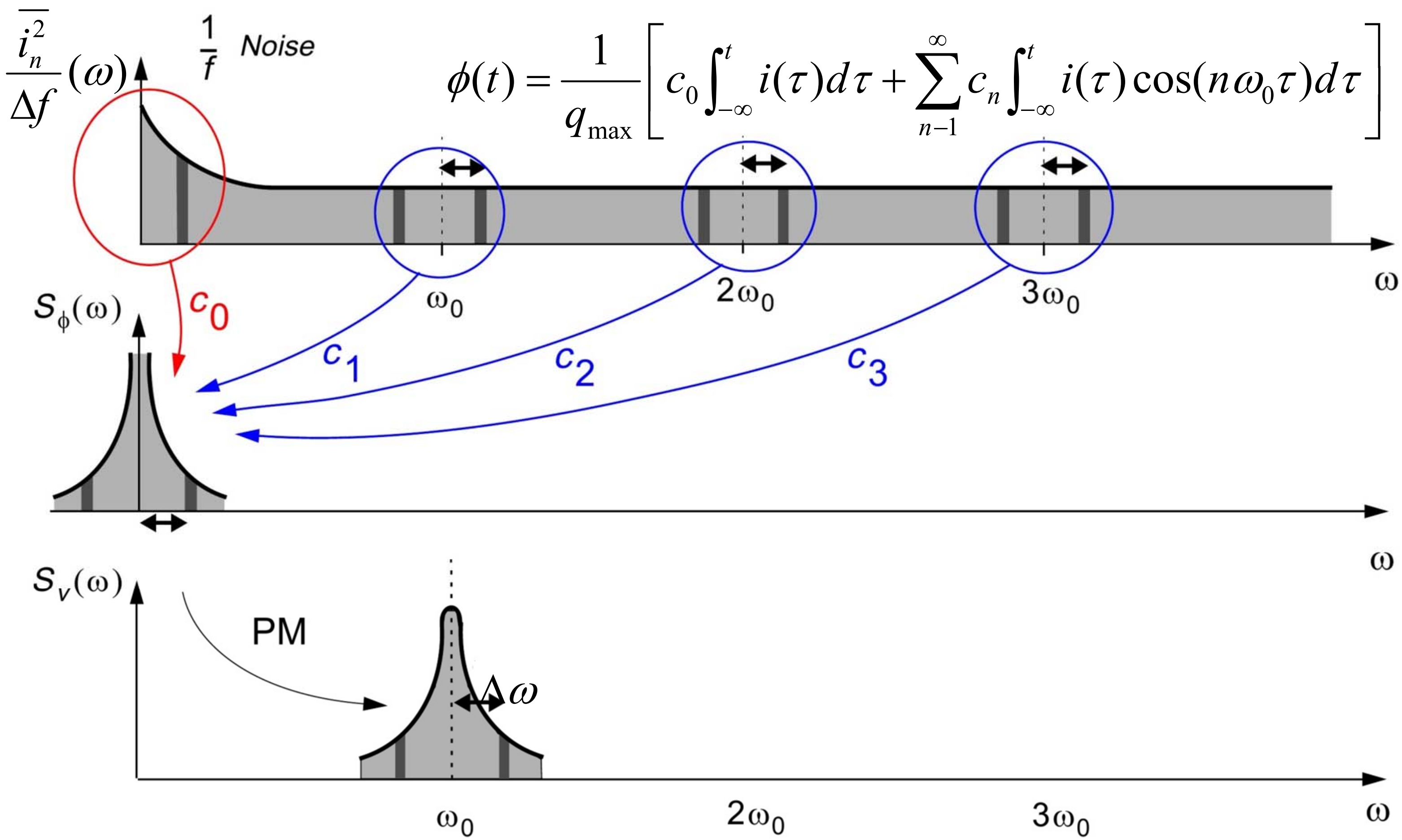
$$\Gamma(\omega_0 t) = c_0 + \sum_{1}^{\infty} c_n \cos(n\omega_0 t + \theta_n)$$

The phase is then as follows:

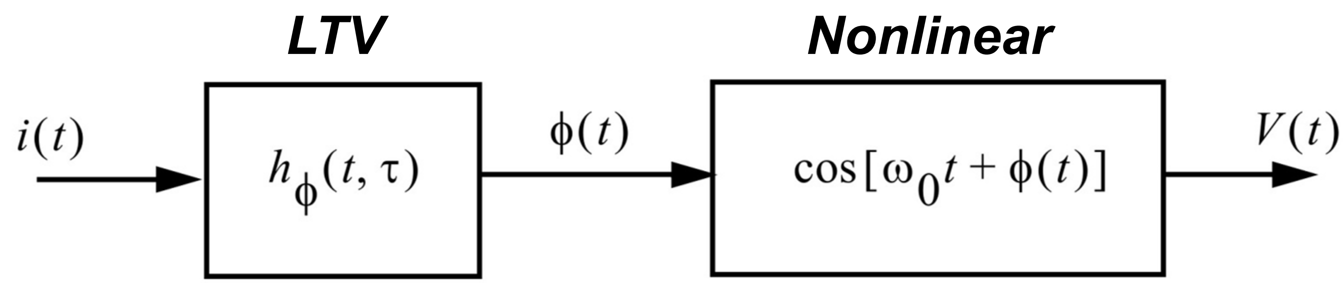
$$\phi(t) = \frac{1}{q_{\max}}\left[c_0 \int_{-\infty}^{t} i(\tau)\,d\tau + \sum_{n-1}^{\infty} c_n \int_{-\infty}^{t} i(\tau)\cos(n\omega_0 \tau)\,d\tau \right]$$



Contributions by noise at $n\omega_0$



Phase noise due to white noise

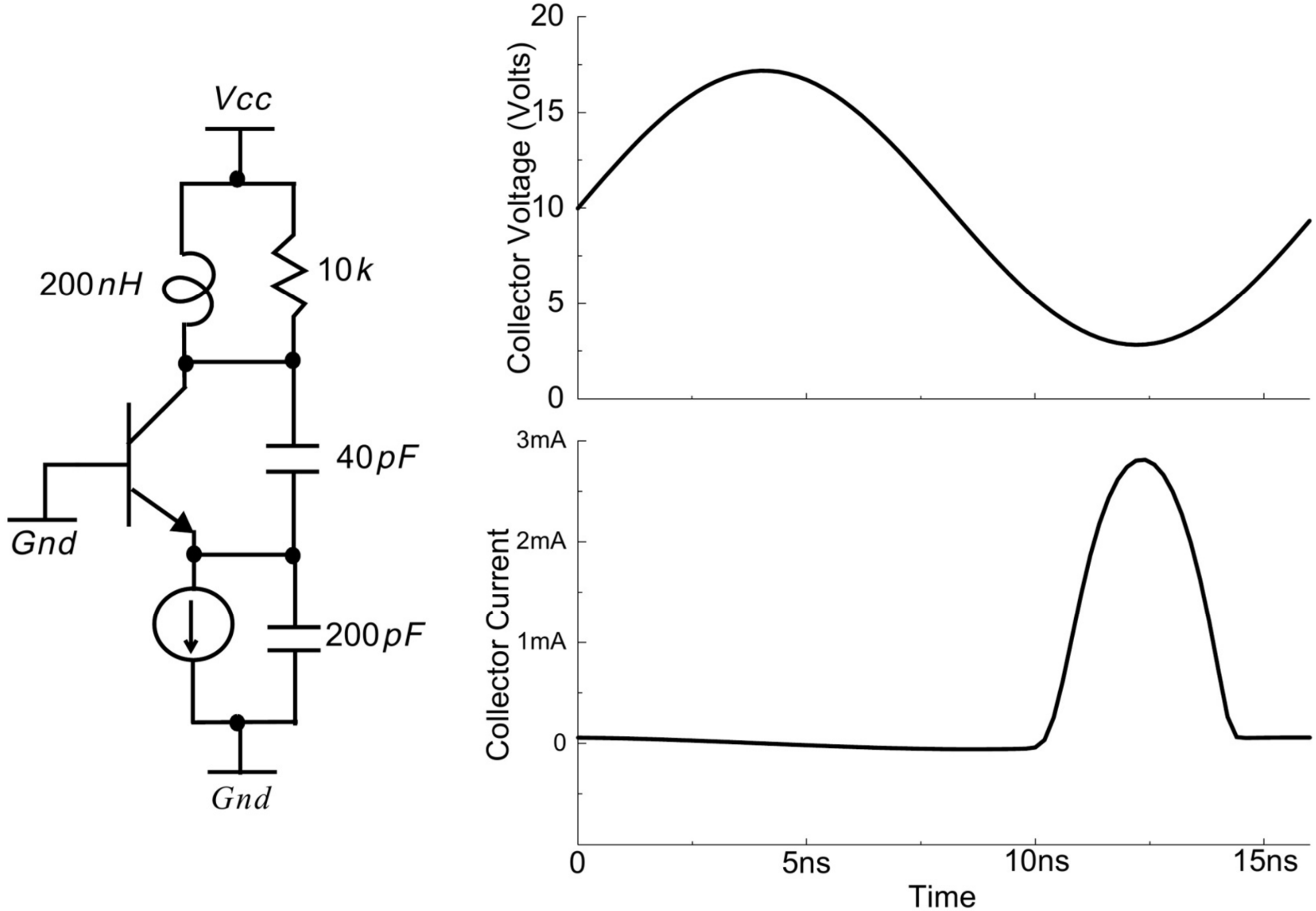


For a white noise input current of spectral density $\dfrac{\overline{i_n^2}}{\Delta f}$,
the phase noise is given by

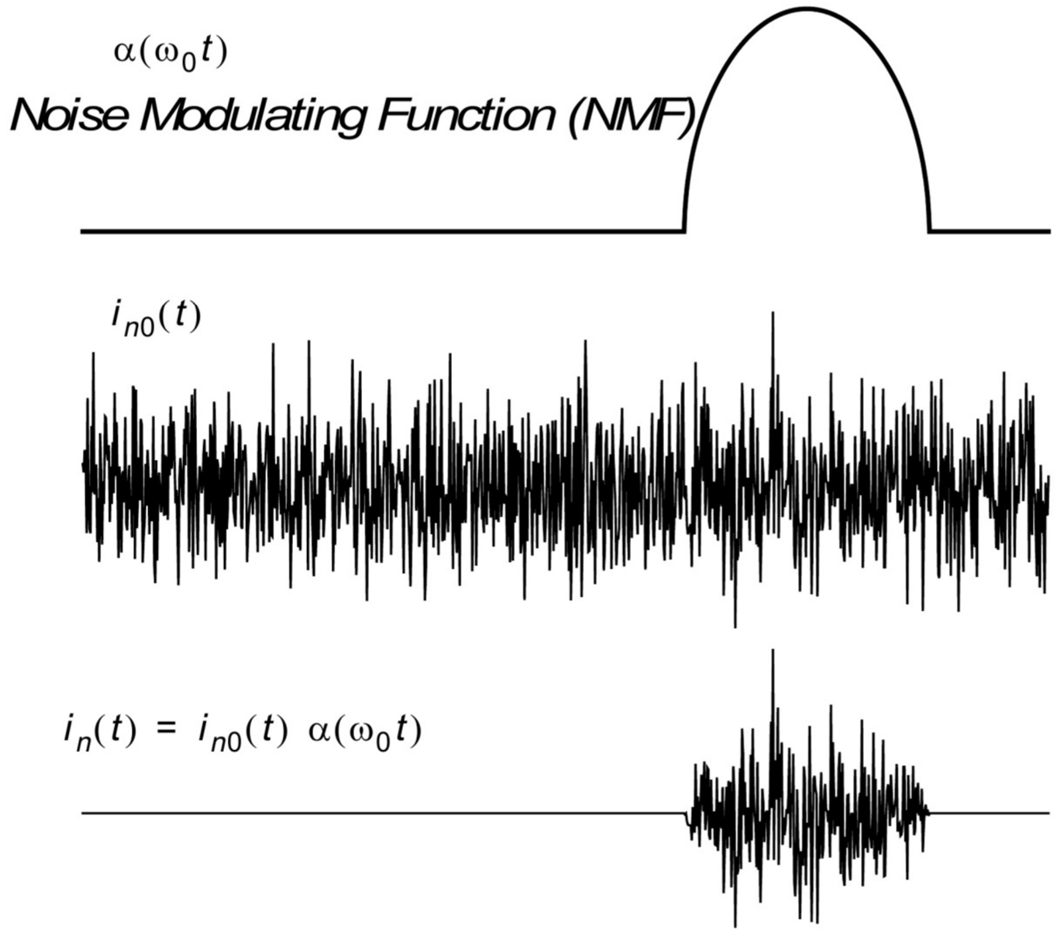
$$\mathcal{L}(\Delta\omega) = 10\log\left[\frac{\Gamma_{rms}^2}{q_{\max}^2}\frac{\overline{i_n^2}/\Delta f}{2(\Delta\omega)^2}\right],$$

where Γ_{rms} ***is the rms value of the ISF.***

Oscillator currents are time-varying



Effects of time-varying currents



$$\phi(t) = \int_{-\infty}^{t} i_n(t) \frac{\Gamma(\omega_0 \tau)}{q_{\max}} d\tau$$

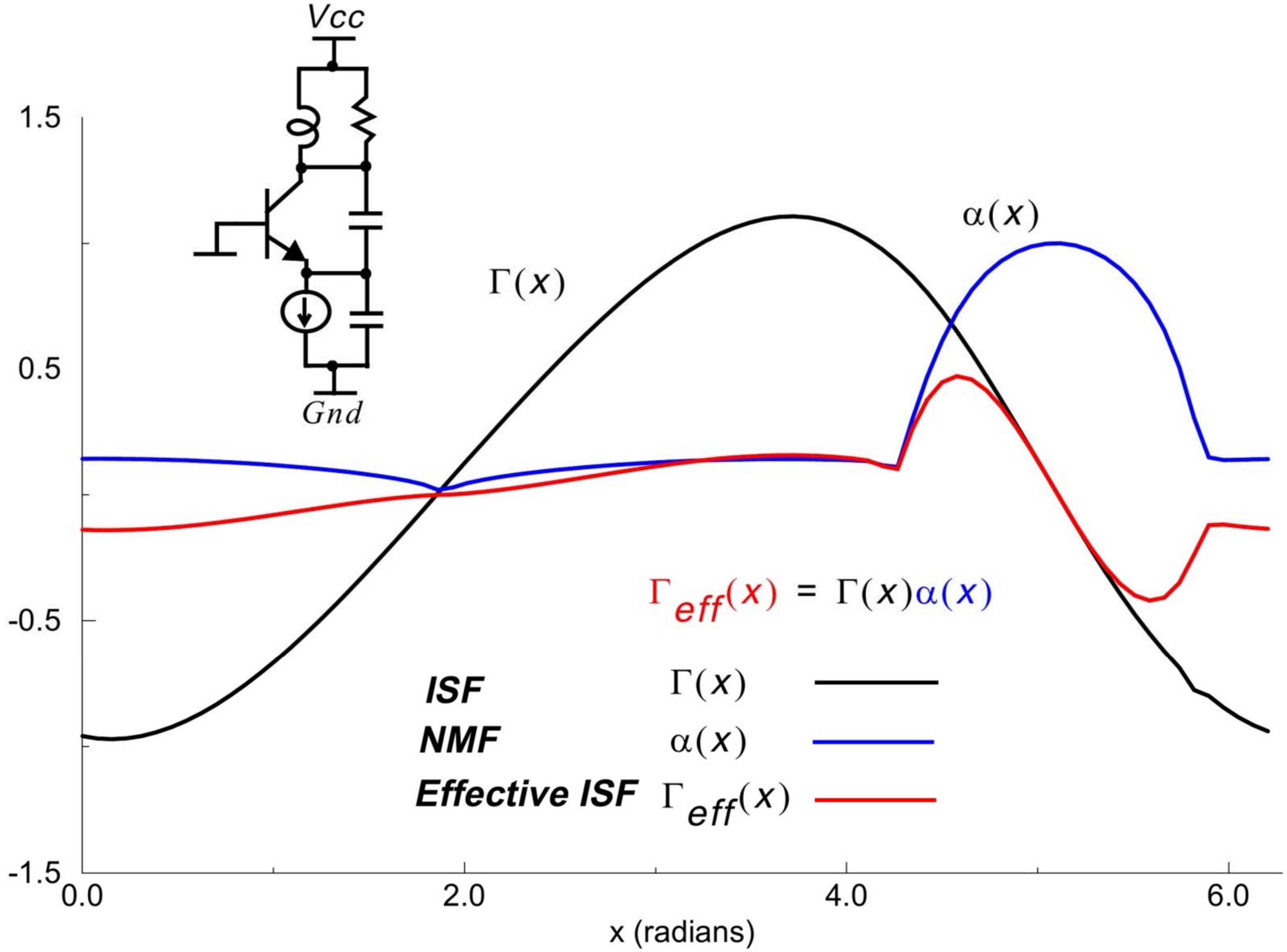
$$= \int_{-\infty}^{t} i_{n0}(t) \frac{\alpha(\omega_0 \tau)\Gamma(\omega_0 \tau)}{q_{\max}} d\tau$$

Effective ISF:

$$\Gamma_{eff}(x) = \Gamma(x) \cdot \alpha(x)$$

Cyclostationary noise can be modeled as stationary with modified ISF.

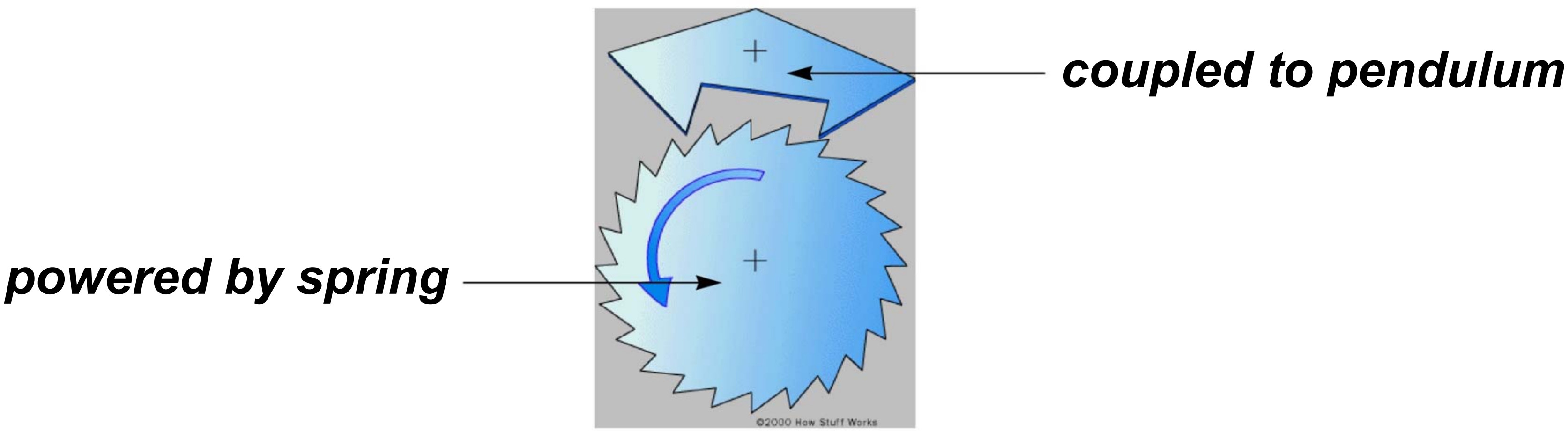
Colpitts



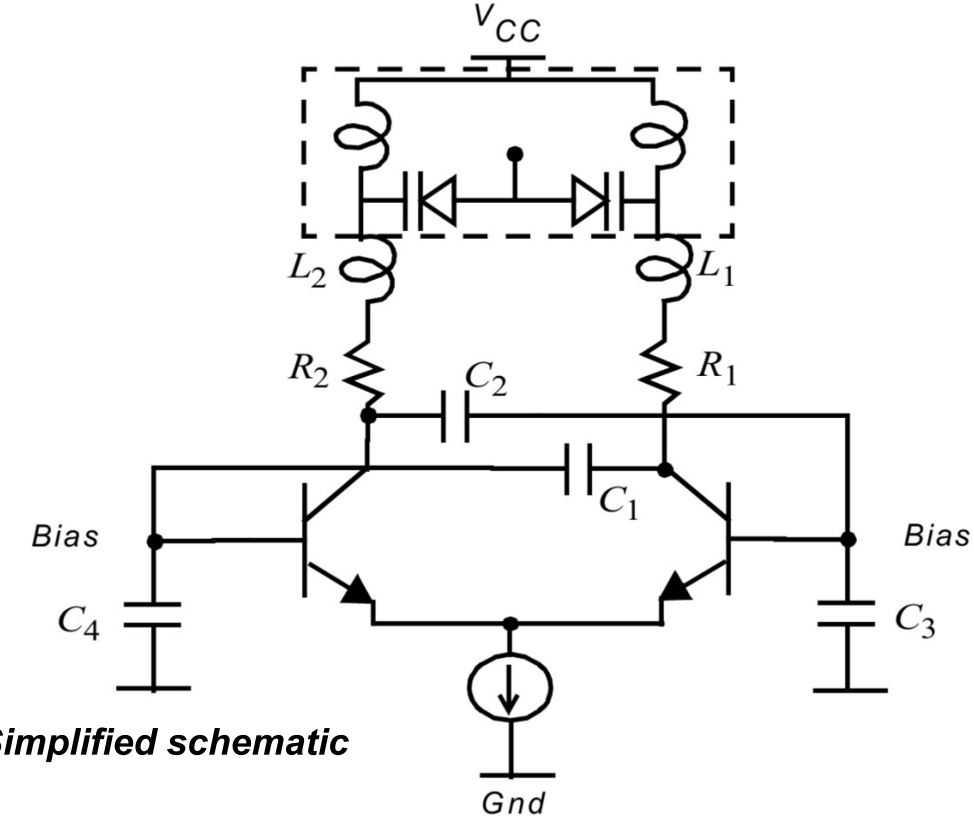
Vcc
Gnd
1.5
0.5
-0.5
-1.5
0.0
2.0
4.0
6.0
x (radians)
Γ(x)
α(x)
Γeff(x) = Γ(x)α(x)
ISF Γ(x)
NMF α(x)
Effective ISF Γeff(x)

Plus ça change...

- To exploit cyclostationarity, deliver energy to the tank impulsively, where the ISF is a minimum.
- This idea is old; an *escapement* in mechanical clocks delivers energy to pendulum impulsively.



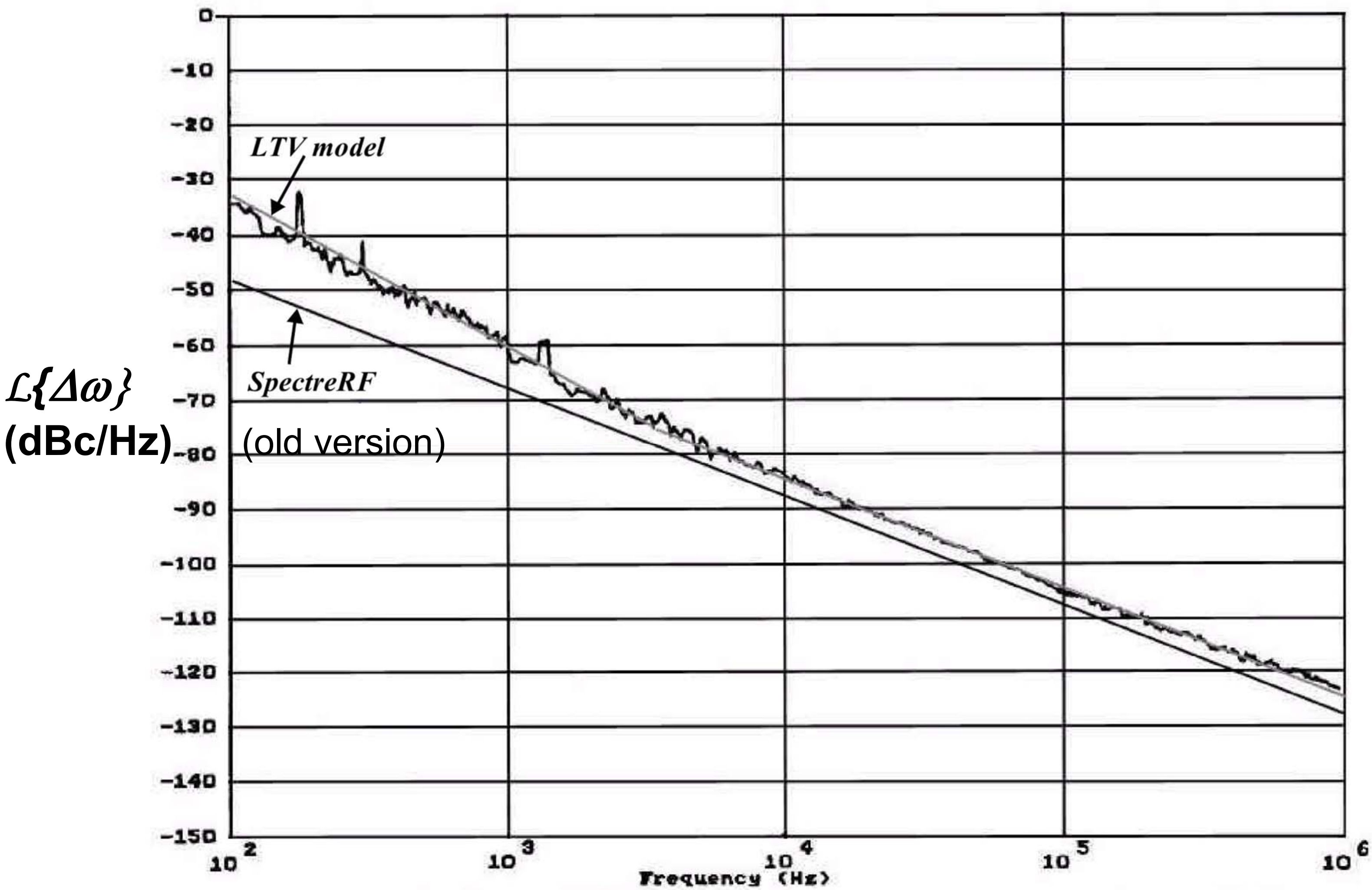
A non-Stanford, non-CMOS example



Simplified schematic

Ref: M.A. Margarit, Joo Leong Tham, R.G. Meyer, M.J. Deen, "A low-noise, low-power VCO with automatic amplitude control for wireless applications," IEEE JSSC, June 1999, pp.761-771.

Theory vs. Measurement vs. Cadence



The Bright Side: "Stupid Amplifier Tricks"

(with apologies to David Letterman)

The myth of GBW limits

- Most circuit design textbooks convey the belief that there is a “gain-bandwidth limit.”
  - Given certain assumptions, yes, G-BW is limited, but that limit isn’t fixed…so it’s not quite as hard a limit.

- Let’s see what happens if we violate those “certain assumptions.”

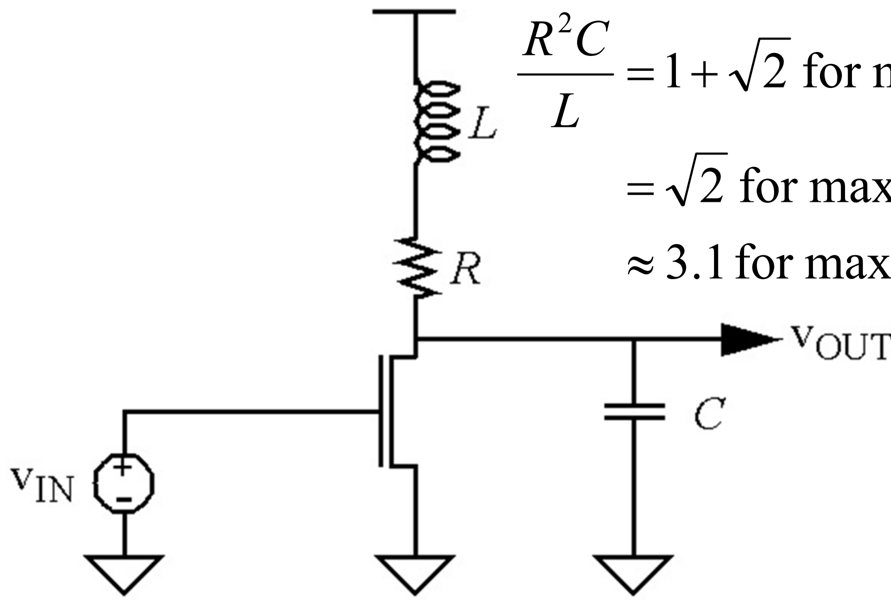
Single-pole, common-source

- *Everyone* knows that the gain is $g_m R_L$, -3dB bandwidth is $1/R_L C_L$, so GBW is just g_m/C_L.
  - Single-pole systems trade gain for bandwidth directly.

- But what about higher-order systems? Can we exploit the implicit additional degrees of freedom to do better?

Shunt peaking

- Ancient idea – dates back to the 1930s.



$$\frac{R^2C}{L} = 1+\sqrt{2} \text{ for maximally flat amplitude}$$

$$= \sqrt{2} \text{ for maximum BW}$$

$$\approx 3.1 \text{ for maximally flat delay}$$

- Adding one low-*Q* inductor nearly doubles BW.
  - 1.85x max. boost; 1.72x boost @ max. flat amplitude, 1.6 boost for max. flat delay.

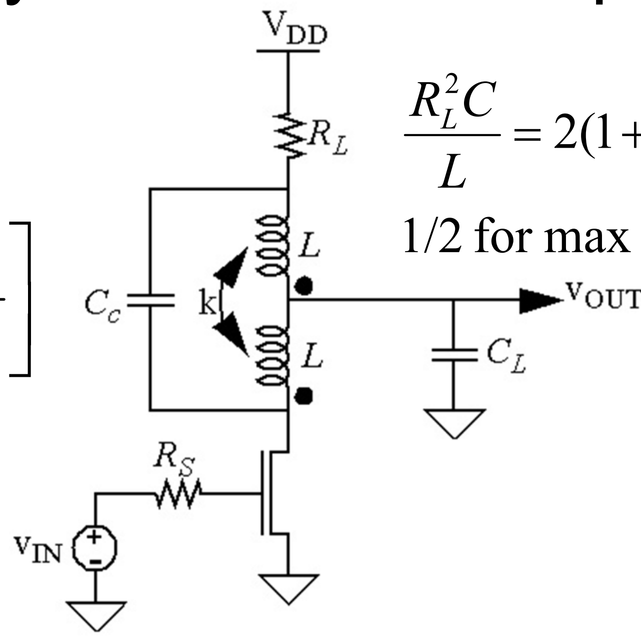
Key idea: Delay is helpful

- Want all of signal current to charge load capacitor.
- Single-pole circuit splits current between resistor and load capacitor.
- Shunt peaking works by delaying current flow through unwanted resistive path.
  - Can we do better than what a single added inductor can do?

Bridged T-coil

- Used extensively in Tektronix ‘scopes.

$$C_C = \frac{C_L}{4}\left[\frac{1-k}{1+k}\right]$$

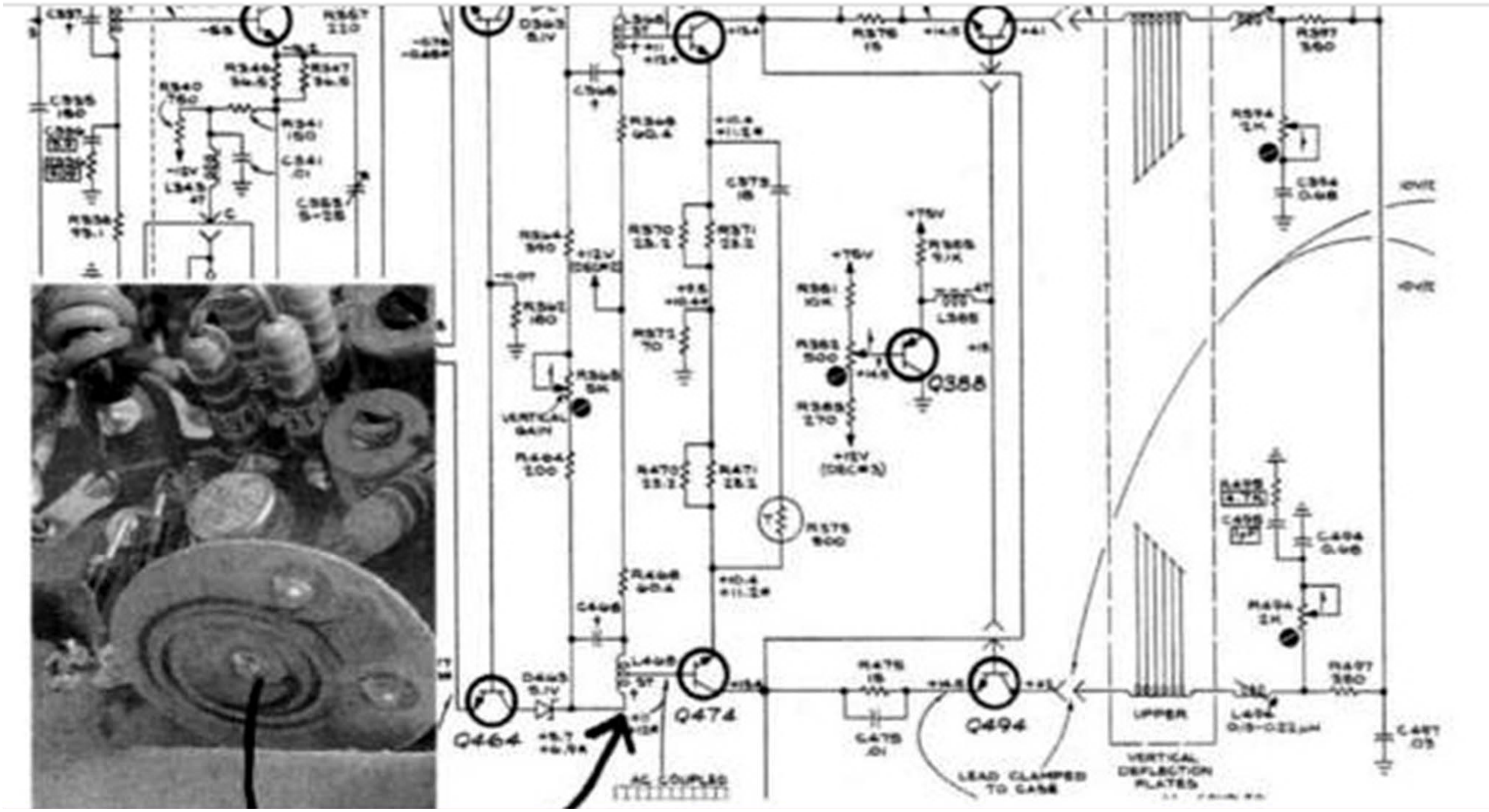


$$\frac{R_L^2 C}{L} = 2(1+k);\ \mathrm{k} = 1/3 \text{ for max. flat gain,}$$

1/2 for max flat delay

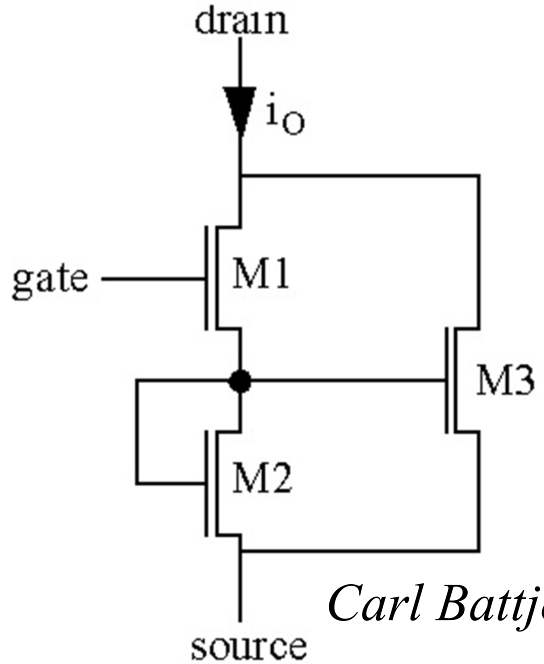
- A small capacitor and low-*Q* transformer nearly triples BW.
    - ~3x max boost; 2.83x boost @ max. flat delay.

Bridged T-coils in Tek 454



1967; last, fastest (150MHz) Tek scope using all discretes

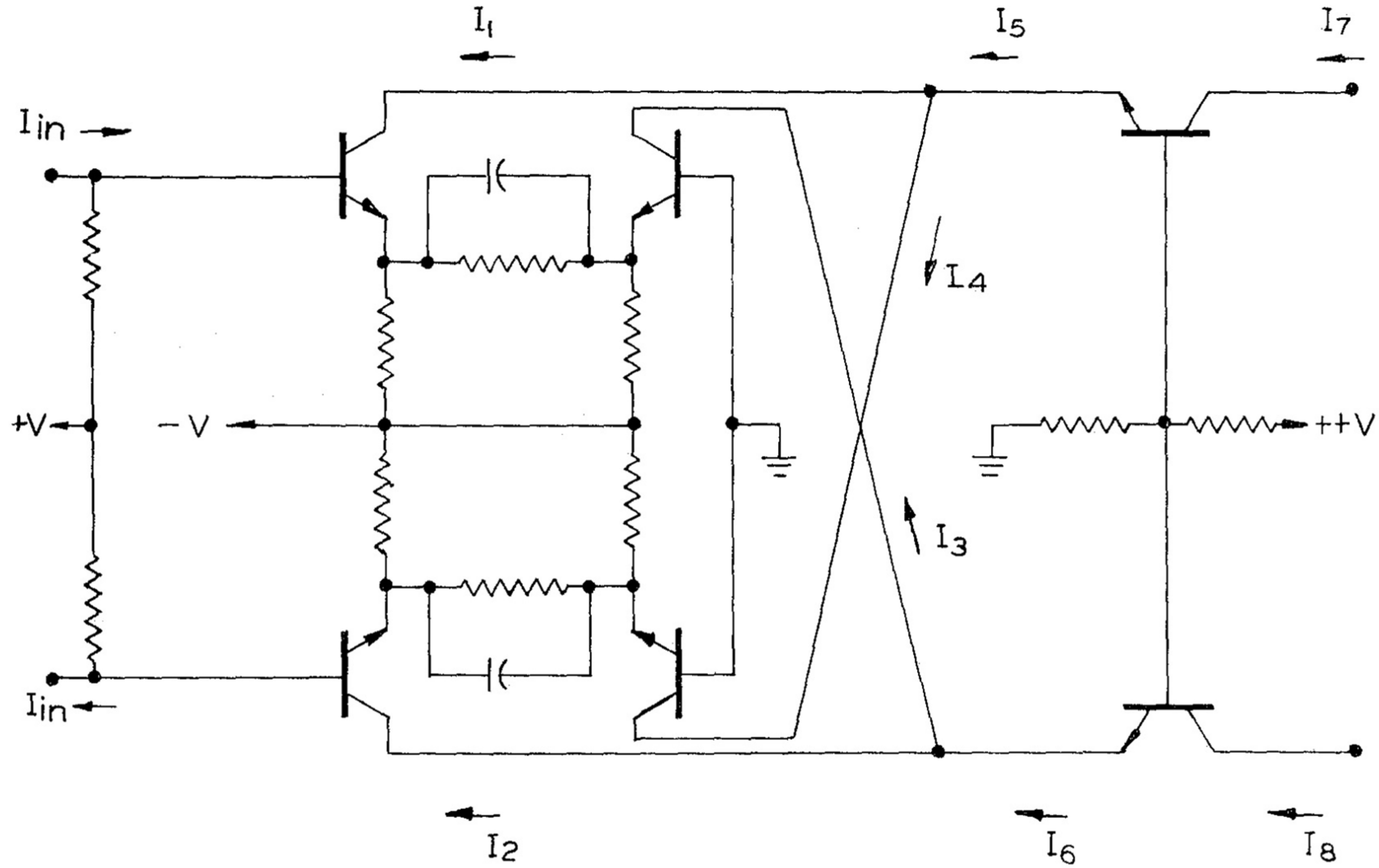
Bright secret: f_T isn't a process limit



Carl Battjes, Tektronix, US Pat. 4236119

- $\omega_T = g_m/C_{in}$, to an approximation. The Battjes doubler reduces C_{in} to provide a 1.5x boost in f_T. Used with T-coils in Tek 7904 scope to get 500MHz system BW using 3GHz transistors.

Bright secret: f_T isn't a process limit



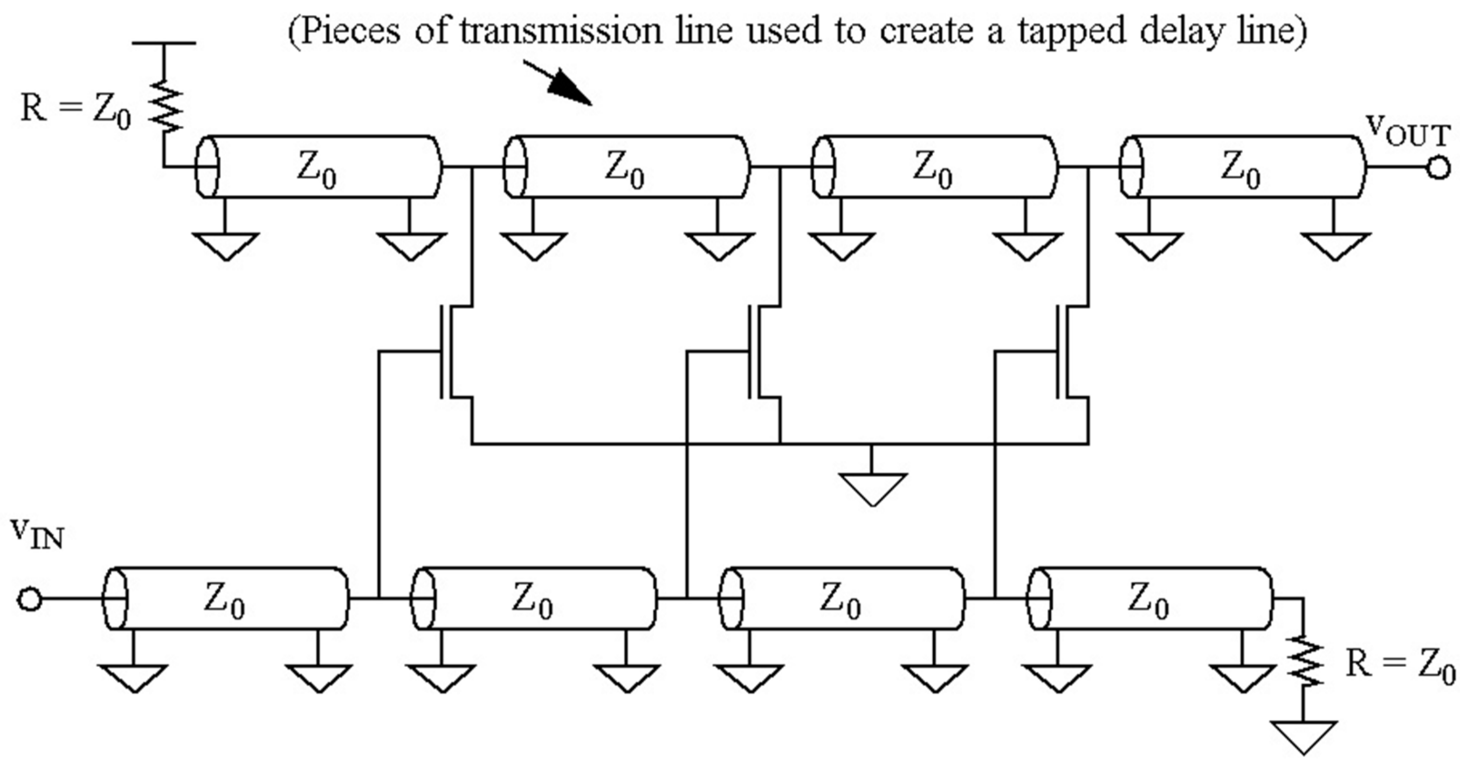
Carl Battjes, Tektronix, US Pat. 3633120

The message?

- More poles = more degrees of freedom = greater ability to effect desirable tradeoffs.
- Expand gain-bandwidth tradeoff to gain-bandwidth-*delay* tradeoff.
  - Delay is frequently unimportant, so trading it for GBW is possible in many practical cases.
  - $G\text{-}BW\text{-}T_D$ tradeoff possible *only* if you can create (and tolerate) excess delay – you can't trade what you don't have. In turn, creating delay over large BW implies many poles.

The distributed amplifier

- First commercialized by Tektronix (after paper by Ginzton, Hewlett, Jasper and Noe).

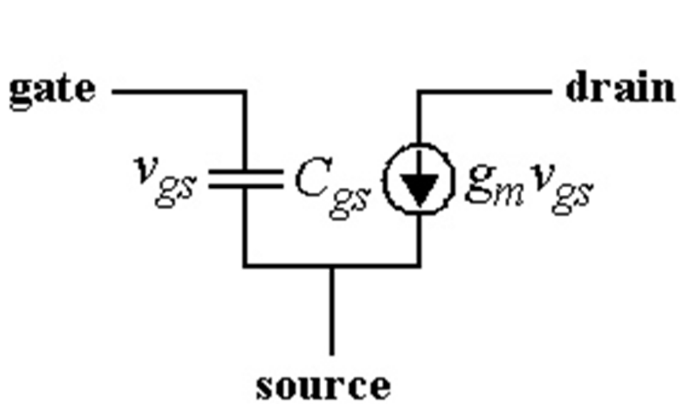


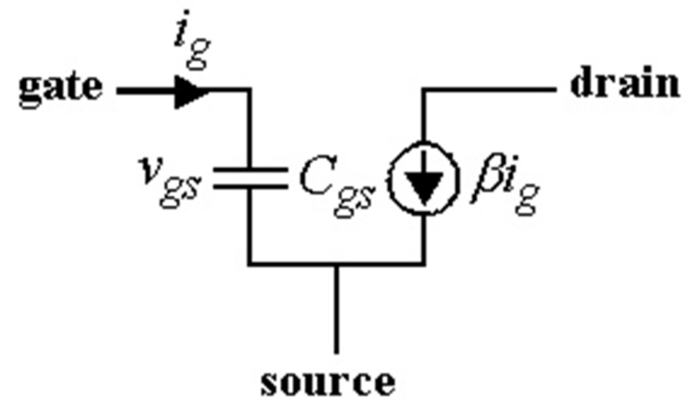
When good amplifiers go bad

Strange Impedance Behaviors

First: Some simple transistor models

- Can use *either* gate-source voltage *or* gate current as independent control variable



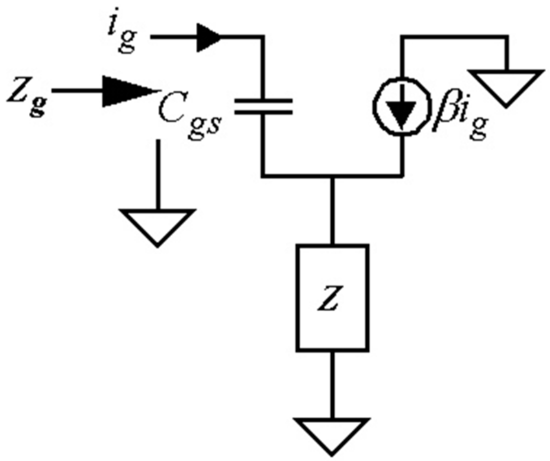


- Models are fully equivalent as long as we choose

$$\beta = \frac{g_m v_{gs}}{i_g} = \frac{g_m}{sC_{gs}} = \frac{\omega_T}{j\omega} = -j\frac{\omega_T}{\omega}$$

View from the gate: Load in source

- Consider input impedance of the following at $\omega << \omega_T$:



$$Z_g = \frac{1}{j\omega C_{gs}} + (\beta+1)Z \approx \frac{1}{j\omega C_{gs}} + (-j\frac{\omega_T}{\omega})Z$$

- The non-intuitive behavior comes from the second term: The impedance *Z* gets multiplied by a (negative) imaginary constant.

What does multiplication by $-j\omega_T/\omega$ do?

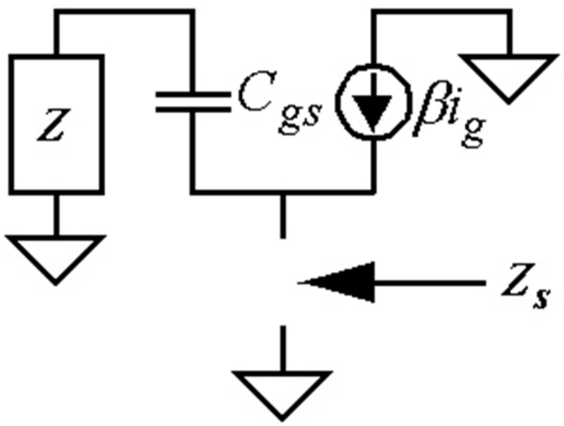
- Turns R into capacitance = $1/\omega_T R$.

- Turns L into resistance = $\omega_T L = g_m(L/C_{gs})$.

- Turns C into *negative* resistance = $-\omega_T/\omega^2 C$
  $= -(\omega_T/\omega)|Z_C|$.

View from the source: Load in gate

- Now consider input impedance of the following:



$$Z_s = \frac{\dfrac{1}{j\omega C_{gs}} + Z}{\beta + 1} \approx \frac{1}{g_m} + \left(j\frac{\omega}{\omega_T}\right)Z$$

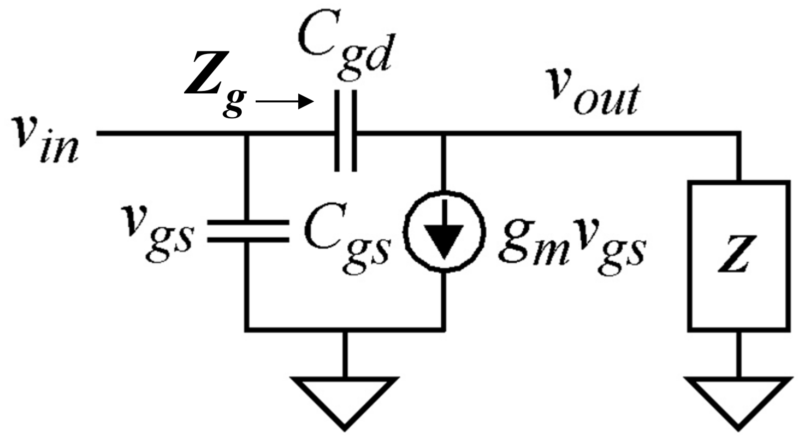
- This time, *Z* gets multiplied by a *+j* factor.

What does multiplication by $+j\omega/\omega_T$ do?

- Turns R into inductance = R/ω_T.
- Turns C into resistance = $1/\omega_T C = (C_{gs}/C)\,(1/g_m)$.
- Turns L into negative resistance = $-\omega^2 L/\omega_T$ = $-(\omega/\omega_T)|Z_L|$.

View from the gate: Load in drain

- Consider (partial) input impedance of the following:



$$Z_g \approx \frac{1}{sC_{gd}(1+g_m Z)}$$

$$Z_g \approx \frac{1}{sC_{gd}\, g_m Z} \quad \textit{if } |g_m Z| >> 1$$

- This is the generalized Miller effect: C_{gd} is multiplied by *complex* gain, when viewed by gate.

What does multiplication by $g_m Z$ do?

- Turns R into capacitance = $C_{gd}(g_m R)$; this is just the classic Miller effect.

- Turns C into resistance = $(C/C_{gd})(1/g_m)$.

- Turns L into negative conductance = $-\omega^2 g_m L C_{gd}$.

Why SIBs are strange

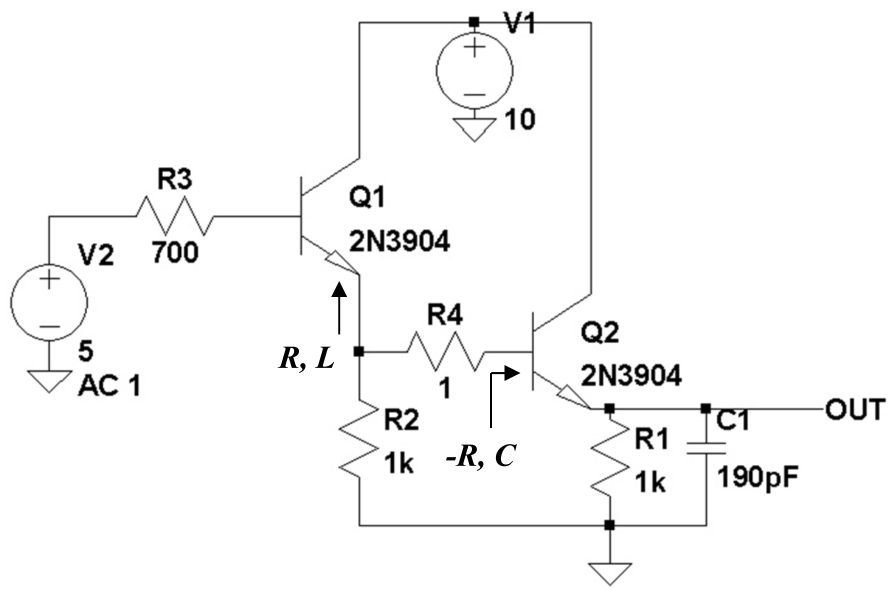
- Apparent weirdness arises because of feedback around complex gains.

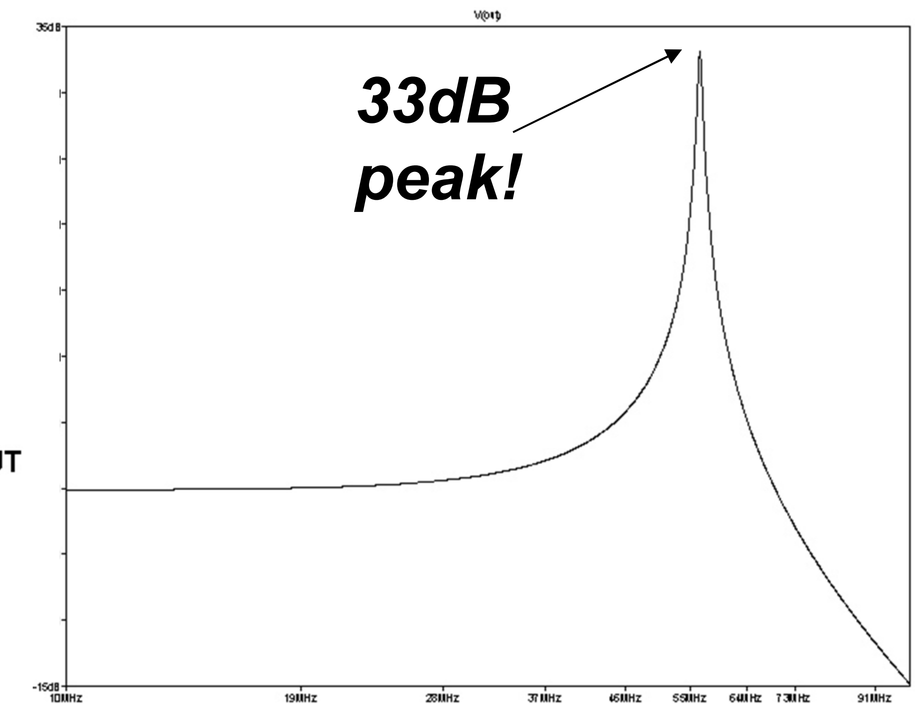
- Phase shift associated with complex gains causes impedances to change *character*, not just magnitude.

- The strangeness evaporates once you spend a little time studying where it comes from.

SIBs example: Cascaded followers

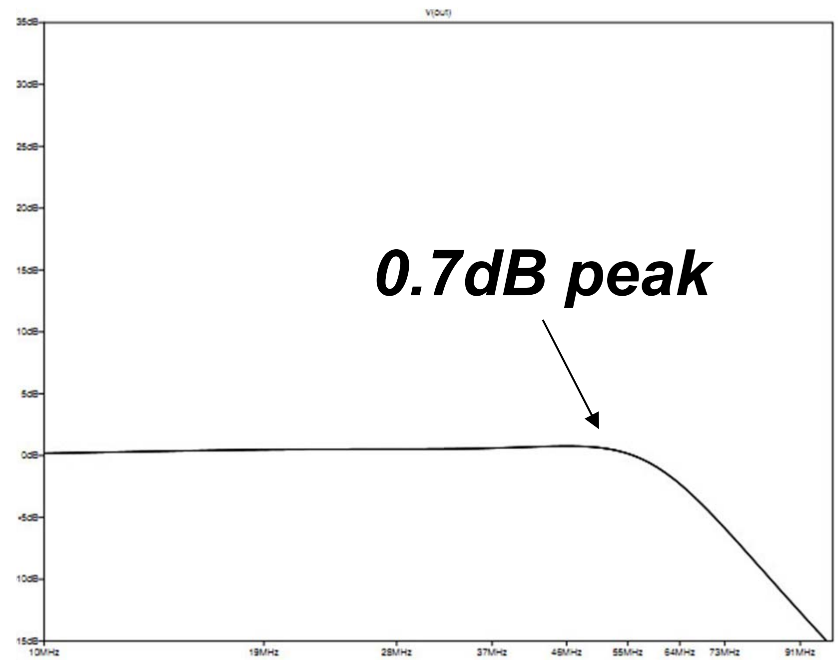
- Familiar circuit has surprising, terrifying but *understandable* behavior:





Cascaded followers: The fix

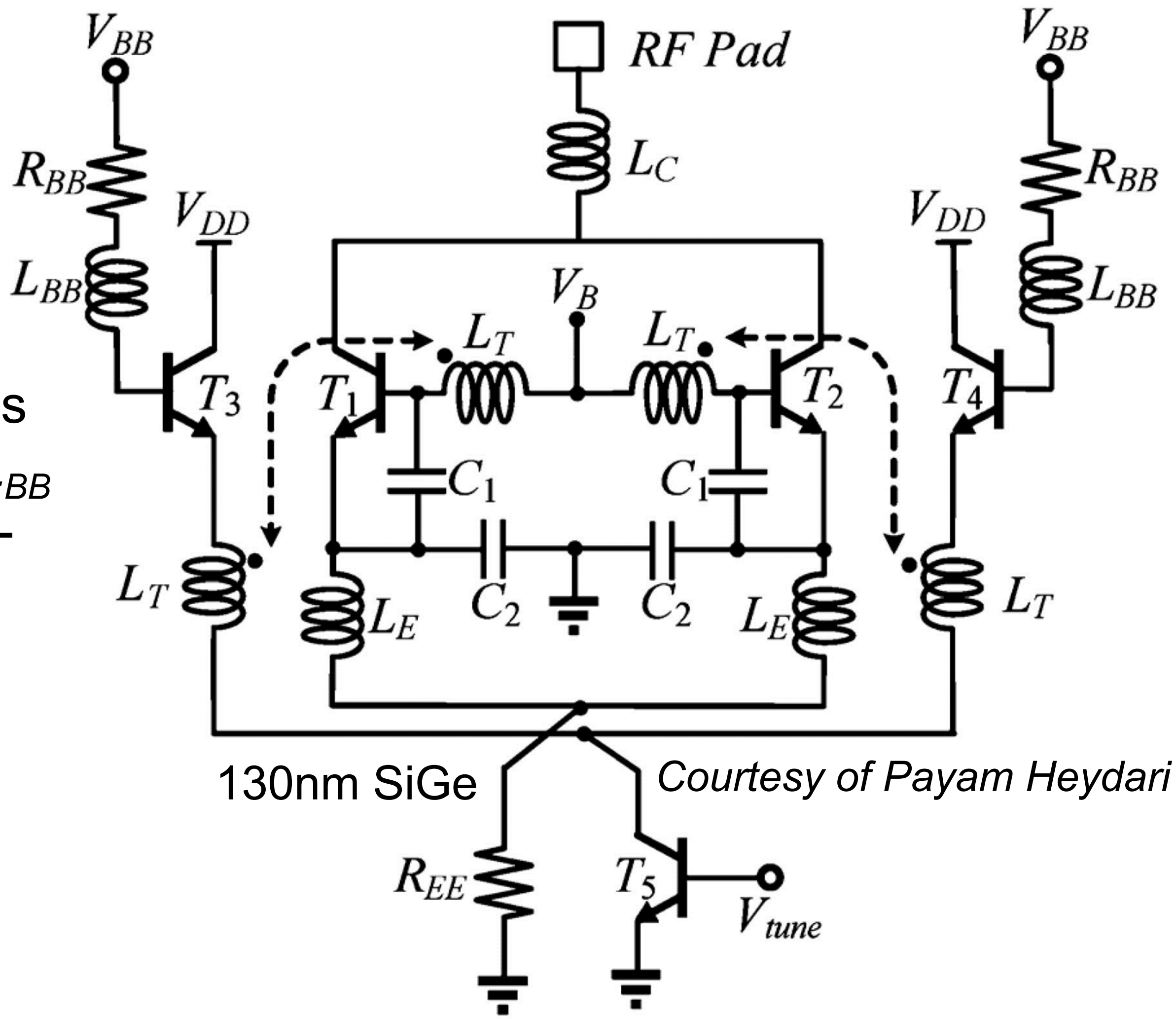
- Problem is negative *R*, so add *R4* (220Ω) to cancel it:



- Shunt R4 with a capacitor to reduce BW loss.

Exploiting SIBs: 200GHz push-push VCO

R_{BB} becomes inductance, L_{BB} becomes Q-enhancing negative R.



130nm SiGe

Courtesy of Payam Heydari

Summary

- RF circuits are certainly complicated, but that shouldn't make us concede defeat.
- Throw away the pointy hat, free the chickens, quit babbling in Latin, and stop weeping uncontrollably.
- Everything is explicable; it's not magic!

References

[Goo] J.S. Goo, *High Frequency Noise in CMOS Low-Noise Amplifiers*, Doctoral Dissertation, Stanford University, August 2001.

[Jones] H. E. Jones, US Pat. #3,241,078, “Dual Output Synchronous Detector Utilizing Transistorized Differential Amplifiers,” issued March 1966.

[Lee] *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd edition, Cambridge U. Press, 2004.

[Shaeffer] D. Shaeffer and T. Lee, “A 1.5-V, 1.5-GHz CMOS Low *Noise* Amplifier,” *IEEE J. Solid-State Circuits*, v.32, pp. 745-758, 1997.

[Terrovitis] M. T. Terrovitis and R. G. Meyer, "Noise in Current-Commutating CMOS Mixers," *IEEE Journal of Solid-State Circuits,* vol. 34, No. 6, June 1999.

Appendix: Mixers

Mixers are supposed to be linear!

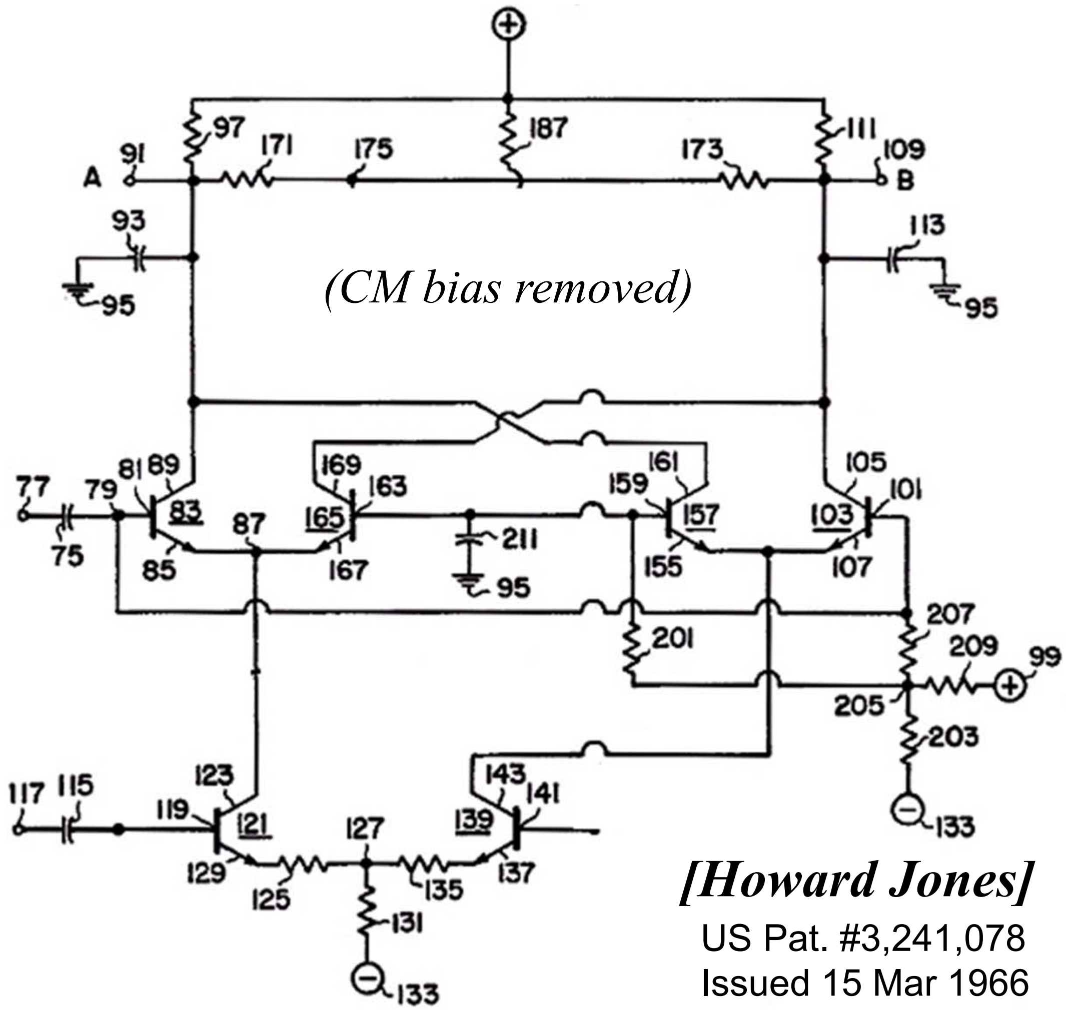
- But they are *time-varying* blocks.
  - Too many textbooks and papers say “mixers are nonlinear…” Mixers are nonlinear in the same way that amplifiers are nonlinear: *Undesirably*.

- Mixers are noisier than LNAs for reasons that will be explained shortly. NF values of 10-15dB are not unusual.

- Main function of an LNA is usually to provide enough gain to overcome mixer noise.

Dark secret: Most "Gilbert" mixers aren't

- This is a *Jones* mixer.
  - Most textbooks and papers (still) wrongly call it a Gilbert cell.
- A true Gilbert cell is a *current-domain* circuit, and uses *predistortion* for linearity.



[Howard Jones]
US Pat. #3,241,078
Issued 15 Mar 1966

The mixer: An LTV element

- Whether Gilbert, Jones or Smith, modern mixers depend on *commutation* of currents or voltages.

- We idealize mixing as the equivalent of multiplying the RF signal by a square-wave LO.
  - Single-balanced mixer: RF signal is unipolar.
  - Double-balanced mixer: RF signal is DC-free.

- Mixing is ideally *linear*: Doubling the input (RF) voltage should double the output (IF) voltage.

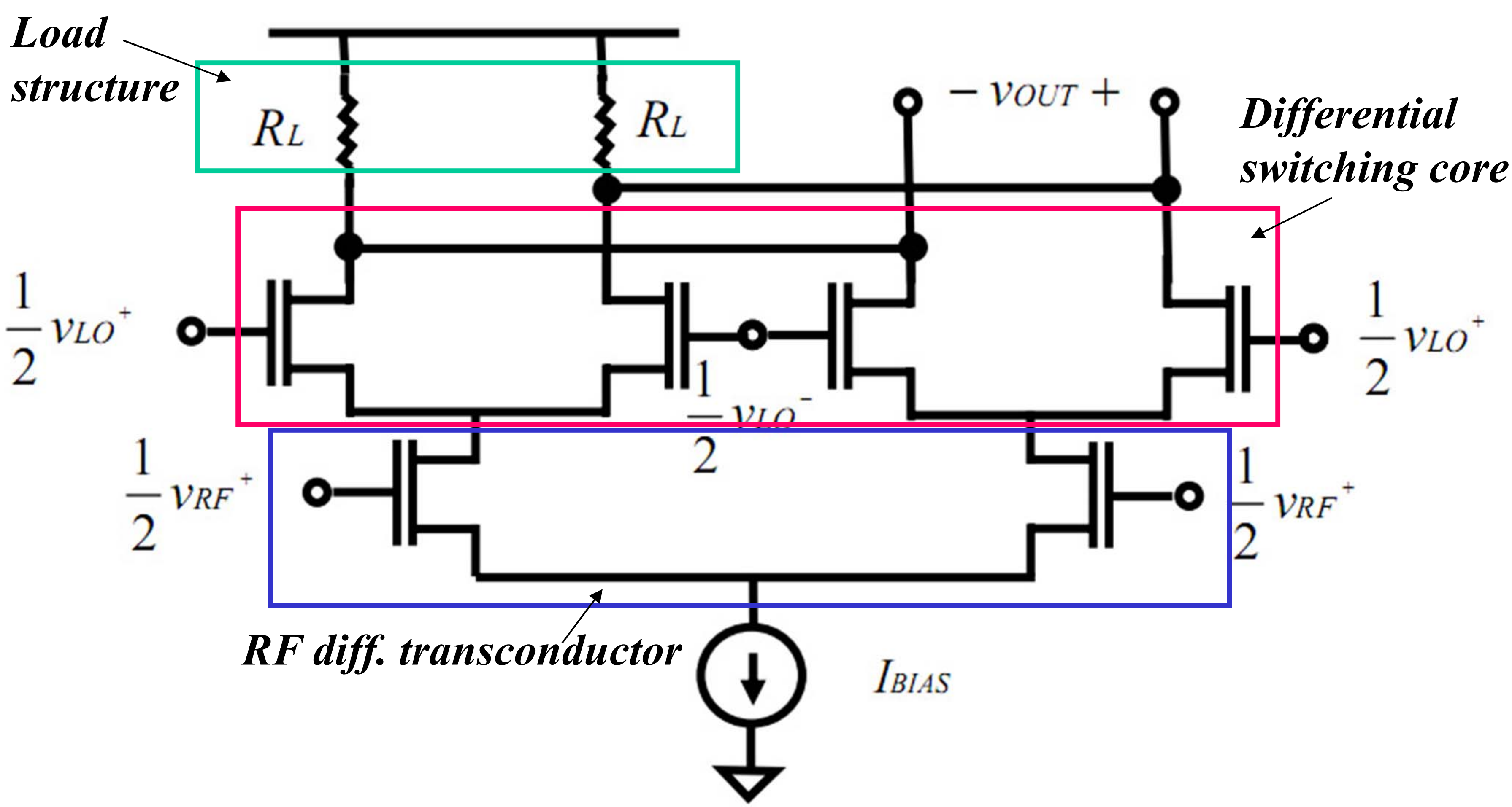
A multiplier is an ideal mixer

- Key relationship is:

$$\boxed{A\cos\omega_1 t}\cos\omega_2 t = \frac{A}{2}[\cos(\omega_1+\omega_2)t + \cos(\omega_1-\omega_2)t]$$

- Can be thought of as an amplifier with a time-varying amplification factor (e.g., term in blue box, above).

Sources of noise in mixers



Mixer noise

- Load structure is at the output, so its noise adds to the output directly; it undergoes no frequency translations.
  - If 1/*f* noise is a concern, use PMOS transistors or poly resistor loads.
- Transconductor noise appears at same port as input RF signal, so it translates in frequency the same way as the RF input.

Dark secret: Switching noise can *dominate*

- Instantaneous switching not possible.
  - Noise from switching core passing through linear region can actually *dominate*.
  - Common-mode capacitance at tail nodes of core can reduce effectiveness of large LO amplitudes.

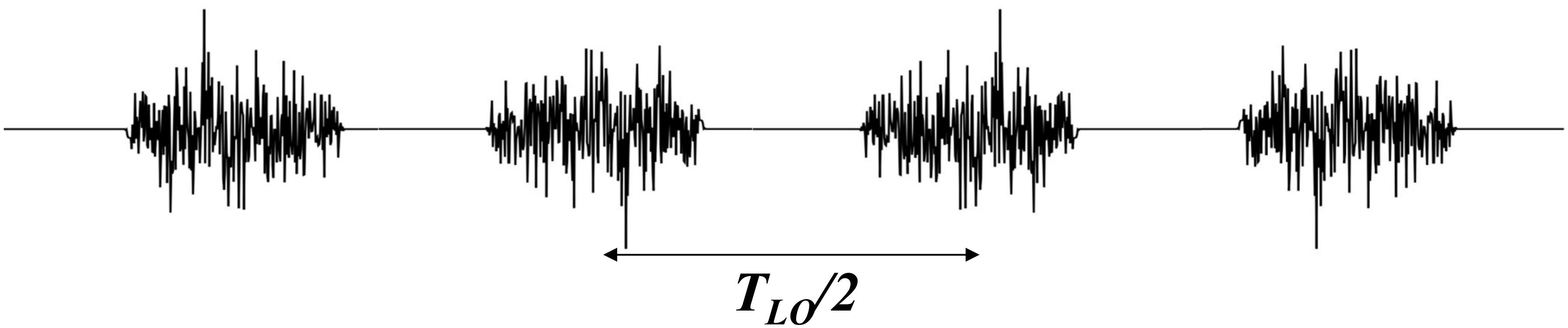
- Periodic core switching is equivalent to windowed sampling of core noise at (twice) the LO rate.
  - Frequency translations occur due to this self-mixing.

Noise contribution of switching core

- As switching transistors are driven through the switching instant, they act as a differential pair for a brief window of time t_s.
  - During this interval, the switching transistors transfer their drain noise to the output.
  - Changing drain current implies a changing PSD for the noise; it is cyclostationary.

Noise contribution of switching core

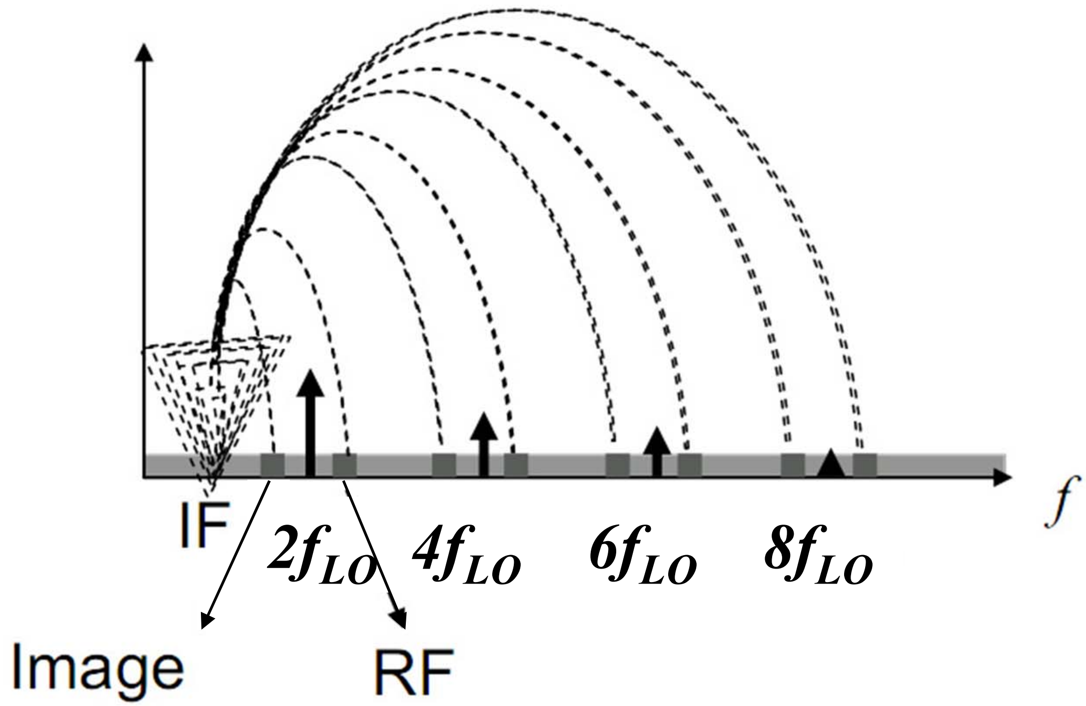
- The noise contributed by the switching core appears roughly as follows:



- Mathematically equivalent to multiplying stationary noise by a shaped pulse train of fundamental frequency $2f_{LO}$.

Noise contribution of switching core

- Noise at $2nf_{LO}$ +/- f_{IF} will therefore translate to the IF. This noise folding partly explains the relatively poor noise figure of mixers.



Terrovitis mixer noise figure equation

- A simplified analytical approximation for the SSB noise figure of a Jones mixer is

important

$$F_{SSB} \approx \frac{\alpha}{c^2} + \frac{2\gamma g_m \alpha + 4\gamma \overline{G} + G_L}{c^2 g_m^2 R_S}$$

- Here, g_m is the transconductance of the bottom differential pair; G_L is the conductance of the load; R_S is the source resistance, and γ is the familiar drain noise parameter.
  - See *[Terrovitis]* for more complete version.

Terrovitis mixer noise figure equation

- The parameter $\overline{G}$ is the time-averaged transconductance of each pair of switching transistors. For a plain-vanilla Jones mixer,

$$\overline{G} \approx \frac{2I_{BIAS}}{\pi V_{LO}}$$

- The parameter α is related to the sampling aperture t_s, and has an approximate value

$$\alpha \approx 1 - \frac{4}{3} t_s f_{LO}$$

Terrovitis mixer noise figure equation

- The parameter *c* is directly related to the effective aperture, and is given by

$$c \approx \frac{2}{\pi}\left[\frac{\sin(\pi t_s f_{LO})}{\pi t_s f_{LO}}\right]$$

- This parameter asymptotically approaches $2/\pi$ in the limit of infinitely fast switching.