Integrated Circuits and Systems for THz Interconnect

Qun Jane Gu

Electrical and Computer Engineering Department
University of California, Davis
Davis, CA 95616, USA
jgu@ucdavis.edu

***Abstract*—THz Interconnect holds the potentials to solve long-standing interconnect issues by leveraging the advantages of both electronics and optics sides due to its unique spectrum position. To support ever-increasing interconnect bandwidth requirement, THz interconnect bandwidth density, energy efficiency, and cost effectiveness need to boost and scale with demands. To achieve that, integrated circuits based on mainstream standard processes are preferred. However, the THz circuit design on mainstream silicon processes impose challenges due to their large parasitics and losses, as well as layout dependent device parameters. This paper presents a design methodology to create layout-aware scalable device model to overcome these challenges and demonstrate it in two circuit design examples.**

Keywords - integrated circuits and systems; interconnect; model; prescaler; THz; VCO

I. Introduction

The rapid development of semiconductor technologies make data processing more and more efficient. For example, energy consumption for each bit processing drops dramatically. It therefore demands increasing data communications within chip and chip-to-chip. However, the energy consumption for data communication are orders of magnitude higher than data processing [1], which forms an increasing gap between processing and communication. Therefore, in the near future, majority of the energy will be consumed by data communication. In addition, the chip I/O constraint requires the communication bandwidth density increase linearly with data rate requirement, which forms a fundamental challenge for conventional electrical interconnect due to its small bandwidth-distance product. The highest data rate per pin reported to date is limited up to 28 Gbps [2, 3].

To solve the issues inherent from conventional electrical interconnect, we have propose THz interconnect, which holds the high promises due to THz unique spectrum position between microwave and optical frequencies by leveraging the advantages from both sides. First, the increasing device speed in silicon processes enable THz signal generation, detection and processing in silicon [4-7]. Standard process based THz interconnect leverages the high volume for extreme low cost and high level of resilience. Second, the THz frequency's quasi-optics feature allows it to use ultra-low loss channels similar as optical communication. This alleviates system link budget requirement to boost the energy efficiency. Furthermore, THz interconnect scales well with processes because deep scaled processes support higher operating frequency for higher data rates and smaller channel size, which leads to higher bandwidth density. In addition, faster device improves the data processing efficiency to improve THz interconnect energy efficiency.

However, one design challenge is the circuit operating speed is highly decreased due to parasitics, which is only a fraction of the device intrinsic speed. The situation gets worse in deep-micron processes due to a large ratio of external parasitics versus intrinsic circuit load. Second, high loss of silicon processes, including both ohmic electric loss and dielectric substrate coupling loss, challenges design efficiency. Even more critical, the models from foundries cannot support THz frequency due to its highly layout dependent feature.

To mitigate these challenges, we have investigated a design methodology to effectively guide the design, which is discussed in Section II. Section III exemplifies two circuit design by using this methodology. Section IV presents measurement results and the conclusion is provided in Section V.

II. Layout Dependnt Scaling Model

Because THz circuit operations are highly dependent on layout, to guide accrate design, we have formed a procedure to build a scalable and layout-aware active device model by adding extrinsic parasitics on top of the device core model from foundries. The additional extrinsic parasitics are extracted from parasitic extraction tools such as Calibre, or EM simulation tools like HFSS and Momentum, and then are modeled based on the device physical layout. In THz circuit design, one of the key concerns is to boost operation speed, which is strongly related to boost device maximum oscillation frequency, f_{MAX}. It is affected by gate resistance Rg, gate drain overlap capacitance C_{OV}, device unit current gain frequency f_T, device output conductance gds [8]. However, these parameters are not changed in the same direction, there exist tradeoffs. To better assist design to achieve optimum performances, we need to conduct quantitative analysis on these parameters and derive the corresponding trend. Figure 1(a) shows a device layout with single gate connection and multiple fingers. The gate resistance can be represented as

$$R_g = R_{acc} + K_{1,Rg} f_n + K_{2,Rg} \frac{R_{cont,p} + R_{via}}{f_n} + \frac{l_{end} + l_{ext}}{l_f f_n} R_{sq,p} + \frac{R_{sq,p}}{3} \frac{w_f}{l_f f_n} \qquad (1)$$

where $R_{cont,p}$, R_{via}, $R_{sq,p}$ are poly to metal1 per contact resistance, metal1 to metal 2 per via resistance and poly gate sheet resistance, and f_n, l_f and w_f are the device number of fingers, finger length and finger width, respectively. The first term R_{acc} in Eq. (1) accounts for the metal access resistance, the second term $K_{1,R_g} f_n$ represents the resistance of the metal parallel with poly and is proportional to the number of fingers f_n, the third term corresponds to the resistance involved with vias and contacts and inversely proportional to f_n. The fourth term represents the poly gate access resistance. l_{end} is the poly length from contact edge to poly extension edge and l_{ext} is the poly extension length. This resistance contribution is also inversely proportional tof_n. The fifth term corresponds to the poly resistance related to the channel, which is typically referred to the gate resistance. The number "3" in the denominator accounts for distributed poly gate resistance.

The source and drain resistance can be represented as

$$R_s = w_f \frac{K_{1,Rs}}{\left|\frac{f_n+2}{2}\right|_n} + \frac{R_{cont,d}}{\left|\frac{w_f}{S_{cn}}\right|_n \left|\frac{f_n+2}{2}\right|_n} + \frac{S_{tg}}{w_f f_n} R_{sq,d} \tag{2}$$

$$R_d = w_f \frac{K_{1,Rs}}{\left|\frac{f_n+1}{2}\right|_n} + \frac{R_{cont,d}}{\left|\frac{w_f}{S_{cn}}\right|_n \left|\frac{f_n+1}{2}\right|_n} + \frac{S_{tg}}{w_f f_n} R_{sq,d} + \frac{R_{via}}{\left|\frac{w_f}{S_{cn}}\right|_n \left|\frac{f_n+1}{2}\right|_n} \tag{3}$$

where $R_{cont,d}$, R_{via}, $R_{sq,d}$ are diffusion area to metal1 per contact resistance, metal 1 to metal 2 per via resistance and diffusion area sheet resistance, respectively. S_{cn} and S_{tg} are the diffusion contact pitch distance and the contact edge to gate distance, respectively, shown in Figure 1. The operation $|x|_n$ is to obtain the round integer number of the inside value x. The first terms in Eqs. (2) and (3) correspond to the metal access resistance, which are proportional to finger width and the number of source or drain regions. The second terms are the resistance related to the diffusion to metal1 contact, which are inversely proportional to w_f and f_n. The third terms are the resistance corresponding to diffusion area resistance, which is proportional to the product of $w_f f_n$. The fourth term in Eq. (3) represents the via resistance from metal1 to metal2 or to the used top metal of the drain. Since $w_f f_n$ is the total device width, this equation indicates that the source resistance is strongly correlated with the total device width, while demonstrates limited adjustable range when changing the number of fingers.

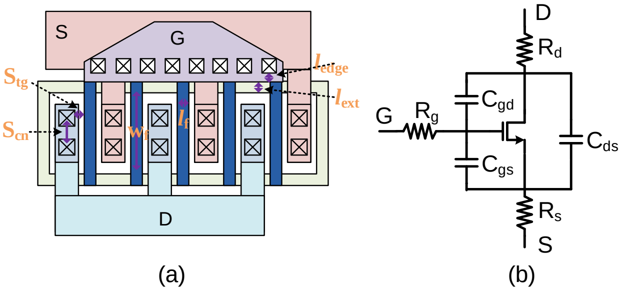
The extrinsic gate drain, gate source, and drain source capacitance can be represented as:

$$C_{gd,ext} = K_{1,Cgd} w_f f_n + K_{2,Cgd} l_f \left|\frac{f_n+1}{2}\right|_n \tag{4}$$

$$C_{gs,ext} = K_{1,Cgs} w_f f_n + K_{2,Cgs} l_f \left|\frac{f_n+2}{2}\right|_n \tag{5}$$

$$C_{ds} = K_{1,Cds} W \tag{6}$$

Both $C_{gd,ext}$ and $C_{gs,ext}$ have two similar terms, where the first term represents the coupling capacitance parallel with the channel and the second term stands for the coupling capacitance at the end of channels. $C_{ds,ext}$ is proportional to device total width W=$w_f f_n$. In Eqs. (3)~(8), all the K values are layout and process dependent, which will be derived through fitting. There are also other parameters such as capacitance between gate to bulk, drain to bulk, and source to bulk which are also related to device layout and should be included in the analysis for more accurate analysis.



(a)

(b)

Figure 1. (a) A layout representation of a single gate connection MOS device and (b) the corresponding scalable model with the annotated extrinsic parasitics to be determined by the proposed method

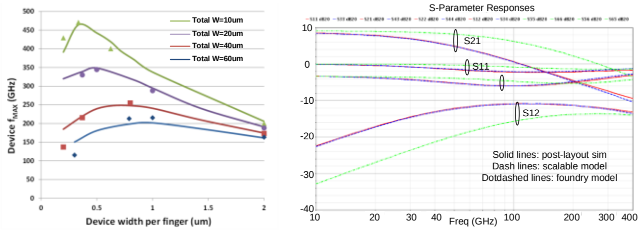
Given these parameters, we can build a scalable and layout-aware model, as shown in Figure 1(b), to assist circuit design. There are two ways to obtain these parameters. Approach one is through direct calculation according to the design manual. This approach is not very accurate at high frequencies due to skin effect and generates distributed parasitic network which is too complex for design insights; Approach two is to leverage post layout extraction and EM tools to derive parasitic values to form scalable models. The second approach is adopted in our design. To derive these variables, different device size layouts have been extracted to form simultaneous equations. After the parameters are extracted, the parameters are then input into the device parasitic components to form a more accurate model. Since this model is scalable with device size, finger number, it can be used similarly as pcells from foundries to assist circuit design. The parasitic inductances are omitted in the above equations assuming the device size is still significantly smaller than the operating signal wavelength.

The steps to extract device parasitics for scalable models are described as follows:

1. With a device layout, we eliminate all the layers except the extended poly layer and metal and via layers. This structure constitutes the extrinsic parasitics enclosing the core device.
2. This structure is then evaluated through EM simulation tools (ADS Momentum or HFSS) to obtain an N-port network S-parameter matrix.
3. To achieve the first order estimation of capacitance, we employ the similar derivative equations as Eqs. (1)~(6) in ref [9]. First order resistance estimation is based on Eqs. (1)~(3).
4. Starting from the first order estimation, we derive the accurate parasitic network by fitting with the EM simulated N-Port network matrix.
5. The steps from (1) to (4) are repeated for two other size devices so that three equations are formed for

each parasitic parameter. Given that, all the unknown fitting numbers (up to three) for each parasitic parameter can be derived.

After that, a scalable device model with an extrinsic parasitic parameter network is constructed. With these extra circuit parameters, optimum device choice can be performed by simulation. To validate the scalable device model, we compare its f_{MAX} with that of actual devices through post layout extraction and EM simulation characterization on the parasitics. Figure 2(a) presents the simulated device f_{MAX} from both the derived scalable model and the extracted post-layout circuit elements. Figure 2(b) shows a broad-band S-parameter simulation results of one device size, 20um/60nm, among the three cases: post-layout simulation results, the proposed scalable model simulation results and the core model from the foundry simulation results. It indicates that the proposed model results agree very well the post-layout extracted results, while the results from the core device model are quite different. This proves the effectiveness to use the scalable model in circuit design. Compared with the design procedure starting from inaccurate core device model then iterating the design and device option with post layout extraction and EM simulation tools, the derived scalable model provides more accurate circuit performance estimation during initial design stage, leads to the right optimization direction, and reduces design number-of-iterations significantly. Figure 2(a) also presents the trend that a smaller device prefers a smaller finger width to achieve a larger f_{MAX}. This trend is consistent with the analysis from Eqs. (1)~(6). Intuitively, multiple finger structure is employed to reduce the poly gate resistance with the price of increased parasitic capacitance. Therefore, an optimum finger number exists. When the device is small, to maintain the optimum finger number leads to a small finger width. When the device is large, optimum finger width becomes larger. In addition, the f_{MAX} versus finger width slope becomes flatter when the device total width increases.



(a) (b)

Figure 2. (a) Device f_{MAX} based on the created scalable model (lines) and direct extraction value from layout (symbols) (b) S-Parameter simulation results comparison among post-layout extraction based device (solid lines), scalable model (dash lines), and the core model (dotdashed lines) from the foundry

III. Fabrication and Measurement Results

A. Terahertz VCO

Using this methodology, we have designed an oscillator with the fundamental frequency higher than device cut-off frequencies [10]. The oscillator schematic is shown in Figure 3. The oscillator consists of a primary tank with L_{tank} shunt between the drains of the bottom cross coupled pair (blue dash circle) and a parallel frequency selective negative resistance (FSNR) tank with L_g shunt between cascode devices (pink solid circle). The FSNR is to provide an equivalent inductor together with a negative impedance. This unique feature occurs at high frequencies and matches with our high operating frequency requirements. Therefore, the overall oscillation frequency is higher than each individual tank oscillation frequency and the overall tank impedance is also boosted by the FSNR's negative resistance. This architecture not only allows large L_{tank} and L_g inductances to facilitate on-chip design, but also combines them via cascode circuit to form a hybrid tank with a low overall inductance for terahertz operations. The boosted overall tank impedance eases the demand on the cross coupled device's g_m to permit smaller devices for further higher oscillation frequencies. In addition, the added FSNR tank vertically shares the same current with the conventional tank and thus does not consume extra power. With the added FSNR, the oscillator can generate a fundamental oscillation frequency higher than device f_T.

To further push the operation frequency into THz range, push-pull structure is adopted to generate the 2nd order harmonic signal at the drain common mode output [11,12]. The second order harmonic signal from the output will feed the on-chip patch antenna to radiate over-the-air.

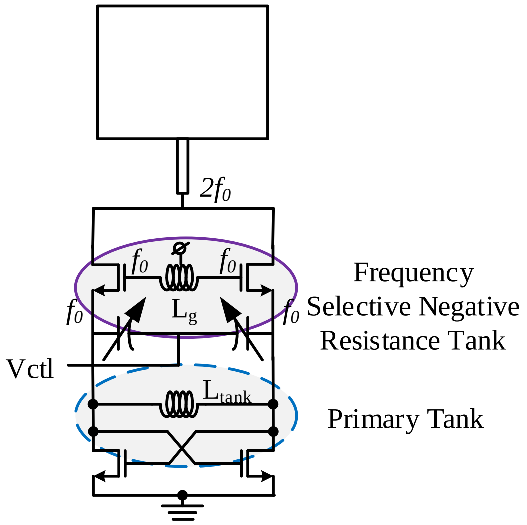


Figure 3. The proposed 450 GHz push-pull VCO with FSNR tank, (a) schematic, (b) the die photo in 65 nm CMOS

B. High Frequency Prescaler

Another key and challenging circuit block is the prescaler after the oscillator, which needs to achieve both high speed and wide bandwidth, which are normally competing parameters. To alleviate the tradeoffs, we have proposed time-interleaved injection locking based circuit design technique as shown in Figure 4 [13]. The input signal is injected to both the top voltage mixing device and the bottom current source device to attain extended injection angles, which leads to a higher oscillation frequency and a wider locking range simultaneously.

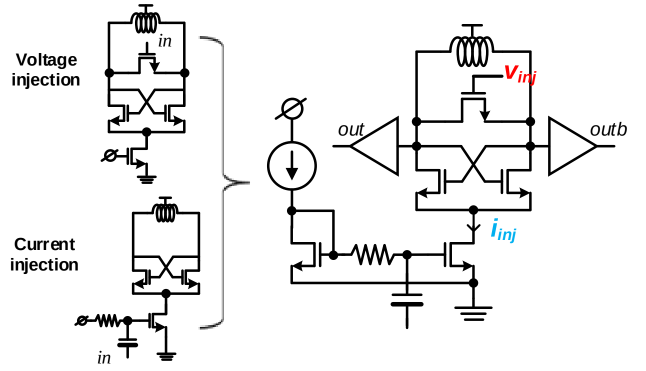


Figure 4. Proposed time-interleaved injection locking scheme based prescaler topology to boost locking range

Voltage and current injection methods, traditionally exclusive from each other, are integrated by this time-interleaved scheme. Through working at different time periods, these two types of injection are complement to each other to boost injection efficiency, as shown in Figure 6. For voltage injection, the input signal V_{in} is injected at the gate of the NMOS mixer that shunts outputs of the cross coupled pair. As the injection voltage increases and the V_{gs} starts to exceed the device threshold, the mixer turns on and introduces a low impedance path to pull its source and drain (or the cross coupled outputs) voltages closer. As a result, when voltage injection occurs at an instance outside of the output voltage crossing time period of the prescaler, the voltage injection tends to pull output voltage toward each other. If the prescaler's natural oscillation frequency is close to half of the injection frequency, such an effect will ultimately align the prescaler's output frequency and its phase with the voltage injection signal, as shown by the orange area in Figure 5 with the defined injection angle of θ_1. On the other hand, a current signal I_{inj} is injected via the current source of the cross coupled pair. During the positive (or negative) current injection cycle, the increased (or decreased) source current would split unequally to the resonant tank and increase (or decrease) the voltage difference between prescaler outputs. When the prescaler's natural oscillation frequency is close to half of the current injection frequency, the prescaler's output maximum (or minimum) points is synchronized with effective current injection time zones, represented by the current injection angle θ_2, blue area in Figure 5. With this proposed time-interleaved dual-injection locking scheme, the overall injection strength is enhanced by two means. First is the added injection strength due to both voltage and current injections. Second, and more importantly, the interleaving injection renders smaller current during voltage injection period that is equivalent to lower the oscillator current I_{osc}, thus increasing the I_{inj} / I_{osc} ratio for an extended locking range.

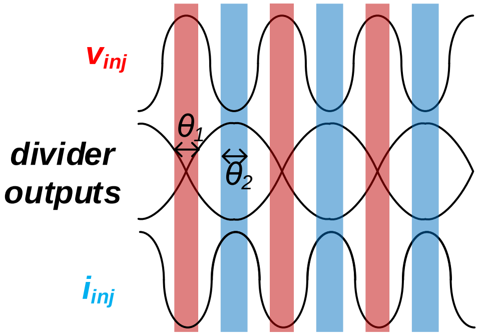


Figure 5. The illustration of boosted injection angles of the time-interleaved injection locking frequency divider, which shows the timing relationship between divider voltage output and injection voltage and current

IV. Measurement Results

Two measurement approaches are used to characterize the VCO: electronic methods and quasi-optic methods. Electronic methods, as shown in Figure 6(a), are used to measure the fundamental oscillation frequency and the tuning range. The oscillator output, mixed with a high order harmonic of an external LO, is down-converted to an IF, which is then fed into a low noise amplifier, corresponding to the IF signal shifting frequency, the used LO harmonic order for down-conversion can be derived, which leads to the measured VCO output frequency [10]. Figure 6(b) presents the VCO tuning range, whose fundamental oscillation frequency is from 225 GHz to 226.7 GHz with 1.7 GHz tuning range, by using a small size varactor of 0.74 um × 0.06 um. This VCO draws 5mA current from 1.4V power supply. The chip photo shown in Figure 6(c).

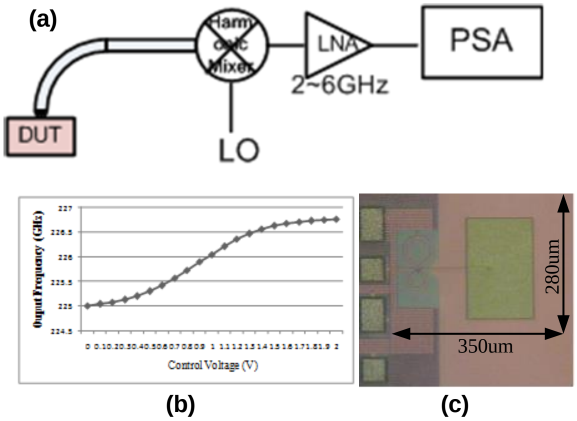


Figure 6. a) The electronic measurement approach testing setup, (b) the measured tuning frequency from 225 GHz to 226.7 GHz, and c) the chip photo

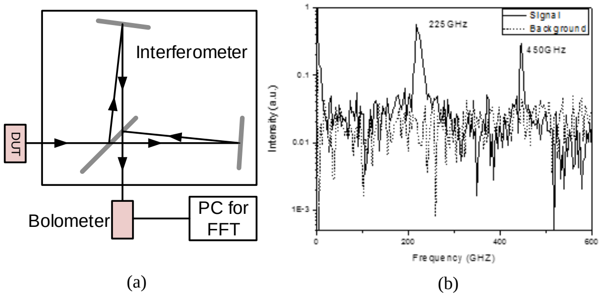


Figure 7. (a) Interferometer-based quasi-optical approaches to measure the 2nd order harmonic output, (b) the measured output spectrum with background noise represented by dotted line

Interferometer-based quasi-optical approaches are adopted for the 2nd order harmonic signal measurement. As shown in Figure 7(a), the signal, radiating from the vertically mounted VCO's on-chip antenna, passes through the interferometer and then is detected by a bolometer. The output signal spectrum, recovered through FFT, is shown in Figure 7(b). The measured spectrum represents un-calibrated power, of which the 2nd order harmonic has a large attenuation from setup and oxygen absorption than the fundamental frequency does. Therefore, the actual output power from the second order harmonic should be large. The fundamental frequency signal radiation may be from the on-chip inductors, which are essentially loop antennas.

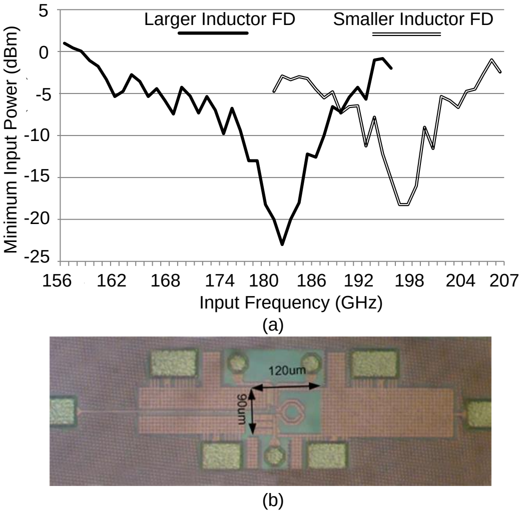


Figure 8. (a) Measured input sensitivities of the two prescalers, (b) die photo of the proposed CMOS prescaler in 65 nm CMOS

Two prescalers with different inductor values (about 120pH and 150pH, respectively) are implemented in 65 nm CMOS technology. The measured input sensitivities of both prescalers are elucidated by drawing the minimum input power versus the input frequency, shown in Figure 8(a). The demonstrated locking ranges are over 37GHz (158GHz~195GHz, or 21%) with < 0dBm input power and 27GHz (181GHz~208GHz, or 14%) with < -1dBm input power, respectively. A chip photo is shown in Figure 8(b) with the core chip area 0.12mm x 0.09mm and the power consumption of 2.4 mW.

V. CONCLUSIONS

This paper presents a systematic active device model and layout optimization approach to guide mm-wave/sub-mm-wave circuit design in CMOS technologies. Specifically, the layout-aware active scalable model assists more accurate design optimization and reduces the number of design iterations between circuit optimization and physical layout. Layout-aware model also facilitates device layout optimization for different circuit blocks. The proposed active device optimization approach is validated by several key mm-wave/sub-mm-wave building blocks in 65 nm CMOS technologies.

ACKNOWLEDGMENT

The authors would like to thank NSF funding support.

REFERENCES

[1] Eli Yablonovitch, SRC Workshop, Asheville, 2005.

[2] U. Singh, A. Garg, B. Raghavan, N. Huang, H. Zhang, Z. Huang, A. Momtaz, J. Cao, “A 780mW 4×28Gb/s Transceiver for 100GbE Gearbox PHY in 40nm CMOS,” ISSCC Dig. Tech. Papers, pp.40-41, Feb. 2014

[3] J. W. Jung, B. Razavi, “A 25Gb/s 5.8mW CMOS Equalizer,” ISSCC Dig. Tech. Papers, pp.44-45, Feb. 2014

[4] U. R. Pfeiffer, Y. Zhao, J. Grzyb, R. A. Hadi, N. Sarmah, W. Förster, H. Rücker, B. Heinemann, “A 0.53THz Reconfigurable Source Array with up to 1mW Radiated Power for Terahertz Imaging Applications in 0.13μm SiGe BiCMOS,” ISSCC Dig. Tech. Papers, pp.256-257, Feb. 2014

[5] Yahya Tousi, Ehsan Afshari, “A Scalable THz 2D Phased Array with +17dBm of EIRP at 338GHz in 65nm Bulk CMOS,” ISSCC Dig. Tech. Papers, pp.258-259, Feb. 2014

[6] P.-Y. Chiang, Z. Wang, O. Momeni, P. Heydari, “A 300GHz Frequency Synthesizer with 7.9% Locking Range in 90nm SiGe BiCMOS,” ISSCC Dig. Tech. Papers, pp.260-261, Feb. 2014

[7] A. Tang, Q. J. Gu, Z. Xu, G. Virbila and M.-C. F. Chang, “A 349 GHz 18.2mW/Pixel CMOS Inter-modulated Regenerative Receiver for Tri-Color mm-Wave Imaging,” 2012 IEEE MTT-S International Microwave Symposium, June 2012

[8] P. H. Woerlee, M. J. Knitel, R. van Langevelde, D. B. M. Klaassen, L. F. Tiemeijer, A. J. Scholten, and A. T. A. Zegers-van Duijnhoven, “RFCMOS performance trends,” IEEE Trans. Electron Devices, vol. 48, no. 8, pp. 1776–1782, Aug. 2001

[9] C. K. Liang and B. Razavi, “Systematic transistor and inductor modeling for millimeter-wave design,” IEEE J. Solid-State Circuits, vol.44, no. 2, pp. 450–457, Feb. 2009

[10] Q. J. Gu, Z. Xu, H.-Y. Jian, X. Xu, F. Chang, W. Liu and H. Fetterman, “Generating Terahertz Signals in 65nm CMOS with Negative-Resistance Resonator Boosting and Selective Harmonic Suppression,” IEEE Symposium on VLSI Circuit, pp. 109-110, June 2010

[11] D. Huang, T.R. LaRocca, L. Samoska, A. Fung, M.-C. Frank Chang, “324GHz CMOS Frequency Generator Using Linear Superposition Technique,” ISSCC Dig. Tech. Papers, pp.476-477, Feb. 2008

[12] E. Seok, C. Cao, D. Shim, D. J. Arenas, et. al., “A 410GHz CMOS Push-Push Oscillator with an On- Chip Patch Antenna,” ISSCC Dig. Tech. Papers, pp.472-473, Feb. 2008

[13] Q. J. Gu, H.-Y. Jian, Z. Xu, Y.-C. Wu, F. Chang, Y. Baeyens, and Y.-K. Chen, “CMOS Prescaler(s) with Maximum 208GHz Dividing Speed and 37GHz Time-Interleaved Dual-Injection Locking Range,” IEEE Transactions on Circuits and Systems-II, vol.58, no.7, pp. 393-397, July 2011