Optimization of Unreleased CMOS-MEMS RBTs

Bichoy Bahr
EECS Department
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: bichoy@mit.edu

Luca Daniel
EECS Department
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Dana Weinstein
School of Electrical and
Computer Engineering
Purude University
West Lafayette, Indiana 47907-2035

Abstract*—In this paper, we present an efficient framework for optimization of MEMS resonators based on model order reduction and memoization to significantly speed-up computations ($\sim 40\times$). Owing to their technological importance and numerous applications, unreleased CMOS resonant body transistors (RBTs) are considered. Their intricate structure requires computationally intensive finite element method (FEM) frequency domain simulations, which hinders their optimization. In this work, numerical optimization combined with a physics-based phononic crystal (PnC) waveguide design enables the realization of unreleased CMOS-RBTs with record breaking performance. The optimized RBTs have been fabricated in IBM 32nm SOI technology, demonstrating a quality factor $Q \sim 11{,}620$ at 3.252 GHz for an $f_\circ \cdot Q \sim 3.8 \times 10^{13}$.

I. Introduction

Unreleased monolithic CMOS Resonant Body Transistors (RBTs) with high quality factors Q and miniature footprint, offer a potential solution for RF frequency synthesizers, filters and timing applications [1]. The true solid-state nature and the lack of a *release* step common to traditional MEMS devices, eliminate any post-processing or packaging requirement for the resulting CMOS die. This reduces the overall fabrication cost and enhances the process yield. More importantly, the CMOS circuits are not affected in this integration scheme. The intimate monolithic integration in the front-end-of-line (FEOL) layers of the CMOS die results in minimal parasitics, a highly desirable advantage for GHz-frequencies circuits. It also allows for large arrays of RBTs and low-power coupled GHz-frequencies oscillators. Monolithically integrated coupled oscillators are suitable for a myriad of applications from fast unconventional signal processing to multi-phase signal synthesis.

For this reason, maximizing CMOS-RBT performance by numerical optimization is highly desirable. Such devices incorporate diverse materials and interfaces, making finite element method (FEM) the option of choice for the simulation thereof. Perfectly matched layers (PMLs) are needed to model acoustic radiation losses, allowing plane waves (radiation modes) to propagate outside the resonance cavity at all frequencies. FEM eigenmode analysis becomes ineffective as it requires numerous eigenmodes to isolate the resonance mode from the radiation modes. Small signal FEM frequency domain analysis overcomes this problem.

Some applications (such as timing for digital circuits) require maximizing Q for better jitter performance with relaxed operating frequency specifications. As device dimensions are varied during optimization, the resonance frequency can change significantly. This represents a challenge for frequency domain analysis: a wide frequency range should be considered with sufficient resolution to capture high-Q resonances. Simulation times become prohibitively long for practical optimization purposes.

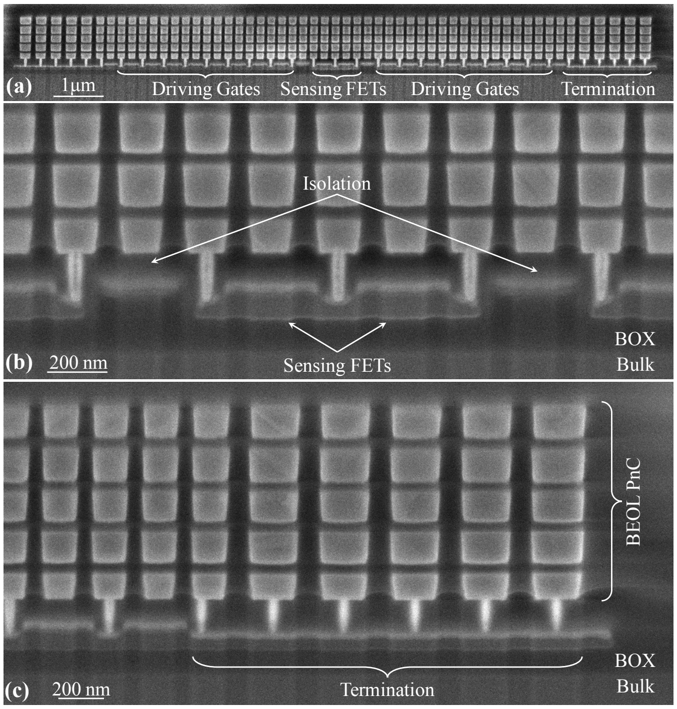

Fig. 1. SEM micrograph showing (a) a cross-section of the unreleased RBT in IBM 32nm SOI technology, (b) zoom-in on sense FETs and isolation gates and (c) the termination waveguide.

In this paper, we present an efficient solution for this problem by using model order reduction (MOR) to significantly speed-up frequency domain simulations. Moreover, memoization is used to store all previous simulation results, allowing the prediction of resonance frequencies and hence reducing the simulation band. Also, gradients are evaluated separately by finite differences over very narrow frequency bands for even more speed-up. This presented framework is generic and can be extended to other MEMS devices as well.

In this work, numerical optimization is combined with physics-based phononic crystal (PnC) waveguide design at the surface of the CMOS die [2], enabling the realization of RBTs with record breaking performance. Optimized RBTs have been fabricated in IBM 32nm SOI technology and RF characterization verifies the benefits of numerical optimization.

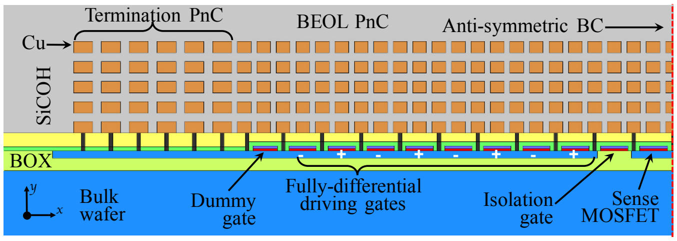

Fig. 2. Half-space cross section of FEM model structure for unreleased CMOS-RBT in IBM 32nm SOI technology.

II. CMOS-RBT Device Structure

Fig. 1 shows an SEM micrograph for the cross section of a CMOS-RBT in IBM 32nm SOI. The FEM model structure is shown in Fig. 2, highlighting the different components of the RBT. The device relies on a back-end-of-line (BEOL) PnC together with fully-differential driving to create a horizontal waveguide at the surface of the bulk CMOS wafer [2]. This waveguide confines the elastic vibration energy vertically to the front-end-of-line (FEOL) layers. In order to achieve horizontal confinement, waveguide sections with mismatched characteristics are used at each side as a termination. Two MOSFETs are used for active FET sensing in the middle of the structure, while eight MOSCAPs are used for fully-differential actuation on each side. Shallow trench isolation separates the driving MOSCAPS from the sensing FETs to reduce electrical feed-through. The device parameters are the Cu metal width and separation of both the main cavity waveguide and termination waveguide, the width of the isolation, sensing and dummy gates.

The termination gates, isolation gates and sensing transistors can all be considered as *perturbations* to the main cavity waveguide, which may induce *scattering*. Scattering to radiation modes represents energy losses, which reduces the quality factor of the RBT. Also, scattering to confined modes at different frequencies ultimately results in spurious modes. For these reasons, it is essential to meticulously select the dimensions of the different waveguide sections to match their characteristics and minimize scattering, while allowing only for *specular* reflection from the termination. The complexity of the structure calls for numerical optimization and prohibits design by phenomenological intuition.

III. Optimization Framework

A. Problem Formulation

The optimization problem can be formulated as

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{c}(\boldsymbol{x}) \le 0 \\ & 0 \le x_i \le 1, \end{array} \tag{1}$$

where $\boldsymbol{x}$ is the resonator N-dimensional normalized parameters vector, $f(\boldsymbol{x})$ is the objective function and $\boldsymbol{c}(\boldsymbol{x})$ is a non-linear function representing the CMOS design rule check (DRC) as well as electromigration constraints. One can choose to maximize the resonator quality factor with $f(\boldsymbol{x}) = -Q(\boldsymbol{x})$, favoring the best energy confinement. Maximizing the electromechanical transconductance can be another objective function choice with $f(\boldsymbol{x}) = -|g_{em}|$, favoring the highest stresses at the transducers. In general, the two designs can be different, a clear distinction to traditional MEMS resonators designs, where the position of the transducers is necessarily optimal by design. It is important to note that there is no need to explicitly include in (1) the matching of the different waveguide sections described in the previous section. This matching requirement is implicitly considered in the abovementioned objective functions and have to be satisfied for an optimal solution. The resulting optimization problem is in general non-convex.

A challenging aspect of unreleased CMOS-RBTs design is compliance with the CMOS DRC constraints, as imposed by the foundry to guarantee successful fabrication with sufficient yield. These constraints include rules about the allowable separations, metal widths, gate lengths, filling densities, etc… DRC constraints are often discontinuous, hence included in (1) as the non-linear function $\boldsymbol{c}(\boldsymbol{x})$. Consider the designed separation between metal lines on the i^{th} layer to be s_i and the DRC required value to be $s_i^{\text{DRC}}(w_i)$, where w_i is the width of the metal line. The corresponding constraints element $c_i(\boldsymbol{x})$ is given by

$$c_i(\boldsymbol{x}) = -\left(s_i - s_i^{\text{DRC}}(w_i)\right) \;,\; c_i(\boldsymbol{x}) \le 0. \tag{2}$$

With this formulation, the actual value of $c_i(\boldsymbol{x})$ is proportional to DRC constraints violations. This will help most optimization algorithms to efficiently pick the next design point, as $\boldsymbol{c}(\boldsymbol{x})$ appears directly in the Lagrangian and the Karush-Kuhn-Tucker (KKT) conditions as $\Lambda^T \cdot \boldsymbol{c}(\boldsymbol{x})$, where Λ is a vector of the Lagrange multipliers λ_i [3], [4]. The resulting Lagrange multipliers λ_i from the optimization are very useful from a CMOS process design point of view. If certain DRC constraints are *tightly* satisfied ($c_i(\boldsymbol{x}) = 0$), the corresponding Lagrange multipliers will be non-zero and indicate which constraint has the most effect on the objective function. Such detailed information can be used by process engineers who may consider relaxing specific DRC rules or optimizing the process differently for better resonator performance.

B. Objective Function: Speed-up by MOR

During the course of numerical optimization, the objective function $f(\boldsymbol{x})$ is evaluated hundreds or even thousands times, depending on the number of parameters. Each objective function evaluation involves a full, frequency domain FEM simulation for the entire RBT structure, over the frequency band $B_\circ$. COMSOL Multiphysics with solid mechanics module is used for these simulations [5]. As demonstrated in the introduction, the large bandwidth and high resolution requirement result in prohibitively long simulations. Efficient objective function evaluation is necessary for practical optimization run-times.

The FEM solution includes the full displacement and stress fields at each point in the structure, for every frequency in the band $B_\circ$. However, the objective function is only concerned about the quality factor, or the peak electromechanical

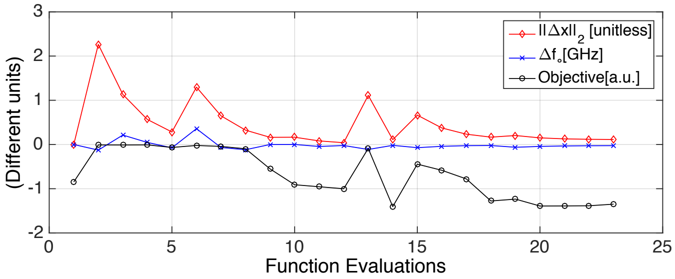

Fig. 3. Euclidean distance ($||\Delta x||_2$) and resonance frequency change ($\Delta f_\circ$) from starting design for different function evaluations.

transconductance g_{em}. Both metrics are calculated from the average stress in the sensing FET channel. This makes the problem at hand an ideal candidate for model order reduction (MOR). A simple rational transfer function approximation with few complex-conjugate pole pairs is sufficient to evaluate $f(\boldsymbol{x})$, without the need for the full FEM solution. Various MOR techniques are available to obtain such approximation. COMSOL Multiphysics readily includes MOR that can be incorporated by enabling the asymptotic waveform evaluation (AWE) feature in COMSOL solver [5].

COMSOL AWE implementation calculates low-order Padé approximation on small frequency intervals for a given output (the average FET channel stress in this case). The Padé approximation of type p/q for the frequency response $H(\omega_\circ+\sigma)$ around frequency $\omega_\circ$ is given by

$$H_{p,q}(\omega_\circ+\sigma) = \frac{b_p\sigma^p + \cdots + b_1\sigma + b_\circ}{a_q\sigma^q + \cdots + a_a\sigma + a_\circ}. \qquad (3)$$

Its Taylor series about $\sigma = 0$ matches that of $H(\omega_\circ+\sigma)$ at least for the first $p+q+1$ terms. The q^{th} order Padé approximation $H_q(\omega_\circ+\sigma)$ (with $p = q-1$) is found by calculating the leading $2q$ moments of $H(\omega_\circ+\sigma)$ as described in [6].

AWE starts with a given frequency interval, where a Padé approximation is calculated both at the start and end of the interval. Next, the result of both approximations is evaluated and compared at several points in that frequency interval. If they match within a certain tolerance, the interval is accepted; otherwise, the interval is bisected and the process repeats.

This technique is very efficient with the wideband sweeps under consideration. Large intervals can be used away from resonance peaks allowing for very fast simulations. When a resonance is encountered, smaller intervals are progressively considered till the simulation tolerance is met. Higher order Padé approximation allows for larger intervals, however, they are not useful beyond $q > 8$, due to moments matrices ill-conditioning [6]. Padé approximations with $q = 5$ were found to provide the most useful speed-up. A typical speed-up of $8\times$ was observed when employing AWE for objective function evaluation.

C. Objective Function: Speed-up by Memoization

The optimization algorithm may often suggest design points $\boldsymbol{x}$ that are close in the parameters space. Since the elastic wave equation is linear under coordinate scaling [1], the resonance frequency $f_\circ$ is bound to change linearly (likely sublinear) with design parameters (given that they represent physical dimensions). For close enough design points, reliable upper bounds for changes in $f_\circ$ can be estimated, eliminating the need for FEM simulation over the entire band $B_\circ$. This becomes more frequent as the optimization approaches convergence (Fig. 3).

This a priori knowledge can be leveraged through *memoization* to speed-up objective function evaluation. The resonance frequency among other metrics are saved as a *side effect* of each objective function evaluation. Whenever $f(\boldsymbol{x})$ is requested for a new design point, all previous design points in the parameters space are searched for the nearest one. If the latter falls within a maximum Euclidean distance, this point is used to estimate an upper bound on $f_\circ$ change. A limited frequency sweep over a bandwidth $B_i << B_\circ$ is used in this case for a significantly faster FEM simulation. Major speed-ups are achieved by memoization on the order of $B_\circ/B_i$. Practically, only few points of the design space end up being evaluated over the full frequency band $B_\circ$.

D. Gradient Evaluation Speed-up

Gradient evaluation speed-up is crucial for optimization algorithms that require gradients. The i^{th} component of the gradient is evaluated with finite difference as:

$$(\nabla f)_i = \frac{f(\boldsymbol{x}+\epsilon\,\hat{\boldsymbol{e}}_i) - f(\boldsymbol{x})}{\epsilon}, \quad \forall i \in [1,N], \qquad (4)$$

where $\hat{\boldsymbol{e}}_i$ is the unit vector in the design parameter space along the i^{th} parameter. The resonance frequency $f_\circ$ can at most change by $\pm\epsilon$, with parameters representing physical dimensions [1]. Thus, it is sufficient to consider a frequency band of $2\,\epsilon\,f_\circ + 2\,BW$, with $BW = f_\circ/Q$ for gradient estimation. Choosing $\epsilon = 0.5\%$, the gradient simulation bandwidth is usually much smaller than the full problem frequency band $B_\circ$. The speed-up in gradient evaluation compared to a naïve finite differencing is $B_\circ/(2\,\epsilon\,f_\circ + 2\,BW) \approx B_\circ/2\,\epsilon\,f_\circ$.

E. Full Optimization Flow

The full optimization flow is shown in Fig. 4. The optimization algorithm is implemented in MATLAB and calls COMSOL Multiphysics for FEM simulations. MATLAB's constrained optimization function `fmincon` with interior-point algorithm was used for this problem [3]. Objective evaluation starts by selecting the simulation bandwidth B_i, based on the memoization search. FEM simulation with Padé approximation is carried out in COMSOL. Rational transfer function fitting is performed on the solution for accurate extraction of the quality factor and peak transconductance value. In case of multiple resonance peaks appearing in the objective function, the one with the best objective value is selected. An optimization geared towards suppressing spurious modes may strongly penalize such designs. Next, the memoization state is saved and the objective value is returned. Gradient evaluation proceeds with finite difference as described above.

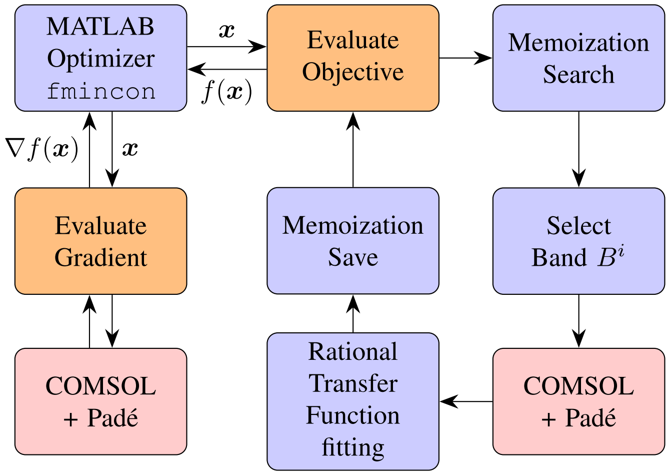

Fig. 4. A flowchart demonstrating the implemented optimization flow, highlighting COMSOL FEM simulation with Padé approximation, along with memoization and finite difference gradient evaluation. 40× optimization runtime improvement is achieved.

IV. Results

Starting from a close enough initial guess, the optimization converges achieving double the starting Q, with 21 function and gradient evaluations. Optimization run-time is 4 hours, marking a $5\times$ improvement over naïve gradient evaluation and $40\times$ improvement if not using our framework. Arbitrary initial guesses will require much longer time, hence the benefit of the presented technique. Artificial losses were introduced in the FEM simulation to limit the quality factor, which simplifies finding the resonance peak.

The optimum design uses PnC copper metal width and separation of 154 nm and 66 nm for the main cavity waveguide, respectively. The length of the driving gates is inferred from these dimensions. Termination waveguide PnC metal width and separation are 214 nm and 98 nm, respectively. The sensing and isolation gate lengths are 314 nm and 312 nm, respectively; which sets the corresponding PnC periods. Simulated stresses and mode shape for the optimized design are shown in Fig. 5. The mode shape clearly shows perfect vertical and horizontal confinement, with apparent lack of scattering. The device occupies an area of $13\,\mu$m $\times$ $4\,\mu$m.

The optimized RBTs have been fabricated in IBM 32nm SOI. Full 4-port RF characterization has been performed using an Agilent N5225A PNA, at room temperature in air. The PNA couplers were reversed for a 15dB improvement in sensitivity, with -5dBm testing power. Sensing FETs were biased in linear regime with $110\,\mu$A each, in order to maximize the channel mobility sensitivity to the mechanical stress. Electromechanical transconductance was extracted from the Y-parameters as $g_{em} = i_{out}/v_{in} = Y_{21} - Y_{12}$, according to transistors π-model. The device response with 1 V driving bias was de-embedded using 0 V driving bias as the open structure. The measured g_{em} is shown in Fig. 6, with a FWHM $Q \sim 11{,}620$ at 3.252 GHz. This marks a $46\times$ improvement over the RBTs of [1] and a record breaking $f_\circ \cdot Q \sim 3.8 \times 10^{13}$.

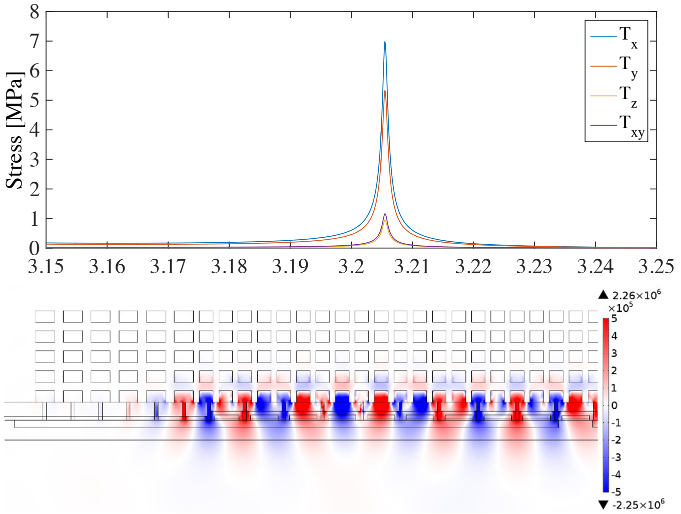

Fig. 5. (a) FEM simulation results for optimized RBT structure (b) Left-half of RBT structure showing T_{yy}-stress at resonance.

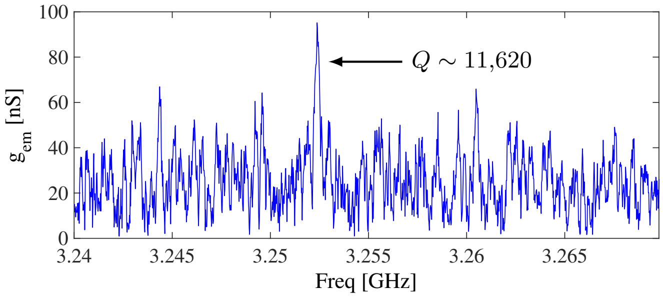

Fig. 6. Measured frequency response of the optimized IBM 32nm SOI RBT, achieving $f_\circ \cdot Q \sim 3.8 \times 10^{13}$ at 3.252 GHz. Resonance frequency is in agreement with simulation within 1.5%.

V. Conclusion

Efficient numerical optimization technique for unreleased CMOS RBTs has been presented. With large speed-ups in objective function and gradient evaluation, optimization of full MEMS resonators relying on FEM frequency domain analysis becomes practical and feasible. Fabricated resonators demonstrated record breaking Q and $f_\circ \cdot Q$ for unreleased CMOS-RBTs, highlighting the benefits of numerical optimization for such complicated device structures. The presented framework is easily extensible to other MEMS devices that requires frequency domain FEM simulation.

References

[1] B. Bahr, R. Marathe, and D. Weinstein, "Theory and design of phononic crystals for unreleased CMOS-MEMS resonant body transistors," *Microelectromechanical Systems, Journal of*, vol. PP, no. 99, pp. 1–1, 2015.

[2] B. Bahr and D. Weinstein, "Vertical Acoustic Confinement for High-Q Fully-Differential CMOS-RBTs," in *Solid-State Sensors, Actuators and Microsystems Workshop (Hilton Head)*, 2016.

[3] MATLAB, *version R2014a*. The MathWorks Inc.

[4] H. R. Byrd, C. J. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming*, vol. 89, no. 1, pp. 149–185, 2000.

[5] COMSOL, Inc.: COMSOL Multiphysics® - http://www.comsol.com.

[6] L. Pillage and R. Rohrer, "Asymptotic waveform evaluation for timing analysis," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 9, no. 4, pp. 352–366, Apr 1990.