Large-scale integration of nanoelectromechanical systems for gas sensing applications

I. Bargatin[1,2†], E.B. Myers[1], J.S. Aldridge[1‡], C. Marcoux[2], P. Brianceau[2], L. Duraffourg[2],
E. Colinet[2], S. Hentz[2], P. Andreucci[2], M.L. Roukes[1]

[1]Kavli Nanoscience Institute and Department of Physics, Caltech, Pasadena, CA
[2]CEA/LETI - MINATEC, Grenoble, France

Fabrication Procedure

The NEMS arrays employed in this work were fabricated from CMOS-compatible materials using state-of-the-art microelectronic lithography and etching techniques with nanoscale alignment. The high-frequency (HF) NEMS arrays were fabricated from a 200-mm SOI wafer with 160-nm-thick silicon layer (resistivity ≈ 10 Ω.cm) and 400-nm-thick buried oxide layer. The metal film — a proprietary alloy fully compatible with CMOS — was deposited by sputtering technique at 175°C. Its thickness varied between 45 and 70 nm depending on the design. Optical deep ultraviolet (248 nm wavelength) lithography was then used to pattern the thin-film metal features: wirebonding pads, lead-frame, and the NEMS array itself. We were able to achieve a better than 200-nm resolution in a reproducible way using a positive resist and a bottom anti-reflective coating (BARC).

The exposed areas of the metal film were etched using reactive ion etching (RIE) in boron trichloride (BCl_3) and argon (Ar) plasma. The resulting metallization layer served as a mask for the subsequent CF_4 plasma etching of the 160-nm-thick silicon structural layer down to the buried oxide. In some designs, additional lithography steps were performed to define bare-silicon (metallization-free) areas on beams or cantilevers before the final silicon etching. In this case, the accuracy of alignment between the lithography levels was better than 30 nm (Fig. S1). In some designs, the metal layer on the bonding pads and lead-frames was thickened to 650 nm to facilitate the wirebonding procedure, decrease the access resistance, and improve the impedance matching.

Finally, the NEMS cantilevers or beams were suspended using a vapor HF etch step that was carefully timed to minimize the undercut of the anchors. The arrays were typically etched

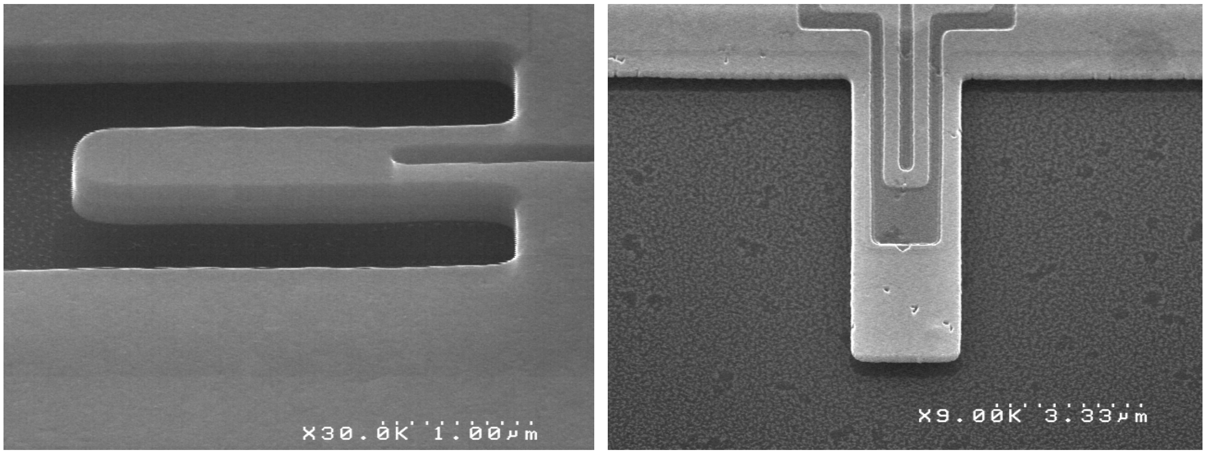

Fig. S1. Scanning electron micrographs of fully released cantilevers with one or two metal loops. The two-loop type of cantilever were not used in the array configuration, but its image illustrates the accuracy of the lithography process. Note that the accuracy of mask alignment (overlay) was better than 30 nm.

for 6 hours at 32 °C in HF vapor concentration of approximately 10%, resulting in silicon dioxide etching rate of 1.2 nm/min. The vapor HF did not significantly attack the metal layer, with the etch rate being only 1nm/hour.

With this process, we were able to produce the first 200-mm wafers of NEMS VLSI, each containing more than 3.5 million NEMS. The arrays described in the main text contained 2 800 NEMS cantilevers and occupied an area 0.14 mm by 1.0 mm. Other arrays we fabricated contained up to 6800 resonant NEMS cantilevers on area of 0.2 mm by 0.6 mm, achieving an integration density of approximately 60 000 NEMS/mm^2, and a functional device yield of approximately 95%.

In future designs and process runs, advanced lithography techniques can be employed to decrease the minimum feature size of the arrays even further. Modern 193-nm DUV dry lithography can potentially achieve resolution of 70-80 nm on 200 mm wafers with 15 nm overlay. Hybrid lithography that combines e-beam and DUV processes can achieve 50-nm minimum feature size while keeping the total lithography process reasonably fast for 200-mm wafers. In this process, the relatively slow e-beam is used to define the smallest features, such as 50-nm-wide metallic lines, while the fast DUV lithography defines all larger features. Variable Shape Beam e-beam lithography has been demonstrated to achieve 35-nm half-pitch size and overlay of between 15 and 7 nm depending on the field size. The 3.5 million NEMS of our current wafers could be written in just 3 hours (simulated writing time). With these advanced lithography techniques, one could achieve integration density exceeding 100.000 NEMS/mm^2.

Balanced two-port measurement scheme for thermoelastic actuation and piezoresistive detection

The combination of thermoelastic actuation and piezoresistive downmixing described in Ref. 1 uses two separate metal loops for actuation and detection—a total of four measurement ports. The same combination of thermoelastic actuation and piezoresistive downmixing can be used when only one loop is available, as for example in the case of a simple two-legged cantilever or an array of such cantilevers. Figure S2 shows the schematic of the resulting two-port measurement setup. The drive voltage oscillating at frequency $\omega_d/2$ creates temperature variations at frequency ω_d, which induces cantilever motion. The bias voltage frequency ω_b is offset from the drive frequency, $\omega_b = \omega_d - \Delta\omega$, typically by less

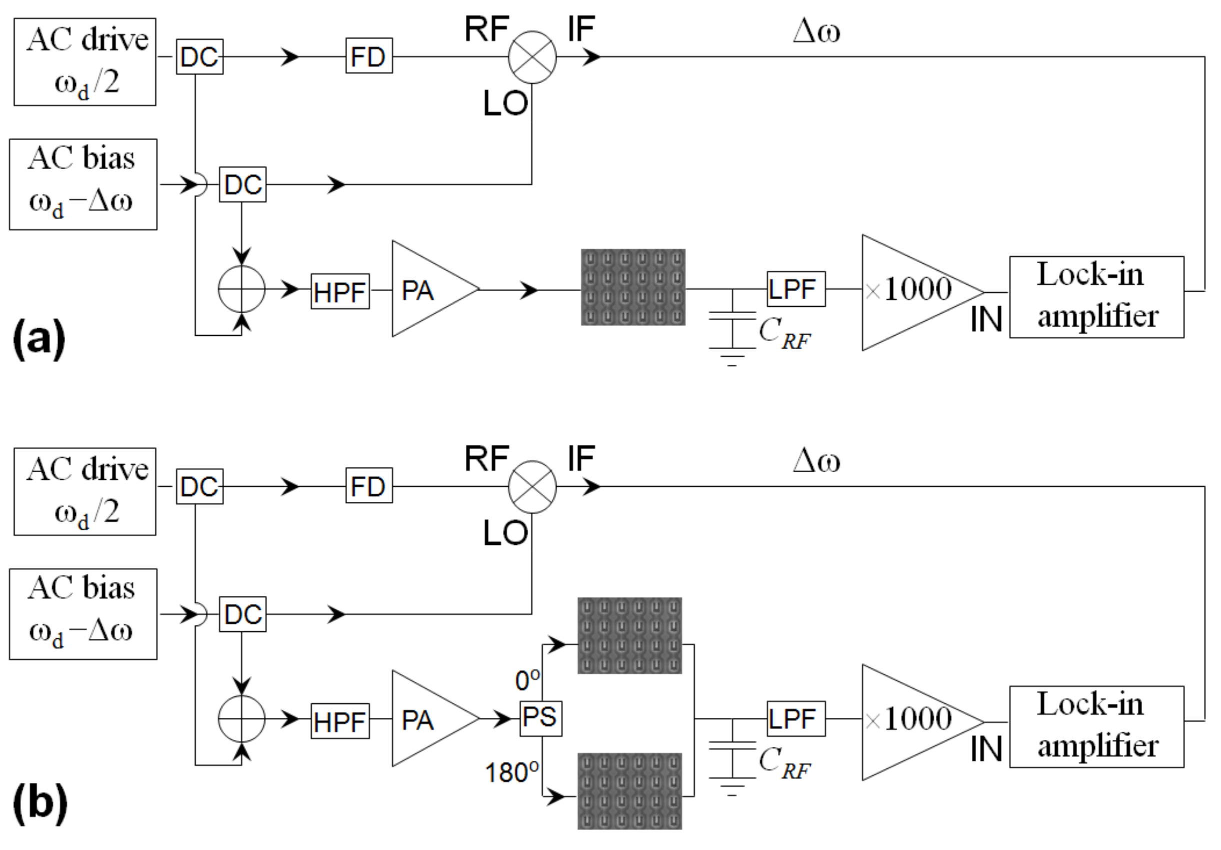

Fig. S2. (a) Schematic of two-port measurement of an array using thermoelastic actuation and piezoresistive detection. DC is the directional coupler, FD is the frequency doubler, HPF is the high-pass filter, LPF is the low-pass filter. (b) Same with balanced detection of two arrays. PS is the 180-degree power splitter.

than 100 kHz. Both the drive and bias voltages are combined using an RF power combiner and sent into the metal loop. On the other side of the loop, a relatively large RF capacitor C_{RF} = 6 nF is connected to ground and therefore creates a virtual ground at high frequencies, f >> $(2\pi\times50\ \Omega\times6\ \text{nF})^{-1} \approx 500$ kHz. This ensures that both the drive and bias voltages primarily drop across the metal loop of the resonator rather than elsewhere in the circuit. A low-pass filter ensures that only the downmixed signal, and not the RF drive and bias voltages are transmitted into the low-noise amplifier.

One difference between the four-port measurement described in Ref.[1] and the two-port measurement is the existence of a significant background in the two-port case. This background arises because the resistance of the piezoresistor depends not only on strain but also on the temperature. Since thermoelastic actuation relies on temperature variations to drive the cantilever, we cannot easily avoid this type of background. This effect produces a background signal at the downmixed frequency because the variations in resistance due to changing temperature and those due to the cantilever motion occur at the same frequency, and therefore mix down to the same frequency $\Delta\omega$. The magnitude of the background signal can be estimated as $V_b \sim V_b\ \alpha_R \Delta T/2$, where α_R is the temperature coefficient of resistance (of the order of 4×10^{-3} K^{-1} for most pure metals) and ΔT is the amplitude of temperature variations. In practice, for resonators with quality factors on the order of one thousand, this background and the resonant signal were roughly of the same order of magnitude. As a result, the two-port measurement is relatively easy to use for vacuum measurements, where the signal is usually comparable to the background, and more difficult in air, where the quality factor can be of the order of 100 or less and resonance signal can be orders of magnitude smaller than the background.

One way to reduce this type of background would be to fabricate the loop from specialty alloys like nichrome, constantan, or manganine, which are designed to have temperature coefficients of resistance up to two orders of magnitude smaller than those of pure metals. Another way to reduce this background is to use two separate, thermally isolated loops for actuation and detection, as illustrated by the cantilever in Fig. S1.

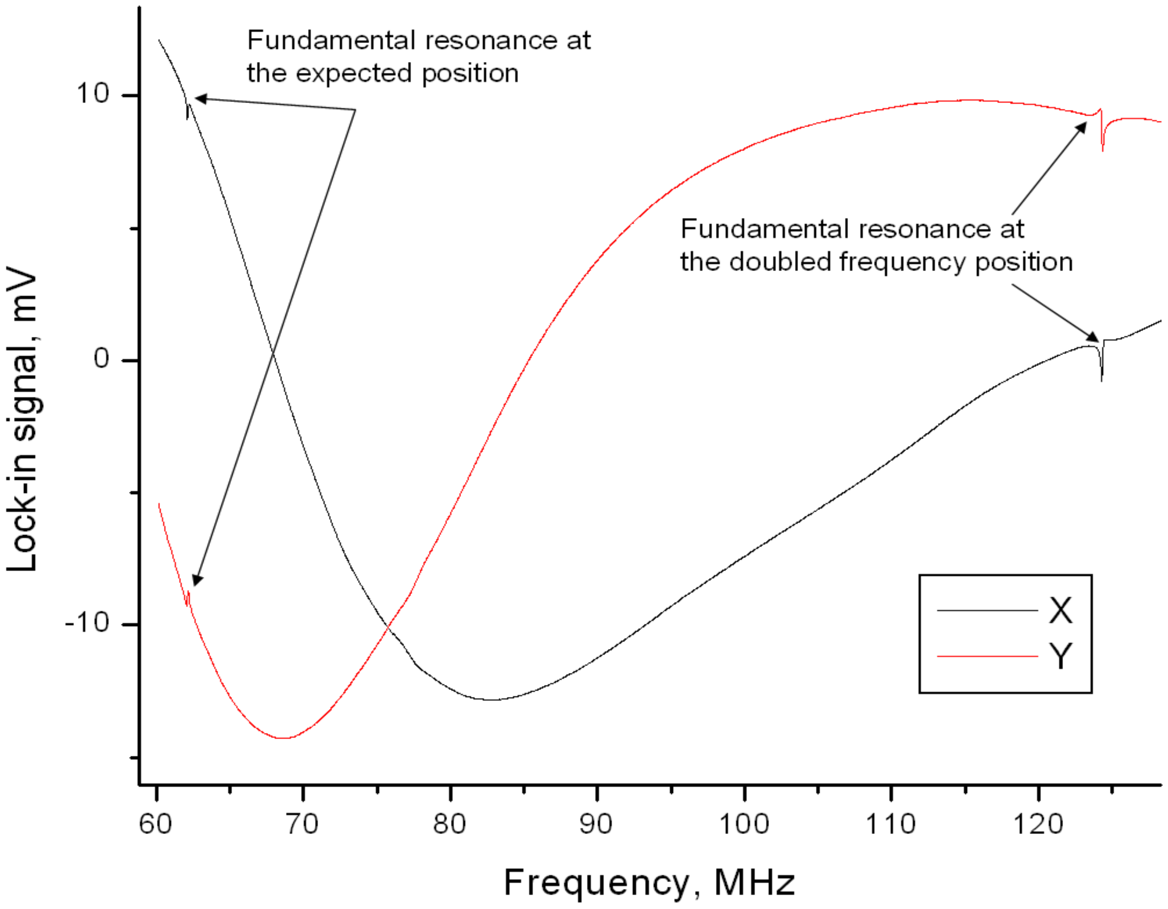

Fig. S3. Wide frequency sweep for a two-port measurement of a single 1.6-μm-long cantilever with the fundamental out-of-plane resonance frequency of 62.11 MHz. The sweep exhibits one resonance peak at the expected resonance frequency and another at roughly twice the expected frequency. The large oscillating background is due to RF cable resonance effects. Lorentzian fits of these two peaks produce resonance frequencies of 62.11 MHz and 124.32 MHz, quality factors of +1000 and -1000, and amplitudes of 0.75 mV and 1.45 mV, respectively. The frequency offset was 44 kHz for this measurement.

A two-port measurement also differs from a four-port measurement in that each mechanical resonance produces not one but two peaks during wide frequency sweeps, as shown in Fig. S3. The first one appears at the expected position, where the voltages are applied at the frequencies $\omega_1 = \omega_d/2 = \omega_R/2$ and $\omega_2 = \omega_b = \omega_R - \Delta\omega$, where ω_R is the resonance frequency. There is, however, a second peak that appears when one voltage is applied at the frequency $\omega_1 = \omega_R + \Delta\omega$ and another at $\omega_2 = 2\omega_R + \Delta\omega$. To understand how this peak forms, note that when we apply both of these voltages to the same loop, they will mix and produce temperature variations at the difference frequency $\omega_2 - \omega_1 = \omega_R$, therefore driving the resonance. The resistance variations at frequency ω_R can then mix with the applied voltage at frequency $\omega_1 = \omega_R + \Delta\omega$ to produce a signal at the expected downmixed frequency $\Delta\omega$. The net result is that another peak appears in the graph at roughly twice the frequency of the expected peak. Figure S3 shows a typical two-port frequency sweep showing a peak at the expected frequency and an additional peak at roughly twice the expected frequency.

Surprisingly, the amplitude of the additional peak is twice bigger than that of the expected peak. This can be explained by considering the algebraic relationships of the various voltage-mixing processes involved. If we apply a sum of two voltages oscillating at frequencies ω_1 and ω_2, the heating is proportional to the square of total voltage:

$$\Delta T = (V_1 \cos\omega_1 t + V_2 \cos\omega_2 t)^2 = V_1^2 \cos^2 \omega_1 t + 2V_1 V_2 \cos\omega_1 t \cos\omega_2 t + V_2^2 \cos^2 \omega_2 t$$
$$= \frac{1}{2} V_1^2 \cos 2\omega_1 t + V_1 V_2 \cos(\omega_1 - \omega_2) t + \frac{1}{2} V_2^2 \cos 2\omega_2 t + \ldots \quad \text{(S1)}$$

Therefore, the temperature variations at frequency $\omega_2 - \omega_1$ are twice as big as those at frequency ω_1, and therefore drive the cantilever motion twice harder. The additional mixing process that produces the downmixed signal at frequency $\Delta\omega$ does not change that conclusion: the amplitude of "double-frequency" peak is twice that of the "regular" peak.

In addition to the twofold difference in amplitude, the phase of the resonance response of the additional peak is flipped with respect to that of the expected peak for reasons similar to those described as in the preceding section. As a result, the regular peaks are fitted with positive quality factors, while additional "double-frequency" peaks, with negative. This turns out to be helpful when analyzing data from wide measurement sweeps that contain peaks from multiple mechanical resonances: if the fitted quality factor Q is positive, then there is indeed a mechanical resonance at the expected frequency $\omega_e = \omega_1 = \omega_2 + \Delta\omega$; however, if the fitted Q is negative, the real mechanical resonance happens at the frequency $\omega_e/2 - \Delta\omega$. In sensing applications, it is often more convenient to work with this "double-frequency" resonance peak since its signal-to-noise ratio is usually twice better for the same amount of heating.

Piezoresistive signal from series-parallel arrays

If we assume that with no excitation all piezoresistors have identical resistances R_0, the resistance of the entire array without excitation is given by $R_{arr} = R_0 \times m/l$, where m is the number of columns in the array, and l is the number of rows. If the resonators are excited into motion, the resistance of a piezoresistor in row i and column j will become $R_{ij} = R_0(1+\delta_{ij})$, where δ_{ij} is the relative change in its resistance due to the motion-related deformation. The resistance of the entire array in this case is given by

$$R_{arr} = \sum_{j=1}^{m} \frac{1}{\sum_{i=1}^{l} R_0 \left(1+\delta_{ij}\right)^{-1}} \approx R_0 \frac{m}{l}\left(1+\frac{1}{lm}\sum_{i=1}^{l}\sum_{j=1}^{m}\delta_{ij}\right), \tag{S2}$$

where the Taylor expansion is justified because the relative changes in the resistance of piezoresistors are generally small, δ_{ij}<<1. When using piezoresistive detection, the signal is proportional to the applied bias voltage V_b:

$$V_{arr} = \frac{1}{2} V_b \frac{\Delta R_{arr}}{R_{arr}} = \frac{V_b}{2lm}\sum_{i=1,\,j=1}^{l,m}\delta_{ij}\,. \tag{S3}$$

If all resonators respond in exactly identical ways, $\delta_{ij}=\delta_0$, the formula for the array signal reduces to that of an individual resonator:

$$V_{arr} = \frac{V_b}{2lm}\sum_{i=1,\,j=1}^{l,m}\delta_{ij} = \frac{1}{2}V_b\delta_0\,. \tag{S4}$$

The maximum drive that can be applied to an individual piezoresistive resonator is limited either by the nonlinearity of mechanical response or the maximum tolerable level of heating. In our experiments, the maximum tolerable temperature increase due to heating was typically on the order of 100 K, corresponding to maximum dissipated power on the order of $P_{max} \sim$ 100 μW for an individual resonator. If this maximum power is applied to each resonator in the array, the total dissipated power will naturally scale as the number of array elements, $N = lm$. In contrast, the bias signal, V_b, and the maximum signal that can be obtained from the array, V_{arr}, will scale as the number of columns, m.

It would seem then that an array consisting of just one row would be the most economical way to leverage the signal of individual resonators. However, having an array of just one row would mean that the array resistance scales linearly with the number of array elements, $R_{arr}=R_0 \times N$, and may reach excessively large values for arrays of thousands of resonators. In experiments, it is often desirable to keep the resistance of the total array close to some fixed value, usually 50 Ohm for high-frequency applications. In addition, a single-row array is very vulnerable to electrical defects since the breaking of the conducting path in just one piezoresistor would render the entire array inoperable. As a result, it is preferable to scale the number of rows proportionally to the number of columns, so that the arrays remain robust with respect to defective individual resonators and have approximately constant resistance. In this case, the piezoresistive signal scales proportionally to m and therefore proportionally to the square root of the total number of array elements, $\sqrt{N}$. At the same time, Johnson noise and thermoelastic noise—the fundamental sources of noise in such measurement—do not depend on N at all. The signal-to-noise ratio then also scales as the square root of the number of elements, $\sqrt{N}$, and, of the total dissipated power, $\sqrt{N}\ P_{max}$. This situation, where the signal-to-noise ratio increases proportionally to the square root of the total dissipated power, is very commonly encountered in electrical engineering.

Note, however, that the scaling of signal as $\sqrt{N}$ is the best-case scenario. In reality, different resonators will not respond to the drive in identical ways for a number of reasons. The first one to consider is that the phase and amplitude of the drive may not be the same for all resonators. For example, in the case of piezoshaker drive, the phases and amplitudes of the surface motion will vary due to the interference of ultrasound waves inside the bulk of the resonator chip. The length scale of such variations is on the order of the bulk acoustic wavelength corresponding to the resonator's frequency, ~ 350 μm for a 25 MHz resonator on silicon substrate, which is smaller than the dimensions of the typical arrays we used in our experiments.

If we assume, for the sake of argument, that the drives for different resonators of the array have completely random phases φ_d but the same amplitude, then the array signal will take the form

$$V_{arr} = \frac{1}{2lm} V_b \sum_{i=1,j=1}^{l,m} \delta_{ij} e^{i\varphi_d} \tag{S5}$$

The addition of such signals with random phase is equivalent to a random walk in the complex plane, which implies that the expected magnitude of the sum will scale as the square root of the total number of terms in the sum:

$$\langle V_{arr} \rangle = \frac{1}{2} V_b \frac{\sqrt{N}}{N} \delta_0 = \frac{1}{2\sqrt{N}} V_b \delta_0 \,. \tag{S6}$$

Since the bias signal V_b normally scales as $\sqrt{N}$, the use of arrays does not offer any advantages with respect to the use of an individual device in this case. It is therefore crucial to keep the drive phase the same for all the resonators in the array. Maintaining such phase coherence of the drive is difficult with piezoshaker drive but is much easier with integrated actuators, such as the thermoelastic actuators.

Even if the drives of all resonators in the array are perfectly in sync, the response of individual resonators may not be the same because they all have slightly different mechanical properties, in particular different resonance frequencies. This effect of the frequency dispersion is considered in the following section.

Effect of frequency dispersion

The finite resolution of e-beam and optical lithography introduces slight variations in the dimensions of the fabricated resonators. As a result, all resonators in the array will have slightly different mechanical properties, and in particular different resonance frequencies. A simple way to judge whether this dispersion in resonance frequency is significant is by comparing it to the natural width of the resonance under typical operating conditions. For example, nanoscale resonators shown in Figure S1 have a typical quality factor on the order of 100 in air at atmospheric pressure, corresponding to a resonance width that is 1% of the resonance frequency. Therefore, if the dispersion of resonance frequency is much smaller than 1%, the individual resonance curves will strongly overlap and the Lorentzian-response term of individual resonators may be treated as the same, leading to the summed array response being Lorentzian as well. In this case, we obtain the maximum possible amplitude of the array response. Conversely, if the dispersion of the frequencies is much larger than 1%, the sum of the individual Lorentzian response curves will be much broader than an individual resonance, and the peak array signal will be much reduced with respect to the maximum possible signal.

In order to quantify this qualitative argument, we need to consider the problem of adding up many Lorentzian resonance curves with randomly distributed resonance frequencies. The normal or Gaussian probability distribution is a common choice in such simulations; however, we have found that in reality the distribution of resonance frequencies has long and "fat" tails, i.e., the probability of finding resonance frequencies far off the mean is much larger than would be expected in a Gaussian distribution. The large number of outliers is illustrated by the data in Figs. 3(b) and 4(a).

To model this large number of outliers, it is more appropriate and convenient to use the Cauchy distribution, which has the probability density function similar to the Lorentzian:

$$p(\omega_R) = \frac{Q_{disrt} \;/\; 2\pi\omega_0}{1+\left(\dfrac{\omega_R-\omega_0}{2\omega_0 \;/\; Q_{distr}}\right)^2}, \qquad \text{(S7)}$$

where ω_0 is the center frequency of the distribution and Q_{distr} characterizes the width of the distribution, similarly to the way that a quality factor characterizes the width of a Lorentzian curve. The Cauchy distribution more accurately describes the long tails of the frequency distribution that we observe in practice and has the added advantage that the expected form of the array response can be calculated analytically, as shown below.

The response of a forced, damped harmonic oscillator is given by

$$s(\omega_D) = A\frac{\omega_R^2/Q}{\omega_R^2-\omega_D^2+i\omega_R\omega_D/Q}, \qquad \text{(S8)}$$

where A is the amplitude of the resonance signal, ω_R is the resonance frequency, Q is its quality factor, and ω_D is the frequency of the drive. If the quality factor of an individual resonator is large, Q >> 1, the response near the resonance can be approximated by the complex Lorentzian

$$s(\omega_D) = A\frac{\omega_R/Q}{\omega_R-\omega_D+i\omega_R/(2Q)}. \qquad \text{(S9)}$$

The expected signal from one array cantilever with a randomly distributed resonance frequency is then a convolution of the complex Lorentzian response with the Cauchy distribution:

$$\left\langle \mathrm{s}(\omega)\right\rangle \approx \int_{-\infty}^{\infty} \frac{\omega_R/Q}{\omega_R-\omega_D+\dfrac{i\omega_R}{2Q}} \times \frac{Q_{disrt}/2\pi\omega_0}{1+\left(\dfrac{\omega_R-\omega_0}{2\omega_0/Q_{distr}}\right)^2}\, d\omega_R = \frac{A\omega_0/2Q}{\omega_0-\omega_D+\dfrac{i\omega_R}{2}\left(\dfrac{1}{Q_{distr}}+\dfrac{1}{Q}\right)}, \qquad \text{(S10)}$$

which is simply the complex Lorentzian response with a new effective quality factor $Q_{eff}= 1/(Q^{-1}+Q_{distr}^{-1})$ and a new effective amplitude $A_{eff}=AQ_{eff}/Q$. The expected signal of the entire array will have the same form, since it is essentially the sum of the expected signals of individual cantilevers and we assume all cantilevers in the array to be described by the same Cauchy probability distribution. In addition, for a large array, N >> 1, the typical response generally will not deviate far from the expected response as the random variations introduced by individual resonances will largely average out. Therefore, the frequency dispersion effectively has the same result as an additional source of damping, corresponding to a quality factor Q_{distr}, which reduces the effective quality factor of the array from Q to $Q_{eff} = 1/(Q^{-1}+Q_{distr}^{-1})$.

To illustrate this effect of the resonance frequency dispersion on the shape of the array response, we have performed numerical simulations for an array consisting of 2800 elements (Fig. S3(a)) In the simulations, all resonators assumed to have a quality factor of

100, corresponding to experimental value in air, and the width of the Cauchy distribution was varied, starting from a distribution width much smaller than the natural width of the resonance, $Q_{distr} >> Q$, and ending with a distribution width much larger than the natural width of the resonance, $Q_{distr} << Q$. As expected, increasing the width of the frequency distribution broadens the resonance peak of the entire array and reduces its amplitude.

In Fig. S3(a), the simulated response curves do not deviate significantly from the perfect Lorentzian curves due to the large number of the elements in the array and their relatively low quality factors. However, the fact that the overall response curves of the array consist of many narrow lines corresponding to individual resonators becomes evident if the quality factors of individual resonances are high enough, as shown in Fig. S3(b). These curves simulate the response of actual arrays used in experiments in air and vacuum. The

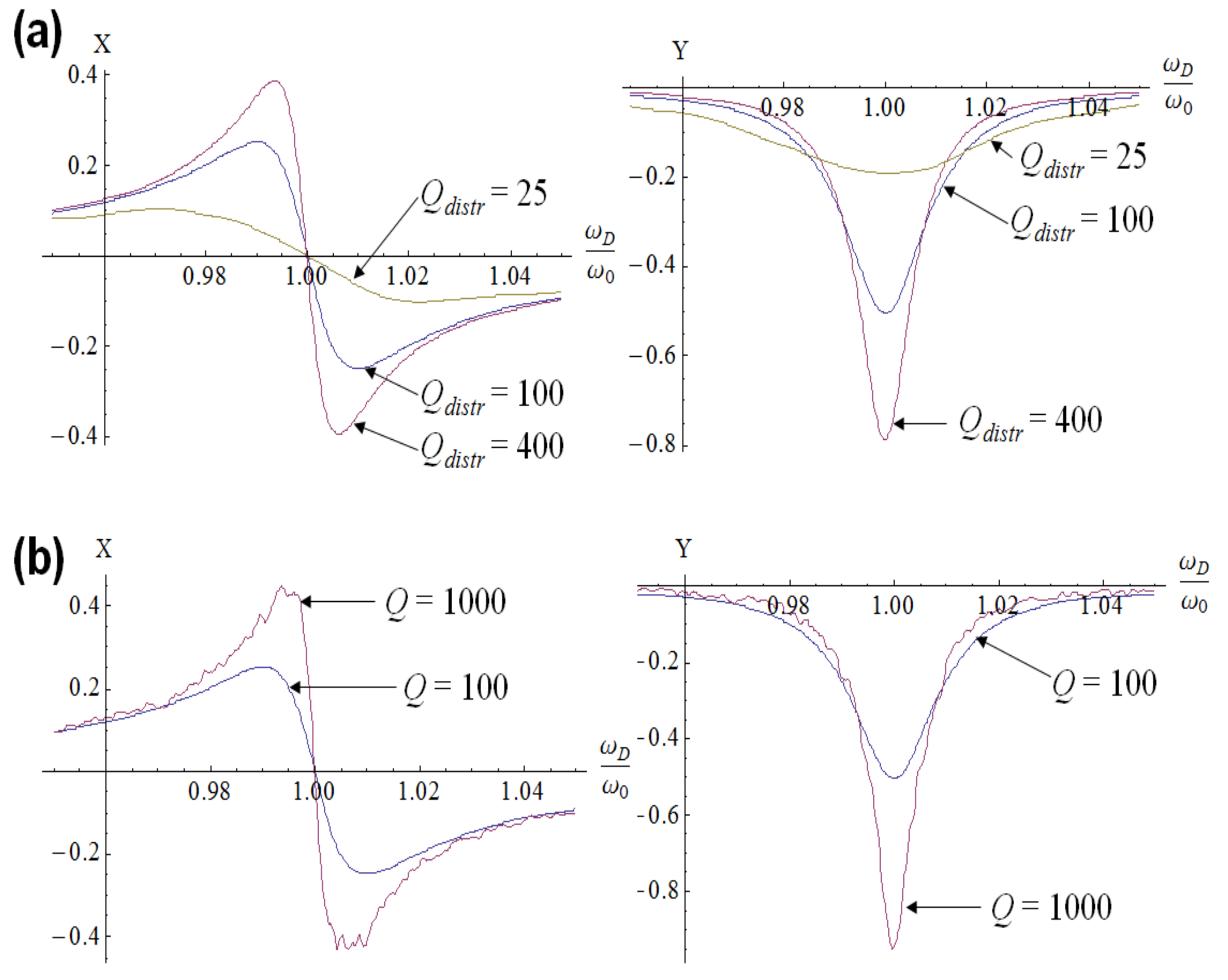

Fig. S4. (a) X and Y quadratures of the simulated response of an array of 2800 cantilevers with Q = 100 and Q_{distr} of 400, 100, and 25. The effective quality factors Q_{eff} in these cases are $(400^{-1}+100^{-1})^{-1} \approx 80$, $(100^{-1}+100^{-1})^{-1} \approx 50$, and $(25^{-1}+100^{-1})^{-1} \approx 20$, respectively. (b) Same for Q_{distr} = 100 and Q of 100 and 1000, which approximately corresponds to the frequency dispersion and quality factors in air and vaccum of actual arrays used in experiments.

frequency distribution width of the typical arrays we have worked with were on the order of 1%, corresponding to Q_{distr} = 100, and the quality factors of individual resonances were on the order of 100 and 1000 in air and vacuum, respectively. As a result, the experimental curves in vacuum had more "fine structure" than those in air due to the response of individual cantilevers.

The differences between different resonators in the array are, of course, not limited to the variations in the drive phase and resonance frequencies. The quality factors and the amplitudes of response of individual resonators will also vary. However, we have found in our experiments that these variations are relatively insignificant and have a negligible effect on the overall response of the arrays.

[1] I. Bargatin, I. Kozinsky, M.L. Roukes, *Efficient electrothermal actuation of multiple modes of high-frequency nanoelectromechanical resonators*, Appl. Phys. Lett. **90**, 093116 (2007).