EUV blank defect and particle inspection with High throughput Immersion AFM with 1 nm 3D resolution

Maarten H. van Es[a] and Hamed Sadeghian[a]

[a]Department of Optomechatronics, Netherlands Organization for Scientific Applied Research, TNO, Stieltjesweg 1, 2628CK, Delft, The Netherlands

ABSTRACT

Inspection of EUV mask substrates and blanks is demanding. We envision this is a good target application for massively parallel Atomic Force Microscopy (AFM). We envision to do a full surface characterization of EUV masks with AFM enabling 1 nm true 3D resolution over the entire surface. The limiting factor to do this is in the sensor itself: throughput is limited by the time that a cantilever needs to adjust its oscillation amplitude to the surface topography while scanning. We propose to use heavily damped cantilevers to maximize the measurement bandwidth. We show that using up to 20.000 cantilevers in parallel we can then reach a throughput of one $152 \times 152\,\mathrm{mm}^2$ substrate per 2 days with 1 nm resolution.

Keywords: AFM, EUV mask, massively parallel, throughput

1. INTRODUCTION

Inspection of EUV blanks is one application where large area nanometer scale inspection is needed. No technique so far can provide good inspection over a full blank at satisfactory speed and resolution.[1] EUV blanks and their glass substrates need to be very high quality to be used in EUV lithography. Critical sizes are expected to reach 1 nm on mask for the 7 nm node.[1] Furthermore, using conventional techniques, some defects may not be detected until the mask is finished or even in use. AFM can detect <1 nm topgraphical defects routinely, but can it be fast enough to inspect a full blank or substrate in a reasonable time? In this paper we explore whether speed of AFM might be increased enough that it becomes a viable candidate for full blank inspection. Even the fastest AFM systems in development today[2,3] are about 4 orders of magnitude too slow for this task. We will review practical and physical limits defining measurement speed of AFM including measurements showing various of the issues discussed. We will also discuss why we think inspection of EUV blanks is a good application to target first with the kind of massively parallel AFM that is needed to provide such a speed increase.

However, before exploring massive parallellism, it is good to explore the speed limit for a single AFM cantilever. In AFM the probe that scans the surface is mounted at the end of a micro cantilever. Often, the cantilever is driven at its first resonance frequency while its response — amplitude, phase or resonance frequency — is modulated by the surface being scanned. This is referred to as dynamic mode and is the preferred method of operation of an AFM as it minimizes damage to sample and tip.[4] The response of the cantilever indicates the height of the surface and the response time is defined as the time the cantilever needs to adjust its response to changes in height. Modeling the cantilever as a simple harmonic oscillator, the response time τ of the cantilever in the amplitude channel is given by $\tau = Q/\pi f_0$. Here f_0 is the resonance frequency of the cantilever and Q is the quality factor, a measure for the sharpness of the resonance in the frequency domain and related to the damping of the cantilever system. A typical cantilever such as the mikromasch NSC35-b has $f_0 = 300\,\mathrm{kHz}$ and $Q = 300$ (in air), so $\tau = 0.3\,\mathrm{ms}$. This cantilever can meaningfully scan only ~ 3000 datapoints per second, or $3\,\mathrm{\mu m\,s^{-1}}$ with 1 nm resolution. This is severely limiting for high speed applications.

We are not the first to recognize this limitation, see for example Mertz *et al*,[5] who reduced Q using force feedback on the cantilever to improve imaging speed. Although this early research on reducing the response time focused on the Q factor, most progress in imaging speed has been made by reducing the cantilever mass and thereby increasing f_0.[6] The smallest cantilevers currently on the market have dimensions of only a few µm and cannot be made much smaller as long as optical methods are the preferred way of detecting deflection. Also, the tip itself is becoming a limit to further reduction of cantilever sizes. These small cantilevers have resonance

Metrology, Inspection, and Process Control for Microlithography XXX, edited by Martha I. Sanchez, Vladimir A. Ukrainstev
Proc. of SPIE Vol. 9778, 97782Z · © 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2219127

frequencies of up to several MHz, speeding up AFM significantly. However, still higher speeds are required, so Q needs to be lowered as well.

Several groups have developed techniques to lower Q. Many have followed the general idea of Mertz *et al*[5] and implemented active feedback on the cantilever motion, see for example Gunev *et al*[7] and Sulcheck *et al.*[8] Fairbain *et al.*[9] used a piezo-electrically coated cantilever in a passive electronic circuit to dampen the cantilever. Miles *et al.*[10] have coated cantilevers with a polymer to increase damping. All methods have their issues. The active damping methods work best with a very clean and direct driving mechanism, which is often not available, and they require additional electronics. Especially for parallel AFM[2] the latter is not desirable. The other methods need custom cantilevers and especially polymer coatings may be difficult to produce reliably without introducing residual, temperature dependent stress. Moreover, all of these methods are limited in performance, *i.e.* in how much they can lower Q. Published numbers in the references quoted above are limited to about one order of magnitude reduction.

Besides all these methods there is one simple method to lower Q, which is immersing the cantilever in a damping environment. In fact, in biophysics research where AFM is a popular tool, it is very common to immerse tip and sample in liquid to preserve biological function of the sample.[6] It is well known that in water the Q factor drops by a factor of about 100, which is significantly more than what has been achieved using the above methods. Even when taking into account the fact that the resonance frequency also drops by a factor 3, this still provides an improvement in speed of between 3 and 10 times.

At this point it should be noted that all the above quoted Q factors are determined with the tip away from the sample surface. However, the tip-sample interaction dissipates energy and hence influences the Q factor. This should be taken into account to realistically determine the benefit of a reduction in Q by any of the above means. Moreover, any realized speed improvement as discussed in literature such as the above papers is based on analysis of specific, usually steep features in topography images, and thus is very dependent on feedback parameters. The feedback loop is seldom described and analysed in full detail in all conditions and as such this is a poorly quantified procedure. Together with the observation that the tip-sample interaction itself changes the Q factor, it is obvious that it is crucial to determine the Q factor while scanning a surface. We developed a method to do this based on the frequency spectrum of the cantilevers amplitude, and we verify the method using different cantilevers in air and water.

While it is clear a significant amount of speed can be gained by further optimization of the cantilever by reducing Q, one cantilever or even a couple handful of cantilevers[2,3] are not going to be fast enough to scan a complete EUV blank. Massive parallellization is needed. EUV blanks are a perfect candidate for this though. EUV blanks are guaranteed to be rather flat, even in the presence of defects. Therefore, large arrays of cantilevers can be brought to the surface and after initial tilt and alignment, no feedback is required, or at most feedback on three points of the array is enough to keep all cantilevers just in contact with the surface. This greatly simplifies most of the issues associated with massive parallellization to a degree that they should become manageable.

2. METHODS

2.1 How to measure the cantilever response speed

To assess the performance of a cantilever we do not measure the Q factor but rather we determine τ directly, and if desired we can get the Q factor from τ instead. We assess the time constant τ of the cantilever to change its amplitude in frequency space, using the transfer function of the cantilever system for variations in the amplitude. Basically, the cantilever will be able to follow slow changes in amplitude, but not fast changes, so that we expect the cantilever to have a transfer function for amplitude modulations similar to a low pass filter. Therefore, we propose to characterize the cantilever analogous to how a low pass filter is characterized in terms of its -3dB point in its frequency dependent transfer function. From a measurement of the -3dB point we can directly find τ. We do this measurement by modulating the amplitude with white noise. We can then do a Fourier transform on the measured amplitude to directly get the transfer function of the system as the white noise simultaneously excites all frequencies. In practice, besides the limited response time of the cantilever we also expect other features of the measurement system to influence the transfer function. Specifically for our measurements these are the limited bandwidth of the lock-in amplifier we used and a limited input noise amplitude at high frequencies.

The white noise of amplitude variations is generated by scanning a rough surface with slow height feedback. The pixel-size for this scan is chosen to be larger than the typical feature-size on the surface so that every measured pixel in the amplitude channel is uncorrelated. Every pixel in the amplitude channel now has a random value and we therefore obtain a white noise modulation on the cantilever amplitude. Because the tip moves over the spatial features in a specified time, the white noise of spatial frequencies is translated into white noise at temporal frequencies depending on the scanning speed. In practice, the noise is not entirely white but has lower amplitude at the highest frequencies. This is because the spatial frequencies available are limited as the tip wears down during the experiment, and subsequent pixels are not entirely uncorrelated. The limited spatial frequency content in the measurement translates to a limited temporal frequency content. This effect can be checked by measuring at several different scanning speeds as the translation of spatial to temporal frequencies is dependent on scanning speed. The other effects in the amplitude spectra are only dependent on the cantilever or scanner system and thus their frequencies do not depend on the scanning speed.

The lock-in amplifier we used is an advanced digital lock-in amplifier with parameters that can be optimized for high bandwidth measurements. In particular the low-pass filter on the amplitude output can be chosen very steeply, up to 48 dB/octave so that the output bandwidth can be a significant part of the modulation frequency. For our experiments, the output bandwidth could be increased to about one third of the modulation frequency before the 2ω component could not be suppressed below the noise anymore. For frequencies above the cut-off of the low-pass filter, the transfer function will show the steep fall-off of this filter.

2.2 Experimental setups

We use two types of cantilevers for both of which we show results. One is a BudgetSensors Multi75-G ($k = 3.4\,\mathrm{N/m}$, $f_0 = 74.1\,\mathrm{kHz}$), the other a MikroMasch NSC35-b ($k = 26.6\,\mathrm{N/m}$, $f_0 = 337\,\mathrm{kHz}$), with these properties determined from their respective thermal spectra in air. The experiments in air and in water are always performed one directly after the other with the same cantilever. For both cantilevers, the tip radius was $<10\,\mathrm{nm}$ according to manufacturers' specifications.

The sample is a flat silicon surface with a naturally grown, rough oxidation layer.

3. RESULTS

After characterizing the cantilevers when not engaged to the surface, we perform measurements according to the method outlined in section 2.1. For this, cantilever response spectra are determined from a Fast Fourier Transform (FFT)) of the measured amplitude data while scanning a rough SiO_2 surface at high speed. Figure 1 shows the data as used for analysis which are obtained in air with the Multi75-G cantilever. The cantilever is not engaged to the surface on the right hand side (so just after the turning point) because feedback is too slow to follow the topography when going from left to right at this high scanning speed. However, the cantilever touches the surface again very soon in the right to left direction and then stays in contact. Oscillations are clearly visible in the amplitude once the cantilever touches the surface. These are caused by a resonance in vertical direction of the piezo-tube scanner which is excited by the fast scanning motion. The data shown was taken at the highest used line speed for this condition, 18.5 lines/s. Data was also taken at 9 lines/s and 1 lines/s, corresponding to sample rates of 74 kSamples/s, 36 kSamples/s and 4 kSamples/s respectively. Raw data in water and for the NSC35-b are very similar.

Figure 2 shows the FFT spectra of amplitude variations for both cantilevers in air and in water. These spectra show the transfer-function of the cantilever system for amplitude variations that we can use to determine the time constant τ of the cantilevers response and its corresponding Q factor. For each setting, the graph shows an average of the spectra of many individual scan-lines to increase the signal to noise ratio. The scanner resonance that we identified in figure 1 is also clearly visible in the spectra taken with a high scanning speed.

4. DISCUSSION

Various effects are visible in the amplitude spectra. First of all, we would like to draw attention to the fact that the fastest curves measured in water start to fall off at a higher frequency than the fastest curves measured in air for both cantilevers. The $-3\,\mathrm{dB}$ points indicated in the graph for air, and the slopes beyond these points

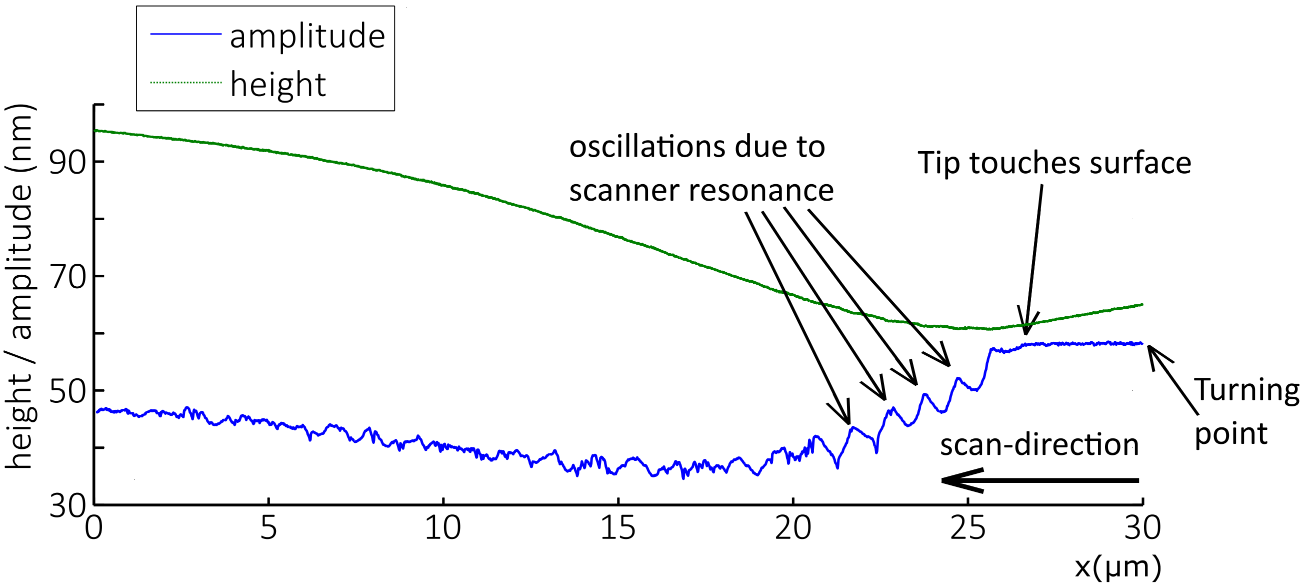

Figure 1. Linetrace of topography and amplitude data as used in the analyses (scan direction right to left). This data has been acquired with 18.5 lines/s, corresponding to 74 kSamples/s. In the amplitude image it can be observed that the cantilever did not touch the surface at the far right due to sample tilt in combination with a relatively slow feedback. After the cantilever gets into contact with the surface, oscillations are visible which are due to a scanner resonance which is excited by the sharp turn-around in the fast scanning direction. Otherwise, the noise in the amplitude trace is the white noise generated by scanning over a rough surface; this line data is Fourier transformed to get the spectra of amplitude variations used in this paper.

correspond to a first order low pass filter as expected for an oscillator limited by its response time or Q factor. For the NSC35-b the -3 dB at about 4.5 kHz point corresponds to a Q factor of about 235, while for the Multi75-G with 2.5 kHz this corresponds to a Q factor of about 93 according to equation $\tau = Q/\pi f_0$.

In water the lock-in limits the response speed of the measurement system with the Multi75-G cantilever. Although we use an advanced lock-in, the lowpass filter limits the bandwidth of the lock-in to about one third of the drive frequency, or 9.6 kHz for this cantilever. Using a peak detector[11] we expect to be able to reach a bandwidth close to the cantilever resonance frequency. We can conclude that in a highly dissipative medium like water, the cantilever Q factor is not a limiting factor for high speed AFM performance.

For the NSC35-b, the lock-in bandwidth was 48 kHz, just beyond what we could measure. The fastest sample rate of 90 kS/s was set by the combination of number of pixels the software could handle (2044) and the linerate the scanner could achieve (22 lines/s). It is clear from the figure that the cantilever is not limiting yet in water at 45 kHz or $1/3rd$ of the drive frequency. Again, the Q factor of the cantilever is not a limiting factor anymore when working in water.

Lock-in amplifiers have originally been developed to measure the amplitude of fast oscillations accurately by converting the signal to a much lower frequency where the measurement is easier and more accurate, although relatively slow. With todays electronics there are however other methods possible which can determine the amplitude of high frequency oscillations more rapidly. One example that has been used in AFM already is peak detection[6] which will allow amplitude detection at a bandwidth up to the cantilever resonance frequency.

Based on current technology, we envision small, encased cantilevers with a resonance frequency up to 5 MHz and high damping, so that their intrinsic measurement bandwidth will also be around 5 MHz. Using peak detection or a similar fast direct amplitude measurement we therefore expect a single AFM cantilever can allow a maximum measurement bandwidth of 5 MHz. If stiffer cantilevers, which might more easily cause damage, are acceptable, this might be extended even by a factor of 3 or more.

For a full surface scan with 1 nm resolution of an EUV blank with a surface area of $152 \times 152\,\text{mm}^2$ an image will need 2.31×10^{16} pixels. With a measurement bandwidth of 5 MHz a single AFM cantilever would

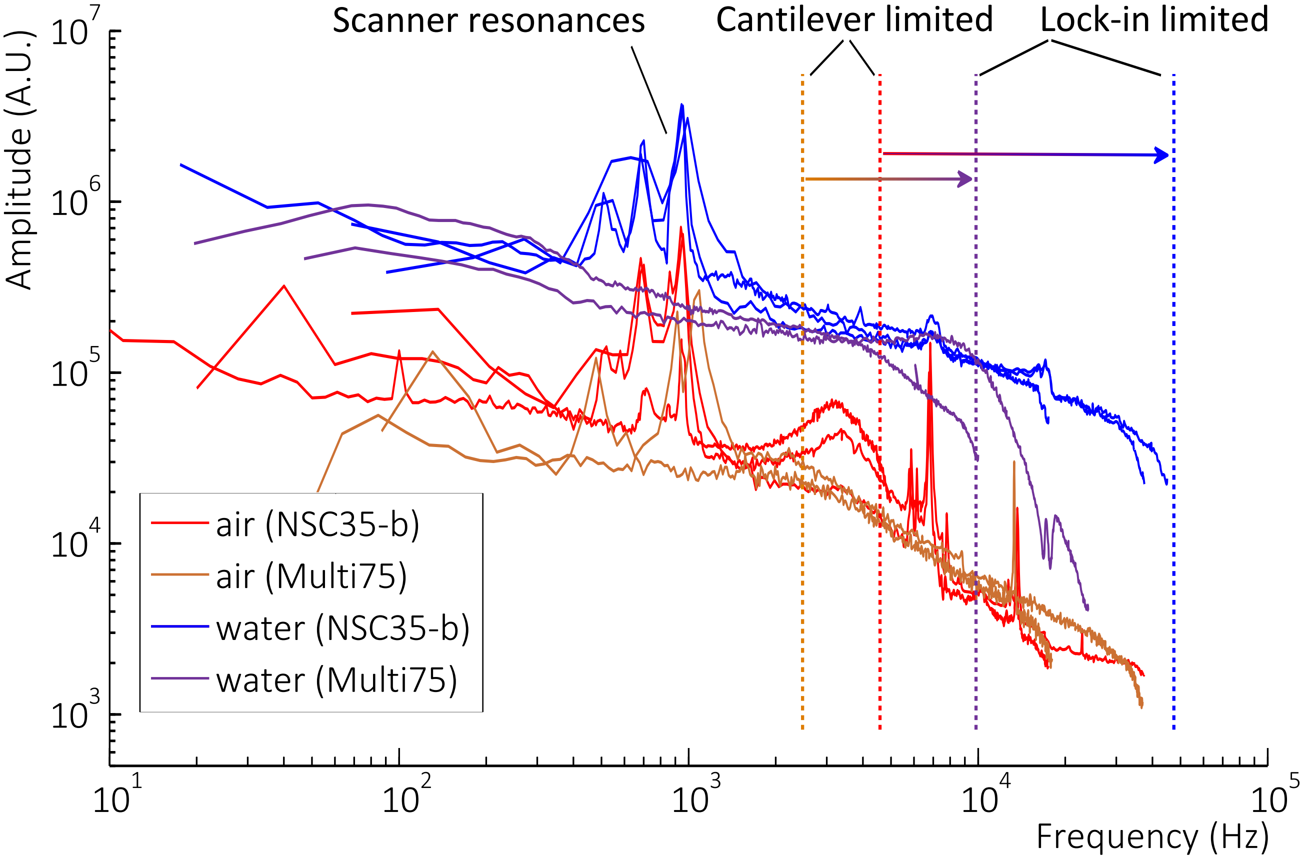

Figure 2. Spectrum of amplitude variations in air and water for both types of cantilever used in this study. For each cantilever and environment, a number of different scan-speeds is used to be able to cover a wider frequency range. This results in two or three curves, here displayed in the same colour for each condition. It can be seen that the individual curves per condition overlap and show the same features. The most prominent features recurring in the curves are the scanner resonance around 1 kHz and the reduction in amplitude beyond a certain frequency. These are indicated in the graph. A more detailed discussion of various features present in the spectra is given in the text.

need 4.62×10^9 s or 1.28×10^6 hours to measure that much data. With 20.000 cantilevers simultaneously, this reduces to less than 3 days, which is starting to become a reasonable number. Is twentythousand cantilevers a reasonable number though? IBM has already shown working arrays consisting of more than 1000 cantilevers for its millipede AFM based storage system in 2000,[12] so 20.000 cantilevers is 'just' a further development of such technology. Twenty thousand cantilevers simultaneously correspond to over $1 \times 1\,\text{mm}^2$ area per cantilever for an array covering an EUV blank, which is very generous considering high frequency, small cantilevers measure only around $10 \times 1\,\mu\text{m}^2$. This suggests that parallellizing to even more cantilevers than we suggest here should be feasible. Deflection readout might be better accomplished by an integrated device on the cantilever, such as using a piezoeresistive effect or an electrostatic effect[13] as there is no space to integrate an optical beam deflection readout for each cantilever. Actually, IBM used a thermal readout method for its millipede AFM based storage system.[12]

Due to the extremely well defined, flat surface topography, feedback does not have to be performed on each cantilever individually, greatly simplifying control and design for this case. Controlling just three points on an array to keep tilt and distance correct will work to keep all cantilevers in contact with the surface, although there may need to be a facility to bring each cantilever in the correct position on first approach to the surface. The best approach might be to have a number of smaller arrays of cantilever each with its own control and coarse positioning, which diminishes differences within arrays simply due to fewer cantilevers per array. Considering the area calculation per cantilever, there is significant space available for the extra hardware this scheme requires. In general, parallel and even massively parallel cantilever schemes have been shown before so the issues mentioned here are not insurmountable.

We conclude that using current AFM technologies it is possible to scan a full EUV blank with 1 nm resolution in a couple of days or even a couple of hours. While there are large technical and engineering challenges to solve, it is fascinating that this is possible. However, this is approaching what is physically possible with reasonably standard AFM techniques, so this is also a reason to start thinking about different AFM approaches or other local sensing techniques.

REFERENCES

[1] Rankin, J., Qi, Z. J., Lawliss, M., Narita, E., Seki, K., Badger, K., Bonam, R., Halle, S., and Turley, C., "Euv photomask defects: what prints, what doesn't, and what is required for hvm," in [*SPIE Photomask Technology*], 96350L–96350L, International Society for Optics and Photonics (2015).

[2] Sadeghian, H., van den Dool, T. C., Crowcombe, W. E., Herfst, R. W., Winters, J., Kramer, G. F. I. J., and Koster, N. B., "Parallel, miniaturized scanning probe microscope for defect inspection and review," *Proc. SPIE* **9050**, 90501B–90501B–7 (2014).

[3] Sadeghian Marnani, H., Dekker, A., Herfst, R., Winters, J., Eigenraam, A., Rijnbeek, R., and Nulkes, N., "Demonstration of parallel scanning probe microscope for high throughput metrology and inspection," *Proc. SPIE* **9424**, 994240O (2015).

[4] Ashby, P. D., "Gentle imaging of soft materials in solution with amplitude modulation atomic force microscopy: Q control and thermal noise," *Applied Physics Letters* **91**(25), 254102–254102 (2007).

[5] Mertz, J., Marti, O., and Mlynek, J., "Regulation of a microcantilever response by force feedback," *Applied Physics Letters* **62**(19), 2344 (1993).

[6] Ando, T., "High-speed atomic force microscopy coming of age," *Nanotechnology* **23**(6), 062001 (2012).

[7] Gunev, I., Varol, A., Karaman, S., and Basdogan, C., "Adaptive q control for tapping-mode nanoscanning using a piezoactuated bimorph probe," *Review of scientific instruments* **78**(4), 043707–043707 (2007).

[8] Sulchek, T., Hsieh, R., Adams, J., Yaralioglu, G., Minne, S., Quate, C., Cleveland, J., Atalar, A., and Adderton, D., "High-speed tapping mode imaging with active q control for atomic force microscopy," *Applied Physics Letters* **76**(11), 1473–1475 (2000).

[9] Fairbairn, M. W., Moheimani, S., and Fleming, A. J., “Control of an atomic force microscope microcantilever: a sensorless approach,” *Microelectromechanical Systems, Journal of* **20**(6), 1372–1381 (2011).

[10] Humphris, A., Miles, M., and Hobbs, J., “A mechanical microscope: high-speed atomic force microscopy,” *Applied physics letters* **86**(3), 034106 (2005).

[11] Ando, T., Kodera, N., Takai, E., Maruyama, D., Saito, K., and Toda, A., “A high-speed atomic force microscope for studying biological macromolecules,” *Proceedings of the National Academy of Sciences* **98**(22), 12468–12472 (2001).

[12] Vettiger, P., Despont, M., Drechsler, U., Dürig, U., Häberle, W., Lutwyche, M. I., Rothuizen, H. E., Stutz, R., Widmer, R., and Binnig, G. K., “The millipedemore than thousand tips for future afm storage,” *IBM Journal of Research and Development* **44**(3), 323–340 (2000).

[13] Michels, T. and Rangelow, I. W., “Review of scanning probe micromachining and its applications within nanoscience,” *Microelectronic Engineering* **126**, 191–203 (2014).