Mixed-Signal Data Interface for Maskless Lithography

Benjamin Warlick, Borivoje Nikolić

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA, USA, 94720-1770

ABSTRACT

A future maskless lithography system that replaces traditional masks with an array of electro-mechanical mirrors relies on a very high rate data interface to achieve the wafer throughputs comparable to today's optical lithography systems. In order to write one layer per minute in 45nm technology node, a throughput of 12Tb/s using 5-bit grayscale data is needed. With EUV light source flash rates limited to below 10kHz, 240 million 1μm x 1μm micromirrors have to be integrated on the writer chip, each driven with 32 possible voltage levels.

This paper explores the system design for various wafer throughputs, with or without data compression. In particular, the design tradeoffs for the mirror interface datapath, implemented on the same silicon die with the writers are discussed. The design of the digital-to-analog converters (DACs) that compensate for the nonlinearity of the mirror transfer function and fit into the required datapath pitch is presented. Extrapolated data from the designs in 0.13μm CMOS technology indicate that DACs will likely limit the throughput to about 30 wafers per hour in 45 nm node.

Keywords: Maskless, lithography, layout, digital-to-analog conversion, mixed-signal design.

1. INTRODUCTION

Mask costs have been rising at an increasing rate. As lithography advances and feature sizes continue to shrink, mask costs will become a major impediment for low-volume ASIC production. In the future, extreme-ultraviolet (EUV) will probably displace optical lithography and further increase mask costs. To tackle this problem, a maskless system has been proposed that can replace masks with an array of individually-controlled micron-scale mirrors [1]. The mirrors are electrostatically tilted to form a pattern and an EUV source reflects the mirror pattern onto the wafer. To completely expose the wafer, new patterns are electronically transferred to the mirror array as the wafer is mechanically scanned under the focused pattern.

The key constraint in the micromirror-array-based EUV lithography system is the achievable light source flash rate. Physical limitations will constrain the flash rate to be below 10kHz in the foreseeable future. High-density layouts require accurate edge positioning, where the features can be placed on a much finer grid than one defined by the lithography. To achieve a 1nm edge placement in the 45nm technology node, the data have to be represented as 5-bit values [2]. Limited EUV source frequency requires analog control of the mirror array, where 32 different analog voltage levels would be used to position the mirrors.

This paper explores the design of the mixed signal interface necessary for the analog control of the mirrors. The systems that are suitable for low-volume and high-volume production are compared, with and without data compression.

2. SYSTEM REQUIREMENTS

With the EUV flash frequency limited to below 10kHz, the system throughput is dictated by the total number of mirrors that can be integrated on the chip and by the speed of the interface electronics. Today's high-volume mask-based systems expose 60 wafers/hour. However, the total volume of a large number of IC designs is only one or a very few wafers, as in the case of prototyping, or in low-volume ASICs. Lower throughputs would be acceptable for manufacturing these products, of the order of 1-6 wafers/hour. We are considering two types of designs in this paper: a high throughput (30-60 wafer/hour) system and a low volume (1-6 wafer/hour) system.

Emerging Lithographic Technologies VIII, edited by R. Scott Mackay,
Proceedings of SPIE Vol. 5374 (SPIE, Bellingham, WA, 2004)
0277-786X/04/$15 · doi: 10.1117/12.536043

The required data throughput for the 60 wafer/hour system is given by:

$$\frac{wafer}{60s} \times \frac{\pi/4 \times (300mm)^2}{wafer} \times \frac{pixel}{(22nm)^2} \times \frac{5bits}{pixel} = 12Tb/s. \qquad (1)$$

Similarly, the 6 wafer/hour system has a throughput of 1.2Tb/s and the 1 wafer/hour system has a throughput of 200Gb/s. The EUV flash rate dictates the number of pixels that have to be exposed in each flash. As there are 144 trillion pixels in a 300-mm wafer, to achieve the throughput of 60 wafers per hour, each flash should expose 240 million pixels. The requirements are summarized in Table 1.

The required throughputs can be met with a combination of several techniques. High-speed I/O can deliver data at several hundred gigabits per second onto a chip. If the I/O speed is the limitation, it can be overcome by compressing the layout data off-chip. The data is then restored by on-chip decompression circuitry. Digital-to-analog converters convert 5-bit grayscale values to analog signals and load them in parallel to the mirrors. The mirrors can take 31 different deflections resulting in different illuminations on the wafer.

The mirrors will be fabricated over the control chip with low-temperature processing techniques that do not harm the CMOS circuitry. Analog memory cells will be electrically connected to the mirrors to enable charge storage. The mirrors have been designed to be around 1μm square. Integrating mirrors on the single chip dictates the structure of the mirror array. To integrate 240 million 1μm × 1μm mirrors on a single chip, an array of approximately 24,000 × 10,000 mirrors, of 2.4cm × 1cm in size has to be formed. The mirror control circuitry would be placed underneath the mirrors and the data would be loaded from two sides of the array. This array would have to be fully loaded with 5-bit analog values in between the EUV light flashes, each separated by about 100μs.

An all-digital system avoids the digital-to-analog conversion, and is therefore much simpler. The mirrors take only two extreme positions, where they either fully expose the wafer spot or fully deflect the light. The interface only handles the digital values. To modulate the edge positions, the same pixel on the wafer has to be exposed up to 31 times. With EUV source flash rate and the minimum size of the mirrors being the key limitations, this approach trades off the simplicity for a reduced throughput (less than 4 wafers/hr with 10kHz flash rate and minimum-sized mirrors). The digital system has a simple mirror interface, consisting only of the binary memories, such as standard SRAM, and provides more opportunities for implementation of redundancy.

To expose even more pixels per flash, or in order to use larger mirrors, multiple writer chips have to be built. Multi-chip solutions are possible, but require very precise mechanical alignment and require large interconnect complexity. The throughput requirements for these systems stay the same as in the single-chip systems.

3. SYSTEM IMPLEMENTATION

3.1. Conceptual System

The chip input bandwidth of approximately 800Gb/s is feasible in the next generation of CMOS technology [3]. This can meet the requirements of the low speed mixed signal system, up to 4 wafers/s excluding any redundancy in the data. The other designs require compression of the layout data off-chip and decompression on-chip [2].

Table 1: Number of pixels exposed in each source flash and required data throughputs for desired wafer throughputs at 10kHz exposure rate with 5-bit pixels.

Wafer throughput	60 wafers/hr	6 wafers/hr	1 wafer/hr
Pixels per exposure	240×10^6	24×10^6	4×10^6
Data rate	12Tb/s	1.2Tb/s	200Gb/s

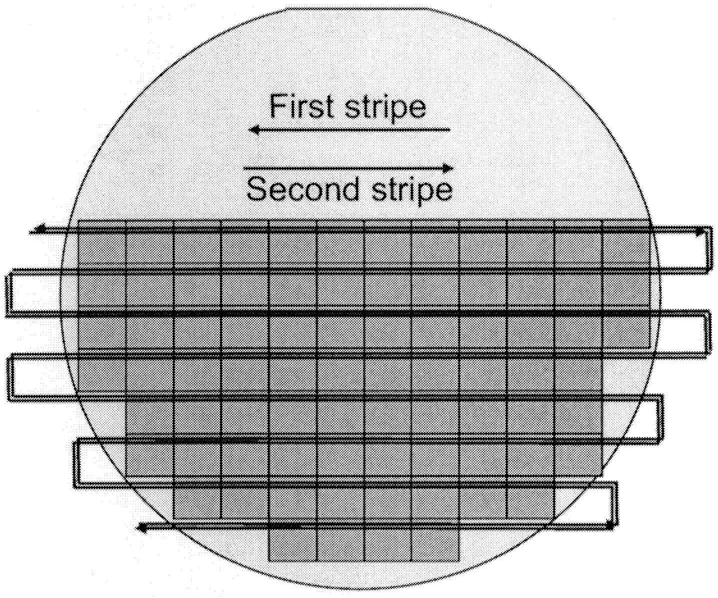

Figure 1: Concept of writing the repetitive data in strips across the wafer.

Each reticle in 45nm technology contains about 800GB (= 6.4Tb) of data per layer. A chip with more than 30 layers contains over 30TB of data, requiring storage on multiple file servers. Large file servers have a limited data throughput that they can output. The largest throughputs available from single disks are up to 2Gb/s, and physical limitations would prevent using massive parallelism, resulting in the maximum throughput below 10Gb/s. Therefore, there exists a throughput gap between the file server output and maximum input data rate to the chip.

The repetitive process of exposing reticles on the wafer presents a natural possibility to bridge this gap. There are approximately 100 reticles on a 300mm wafer, providing an opportunity to expand the throughput by two orders of magnitude. By introducing the interface board that would store the data downloaded from the file server, this data could be repetitively transferred to the mirror array. Since the amount of memory that can be present on the board is much less than 800GB, which is the amount necessary to store the whole reticle, the reticles would be exposing the area narrower than the reticle, as illustrated in Figure 1. With a realistic limit of about 4GB of memory per board, each board would be able to write a 0.1mm wide stripe over the wafer. At least two boards would be needed to achieve the continuous operation: while one is downloading the data from the file server, the other one would be repetitively exposing the wafer with the stored data. Multiple boards can be introduced for a faster throughput. The concept is illustrated in Figure 2, using 4 interface boards.

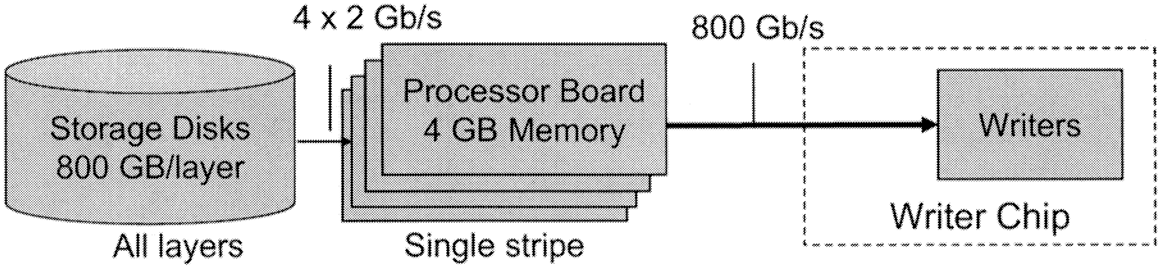

Figure 2: The writer system, that uses the processor board that interfaces to the writer chip.

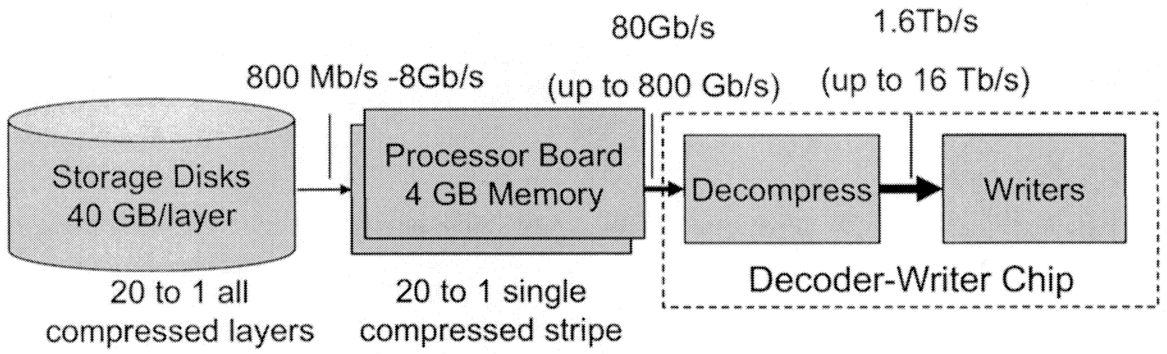

Figure 3: Concept of a system that stores the compressed data on the file server and decompresses the layout on the writer chip.

This conceptual system is feasible, but pushes various components to the limit, from the file server's capacity and the output data rate to the maximum chip input data rate. Introduction of data compression dramatically relaxes the demands on the system implementation. The example system shown in Figure 3 relies on the compression rate of about 20x to store the rasterized layouts of all chip layers on the file server. The volume of data is dramatically reduced: instead of the requirement for multiple file servers, it can fit on a single shelf of disk-drives. Two processor boards are used to interface to the writer chip. Depending on the targeted wafer throughput these boards could produce from 80Gb/s of data to 800Gb/s of data. The data is then decompressed on the writer chip. This is feasible, using a large number of parallel decompressing paths [4].

4. INTERFACE CIRCUITRY

Two key parts of the interface are the design of the digital-to-analog converters and the memory cells that will be storing the analog control voltages. To evaluate their feasibility, this work investigates their implementation in a 0.13μm CMOS technology, and relates it to the system design.

4.1. Digital-to-Analog Conversion

To achieve the required throughputs, it is necessary to integrate the digital-to-analog converters (DACs) on the writer chip. Using off-chip DACs would dramatically reduce the throughput because of the limited pin count. The intensity of light is a non-linear function of the voltage applied to the mirror, and the DACs have to be able to adjust to that. The requirements for the integrated DACs are:

1) the pitch of the DAC is about 1μm, equaling the pitch of the micro-mirrors,

2) the DAC is 6-bit, to allow for compensation for the nonlinearity in the mirror transfer function, and

3) the DAC is highly linear.

The most challenging feature is the implementation of a very narrow pitch layout for the DAC. To explore the feasibility of this approach, we implemented the design in a general-purpose 130nm CMOS technology. Assuming appropriate scaling in the matching requirements from 130nm design rules to 65nm design rules, we designed the DACs with a 1.5μm pitch. The memory cells are designed with size 2μm x 2μm. By scaling the feature sizes and the matching requirements in the future, a similar layout will fit a 1-μm pitch.

A simple voltage driver DAC is shown in Figure 4.a. An n-bit unit element DAC requires 2^n-1 identical current sources. The 2-bit DAC in the figure has 3 current sources, represented by a single transistor and switch. The voltage generated is a function of the output of each current source, I, the resistance, R, and the number of current sources switched on, M:

$$vout = M \cdot I \cdot R. \tag{2}$$

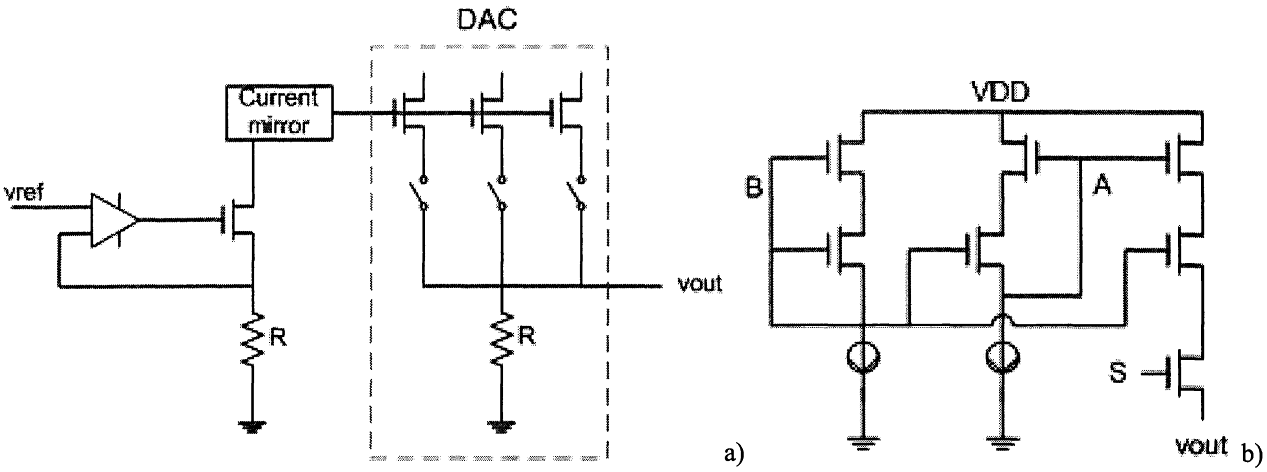

Figure 4: a) Simplified DAC, and b) bias network and single current source.

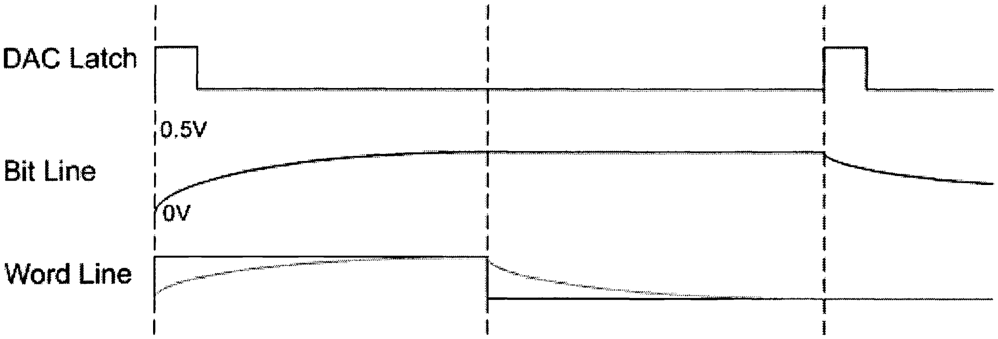

Figure 5: Timing diagram of the DAC.

The unit element DAC is highly linear if the output impedance of the current sources is high and the current sources are well matched. High output impedance is achieved with a cascode current source. A single high-impedance current source and an appropriate bias network are shown in Figure 4.b. The reference voltages *A* and *B* drive the sample current source, where the signal *S* switches the current source on or off. The high output impedance results in the current through the current source, *I*, being independent of the voltage at the output, *vout*, which is necessary for a linear response. The current *I* requires an external calibration system. This calibration system can be made relatively accurate by adjusting itself with an external voltage source that can be set precisely.

The linearity of a DAC is dependent on current source matching. In addition, if the current sources are well matched, a single calibration system can be shared across several DACs. The unit element DAC has a great advantage because the large number of current sources "average out" mismatch.

Even with perfect calibration, the accuracy of the DAC is still limited by interconnect delay. The bit line voltage exponentially approaches, but never actually reaches the DAC output voltage. Thus, the interconnect results in a finite error, dependent on the settling time *t* and time constant *RC*. For large micro-mirror arrays, this settling time will be the throughput-limiting constraint. Depending on the scaling of transistor and wire capacitances, the DAC will take several microseconds for each thousand of micro-mirors it drives. The timing of the DAC is demonstrated in Figure 5 and the narrow pitch layout design of one cascode current source is shown in Figure 6.

Most error in the system arises during the transfer of the analog values to the micro-mirrors. These errors are considered in the next subsection.

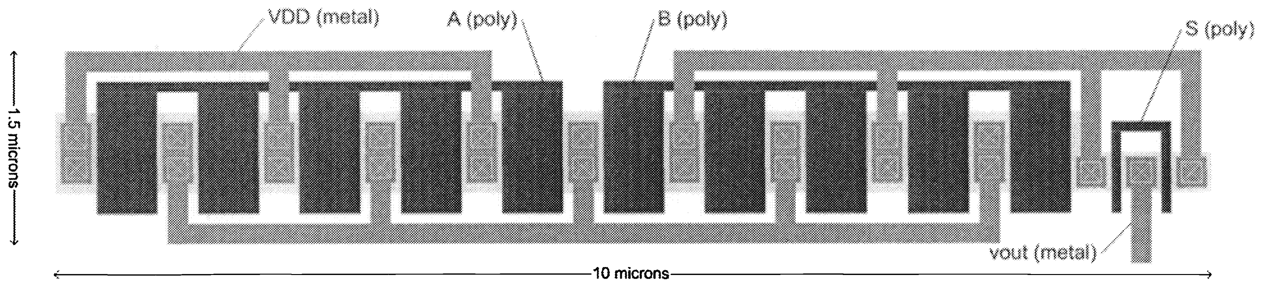

Figure 6: DAC current source layout.

4.2. Memory Cell

The second critical part of the interface is the analog memory cell that will be placed directly underneath the micro-mirrors. There are five sources of error on the memory cell: sampling noise, word line feedthrough, charge leakage, charge injection, and bit line coupling.

The noise on the memory cell is sampling noise, expressed as kT/C, where k is Boltzmann's constant, T is temperature, and C is capacitance. For the projected capacitances and the required 6-b precision with 1V supplies, this noise is in the microvolt range and can be safely ignored.

Capacitive coupling between the word line and the memory cell can cause voltage feedthrough. At the end of a writing cycle, the word line voltage falls to turn off the sample transistor. As it undergoes the 1-to-0 transition, capacitive coupling between the word line and memory cell tends to drop the voltage on the memory cell. This drop is significant, about 4%, corrupting the least significant bit. In order to compensate for this effect, an inverse word line is placed in parallel in the layout adjacent to the word line. In simulations, coupling from both lines cancels out to a better than 7-bit level.

A complete memory array write cycle is 100μs with a 10kHz EUV source. The DAC feeds all the cells in one row, in a sequential order. Consequently, the first memory cells written must retain their precise values for that time period. However, charge may leak from the memory cell to the bit line or from the bit line to the memory cell, causing a significant change in voltage. There are two mechanisms explored in this project to reduce charge leakage: body bias and negative gate voltage. The leakage current through a MOS transistor drops exponentially as a function of threshold voltage [5]. Imparting a body bias increases the threshold voltage and thereby decreases leakage. By raising the threshold voltage, leakage drops. However, due to reduced effects of body bias in deeply scaled CMOS technology, the leakage reduction from negative body bias does not satisfy the design constraints. Lowering the gate voltage below zero also decreases leakage. A negative gate voltage reduces leakage substantially (to about 0.4% of the full-scale voltage over 100μs). This leakage may be corrected by algorithms that pre-correct the data for this effect, by knowing a-priori the sequence of writing the cells.

When the memory cell transistor is on, a small charge is present in the depletion region. After the transistor switches off, that charge exits the transistor through the source or drain junction. In the maskless system, this charge would be injected either to the memory cell or onto the bit line. It is very difficult to predict or model which direction the charge will take; it depends in part on the relative impedance of the two pathways. The worst-case error model assumes that the entire charge is injected on the memory cell. The resultant change in voltage on the memory cell is a function of the ratio between the sample transistor gate capacitance and the memory cell capacitance. This ratio is approximately 1 to 100. Thus, charge injection could result in an error of up to 1%.

During the complete memory array write cycle, the bit line adjacent to a memory cell will frequently change values as other memory cells along the bit line are written. A major design goal is to protect the memory cell from capacitive coupling. A change in the adjacent word line (WL) or bit line (BL) induces a change in the stored value on the memory cell (C_{mem}). For example:

$$\Delta V_{mem} = \Delta V_{BL} \frac{C_{BL-mem}}{C_{BL-mem} + C_{mem}}. \tag{3}$$

The equation demonstrates that a change in voltage on the BL results in a voltage change on the memory cell. This change is a function of both the coupling capacitance between the BL ($C_{BL\text{-}mem}$) and memory cell and the capacitance of the memory cell (C_{mem}). If $C_{BL\text{-}mem}$ is small relative to C_{mem}, the capacitive coupling will be small. However, the changes in the BL have a cumulative effect on the memory cell:

$$\Delta V_{mem} = \sum \Delta V_{BL} \frac{C_{BL-mem}}{C_{BL-mem} + C_{mem}}. \tag{4}$$

Thus, if the BL changes many times, the effect on the memory cell will not be negligible. The conceptual design of the 60 wafer/hour system reveals that this is the case. Each row of the array has 10,000 memory cells and each row is written from two sides. The driver for each side, therefore, writes 5,000 memory cells every cycle. After an analog value is stored in the first cell, the BL running over that cell will fluctuate 4,999 times before the end of the cycle. By that time, capacitive coupling could result in a random value stored on the memory cell.

Equations 3 and 4 show that coupling can be reduced by maximizing C_{mem} and minimizing $C_{BL\text{-}mem}$. The design uses the gate capacitance of an NMOS transistor as the memory capacitance. The drain and source are connected to ground. Both calculations and subsequent simulations reveal that a capacitance of up to 10fF can be made in the area under a micromirror. For a maximum ΔV_{mem} of 2%, and assuming normal changes in the BL, the maximum allowable $C_{BL\text{-}mem}$ is around 0.1aF, which is very low.

The effect of bit-line coupling can be eliminated almost completely. Equation 4 indicates that if an inverse delta is applied to the bit line for every delta, the net change on the bit line will be zero. This can be done by simply draining the bit line after each write cycle.

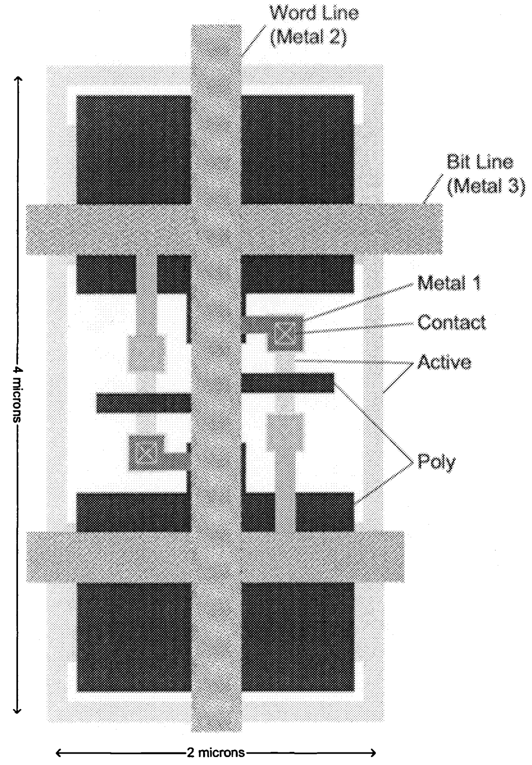

Figure 7: Layout of the two memory cells.

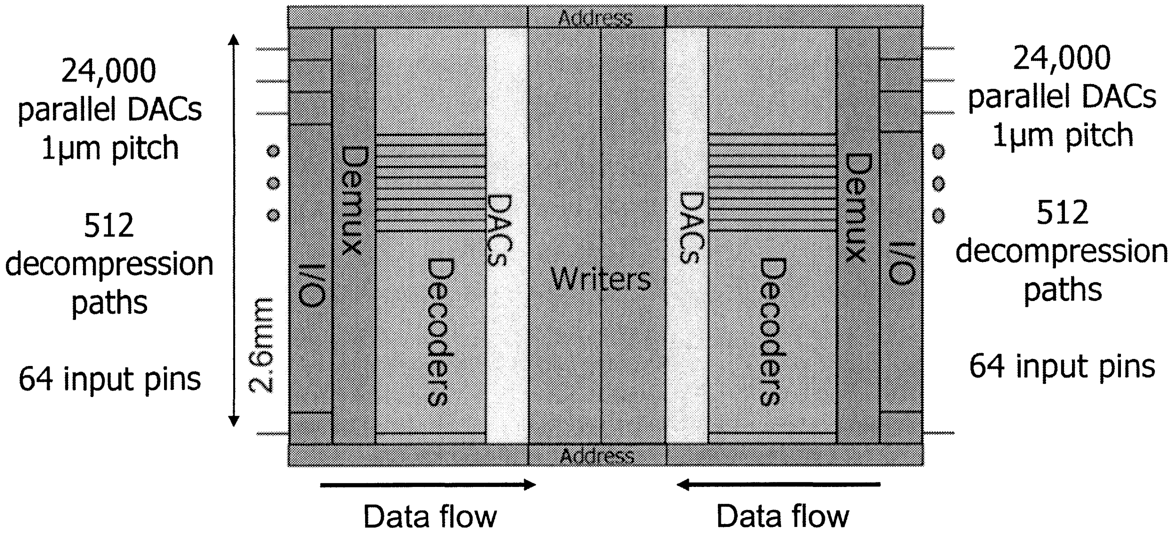

Figure 8: Example high-throughput maskless writer chip.

4.3. Data Decompression

Data compression can be used to increase the effective data throughput and reduce the volume of stored and transferred data. The data is rasterized and compressed off-line, before being stored. The data would be decompressed on the writer chip, before the digital-to-analog conversion, to overcome the chip I/O bottleneck. The implementation of the on-chip data decompression must support the required throughputs, and must be area and power efficient. The area and power budget on the chip are very limited because of integration with high-speed data I/O, digital-to-analog converters, and the large mirror array. To achieve the required throughput a large number of parallel decompression paths have to be used.

In the related work [4] we investigate the implementation of the lossless data compression and decompression. Lempel-Ziv (LZ77) decoding algorithm can be designed to match the required pitch of the DACs and micro-mirrors.

5. EXAMPLE SYSTEMS

As derived in Section 2, the one wafers per minute maskless lithography system has a data throughput requirement of approximately 12Tb/s. To achieve this throughput, the layout data has to be compressed on the storage disks, and then decompressed on the writer chip. The mirror array size is 240 million mirrors. This can be implemented as an array of 24,000 mirrors by 10,000 mirrors. The writer chip will have integrated 48,000 parallel digital-to-analog converters on both sides of the mirror array. The maximum wire length from DAC to the mirror element is about 5mm. The requirement for writing precise analog values is the key limitation of this system, limiting the throughput to several tens of wafers per hour.

The four wafer per hour analog system has a data throughput requirement of 800Gb/s. This requirement can be met with high-speed I/O, thus avoiding the need for decompression hardware. The 16 million mirrors can be implemented as an array of 24,000 by 670 mirrors. The wire length is reduced to less than 1mm. With such short wires, the rate of loading analog values onto the mirrors is no longer limited by RC wire delays, and even fewer DACs could be used to achieve the same rate.

In this study, it is assumed that individual transfer characteristics of each mirror are time invariant, have been characterized in advance, and are incorporated in the layout image stored on the file server. Also, the layour is pre-compensated for any possible non-functional mirrors.

6. CONCLUSIONS

This paper demonstrates several different data interfaces for maskless lithography systems with various throughputs. A high speed system is feasible with a comparable wafer throughput to the conventional lithography. Similar, but lower wafer throughput systems are feasible with reduced design constraints. The most challenging component of the interface is the digital-to-analog converter which needs to write 32 different voltage levels to an array of pitch-matched micro-mirrors, between the EUV light flashes.

ACKNOWLEDGEMENT

This research is conducted under the Research Network for Advanced Lithography, supported jointly by the funding of the Semiconductor Research Corporation (01-MC-460) and the Defense Advanced Research Project Agency (MDA972-01-1-0021).

REFERENCES

[1] Y. Shroff, Y. Chen, W.G. Oldham, "Fabrication of parallel-plate nanomirror arrays for EUV maskless lithography", *Journal of Vacuum Science and Technology*, Nov. 2001.

[2] Vito Dai, Avideh Zakhor, "Lossless Layout Compression for Maskless Lithography" in *Emerging Lithographic Technologies IV, Proceedings of SPIE*, Vol. 3997, 467-477, 2000.

[3] Rambus Redwood Interface, http://www.rambus.com/products/redwood/

[4] B. Nikolić, B. Wild, B. Warlick, V. Dai Y. Shroff, A. Zakhor, W.G. Oldham, "Lossless compression techniques for maskless lithography data" in *Emerging Lithographic Technologies VI, Proceedings of the SPIE*, vol. 5374, San Jose, California, February 2004.

[5] J.M. Rabaey, A. Chandrakasan, B. Nikolić, *Digital Integrated Circuits: A Design Perspective*, Prentice-Hall, 2003.