GPU-based MRC Methods for Overlapping eBeam Shots

Kokoro Kato, Ph.D.
Hitachi High-Tech Science Corporation
1. Model Based MDP (MB-MDP)
2. GPU computing and its benefits
3. Key points of GPU computing
4. Experimental results
5. Conclusion
1. Model Based MDP (MB-MDP)

- Model Based MDP (MB-MDP) has been used for EB shot reduction and speed-up of mask writing time.
- By allowing overlapping shots, EB shot count can be significantly reduced.

From http://www.ebeam.org/home
Problem of overlapping shot verification

- Usually, EB shot data verification is done before mask writing
  - This process is called Mask Rule Check (MRC)
- However, MB-MDP shots are not identical to mask target. So, new methods are needed to verify the MB-MDP shots.

![Diagram showing the relationship between wafer target, mask target, conventional shots, and MB-MDP shots, highlighting the new necessity to verify between MB-MDP shots and mask target.]
In this presentation, we introduce two methods for verification of overlapping EB shots using GPGPU.

- Comparison of simulated contour images between conventional shots and MB-MDP shots
- Overdose detection

**Conventional shots**

**MB-MDP shots**

**Simulation image by GPU**

Differences (red) are errors

Overdose error

Threshold
2. GPU computing and its benefits

- One of the greatest merits of GPU computing is its speed.
  - Currently many technologies of high speed computation based on GPUs have been reported.
  - Especially, an excellent program development environment called CUDA enables to build such high performance programs easily.

CPU has a single cache memory and controller and no more than 4 or 6 cores. On the other hand, GPU can have multiple cache memory and controllers, and more than one hundred cores.
Another major merit of GPU is low power consumption.

- Comparison of GFLOP per Watt
  - GPU is much better than CPU

Computational lithography technologies need great amount of CPU powers.

- Total reduction of power consumption is needed.

In this study, we develop GPU computing techniques for faster and greener MRC.
3. Key points of GPU computing

- Need to pay high attention to optimal memory usage.
  - Shared memory – small (16k), but very fast
  - Global memory – large (1G), but very slow

- We applied multiple Gaussian filtering method for energy distribution calculation.
Contour generation

Contour generation flow

- Smoothing is needed

(a) Energy distribution and cut line
(b) Intensity at cut line
(c) Generated contour lines

Before smoothing

After smoothing
We use the API (Application Program Interface) function of SmartMRC as the fundamental framework of the software development and evaluation.

By using the API function of SmartMRC, programmers can access to the mask layout data easily without knowing the details of the mask data format.
4. Experimental results

- We evaluated new MRC system with 2 types of data.
- The specs of the system is as follows:

<table>
<thead>
<tr>
<th>CPU</th>
<th>Intel® Xeon® CPU E5645 @ 2.40GHz X 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core number</td>
<td>24 logical cores</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPU</th>
<th>Tesla C2050 / C2070</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core number</td>
<td>448 CUDA Cores</td>
</tr>
<tr>
<td>Global-memory size</td>
<td>2.64 Gbyte</td>
</tr>
<tr>
<td>GPU Clock-speed</td>
<td>1.15 GHz</td>
</tr>
<tr>
<td>Constant-memory size</td>
<td>64 Kbyte</td>
</tr>
<tr>
<td>Shared-memory per block</td>
<td>48 Kbyte</td>
</tr>
<tr>
<td>Register count per block</td>
<td>32,768</td>
</tr>
</tbody>
</table>
Dumbbell type data (DB)

- Dumbbell type data – array of dumbbell shape with height=400nm and width= 40/160nm.

<table>
<thead>
<tr>
<th>Type</th>
<th>Figure count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>125,000</td>
</tr>
<tr>
<td>Overlapping</td>
<td>875,000</td>
</tr>
<tr>
<td>Conventional</td>
<td>6,875,000</td>
</tr>
</tbody>
</table>

(a) Input

(b) Conventional shots

(c) Overlapping shots
Angled-line type data (AL)

- Dumbbell type data – array of lines with angle=30degree and width=100nm.

<table>
<thead>
<tr>
<th>Type</th>
<th>Figure count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>5,600</td>
</tr>
<tr>
<td>Overlapping</td>
<td>417,600</td>
</tr>
<tr>
<td>Conventional</td>
<td>974,400</td>
</tr>
</tbody>
</table>

(a) Input

(b) Conventional shots

(c) Overlapping shots
Results of DB

(a) Overlapping shots
(b) Conventional
(c) Zoomed image
(d) Difference
(e) Difference zoomed
(e) Overdose detection

Result images

<table>
<thead>
<tr>
<th>Type</th>
<th>Processing time</th>
<th>Power consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>1,258 s</td>
<td>100 Wh</td>
</tr>
<tr>
<td>CPU</td>
<td>6,130 s</td>
<td>340 Wh</td>
</tr>
</tbody>
</table>

Results

- 5 times faster
- 3.4 times smaller power consumption
Results of AL

<table>
<thead>
<tr>
<th>Type</th>
<th>Processing time</th>
<th>Power consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>1,258 s</td>
<td>130 Wh</td>
</tr>
<tr>
<td>CPU</td>
<td>5,460 s</td>
<td>300 Wh</td>
</tr>
</tbody>
</table>

-4 times faster
- 2.3 times smaller power consumption
5. Conclusion

- We have proposed novel verification methods for MB-MDP shots using GPU computing techniques.
- We have achieved X4-5 faster speed and X2.3-3.4 smaller power consumption compared to conventional CPU methods.

Future plans

- System development for actual mask production
- Further improvement of basic algorithm for faster calculation

This study is a joint work of AIST (Advanced Industrial Science and Technology) and Hitachi High-Tech Science Corporation.