Virtual elastic computing use case

Daria Negri, OMD
Evgeny Ostrovsky, OMD
Yuval Lubashevsky, OMD
Amnon Manassen, OMD
Rajiv Gupta, Corporate IT
Joris Vandermeulen, Corporate IT

Metrology challenges require an electromagnetic behavior analysis of customer wafers at various process stages. Moreover, detailed study of process variations and its effects on optical properties of our measurement targets are vital for selection of technological means. Electromagnetic simulations require high computation power, which is not always available. This paper represents simulation cycle time optimization by using elastic computing.

I. Introduction

Nowadays metrology challenges require in-depth analysis of physical behavior of customer wafers at various process stages. Detailed electromagnetic simulations of various target parameters are required to predict the light matter interaction of the metrology process.

The simulations we use require high computation power that is increased continuously. In general, we optimized the available computation power to our needs. However, sometimes there are tasks that require a much faster response than what can be achieved using full capacity of the available resources. For these cases the most rapid and cost effective solution we found is performing computations in parallel by using elastic computing – EC2.

II. Parallel computing by EC2

Detailed study of process variations and their effect on optical properties of our measurement targets require tens of simulation runs with varying independent parameters. Run-time estimation based on existing computation capabilities of the specific short-term tasks is presented in Fig. 1 by the green and purple curves. The occasional rapid response simulations mentioned above are described by the red.

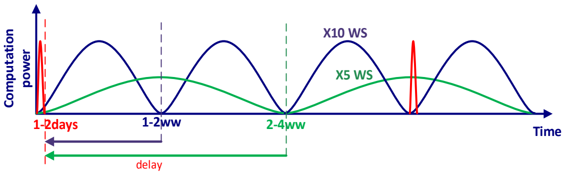

Fig. 1: Typical simulation workflow. The green and the purple curves describe the response time based on x5WS, x10WS respectively. The red curve represents the short-term simulation peak required from time to time.

In order to use EC2 for handling the spikes several challenges had to be addressed.

The VMs were opened and prepared for use by KT IT team.

The electromagnetic simulations are

Matlab based, therefore in order to avoid multiple licenses we compiled the simulator to execute version and run it on the Virtual Machines (VMs) through a single primary desktop computer as described in Fig.2.

To enable the execute version run, each VM should have the appropriate version of MATLAB Runtime application (free and available on Mathworks site) that includes standalone set of shared libraries that enables the execution of compiled MATLAB. The remote activation of the simulator was done through cmd on the VMs, which is activated by the MATLAB session on the primary desktop. The output of the simulations is automatically sent to the primary desktop, so data storage on VMs is avoided.

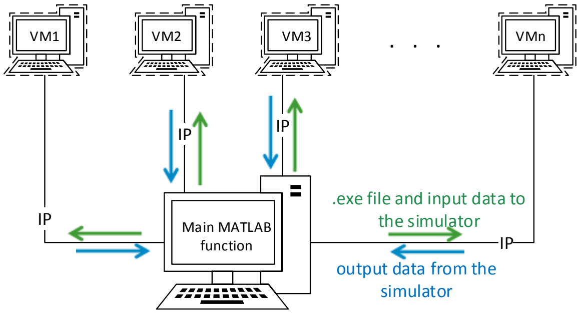
VM1
VM2
VM3
. . .
VMn
IP
IP
IP
IP
Main MATLAB function
.exe file and input data to the simulator
output data from the simulator
IP

Fig. 2: Remote communication scheme between the main desktop and the virtual machines through IP address.

The sequence of the work included the following steps:

1. Risk identification – lack of computation power for the required simulation to meet the challenging schedule.

2. Definition of the necessary resources for parallel run to accomplish the peak simulations on time.

3. VM configuration – by IT group and enabling the option of RUN/STOP to the user.

4. MATLAB Runtime installation on all of the VMs.

5. Compilation of the simulator to execute version.

6. MATLAB code generation that remotely connects the VMs, send the simulator compiled version with the required input data and executes it.

7. When the simulations are finished, turn the VMS from RUN to STOP mode. This option is preferred over the TERMINATE mode because the VMs remain available for the user with minimal fee and step #4 can be skipped for the next time.

III. Conclusions

The advantage using elastic computing can be expressed by comparing the runtime of simulation process. Without using EC2 the process would typically takes 6-7 days and by using EC2 the process takes 0.5-1 day. Due to reduction in cycle time OMD response time for analyzing issues requiring complex simulations were reduced considerably.

EC2 availability in KT as an IT service

Amazon Web Services elastic cloud computing (EC2) is on-demand delivery of compute resources with pay-as-you-go pricing. Amazon EC2 reduces the time required to obtain and boot new Windows and Linux instances to minutes, allowing KLA Tencor to scale capacity both up and down, as the computing requirements change. There are many EC2 instance types offered by Amazon that range from fractional core up to 36 cores per instance. These instances can then be grouped together over a low latency network connection to form a horizontally scaled super compute cluster ranging to several hundred or thousand cores. Pricing for the compute instances are hourly and we only pay for the duration compute resources are running.

KLA Tencor has securely extended Leuven DataCenter to Amazon Web Services. Virtual machines provisioned in AWS cloud get a KLA Tencor IP address and can be seamlessly joined to KLA Tencor Active Directory for security and manageability. Additionally, these instances can only be accessed while on KLA network and cannot be accessed from the Internet. From an end user functionality/ performance perspective they will not notice a difference if Windows and Linux instances are running at KLA Datacenter or in AWS cloud. After IT provisions the compute resources, Engineers can start and shutdown the instances via Self Service to keep the costs in control.