Reducing Spread-sheet Calculation Time by using High Performance Computing Cluster

May Myint Thwe*

Abstract

High-performance computing (HPC) encompasses advanced computation over parallel processing, enabling faster execution of highly compute intensive tasks. HPC cluster is an interconnected configuration of servers, typically having 2 or more processing cores per server or node. HPC uses a cluster of servers to split complex computations into smaller tasks that can be calculated in parallel on multiple servers. HPC Services for Excel supports a variety of new programming models to run Excel calculations on an HPC cluster. HPC Services for Excel include service-oriented architecture (SOA) clients and services that enable developers to quickly convert workbooks to run on a cluster. When a workbook runs iteratively, the best option for parallelizing of the calculation is to run the entire workbook in the cluster. Workbooks in Microsoft Excel not only record and display data but also perform computations and simulate complex models. However, as these computations and models increase in size and complexity, the time required for calculation grows and workbook use becomes cumbersome due to long waiting for calculation results. The HPC system is developed to provide users speed-up in Excel calculations on it by reducing in size, complexity and calculation time. This work describes how an HPC cluster can be utilized to run Excel calculations in parallel. Excel includes Visual Basic for Applications (VBA) language. VBA is the tool to develop programs that control Excel. The system is operated by Windows® HPC Server 2008 R2 and Office Excel 2010 process is used to calculate the workbook in the cluster.

Keywords: High Performance Computing, HPC cluster, Windows® HPC Server 2008 R2, SOA clients, Microsoft Office Excel 2010

Introduction

High Performance Computing (HPC) is a branch of technical computing that uses supercomputers and cluster computing to solve computationally intensive problems. HPC is the use of parallel processing for running advanced application programs efficiently, reliably and quickly. The key to any HPC solution is parallelization. Parallelization refers to split a complex computation into component parts that can be run simultaneously (or in parallel) on multiple servers. Large problems are to divide into smaller parts and distribute among the other many computers. Cluster computing involves utilizing the physical resources of multiple computers in order to perform a computationally intensive task. Clusters are just multiple computers connected together to solve a specific problem. A cluster has three basic elements—a collection of individual computers, a network connection to those computers, and software that enables a computer to share work among the other computers via the network.[1, 3]

To run an Office Excel 2010 workbook in a Windows HPC cluster, some part of the calculation must be able to run in parallel. The actual performance improvement in the application will depend on how many calculations can be parallelized, the overhead involved in the cluster calculation, and the size of the cluster and its available resources. Windows HPC Server 2008 R2 enables running multiple instances of Office Excel 2010

* Lecturer, Department of Computer Studies, Yadanabon University

in a Windows HPC cluster, where each instance is running an independent calculation or iteration of the same workbook with a different dataset or parameters. Windows HPC Server 2008 R2 also supports one or more Windows Communication Foundation (WCF) Broker nodes, which enable running WCF services in the cluster. WCF enables and simplifies building service-oriented applications and Windows HPC Server 2008 R2 supports Service Oriented Architecture (SOA) model that enables running WCF services in a Windows HPC cluster. WCF and SOA provide a simple framework for building service applications to run on a cluster.

In a SOA system, distinct computational functions are packaged as software modules called *services*. Developers can create cluster-SOA client applications to provide access to services that are deployed to a Windows HPC cluster. The client application submits a job that contains a Service task to the cluster, initiates a session with the broker node, and sends service requests and receives responses (calculation results). This system addresses how the calculation time required for long-running workbooks can be reduced to give users faster access to critical information.

Developing HPC Cluster for Excel

High Performance Computing(HPC)

HPC is used to describe computing environments which utilize supercomputers and computer clusters to address complex computational requirements, support applications with significant processing time requirements, or require processing of significant amounts of data. Supercomputers have generally been associated with scientific research and compute-intensive types of problems, but more and more supercomputer technology is appropriate for both compute-intensive and data-intensive applications. Supercomputers utilize a high-degree of internal parallelism and typically use specialized multi-processors with custom memory architectures which have been highly-optimized for numerical calculations[2]. Supercomputers also require special parallel programming techniques to take advantage of its performance potential. Today a higher-end desktop workstation has more computing power than the supercomputers which existed during the early 1990's. This has led to a new trend in supercomputer design for high-performance computing: using clusters of independent processors connected in parallel[5]. Many computing problems are suitable for parallelization, often problems can be divided in a manner so that each independent processing node can work on a portion of the problem in parallel by simply dividing the data to be processed, and then combining the final processing results for each portion. This type of parallelism is often referred to data-parallellism, and the data parallel applications are a potential solution to petabyte scale data processing requirements[1, 3, 4]

Hardware Development

HPC uses a cluster of servers to split complex computations into smaller tasks that can be calculated in parallel on multiple servers. A Windows HPC cluster consists of several networked servers.

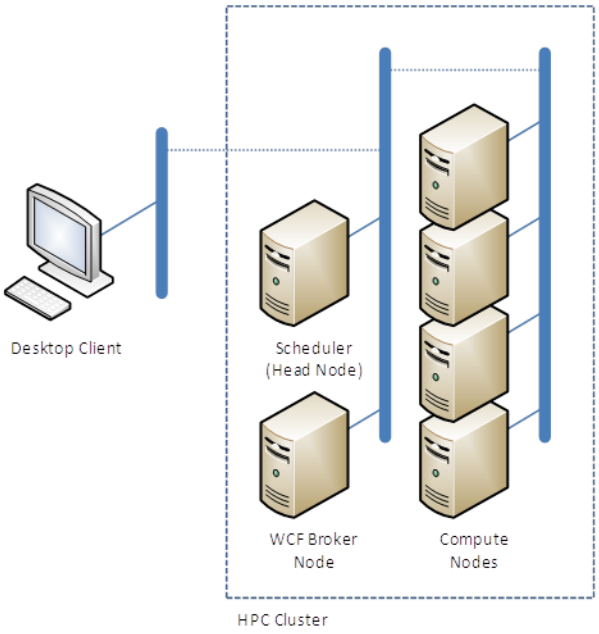

Figure 1. Windows HPC cluster topology

Each server in the cluster performs one or more specific roles. A Windows HPC cluster uses a Scheduler (hosted on a *Head node)* to manage computation tasks and Compute nodes (that is, servers) to perform the actual work. In a typical cluster calculation, a client computer—such as a desktop computer running the Windows operating system—connects to the Scheduler over the network to submit a job. The Scheduler handles the work of managing the Compute nodes, locating available resources, and running the computation. When the Compute nodes complete units of work, the Scheduler sends calculation results to the client computer.

Software Requirements

Windows® HPC Server 2008 R2

Windows® HPC Server 2008 R2 integrates with Microsoft® Excel® 2010 to help to run workbooks and user-defined functions (UDFs) faster by offloading calculations to the cluster. If a workbook contains independent units of calculation, multiple compute nodes can perform the calculations simultaneously. Parallel computation can significantly reduce workbook calculation time and make calculations across larger data sets more feasible. Many complex and long-running workbooks run iteratively—that is, they perform a single calculation many times over different sets of input data. These workbooks might contain complex Microsoft Visual Basic for Applications (VBA) functions or calculation intensive XLL Excel Library add-ins. This type of workbook is suitable for cluster acceleration.

Windows HPC Server 2008 R2 now enables running multiple instances of Office Excel 2010 in a Windows HPC cluster, where each instance is running an independent calculation or iteration of the same workbook with a different dataset. Many complex and long-running workbooks run *iteratively*—that is, they perform a single calculation many times over different sets of input data. These workbooks might include intensive

mathematical calculations contained in multiple worksheets, or they might contain complex Microsoft Visual Basic for Applications (VBA) functions.

When a workbook runs iteratively, the best option for parallelizing the calculation is to run the entire workbook in the cluster. In this model, individual calculations need not to be split into component parts, but the overall calculation—generating the results from many individual calculations—can be run in parallel.

Every application that benefits from this solution has three parts: the *workbook*, a *service,* and a *client.* Office Excel 2010 and Windows HPC Server 2008 R2 must be installed on each cluster server. Microsoft Office Excel 2007 or Office Excel 2010 must be installed on the client computer. The client computer can run the Windows XP with Service Pack 3, Windows Vista®, or Windows 7 operating system.

The workbook

The workbook refers to a standard Excel workbook. This solution runs multiple instances of Office Excel 2010 on cluster servers, meaning that it supports workbooks that use VBA or XLL add-ins or an Excel Add-in (XLA) as well as external resources (provided these resources are accessible from the servers).

In some cases, workbooks may need to be modified to work with this solution. When Office Excel 2010 runs on the server, it does not support user interaction. Windows HPC Server 2008 R2 includes a comprehensive pop-up manager that can handle occasional dialog boxes and pop-up messages, but it is not designed to support interactive Office Excel 2010 features: Users cannot create a PivotTable when running on the server, for example, because doing so requires user interaction.

When a workbook is used in this scenario, it is important to identify the input values and the output or result of the calculation. The input values might be cells within a worksheet in which the user enters a value, or they might be parameters to a VBA function. The output might be a second set of cells within a worksheet or the result of a VBA function. Identifying the workbook input and output values are important when developing the service, which will run on the cluster servers and execute the workbook calculation.

The Service

The service is a WCF service that controls the execution of the Office Excel 2010 workbooks on the cluster servers. The service starts Office Excel 2010, calculates a workbook, and returns results.

Windows HPC Server 2008 R2 offers an SOA model that enables running WCF services in the cluster. In a Windows HPC cluster, the WCF Broker node handles managing and hosting the service library. The Scheduler handles assigning and managing compute resources. From the standpoint of the developer, using WCF removes all the complexity of hosting and managing the service. The developer need only build the calculation functions.

Once the service has started Office Excel 2010 on a server, it can interact with the workbook in either of two ways to provides access to the Excel object model (similar to using Excel automation on a client computer): Using the Excel object model, the service can read from and write to the workbook, and it can trigger calculation of the workbook.

For example, in a simple workbook, the service might write values into some input cells, recalculate the workbook, and then read the values of some output cells.

The Client

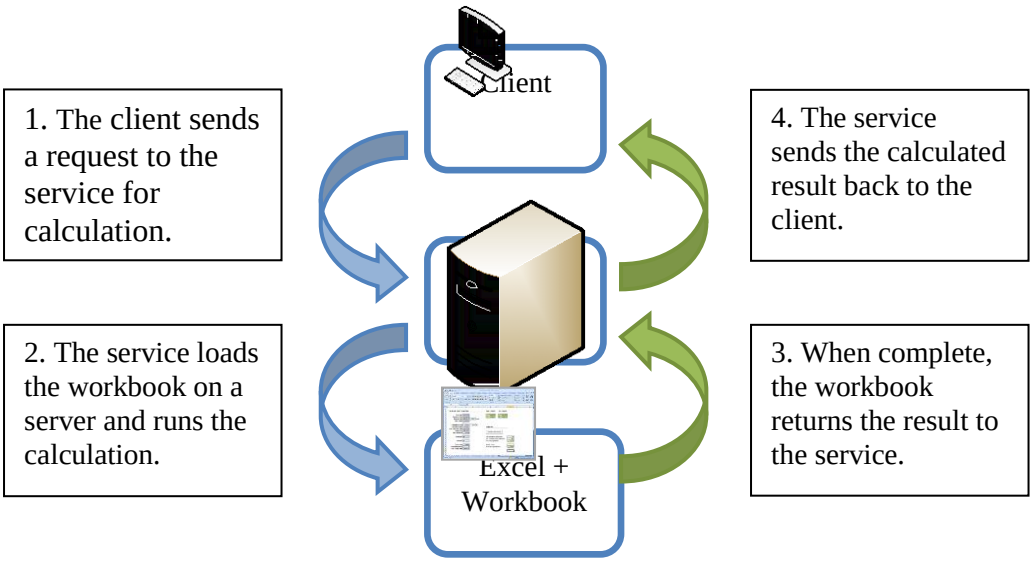

Figure 2. Application flow for a single service call of running workbooks on a cluster

The client is a program that controls the overall calculation. It is designed to work with the WCF service, and it tells the service which workbook to calculate along with the parameters or options to use, and it receives results when the calculation is complete.The client typically runs on a user's client computer. Client programs can be written in any language supported by Microsoft .NET or COM, and the client can be a Windows application or a command-line program.

Service-Oriented Architecture

Service oriented architecture (SOA) is an approach to building distributed, loosely coupled systems. HPC Services for Excel uses a SOA infrastructure to run Excel jobs on a cluster. HPC Services for Excel includes ready-made SOA clients and services that enable developers to quickly convert workbooks to run on a cluster. The HPC Services for Excel features support a number of programming models that enable Excel workbooks to run in parallel on a cluster. The features include SOA clients and services that enable developers to quickly convert workbooks to run on a cluster. In a SOA system, services can be distributed across a network and accessed by other applications. A *client* application provides an interface for the end-user to access the functionality of one or more services. The client-side and developer features are available in the Microsoft® HPC Pack 2008 R2 client utilities and software development kit (SDK). The cluster-side features are available with the Enterprise Edition of HPC Pack 2008 R2.

SOA Clients

A client application provides an interface for the end-user to access the functionality of one or more services. Developers can create cluster-SOA client applications to provide access to services that are deployed to a Windows HPC cluster. The client-side and developer features are available in the Microsoft® HPC Pack 2008 R2 client utilities and software development kit (SDK). The cluster-side features are available with the Enterprise Edition of HPC Pack 2008 R2. On the back end, the client application submits a job that contains a Service task to the cluster, initiates a session with the broker node, and sends service requests and receives responses (calculation results).

SOA Services

SOA runs on the cluster as shown in Figure 3. It is a software design and software architecture design pattern based on structured collections of discrete software modules, known as services that collectively provide the complete functionality of a large software application. In a SOA system, distinct computational functions are packaged as software modules called *services*. Services can be distributed across a network and accessed by other applications. If applications perform repeated parallel calculations, the core calculations can be packaged as services and deployed to a cluster. This allows developers to solve embarrassingly parallel problems without rewriting the low-level code and to rapidly scale out applications. Applications can run faster by distributing core calculations across multiple service hosts (compute nodes). End-users run the application on their computers, and cluster nodes perform calculations. The job scheduler on the head node allocates resources to the service job according to the job scheduling policies. An instance of the **Service** task runs on each allocated resource and loads the SOA service. The job scheduler tries to adjust resource allocation based on the number of service requests.

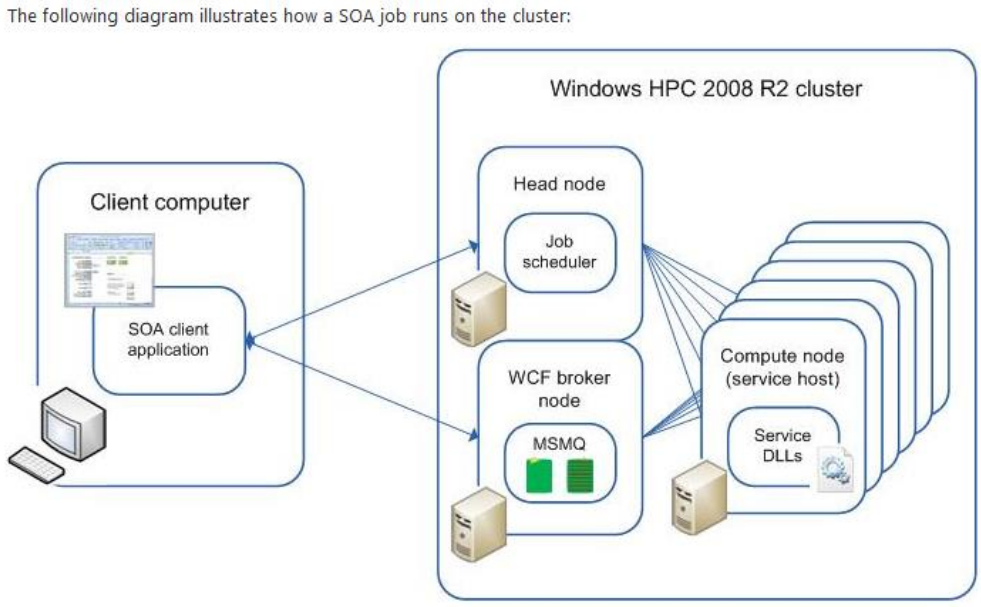

Figure 3. SOA job running on the cluster

Experimental Environment

The test platform was built on a cluster with one head node, one broker node and three compute nodes. The system was operated by Windows® HPC Server 2008 R2 and Office Excel 2010 process was used to calculate the workbook in the cluster. A spreadsheet application with VBA was performed on HPC cluster to calculate the interest on saving deposit of bank for one year. There are 2 command buttons such as *Local Run* and *Run HPC* to call the macros. Input boxes such as cells of D5 to D505 (Kyats amount) and cells of E4 to P5(months) were filled with the values to be calculated. The cells of Interest_Table are used to output calculation results.

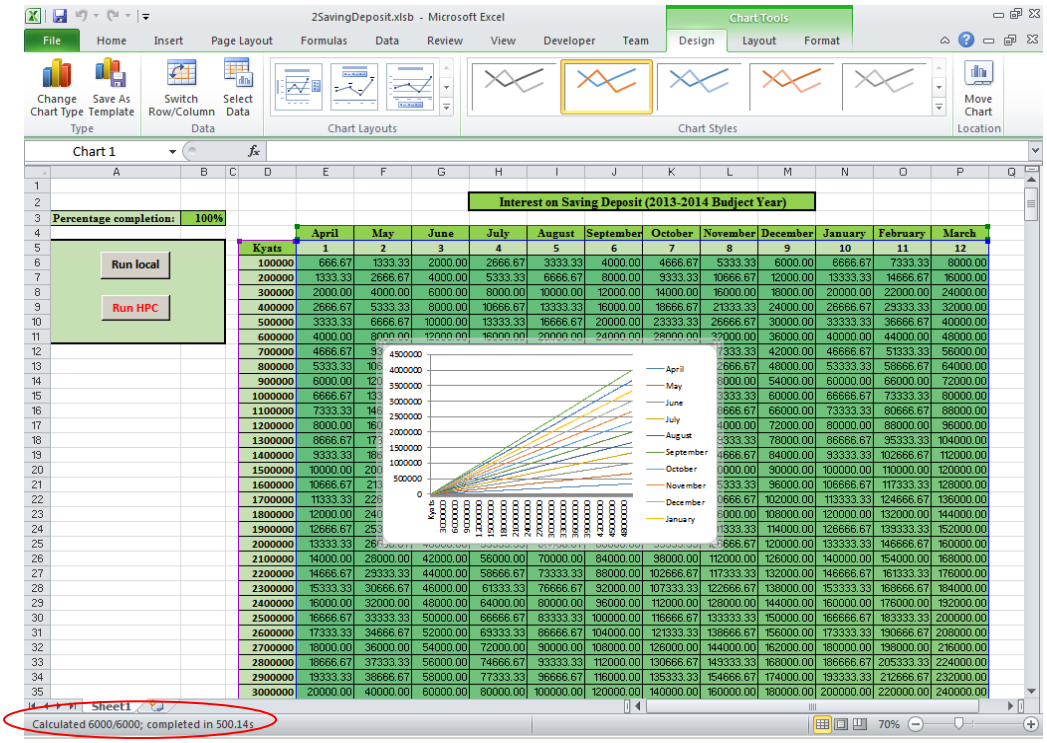

Figure 4. Execution time on Local with 8Cores

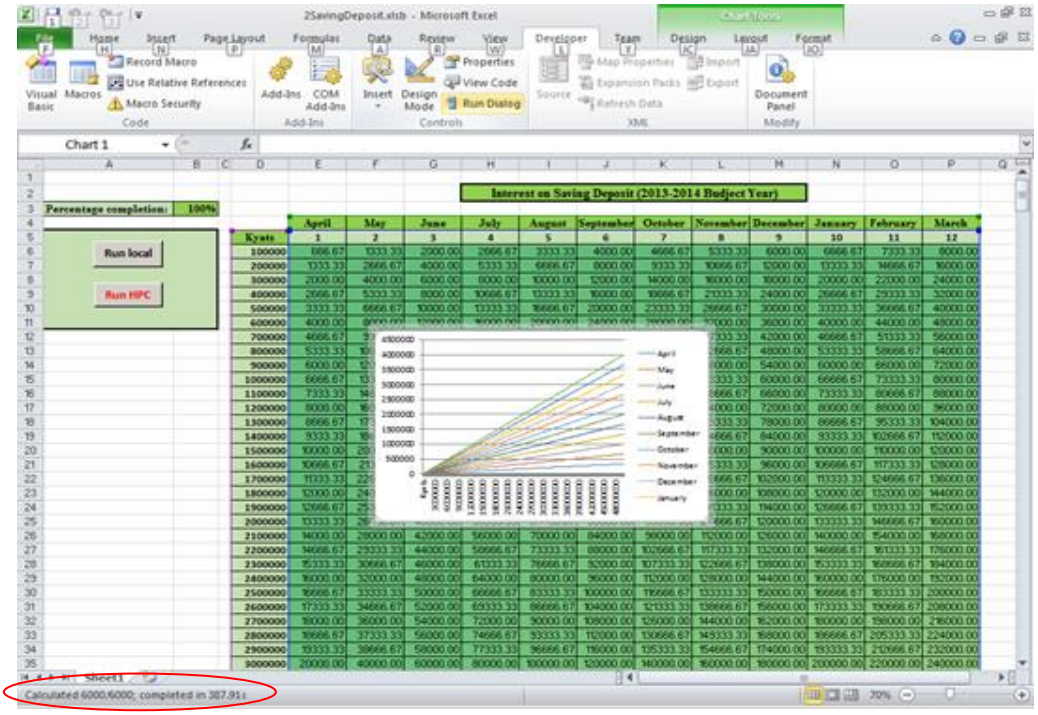

Figure 5. Execution time on HPC with 8Cores

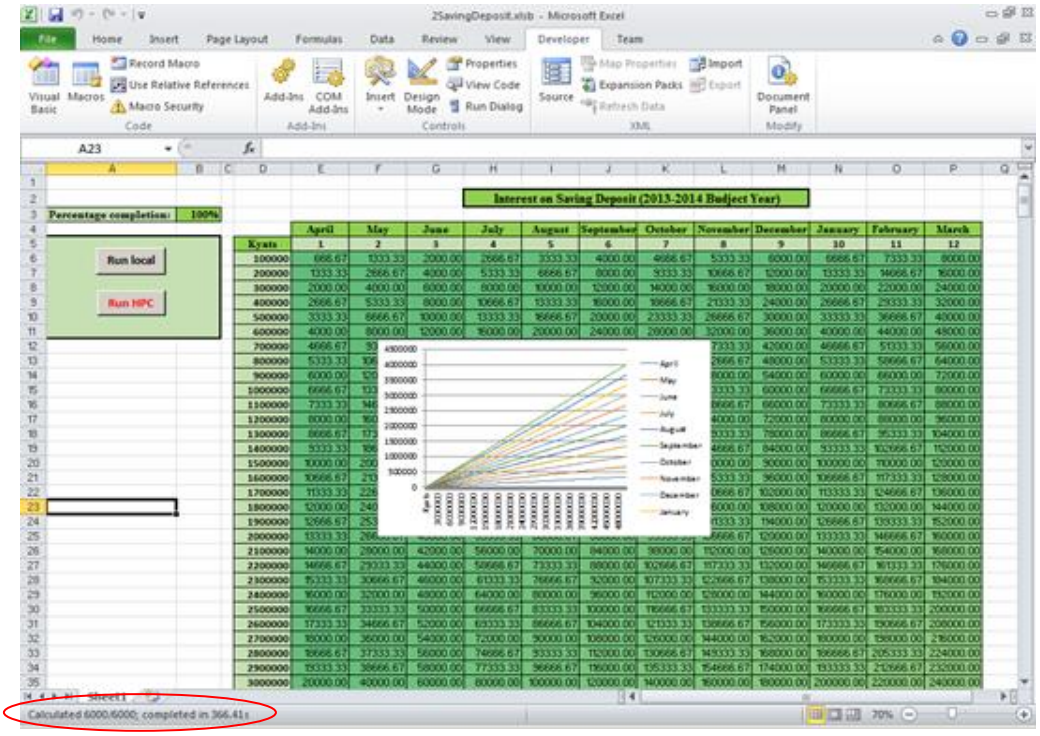

Figure 6. Execution time on HPC with 16Cores

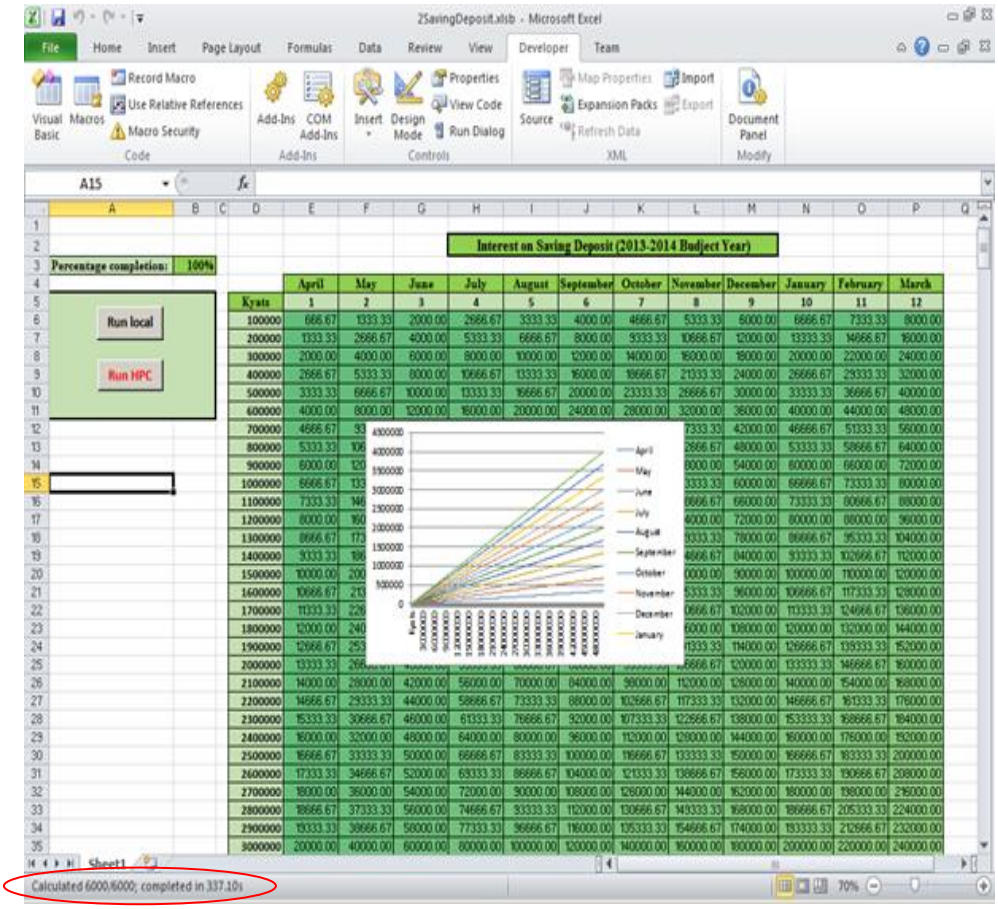

Figure 7. Execution time on HPC with 32Cores

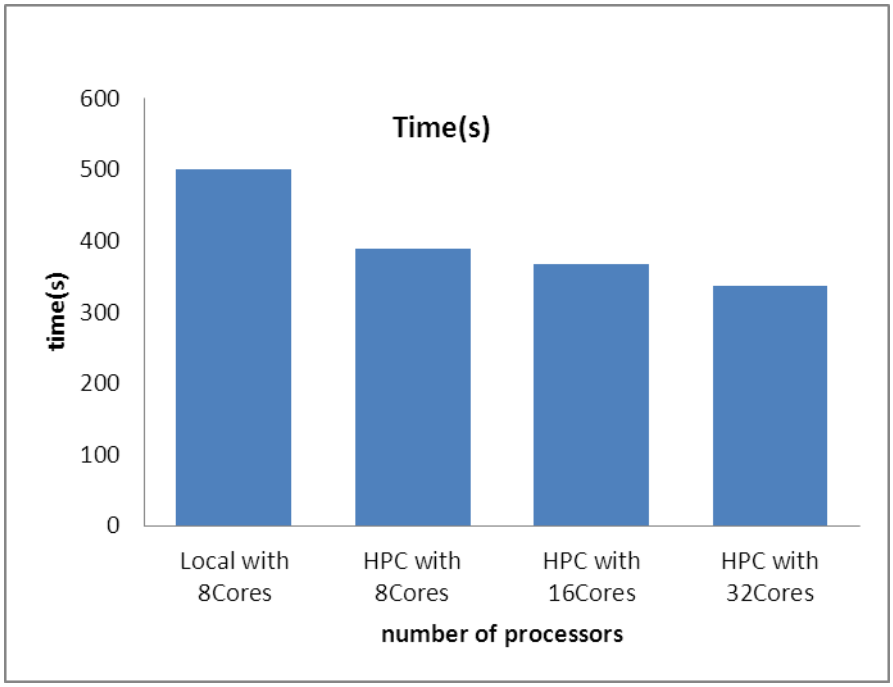

Figure 8. Comparison of execution time for local and HPC with different processors

Result and Discussion

Figure.8 shows the execution times taken by Local or HPC with different number of processors for the same number of records HPC with 8Cores is faster than Local with 8Cores by 112.23s. HPC with 16Cores is faster than HPC with 8Cores by 21.5s and faster than Local with 8Cores by 133.73s. HPC with 32Cores is faster than HPC with 8Cores by 50.81s, HPC with 16Cores by 29.31s and Local with 8Cores by 163.04s. HPC with 32Cores can make Excel workbook run 1.5 times faster than Local with 8Cores. Therefore, HPC Services for Excel with a Windows HPC cluster can improve calculation performance.

Conclusion

The basic objective of this system is to reduce the calculation time required for long-running Spread-sheet Calculations model in a Windows HPC cluster. Windows HPC Server 2008 R2 enables multiple instances of Office Excel 2010 run in a Windows HPC cluster. Multiple compute nodes can perform the calculations simultaneously if a workbook contains independent units of calculation. In a Windows HPC cluster, parallel computation can significantly reduce workbook calculation time and make calculations across larger data sets more feasible. In addition, it also supports one or more Windows Communication Foundation (WCF) Broker nodes and Service Oriented Architecture (SOA) model that enables WCF services operate on a Windows HPC cluster. By using Windows HPC Server 2008 R2, the set-up of an HPC cluster can reduce calculation times for Excel 2010 workbooks. This system is able to remove all the complexity of hosting and managing the service and quickly make workbooks to be run on a cluster by applying WCF service and SOA model on a Windows HPC cluster.

Acknowledgments

I wish to express my sincere thank to Rector, Dr Aye Kyaw and Pro-rector, Dr Aung Aung Min, University of Yadanabon for their kind permission for the submission of this research paper.

I would like to express my sincere gratitude to Dr Pho Kaung, Acting-rector, University of Yangon for his enthusiastic guidance and various suggestions.

I also wish to thank Daw Wai Wai Myint, Lecturer and Head of Department of Computer Studies, University of Yadanabon for her encouragement and support.

References

Anthony M. Middleton, Ph.D. LexisNexis Risk Solutions Date: May 24, 2011 "Introduction to HPCC (High-Performance Computing Cluster)"

E. Joseph Billo - Excel for Scientists and Engineers - Numerical Methods (Wiley, 2007)

Grama A, Gupta A, Karypis G and Kumer V 2003 “Introduction to Parallel Computing” (NewYork: Benjamin/Cummings)

K. Dowd, and C. Severance, “High Performance Computing,” O’Reilly and Associates, Inc., 1998.

Karniadakis G E and Kirby R M 2008 “Parallel Scientific Computing in C++ and MPI” (Cambridge: CUP)

L.S. Nyland, J.F. Prins, A. Goldberg, and P.H. Mills, “A Design Methodology for Data-Parallel Applications,” IEEE

Microsoft Corporation 2010 "THEWinServer2008_R2StepByStep"

Peiris, C., Mulder, D., Cicoria, S., Bahree, A., Pathak,. Pro WCF: Practical Microsoft SOA Implementation [M]. *Springer-Verlag New York, Inc.*2007 pp.60-67

R. Buyya, “High Performance Cluster Computing,” Prentice Hall, 1999.

Transactions on Software Engineering, Vol. 26, No. 4, 2000, pp. 293-314.