Mo.net Financial Modelling Platform
Unlocking Elastic Scalability with VMSS and Mo.net

January 2024
Revision 8

Purpose

This paper outlines how Azure-based Virtual Machine Scale Sets ("VMSS") can be used to address the unpredictable and demanding workloads of financial modelling & other calculations across the end-to-end insurance enterprise. It describes how the Mo.net Worker Service and Mo.net Quotations Service can be used in conjunction with VMSS to provide efficient and cost effective scaling without having to invest in additional products, complex integrations, & dedicated support personnel.

Background

Since the dawn of desktop-based financial modelling 25 years ago, the modelling community (and their customers) has continuously demanded faster results, even though the models & calculations have got ever-more complex, and the volume of data has increased. These challenges have been partially addressed by the increasing speed of computers over the period, but other techniques have also been required.

The first approach, which was bundled with the first generation of financial modelling platforms was to simply divide up the policy or scenario data into a series of splits and process each split on a different worker or slave computer before aggregating the results from each split back into a coherent set. These proprietary so-called "master worker" solutions worked well enough and are still being used to this day, but often suffer from resilience issues, especially when running at any significant scale (e.g. more than 100 computers).

The next approach, which first appeared in the early 2000s, was to use third-party high performance computing ("HPC") or grid computing solutions to handle the orchestration and management of distributed modelling runs across a cluster of usually server-based machines. These solutions, which have been proven to be successful when running at large scale (hundreds or thousands of compute nodes) work well, but require dedicated integration points between the modelling

software and the HPC environment. HPC-based solutions were typically only adopted by those insurers who could justify the investment in both the software & infrastructure to support such technology – i.e., those with many product lines / models and high volumes of data.

Unfortunately, these solutions both suffer from the need to provide dedicated infrastructure to meet peak demand, rather than having compute provision which adapts to the cyclical & variable needs of the modelling community.

The Promise of Cycle Stealing

In the late 1990s, when distributed computing was in its infancy and access to high-performance computing power was very limited and cost prohibitive for most, several initiatives were started to harness the spare capacity of internet-connected home PCs. Projects such as SETI@home (looking for extra-terrestrial life) and ClimatePrediction.net (to understand the reality of global warming) allowed members of the public to run a relatively simple model in the background while they were using their word processors or spreadsheets. Each home computer processed a small element of a huge dataset and then sent results back to a central repository for aggregation and analysis.

While these projects were rather novel and quite interesting for members of the public, the application of similar cycle-stealing solutions in the financial modelling arena never really materialised into anything substantial. The complexity of financial models and the size of data, meant that any machine used to run a model was effectively monopolised by it, leaving no cycles for anything else. And with the ability to cycle steal effectively limited to machines inside the corporate firewall, its appeal quickly faded.

More recently cloud-based compute has become accessible to all, but the legacy architecture of most financial modelling platforms has prevented insurers from unlocking the full potential of cloud. Most insurers have simply lifted on-premise products & workloads and dropped them into the cloud, without being able to take advantage of features such as elastic scalability and utility-based charging models. With the cloud often acting as little more than a hosted infrastructure service, and with many vendors charging significant margins simply for the privilege of using the cloud, clients are now looking to get more flexibility from their cloud-based modelling environments and pay less for it.

So how can financial modellers take advantage of the tangible benefits of cloud computing, without making fundamental changes to how they work? Virtual Machine Scale Sets might provide the answer.

What is a Virtual Machine Scale Set?

A Virtual Machine Scale Set is a new feature of the Microsoft Azure platform which supports the creation and management of a set of identical, manual, or auto-scaling Virtual Machines ("VMs"). The number of VM instances can automatically increase or decrease in line with a specific schedule or a set of rules, such as CPU consumption, memory usage, or network traffic.

VMSS solutions are especially beneficial for applications with variable or unpredictable workloads, like financial modelling, since it can automatically alter the number of VM instances based on demand rather than having to use guesswork or manually monitor the workload. This helps to ensure that financial models remain available and responsive to users even during periods of high

traffic or increased activity, but without providing expensive redundant infrastructure when workloads are reduced.

Key Features of VMSS

There are three main features of VMSS which have a direct application in the financial modelling arena. Each of these is discussed in detail below.

Auto-Scaling

Auto-scaling is a way to automatically scale up (and down) or out (and in) the number of compute resources that are being allocated to an application / model based on its needs at any given time. Auto-scaling rules can either be:

- Metric-based - Application load is continuously monitored and VM instances are added or removed based on one or more metrics. For example, add an instance in a scale set when CPU usage is above 80%.
- Time-based - Application load follows a time-based or cyclical pattern. For example, add additional VM instances for each weekend during January and February (to support year-end processing).

Auto-scaling rules are subject to maximum and minimum constraints of the environment / subscription.

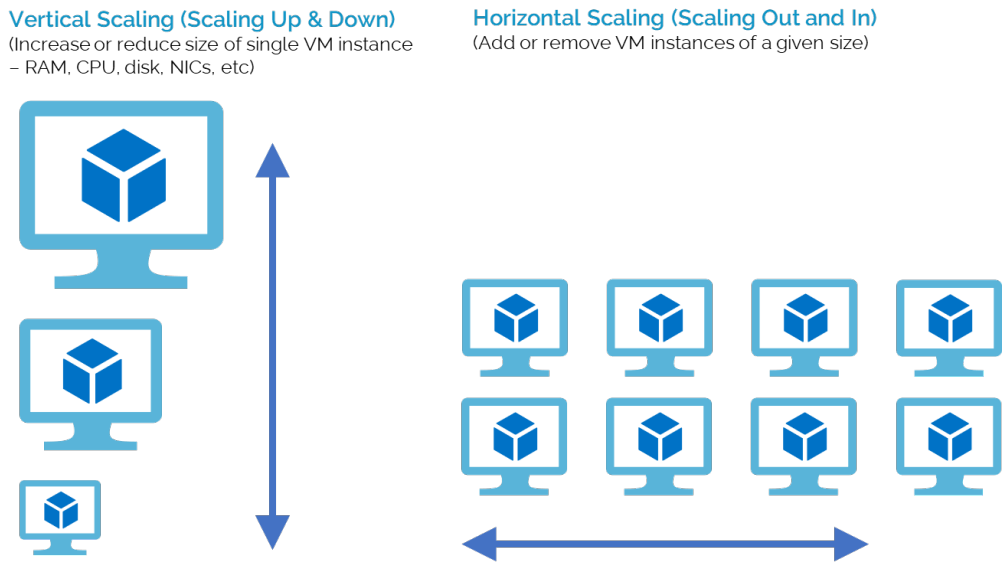

Figure 1 - Vertical and Horizontal Scaling

Vertical Scaling:

Vertical scaling, also known as scaling up & down, is an attempt to increase or decrease the computational power or capacity of a single virtual machine (VM) instance within the VMSS. Scaling up often entails adding more resources to a single VM, such as increasing the number of CPU cores or RAM, to improve performance or accommodate rising workload needs. Vertical scaling is used when you need to react quickly to fix a performance issue that can't be solved in more conventional ways.

Vertical scaling can be done manually or automatically depending on predefined rules and metrics, such as CPU load or available memory.

Scaling down is simply the process of reducing the capacity of virtual machines to match the current demand of the application.

Scaling up requires downtime, as the physical machine's configuration needs to change. Once any configuration change has been made a restart of all affected machines is required.

Horizonal Scaling

Horizonal scaling, also known as scaling-out & in, is the process of adding extra resources, typically VMs, to an existing scale set to manage an increase in demand for an application or model.

Scaling out a VMSS increases the number of operating VM instances, letting the application or model handle additional workload. This is accomplished by simply cloning an existing VM instance. All VM instances are provisioned and configured in an identical manner, ensuring that they work in concert together.

Scaling in helps optimise resource utilisation and minimise expense by deleting unnecessary VM instances. VM instances are removed based on the scale-in rules configured for the VMSS. When a scale in event occurs, one or more VM instances are removed from the VMSS, reducing the system's total capacity.

Scaling out does not require any downtime as new instances of a VM are being created and doesn't "touch" any existing VM instance.

Load Balancing

VMSS also has built-in load balancing technology that can balance traffic across the scale set's VM instances, ensuring that each instance is used efficiently & effectively and that the burden is divided equitably across the set. This is particularly useful for rapidly changing workloads common in the financial modelling & insurance calculations arena.

Configuration

With VMSS a group of identical VM instances can be deployed with the same configuration, ensuring consistency, and simplifying management. Changes to configurations can be made centrally to the base VM image and deployed across the scale set with the minimum of downtime. This significantly reduces the support overheads of the system and reduces the risk of system outages.

Using Mo.net with VMSS

The open and service-oriented architecture of the Mo.net platform is particularly well-suited to distributing workloads with VMSS.

Mo.net Worker Service

Using VMSS with the Mo.net Worker Service allows the performance of traditional valuation or reserving workloads to be vastly improved, but without the user community (or their IT support) having to configure a fixed worker pool in advance. VMSS allows for modelling runs to be performed as quickly as a user requires the results (subject to any cost constraints). Auto-scaling rules can be designed to maximise throughput or to minimise cost, depending on the particular workload.

Mo.net Quotations Service

VMSS provides an alternative to the existing Mo.net Quotations Service Router Service and provides elastic scalability for the Mo.net Quotations Service. This is particularly relevant when the Mo.net Quotations Service is being used behind a front-office portal or in conjunction with a policy administration system, where workloads can change rapidly. Most clients currently deploy a number of services to cope with a peak aggregate demand, which means excess resources are available in times of low demand, but insufficient resources may be available during periods of high demand. VMSS allows clients to scale their Mo.net Quotation Service instances to align exactly with real-time demand.

Key Benefits

The main advantages of using Virtual Machine Scale Sets for financial modelling & other insurance calculations are as follows:

- **Scalability**
  VMSS offers automatic scaling of VM instances based on the demand of the model or application.
- **High availability**
  VMSS provides fault-tolerant and redundant deployment options to assure model or application availability.
- **Unattended operation**
  Auto-scaling and availability rules allow the system to grow & contract unattended, with monitoring limited to traditional infrastructure failures.
- **Cost-effectiveness**
  VMSS allows for more efficient resource usage, which can result in cost savings.
- **Load balancing**
  VMSS has built-in load balancing capability to optimise traffic distribution across VM instances.
- **Consistency**
  VMSS enables the consistent deployment and control of identical VM instances with the same configuration.
- **Integration**
  VMSS also works in conjunction with other Azure services such as Azure Load Balancer and Azure Application Gateway.

Summary

Virtual Machine Scale Sets is a particularly interesting new feature of the Azure platform which provides a relatively easy & cost effective pathway to true elastic scalability for the Mo.net Worker Service and Mo.net Quotation Service. The integration between VMSS and the Mo.net Platform can usually be implemented in a few hours, subject to specific client constraints, providing a rapid solution for clients regardless of their workloads.

Contact Us

For more information regarding the use of Virtual Machine Scale Sets with the Mo.net platform, please get in touch:

Software Alliance Limited
30 Stamford Street, London, SE1 9LQ
Tel: +44 (0) 20 3964 2755
www.softwarealliance.net