AzureScaleMe

Technical Overview

Introduction

AzureScaleMe is a simple tool to scale Windows Azure instances according to pre-defined metrics. The application is designed to be simple, configurable and extensible, so you can plug your own code in to customise its behaviour.

Key features:

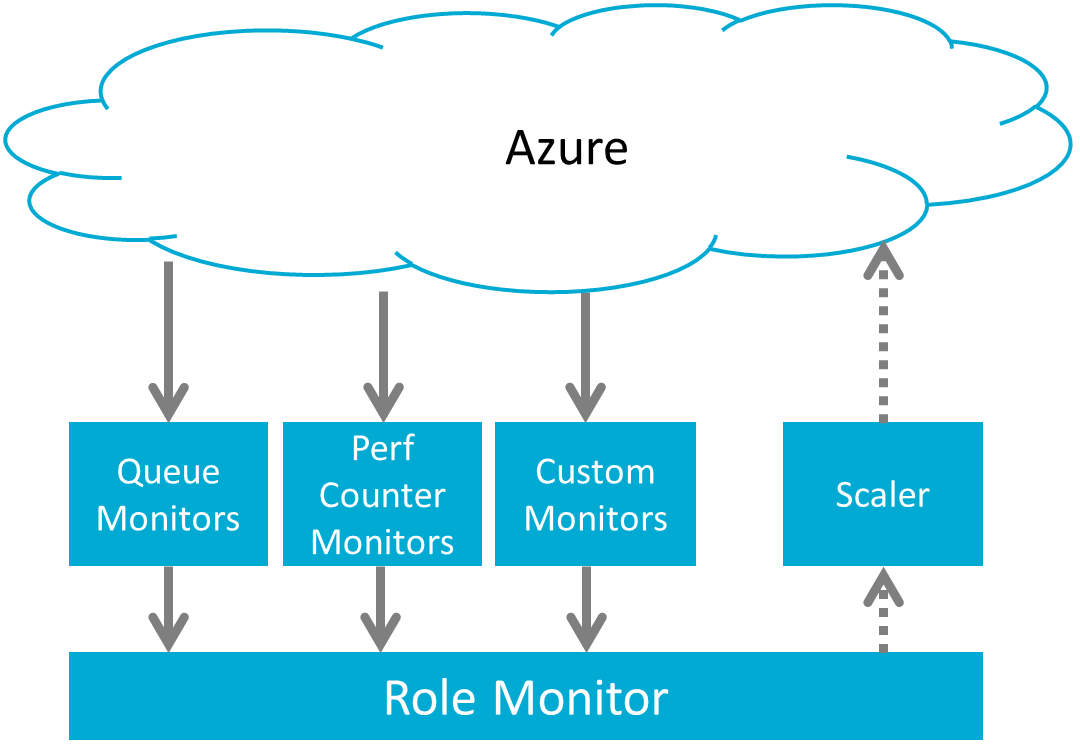
* No changes needed in your application.
* AzureScaleMe will run in the cloud, or in your premises.
* Configure a combination of ‘out of the box’ metrics, or plug in your own.

Examples scenarios:

* Automatically scale the number of Worker Role instances according to the length of an Azure Queue.
* Automatically scale the number of instances of a Web Role according to the CPU utilisation averaged across all instances.
* Scale instances according to a combination of performance counters, including custom counters.
* Build your own plugin, to monitor something specific to your application (for example, the number of records in a SQL Azure table).
* Build your own plugin, to send an email instead of scaling.

Architecture

AzureScaleMe supports any number of Role Monitors, which is responsible for coordinating the collection of performance information, scheduling a scaling event.



1. The Role Monitor collects measurements from your application running in Azure, using any combination of Queue Monitors, Performance Counter Monitors and your Custom Monitors, which can be plugged into AzureScaleMe.
2. Each monitor returns a vote on whether the application should be scaled up, down or maintained, according to individual settings they hold regarding the performance profile of your application.
3. The Role Monitor combines these votes (there are options for how these votes are counted) and decides whether the instance count should be adjusted. It records the time since the last scale, and will withhold a scale operation if this timespan hasn’t reached a minimum period.
4. If a scale is required, this request is passed to a Scaler, which will enact the scale. You can plug your own custom scaling logic in, if the default Scaler is unsuitable.

Deployment

AzureScaleMe will run as a console application, an Azure Worker Role, or can be hosted within another application, using a simple API to start monitoring.

The application is configured using an XML file, details of how to adjust the settings are provided with the tool. Any 3rd party plugins can be loaded by referencing them in the configuration. The tool will write out debug information, making it easy to understand the readings it’s taking from your application, and the decisions it’s making on how to scale.

Prior understanding of how your application behaves under load is essential for tuning the scaling parameters. It’s important that the correct values are chosen, so your application always remains responsive, but have over-provision during periods of low activity. It may take time and close monitoring to find the perfect balance.