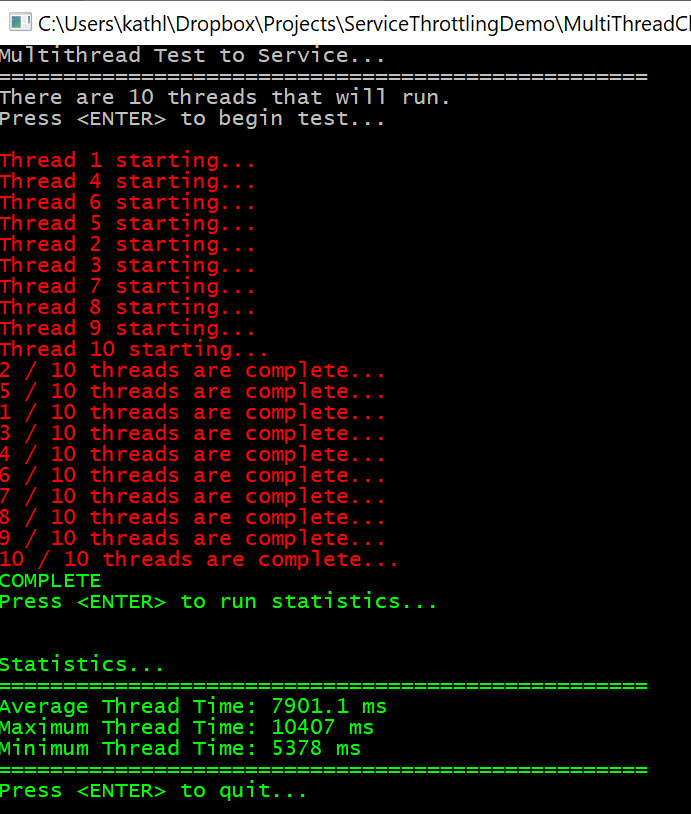
# Service Throttling in WCF Services – A Demo of Service Concurrency and Instance Throttling Behaviors with Multi-threading Clients

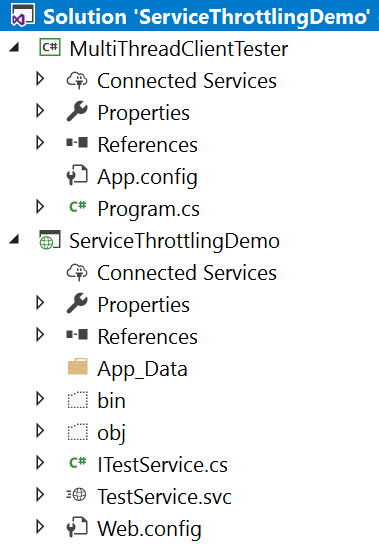
## 

## About

This project presents a simple Demo WCF Service and “Tester” Client Application demonstration that implements concurrency and instancing behaviors on a service with multiple client thread calls to a method on the service. The project also demos throttling service behaviors that are in the service configuration settings. Service throttling limits the client calls that could otherwise drain or slow down its service to other clients if too many are calling simultaneously. The Demo Service is a standard template WCF service application hosted by the development IIS. The service features one simple method… a test method that simulates a long running process (it sleeps for 5 seconds). The client “tester” is a simple console application that creates multiple threads that access the service and report back on the results. The objective of this project was not to demo setup and hosting of a service, nor the client interface, but retrieve and display results of service behaviors with respect to multi-threaded access. Discussion regarding the hosting and setup of the simple IIS hosted service application will be skipped in this project article.



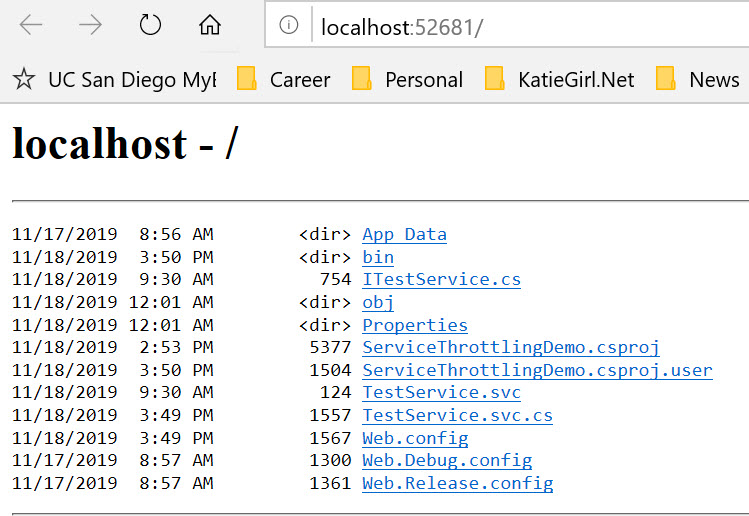
## Architecture



The demo project consists of these simple component topics:

* WCF Service (Hosted by IIS) Application Project “ServiceThrottlingDemo”
  + ITestService – Interface of Service and Operational Contracts
  + TestService – Implements the Service Interface
  + Web.config (Configuration for Service Hosted on IIS Express)
* Client “Tester to Service” Windows Console Application “MultiThreadClientTester”
  + *Service Reference Proxy (WSDL) to WCF Service*
  + Program – Creates multiple threads that call the service, records time spent in threads, and reports statistics on the final results in a console window.

## Service Application



A service template application project (*ServiceThrottlingDemo*) was added to my Visual Studio solution. The code is available on GitHub [here].

### ITestService (Interface for Service)

This interface is a simple test service that simulates a long running process. The ServiceContract for the simple demo service contains only one operation contract. The method simulates a long running process on a service and returns the TotalMilliseconds (int) spent in the service method. The code is available on GitHub [here].

[insert code here]

### Definitions

#### 

#### Concurrency Multiple

#### Each service instance processes multiple client method calls concurrently. The service implementation must be thread-safe to use this concurrency mode. This is ideal for multiple threads accessing the service method at the same time, however, resources must be thread-safe. The use of concurrency is related to the instancing mode. In PerCall instancing, concurrency is not relevant, because each message is processed by a new service instance.

#### Concurrency Single

#### Each service instance processes one client call to a method at a single time. This is the default concurrency mode (if not manually configured with attributes). If there is a multi-threaded client having multiple threads call this service, then the client call thread requests will be processed in series. Each service caller/thread must wait until the service has completed the current caller it is processing. The use of concurrency is related to the instancing mode. In PerCall instancing, concurrency is not relevant, because each message is processed by a new service instance.

#### Instance Single

In this mode, only one service InstanceContext object is used for all incoming client calls and is not recycled after the calls. If a service object does not exist, a new one is created when the service is called for the first time.

#### Instance PerCall

In this mode, a new service InstanceContext object is created prior to and recycled subsequent to each client call.

### TestService.svc.cs (Code that Implements the Service Interface)

The service implementation code has service-level attributes configured to control the concurrency and the instance behaviors of the service. This code implements the contracts for the ITestService interface. The code is available on GitHub [here].

There are four modes that were available for configuration on the service:

* Concurrency: Multiple | Instance: PerCall
* Concurrency: Multiple | Instance: Single\*
* Concurrency: Single | Instance: PerCall
* Concurrency: Single | Instance: Single

*\* This is the current configuration on the code in the repo.*

The use of concurrency is related to the instancing mode. In PerCall instancing, concurrency is not relevant, because each client call to the service is processed by a new service instance. However, in throttling behaviors and settings, the concurrency setting is relevant and can also affect the throttling ability of PerCall instancing (see Demo video).

[insert code here of the attributes]

#### Service Interface Implementation Methods

The service implementations are described below.

##### TestMethod

This is the only method available on the service. When a client calls this method, it simulates a long running process by sleeping for five seconds and then returns the time spent in the method (representing milliseconds) as an integer.

[insert code here]

### Web.Config and Service Throttling Behavior Configuration

The configuration file on the service application project directs the service settings, including throttling. This demo project simulated various scenarios that involved manipulating the service behavior attributes for throttling and then to observe the differences in the client call processing. Multi-threaded client calls to the service were throttled and the results of the experiment are shown and discussed in a video (see Demo section).

There were two types of throttling considered in my project demo and experiment: concurrent calls and instances. Some definitions and explanations are shown below along with a sample code showing the configuration for setting up throttling.

#### maxConcurrentCalls

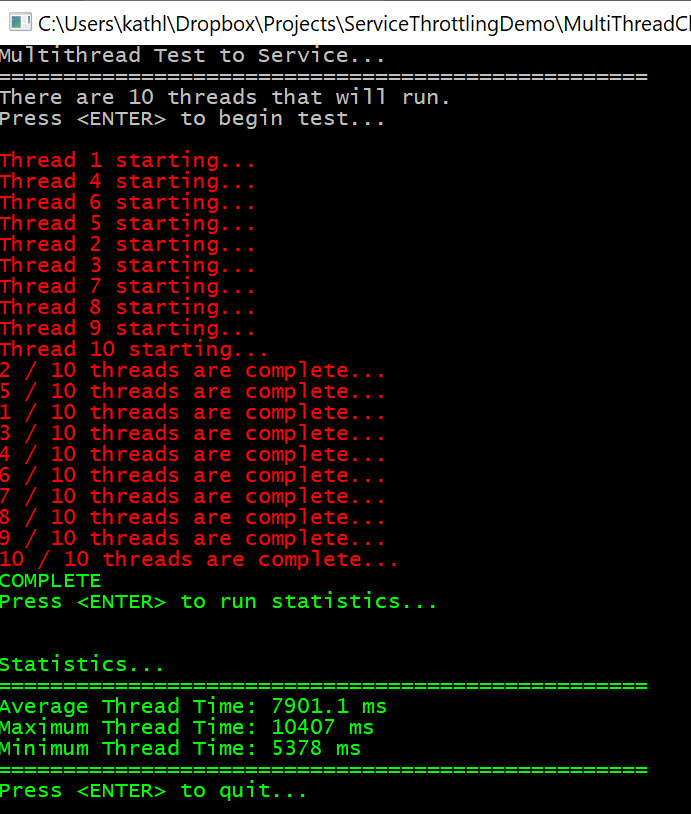
The *MaxConcurrentCalls* is a property that limits the number of messages actively processing on a ServiceHost object. This property is usually set in the configuration file as an xml attribute.

#### maxConcurrentInstances

The *MaxConcurrentInstances* is a property that limits the total number of InstanceContext objects that can run with the service. If the service instance mode is configured as PerCall, the resulting value is the number of concurrent calls (which can also be throttled or set to a limit). While in PerCall configuration, if a message arrives after the InstanceContext objects are throttled (ex: reached a maximum), the message is held until an instance completes its PerCall process and closes, effectively freeing up another available InstanceConext object. This property is usually set in the configuration file as an xml attribute.

[insert code here]

## Client “Tester to Service” Windows Console Application



The client “tester to service” is a simple windows console application project (*MultiThreadClientTester*) in the same solution that connects to the “Demo Service” by use of a Service Reference proxy generated by the Add Service Reference Wizard. The client program will use this proxy to test the OperationContract or methods available in the ServiceContract and return the results to the user on the console window.

The project code is available on GitHub [here].

### Features

The client application was meant to only demonstrate consumption of the Demo service, so this client has limited scope of features. Some of the client application features are:

#### Simple Design

The user interface is simplified to do basic demo, testing, and reporting of the multi-thread calls to the service with its configured behaviors.

#### Multi-threading

The program creates and runs threads while each one calls and waits on a call to the service in parallel with other threads. Each thread asynchronously updates its completion on the console window after it completes the service call. The program flow is actually asynchronous but waits for the work to be complete before running the stats.

#### Completion Status

The total completion status is updated asynchronously on the console window so that the tester can observe the effects of service throttling in real-time with their threads. For example, if throttling limits a processing of only 5 threads, the user will see threads completed in blocks of 5.

#### Stats

After the thread work is complete, the user presses enter to displays simple stats on the threads including the average, max, and min run-times of the threads.

### Program Code

The code behind file is for the client tester and manages the application in one file. The code is available on GitHub [here].

#### Fields

Private fields allow the client to work with the following:

* Proxy for the Service Reference (Generated by the SVCUTIL Wizard)
* List of integers representing results of thread run-times
* Total number of threads to run in parallel
* Completion status of thread simulation

[insert code here]

#### Methods

##### Main

This main method is the entry point of the client tester application. It first creates a proxy object (instance of the reference created by SVCUTIL), then creates and runs the set number of threads in parallel. It then asks the user to wait until thread completion so they can see the individual results on the console window and then display the stats.

[insert code here]

##### ServiceThread

This is the main thread that implements the *“async await”* pattern of asynchronous processing of the thread. It also accesses the service using the async version of the generated proxy methods (proxy.TestMethodAsync()). It calls the service and awaits the completion, then adds its total thread time information to the list of thread times.

[insert code here]

##### Print Statistics

This method runs and prints basic statistics (average, maximum, minimum) on the List of thread times.

[insert code here]

## Demo

[insert youtube video]

## Code

The entire project code repository is available on GitHub [here].