

# Queue-Based Load Leveling

☰ Tags	
🕒 Date Created	@February 27, 2025 11:28 AM
➦ Class	<u>Ingesoft V</u>
☑ Organized?	<input type="checkbox"/>

Se usa como un buffer intermediario entre una tarea y un servicio, de tal manera que si la tarea es muy pesada, poder tener una mejor gestión de la carga para evitar que el servicio que está siendo ejecutado se caiga o que ocurra un time out en la task.

Trigger: intermittent heavy loads → Causa problemas de performance o realability.



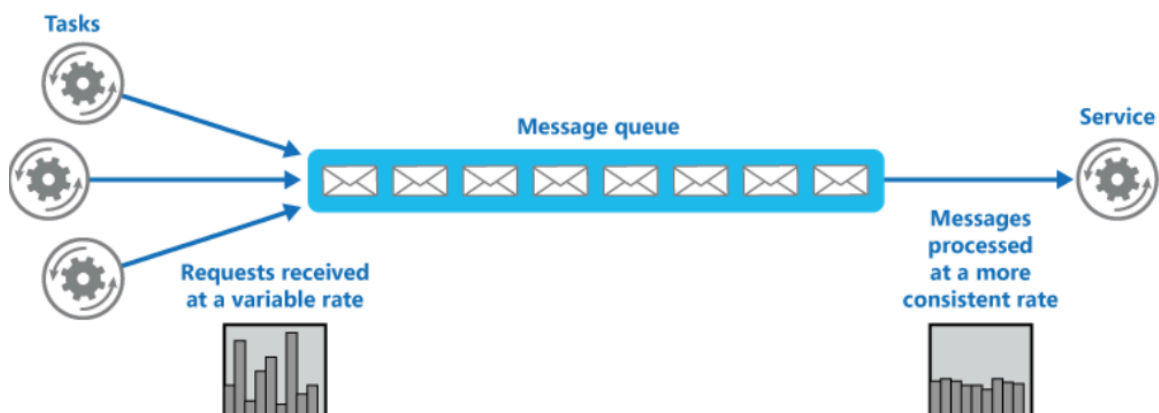
Objetivo: *smooth intermittent heavy loads*

## Características

1. La task y el service corre de forma asincronica.
2. La tarea publica un mensaje que contiene los datos requeridos por el servicio en una cola.
3. La cola actúa como un búfer, almacenando el mensaje hasta que sea recuperado por el servicio.
4. El servicio recupera los mensajes de la cola y los procesa.

## Resultados:

Las solicitudes de varias tareas, que pueden generarse a una tasa altamente variable, pueden enviarse al servicio a través de la misma cola de mensajes. Esta figura muestra el uso de una cola para nivelar la carga en un servicio.



## Beneficios

1. Maximize availability: los delays de los servicios no tienen impactos en el funcionamiento de la aplicación, ya que, a pesar de que ocurra una caída o no esté trabajando correctamente, las task serán almacenadas y próximamente procesadas.
2. Maximize scalability: el numero de colas y de servicios, varian en función de la demanda.
3. Control cost: te ayuda a tener una métrica para definir el numero de servicios necesarios, comúnmente teniendo como referencia el average de la cantidad de tareas de la cola (no usar el peak). Además de lo anterior, usualmente se usa una técnica llamada **throttling**, la cual se describe a continuación:

- **Control de demanda (Throttling)**

Algunos servicios implementan un mecanismo llamado **throttling** cuando la demanda alcanza un umbral crítico. Este umbral es el punto en el que el sistema corre el riesgo de fallar debido a una sobrecarga de solicitudes.

- **Reducción de funcionalidad**

Cuando se activa el **throttling**, el sistema puede **limitar o reducir ciertas funcionalidades** para evitar el colapso. Por ejemplo, podría rechazar solicitudes, ralentizar respuestas o deshabilitar temporalmente algunas características.

- **Nivelación de carga (Load Leveling)**

Para evitar que el sistema llegue a este punto crítico, puedes implementar una estrategia de **nivelación de carga**. Esto implica distribuir las solicitudes de manera más uniforme en el tiempo, evitando picos bruscos de demanda que puedan activar el **throttling**.

## Cómo implementarlo?

1. El sistema está pensado para ser unidireccional, si se desea una respuesta, entonces se debe usar otro recurso.
2. Tener cuidado al momento de usar auto-scaling en función de las task en la cola, pues puede aumentar recursos innecesariamente, sin dejar que se cumpla la función de la cola.
3. Tener en cuenta la tasa de trabaja y la tasa de resolución, pues, si la tasa de task es mucho mayor a la resolutive entonces se generara una cola infinita.
4. Si la cola falla o se llega a los limites de la cola, puede que haya perdida de información.

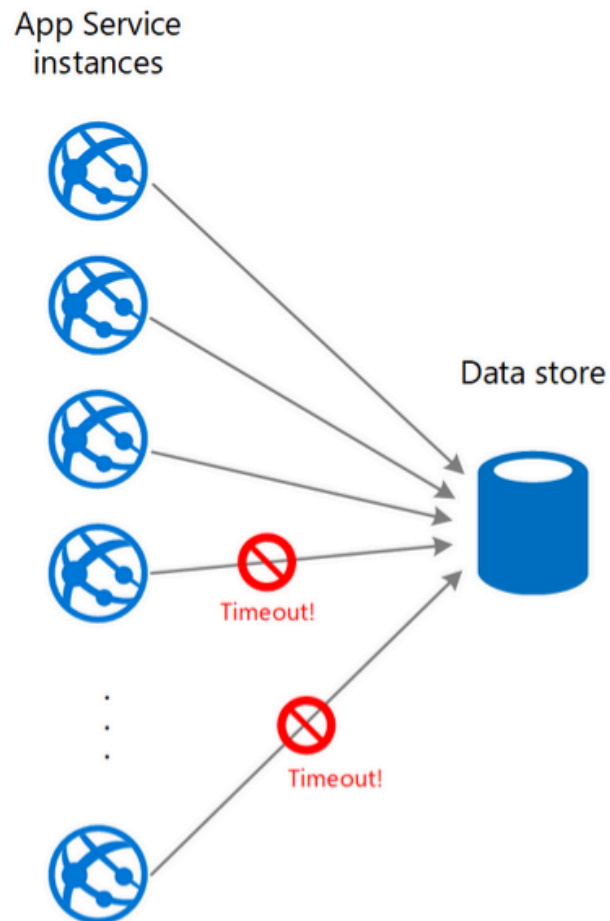
## Cuando usarlo?



**This pattern is useful to any application that uses services that are subject to overloading.**

**This pattern isn't useful if the application expects a response from the service with minimal latency.**

## Problem example



---

## On azure

Queue: Service bus queue - Azure queue storage

### **azurerm\_servicebus\_queue:**

El recurso `azurerm_servicebus_queue` en Terraform se usa para **crear una cola en Azure Service Bus** dentro de un **Namespace de Service Bus**.


Las colas en **Service Bus** permiten enviar y recibir mensajes de manera asíncrona, garantizando que los mensajes lleguen en orden y puedan ser procesados de manera confiable por diferentes aplicaciones o servicios, como **Azure Functions**.

## Resources

1 hour+ about this pattern.

## Azure Cloud Streak 3.0 - Part 1 - Queue based load leveling with Azure Service Bus

Azure Cloud Streak 3.0


Our guest speaker, Kamal Rathnayake (Associate Tech Lead - 99X) conducted the session on  
 <https://www.youtube.com/watch?v=neWXUwRSR3E>



## How to implement a service bus:

### Use the Azure portal to create a Service Bus queue - Azure Service Bus

In this quickstart, you learn how to create a Service Bus namespace and a queue in the namespace by using the Azure portal.

 <https://learn.microsoft.com/en-us/azure/service-bus-messaging/service-bus-quickstart-portal>

