Previous topic

Security

Next topic

Debugging

This Page

Show Source

Quick search

Go

This document describes the current stable version of Celery (4.2). For development docs, go here.

Optimizing

Introduction

The default configuration makes a lot of compromises. It's not optimal for any single case, but works well enough for most situations.

There are optimizations that can be applied based on specific use cases.

Optimizations can apply to different properties of the running environment, be it the time tasks take to execute, the amount of memory used, or responsiveness at times of high load.

Ensuring Operations

In the book Programming Pearls, Jon Bentley presents the concept of back-of-the-envelope calculations by asking the question;

> ❝ How much water flows out of the Mississippi River in a day? ❞

The point of this exercise [*] is to show that there's a limit to how much data a system can process in a timely manner. Back of the envelope calculations can be used as a means to plan for this ahead of time.

In Celery; If a task takes 10 minutes to complete, and there are 10 new tasks coming in every minute, the queue will never be empty. This is why it's very important that you monitor queue lengths!

A way to do this is by using Munin. You should set up alerts, that'll notify you as soon as any queue has reached an unacceptable size. This way you can take appropriate action like adding new worker nodes, or revoking unnecessary tasks.

General Settings

librabbitmq

If you're using RabbitMQ (AMQP) as the broker then you can install the librabbitmq module to use an optimized client written in C:

```
$ pip install librabbitmq
```

The 'amqp' transport will automatically use the librabbitmq module if it's installed, or you can also specify the transport you want directly by using the `pyamqp://` or `librabbitmq://` prefixes.

Broker Connection Pools

The broker connection pool is enabled by default since version 2.5.

You can tweak the **`broker_pool_limit`** setting to minimize contention, and the value should be based on the number of active threads/green-threads using broker connections.

Using Transient Queues

Queues created by Celery are persistent by default. This means that the broker will write messages to disk to ensure that the tasks will be executed even if the broker is restarted.

But in some cases it's fine that the message is lost, so not all tasks require durability. You can create a *transient* queue for these tasks to improve performance:

```
from kombu import Exchange, Queue

task_queues = (
    Queue('celery', routing_key='celery'),
    Queue('transient', Exchange('transient', delivery_mode=1),
          routing_key='transient', durable=False),
)
```

or by using `task_routes`:

```
task_routes = {
    'proj.tasks.add': {'queue': 'celery', 'delivery_mode': 'transient'}
}
```

The `delivery_mode` changes how the messages to this queue are delivered. A value of one means that the message won't be written to disk, and a value of two (default) means that the message can be written to disk.

To direct a task to your new transient queue you can specify the queue argument (or use the `task_routes` setting):

```
task.apply_async(args, queue='transient')
```

For more information see the routing guide.

Worker Settings

Prefetch Limits

Prefetch is a term inherited from AMQP that's often misunderstood by users.

The prefetch limit is a **limit** for the number of tasks (messages) a worker can reserve for itself. If it is zero, the worker will keep consuming messages, not respecting that there may be other available worker nodes that may be able to process them sooner [†], or that the messages may not even fit in memory.

The workers' default prefetch count is the `worker_prefetch_multiplier` setting multiplied by the number of concurrency slots [‡] (processes/threads/green-threads).

If you have many tasks with a long duration you want the multiplier value to be *one*: meaning it'll only reserve one task per worker process at a time.

However – If you have many short-running tasks, and throughput/round trip latency is important to you, this number should be large. The worker is able to process more tasks per second if the messages have already been prefetched, and is available in memory. You may have to experiment to find the best value that works for you. Values like 50 or 150 might make sense in these circumstances. Say 64, or 128.

If you have a combination of long- and short-running tasks, the best option is to use two worker nodes that are configured separately, and route the tasks according to the run-time (see Routing Tasks).

Reserve one task at a time

The task message is only deleted from the queue after the task is acknowledged, so if the worker crashes before acknowledging the task, it can be redelivered to another worker (or the same after recovery).

When using the default of early acknowledgment, having a prefetch multiplier setting of *one*, means the worker will reserve at most one extra task for every worker process: or in other words, if the worker is started with `-c 10`, the worker may reserve at most 20 tasks (10 unacknowledged tasks executing, and 10 unacknowledged reserved tasks) at any time.

Often users ask if disabling "prefetching of tasks" is possible, but what they really mean by that, is to have a worker only reserve as many tasks as there are worker processes (10 unacknowledged tasks for `-c 10`)

That's possible, but not without also enabling late acknowledgment. Using this option over the default behavior means a task that's already started executing will be retried in the event of a power failure or the worker instance being killed abruptly, so this also means the task must be idempotent

See also:

Notes at Should I use retry or acks_late?.

You can enable this behavior by using the following configuration options:

```
task_acks_late = True
worker_prefetch_multiplier = 1
```

Prefork pool prefetch settings

The prefork pool will asynchronously send as many tasks to the processes as it can and this means that the processes are, in effect, prefetching tasks.

This benefits performance but it also means that tasks may be stuck waiting for long running tasks to complete:

```
-> send task T1 to process A
# A executes T1
-> send task T2 to process B
# B executes T2
<- T2 complete sent by process B

-> send task T3 to process A
# A still executing T1, T3 stuck in local buffer and won't start until
# T1 returns, and other queued tasks won't be sent to idle processes
<- T1 complete sent by process A
# A executes T3
```

The worker will send tasks to the process as long as the pipe buffer is writable. The pipe buffer size varies based on the operating system: some may have a buffer as small as 64KB but on recent Linux versions the buffer size is 1MB (can only be changed system wide).

You can disable this prefetching behavior by enabling the `-Ofair` worker option:

```
$ celery -A proj worker -l info -Ofair
```

With this option enabled the worker will only write to processes that are available for work, disabling the prefetch behavior:

```
-> send task T1 to process A
# A executes T1
-> send task T2 to process B
# B executes T2
<- T2 complete sent by process B

-> send T3 to process B
# B executes T3

<- T3 complete sent by process B
<- T1 complete sent by process A
```

Footnotes

[*] The chapter is available to read for free here:The back of the envelope. The book is a classic text. Highly recommended.

[†] RabbitMQ and other brokers deliver messages round-robin, so this doesn't apply to an active system. If there's no prefetch limit and you restart the cluster, there will be timing delays between nodes starting. If there are 3 offline nodes and one active node, all messages will be delivered to the active node.

[‡] This is the concurrency setting; **`worker_concurrency`** or the **`celery worker -c`** option.