Gurobi Instant Cloud Guide

Nov 07, 2024

Revision: 98c472f3d

CONTENTS

This is the guide for using the Gurobi^{TM} Instant Cloud, which provides Gurobi Remote Services via cloud computing. Gurobi Instant Cloud is designed specifically to streamline the use of Cloud resources for the Gurobi Optimizer. This is the easiest way to use the Gurobi in the Cloud; no prior experience with cloud computing is needed. Gurobi Instant Cloud also provides additional features such as:

- Fully automated machine provisioning (AWS or Azure)
- Web interface to manage your machines and pools
- Dashboard to monitor your optimization jobs
- Machine and Job history
- REST API

If you are new to Gurobi Instant Cloud, you can quickly start your first solve or tuning sessions by following the *Getting Started* guide. You will also find detailed descriptions of the *Instant Cloud Manager* used to configure your Cloud environment. Finally, if you need to integrate Instant Cloud in a custom solution or framework, the *REST API* will give you all the tools to automate your processes.

Please check this document periodically to ensure you have the latest instructions for the Gurobi Cloud. Other cloud options exist. Please contact sales@gurobi.com to discuss other options.

CHAPTER

ONE

GETTING STARTED

You can get started with a few simple steps as Gurobi Instant Cloud comes with a predefined configuration. Let's review some typical use cases.

In order to access Gurobi Instant Cloud you need to register and request a Cloud license from your sales account manager, or sign up for a trial if you are eligible.

You also need to install the latest Gurobi Optimizer.

Finally, we recommend that you subscribe to updates from the Gurobi status page to receive notifications about scheduled maintenance and incidents.

1.1 First Solve in the Cloud

You can perform your first solve in a few simple steps:

1.1.1 1. Open the Instant Cloud Manager

Go to cloud.gurobi.com. If you are not logged in, you will be prompted for your credentials. If you do not have an account, please register and contact Gurobi to request an evaluation license.

1.1.2 2. Download the default license file

The list of licenses is displayed in the Instant Cloud Manager and your default license file is ready to be downloaded with the following button.

⭳

The license file contains the default access ID and secret key for the selected license. You just have to place this file in your home directory which takes precedence, or in one of the following shared locations:

- C:\gurobi\ on Windows
- /opt/gurobi/ on Linux
- /Library/gurobi/ on Mac OS X

In case you previously had a license file installed, please make sure to replace it, or set the environment variable `GRB_LICENSE_FILE` to point to the cloud license file, it will override the default locations.

1.1.3 3. Solve

You can try to solve any predefined MPS file provided with the Gurobi distribution. Here is an example on Mac OS X:

```
$ gurobi_cl /Library/gurobi1001/macos_universal2/examples/data/afiro.mps
Waiting for cloud server to start..........
Capacity available on '999999-default' cloud pool - connecting...
Established HTTPS encrypted connection with Compute Server

Gurobi Optimizer version 10.0.1 build v10.0.1rc0 (mac64[x86])
Copyright (c) 2023, Gurobi Optimization, LLC

Read MPS format model from file /Library/gurobi1001/macos_universal2/examples/data/afiro.
↪mps
Reading time = 0.12 seconds
AFIRO: 27 rows, 32 columns, 83 nonzeros
Optimize a model with 27 rows, 32 columns and 83 nonzeros
Coefficient statistics:
  Matrix range     [1e-01, 2e+00]
  Objective range [3e-01, 1e+01]
  Bounds range     [0e+00, 0e+00]
  RHS range        [4e+01, 5e+02]
Presolve removed 18 rows and 20 columns
Presolve time: 0.73s
Presolved: 9 rows, 12 columns, 32 nonzeros

Iteration    Objective       Primal Inf.    Dual Inf.      Time
       0   -4.8565680e+02   1.363638e+02   0.000000e+00      1s
       3   -4.6475314e+02   0.000000e+00   0.000000e+00      1s

Solved in 3 iterations and 0.74 seconds
Optimal objective -4.647531429e+02
```

As you can see in the log, the client automatically starts the pool and connects to it. If you wish, you can check the status of the machine using the Instant Cloud Manager. If you run again a new solve, you will notice that it can start right away because the machine is already available.

1.1.4 4. Terminate the pool (optional)

The machine will auto-terminate once it stayed idle for a duration limit called the idle shutdown. The default idle shutdown is 60 minutes, and it can be changed in the settings of the pools and your preferences. Otherwise, you can terminate the pool manually in the Instant Cloud Manager, by selecting the default pool and clicking on the terminate button.

1.2 First Tuning in the Cloud

You can perform your first tuning in a few simple steps. If you already installed your default cloud license file, you can go directly to step 3.

1.2.1 1. Open the Instant Cloud Manager

Go to cloud.gurobi.com. If you are not logged in, you will be prompted for your credentials. If you do not have an account, please register and contact Gurobi to request an evaluation license.

1.2.2 2. Download the default license file

The list of licenses is displayed in the Instant Cloud Manager and your default license file is ready to be downloaded with the following button.

⤓

The license file contains the default access ID and secret key for the selected license. You just have to place this file in your home directory which takes precedence, or in one of the following shared locations:

- C:\gurobi\ on Windows
- /opt/gurobi/ on Linux
- /Library/gurobi/ on Mac OS X

In case you previously had a license file installed, please make sure to replace it, or set the environment variable `GRB_LICENSE_FILE` to point to the cloud license file, it will override the default locations.

1.2.3 3. Tune

You can try to tune a MIP MPS file provided with the Gurobi distribution. Here is an example on Mac OS X:

```
$ grbtune /Library/gurobi1001/macos_universal2/examples/data/misc07.mps
Waiting for cloud server to start.............
Capacity available on '999999-default' cloud pool - connecting...
Established HTTPS encrypted connection

grbtune version 10.0.1 build v10.0.1rc0 (mac64[x86])
Copyright (c) 2023, Gurobi Optimization, LLC

Read MPS format model from file /Library/gurobi1001/macos_universal2/examples/data/
↪misc07.mps
Reading time = 0.56 seconds
MISC07: 212 rows, 260 columns, 8619 nonzeros

Solving model using baseline parameter set with TimeLimit=3600s

Solving with random seed #1 ...
Optimize a model with 212 rows, 260 columns and 8619 nonzeros
Variable types: 1 continuous, 259 integer (0 binary)
    [...]
```

As you can see in the log, the client automatically connects to the Instant Cloud server and checks for the pool status. As the machines are not launched yet, Instant Cloud starts the machines and the client reports that it is waiting until capacity is available. Then, it starts the tuning process.

1.2.4 4. Terminate the pool (optional)

The machine will auto-terminate once it stays idle for a duration limit called the idle shutdown. The default idle shutdown is 60 minutes, and it can be changed in the settings of the pools and your preferences. Otherwise, you can terminate the pool manually in the Instant Cloud Manager, by selecting the default pool and clicking on the terminate button.

■

1.3 First Distributed Optimization in the Cloud

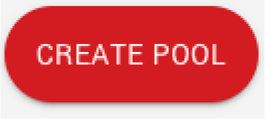
The Gurobi Instant Cloud makes it easy to launch a cluster of machines for distributed optimization. This guide will walk you through the process of completing your first distributed solve in the Cloud.

1.3.1 1. Open the Instant Cloud Manager

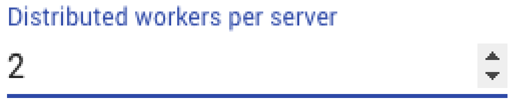
Go to cloud.gurobi.com. If you are not logged in, you will be prompted for your credentials. If you do not have an account, please register and contact Gurobi to request an evaluation license.

1.3.2 2. Create a pool with distributed workers

In the Instant Cloud Manager, go to the ‘Pools’ section and click on the create pool button:



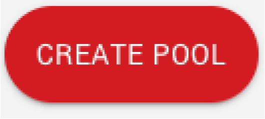
Then, select the ‘License’ tab and set the number of workers to 2.



Distributed workers per server

2

Finally, create the new pool. Note that a default name is assigned for you such as pool1.



1.3.3 3. Download the pool license file

The list of pools is displayed in the Instant Cloud Manager and your license file is ready to be downloaded with the following button.

⬇

The license file contains the default access ID and secret key for the selected pool. You just have to place this file in your home directory which takes precedence, or in one of the following shared locations:

- C:\gurobi\ on Windows
- /opt/gurobi/ on Linux
- /Library/gurobi/ on Mac OS X

In case you previously had a license file installed, please make sure to replace it, or set the environment variable `GRB_LICENSE_FILE` to point to the cloud license file, it will override the default locations.

1.3.4 4. Solve

You can try to solve a MIP MPS file provided with the Gurobi distribution. Here is an example on Mac OS X:

```
$ gurobi_cl /Library/gurobi1001/macos_universal2/examples/data/misc07.mps
Waiting for cloud server to start..........
Capacity available on '999999-pool1' cloud pool - connecting...
Established HTTPS encrypted connection

Gurobi Optimizer version 10.0.1 build v10.0.1rc0 (mac64[x86])
Copyright (c) 2023, Gurobi Optimization, LLC

Read MPS format model from file /Library/gurobi1001/macos_universal2/examples/data/
↪misc07.mps
Reading time = 0.47 seconds
MISC07: 212 rows, 260 columns, 8619 nonzeros
Optimize a model with 212 rows, 260 columns and 8619 nonzeros
Coefficient statistics:
  Matrix range    [1e+00, 7e+02]
  Objective range [1e+00, 1e+00]
  Bounds range    [1e+00, 1e+00]
  RHS range       [1e+00, 3e+02]

Starting distributed worker jobs...

Using Compute Server as first worker - running now
Started distributed worker on ip-52-91-137-123
Started distributed worker on ip-54-159-77-110

Distributed MIP job count: 3

    Nodes    |    Current Node    |     Objective Bounds      |     Work
 Expl Unexpl |  Obj  Depth IntInf | Incumbent    BestBd   Gap | ParUtil Time

H    0                          4155.0000000          -      -           3s
```

(continues on next page)

(continued from previous page)

```
H    0                          3610.0000000          -      -          3s
H    0                          3500.0000000 1415.00000  59.6%          3s
H    0                          2940.0000000 1415.00000  51.9%          3s
H    0                          2810.0000000 1415.00000  49.6%          4s
    24    22                      2810.00000 1544.28571  45.0%   99%    4s
  1114   475                      2810.00000 1926.66667  31.4%   99%    5s

Ramp-up phase complete - continuing with instance 1 (best bd 2175)

  7533   931 1492.85714    0   48 2810.00000 2175.00000  22.6%   99%    7s
 15311     0 2785.00000   21   13 2810.00000 2810.00000  0.00%   93%    9s

Cutting planes:
  Cover: 2
  Clique: 4
  MIR: 17
  Zero half: 10

Runtime breakdown:
  Active:   8.09s (88%)
  Sync:     0.81s (9%)
  Comm:     0.28s (3%)

Explored 15311 nodes (152346 simplex iterations) in 9.17 seconds
Distributed MIP job count: 3

Optimal solution found (tolerance 1.00e-04)
Best objective 2.810000000000e+03, best bound 2.810000000000e+03, gap 0.0%
```

Within this log, we have highlighted in bold some important steps. First, the client automatically connects to the Instant Cloud server and checks for the pool status. As the machines are not launched yet, Instant Cloud starts the machines and the client reports that it is waiting until capacity is available.

Then the Gurobi Optimizer detects that the pool is setup with 2 distributed workers. So it automatically starts the solve in distributed mode with 3 workers (the master compute server counts as one worker as well).

1.3.5 5. Terminate the pool (optional)

The machine will auto-terminate once it stays idle for a duration limit called the idle shutdown. The default idle shutdown is 60 minutes, and it can be changed in the settings of the pools and your preferences. Otherwise, you can terminate the pool manually in the Instant Cloud Manager, by selecting the created pool and clicking on the terminate button.

1.4 First Distributed Tuning in the Cloud

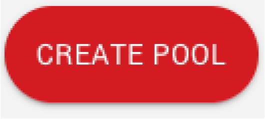
The Gurobi Instant Cloud makes it easy to launch a cluster of machines for distributed tuning. If you already installed a cloud license file for a pool with distributed workers, you can go directly to step 4.

1.4.1 1. Open the Instant Cloud Manager

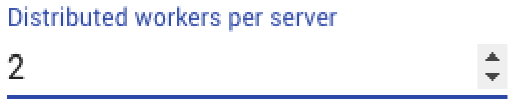
Go to cloud.gurobi.com. If you are not logged in, you will be prompted for your credentials. If you do not have an account, please register and contact Gurobi to request an evaluation license.

1.4.2 2. Create a pool with distributed workers

In the Instant Cloud Manager, go to the 'Pools' section and click on the add new pool button:



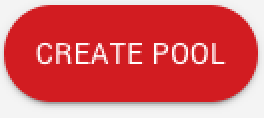
Then, open the 'License' tab and set the number of workers to 2.



Distributed workers per server

2

Finally, create the new pool. Note that a default name is assigned for you such as pool1.



1.4.3 3. Download the pool license file

The list of pools is displayed in the Instant Cloud Manager and your license file is ready to be downloaded with the following button.

⤓

The license file contains the default access ID and secret key for the selected pool. You just have to place this file in your home directory which takes precedence, or in one of the following shared locations:

- C:\gurobi\ on Windows
- /opt/gurobi/ on Linux
- /Library/gurobi/ on Mac OS X

In case you previously had a license file installed, please make sure to replace it, or set the environment variable `GRB_LICENSE_FILE` to point to the cloud license file, it will override the default locations.

1.4.4 4. Tune

You can try to tune a MIP MPS file provided with the Gurobi distribution. Here is an example on Mac OS X:

```
$grbtune /Library/gurobi1001/macos_universal2/examples/data/misc07.mps
Waiting for cloud server to start...........
Capacity available on '999999-pool1' cloud pool - connecting...
Established HTTPS encrypted connection

grbtune version 10.0.1 build v10.0.1rc0 (mac64[x86])
Copyright (c) 2023, Gurobi Optimization, LLC

Read MPS format model from file /Library/gurobi1001/macos_universal2/examples/data/
↪misc07.mps
Reading time = 0.26 seconds
MISC07: 212 rows, 260 columns, 8619 nonzeros

Distributed tuning: launched 3 distributed worker jobs

Solving model using baseline parameter set with TimeLimit=3600s

Solving with random seed #1 ...
Optimize a model with 212 rows, 260 columns and 8619 nonzeros
Variable types: 1 continuous, 259 integer (0 binary)
    [...]
```

Within this log, we have highlighted in bold some important steps. First, the client automatically connects to the Instant Cloud server and checks for the pool status. As the machines are not launched yet, Instant Cloud starts the machines and the client reports that it is waiting until capacity is available.

Then the Gurobi Optimizer detects that the pool is setup with 2 distributed workers. So it automatically starts the tuning in distributed mode with 3 workers (the master compute server counts as one worker as well).

1.4.5 5. Terminate the pool (optional)

The machine will auto-terminate once it stayed idle for a duration limit called the idle shutdown. The default idle shutdown is 60 minutes, and it can be changed in the settings of the pools and your preferences. Otherwise, you can terminate the pool manually in the Instant Cloud Manager, by selecting the created pool and clicking on the terminate button.

■

1.5 Using Instant Cloud in a Program

Using a cloud license file will work seamlessly with any program or supported environment: C++, Python, MATLAB, Java, .Net, C or R. The cloud license file can be easily downloaded from a license, or a pool using the Instant Cloud Manager.

In addition, when programming in C, C++, Python, Java or .Net, the Gurobi client libraries provide you with dedicated environment constructors to specify the access ID, the secret key and optionally the pool. If the pool is not provided, your job will be launched in the default pool associated with your cloud license. Please refer to the Gurobi Optimizer Reference Manual.

Each license comes with a predefined pool called 'default'. You can edit the configuration of pools in the Instant Cloud Manager or create new ones. The updated configuration will be effective for newly launched machines only. So if the machines of a pool are already running, please make sure to terminate them so that new configuration will be taken into account.

One of the important configuration options is the idle shutdown time. When a client program requests a cloud server, it takes some time (usually 1-2 minutes) to launch that server. Rather than forcing client programs to incur this delay each time they run, the Gurobi Instant Cloud leaves a server running until it has been idle for the specified idle shutdown time. In this way, a second client program may find a cloud server already available. You can set this to a small value if you want your server to shut down immediately after your job finishes, or to a very large value if you want your server to always be available.

1.6 Security and Network Settings

Our goal is to provide a secure environment to our customers, and we are continuously monitoring and improving our architecture and processes. In this section, we will review the security features and the required network settings to operate Gurobi Instant Cloud.

1.6.1 Accessing the Cloud Manager

The Cloud Manager is designed to streamline the control of the Gurobi Optimizer on the Cloud. With the Cloud Manager, Gurobi manages AWS EC2 or Azure instances. The Cloud Manager consists of the website **cloud.gurobi.com** and a REST APIs. The main functions of the Cloud Manager are about configuring, controlling and monitoring Gurobi compute servers. No optimization model data is communicated with the Cloud Manager.

When accessing the website, users must be authenticated with their Gurobi accounts. When using the *REST API*, the clients are authenticated with the API key and API secret related to a user account. The communication is secure using the HTTPS protocol (minimum of TLS 1.2) and the Cloud Manager database is encrypted at rest. Access to **cloud.gurobi.com** is also protected by a Web Application Firewall. For security purposes, Gurobi records and monitors the metadata of HTTPS communication.

For better availability and scalability, the Cloud Manager is hosted in different regions of the world. The clients will be routed to the most appropriate available server using a latency based routing. Each region may also provide several instances of the servers. Clients should not hardcode IP addresses to access the Cloud Manager, and should always make sure to use the latest DNS resolution.

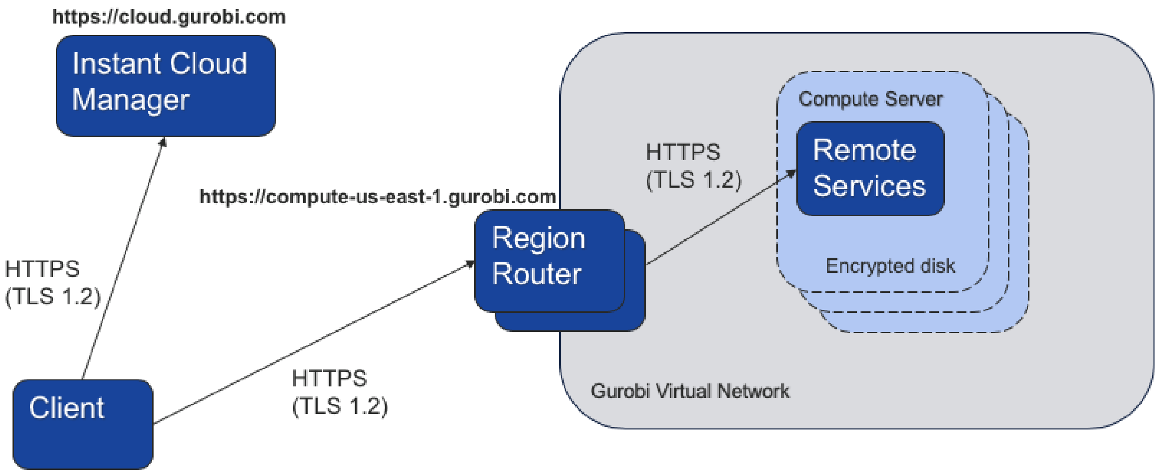
The Gurobi client (gurobi_cl, grbtune, Gurobi library…) will first connect to the Cloud Manager using the secure REST API to check the pool status and launch the compute servers as necessary. In order to enable this connection, the client firewalls must be configured to open the standard HTTPS port 443 to host **cloud.gurobi.com**.

1.6.2 Accessing the Compute Servers

When a Compute Server is started, you get a new EC2 or Azure virtual machine that is not shared with any other Gurobi customers, it is always dedicated. Access to each machine is authenticated with API keys and secured with end-to-end encryption. Machine disks are also encrypted. When the machine is terminated, all optimization data are discarded from memory and disk.

Once the compute server has been launched, optimization commands are exchanged between the client and the server. The communication is secure using end-to-end encryption with HTTPS (minimum of TLS 1.2). The region router consists of a load balancer and a region router. The region router is a reverse proxy that will forward the communication to the appropriate compute server within the Gurobi private virtual network. The load balancer, the region routers and the compute servers all use encrypted HTTPS communication.

The started machines are not accessible directly and passing through the region router is enforced. The diagram below summarizes the architecture with AWS.



As shown below, each region provides a different URL address to its router. Clients should not hardcode IP addresses to access the region routers, and should always make sure to use the latest DNS resolution. In order to enable this connection, the client firewalls must be configured to open the standard HTTPS port 443 to the following hosts depending on the region.

Provider	Region	Router
AWS	us-east-1	https://compute-us-east-1.gurobi.com
AWS	us-west-1	https://compute-us-west-1.gurobi.com
AWS	eu-central-1	https://compute-eu-central-1.gurobi.com
AWS	ap-northeast-1	https://compute-ap-northeast-1.gurobi.com
AWS	ap-southeast-2	https://compute-ap-southeast-2.gurobi.com
Azure	eastus	https://compute-eastus-azure.gurobi.com
Azure	westus2	https://compute-westus2-azure.gurobi.com
Azure	westeurope	https://compute-westeurope-azure.gurobi.com

1.6.3 Managing API keys

The Cloud Manager website is the only place where the API keys can be generated. Multiple API keys can be generated so that keys can be replaced in case one of them has been compromised. Each key is owned by a user. Before disabling a user by contacting the Gurobi support or before deleting an API key, please make sure that you have migrated your applications to new API keys.

1.6.4 Proxies

The architecture is compatible with standard proxy settings using environment variables HTTP_PROXY and HTTPS_PROXY. HTTPS_PROXY takes precedence over HTTP_PROXY for https requests. The values may be either a complete URL or a “host[:port]”, in which case the “http” scheme is assumed.

CHAPTER

TWO

INSTANT CLOUD MANAGER

The Gurobi Instant Cloud Manager is the web application that lets you manage your Gurobi Cloud environment. Please refer to the on-line documentation.

CHAPTER

THREE

INSTANT CLOUD REST API

The Instant Cloud REST API lets you build custom solutions or frameworks when you need to create, start and stop machines or pools automatically. The API follows standard REST principles and can be used in various languages and tools (Java, Python, Node.js, curl…)

3.1 REST API v2

The API v2 provides a set of endpoints to perform the following actions:

- list licenses,
- list, launch and terminate machines,
- list, create and delete machine pools,
- launch and terminate machine pools.

3.1.1 API keys

In order to use the API, you will need an API key consisting of an **API access ID** and a **secret key**. Please refer to the API key documentation to review the steps to retrieve the keys in the Instant Cloud Manager.

If a key has been created for a specific license, only the machines and pools related to this license will be accessible. If it is a global account key, all the information related to the account will be accessible.

The secret key is a like a password and it should never be shared with others and should not be sent by emails.

The access ID and the secret key must be passed with each request in the following HTTP headers:

- `X-GUROBI-ACCESS-ID`
- `X-GUROBI-SECRET-KEY`

3.1.2 Base URL

The API can be accessed at the following base URL:

```
https://cloud.gurobi.com/api/v2
```

HTTPS must be used to ensure that the communication is encrypted. Using HTTP will return an error.

3.1.3 Reference Documentation

The reference documentation provides all the details about the API by listing the endpoints and specifying the parameters as well as the input and output data. It is also an interactive API playground to let you try out each endpoint.