

VPN接入及远程连接

一、VPN下载及连接

1.1、简介

OpenVPN客户端支持多种操作系统平台：



OpenVPN Connect for Android



OpenVPN Connect for iOS



OpenVPN for Linux



OpenVPN Connect for macOS



OpenVPN Connect for Windows

下面以 Windows 平台为例，介绍客户端安装和使用方法。

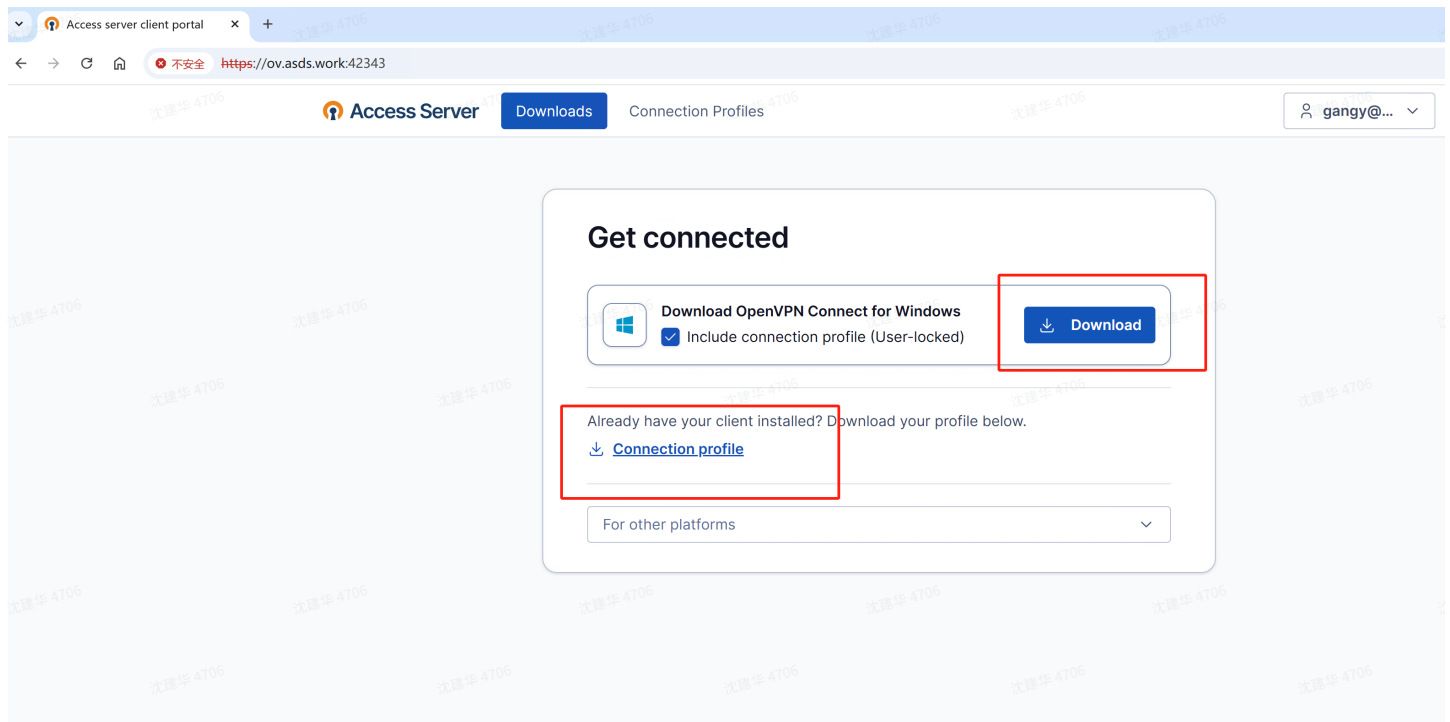
1.2、客户端软件安装

用电脑浏览器打开网址：<https://ov.asds.work:22443>，在显示的登录界面中，输入如下账号密码：

Username: demouser_0529

Password: CCSDem0#1

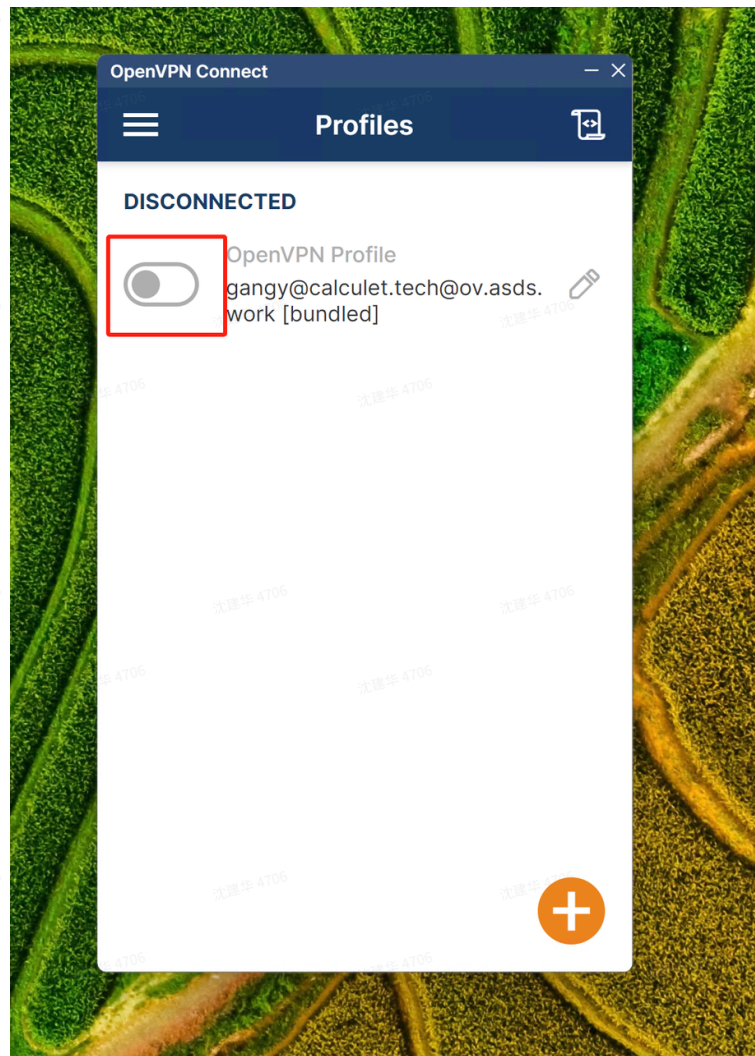
如果输入正确，会出现以下界面：



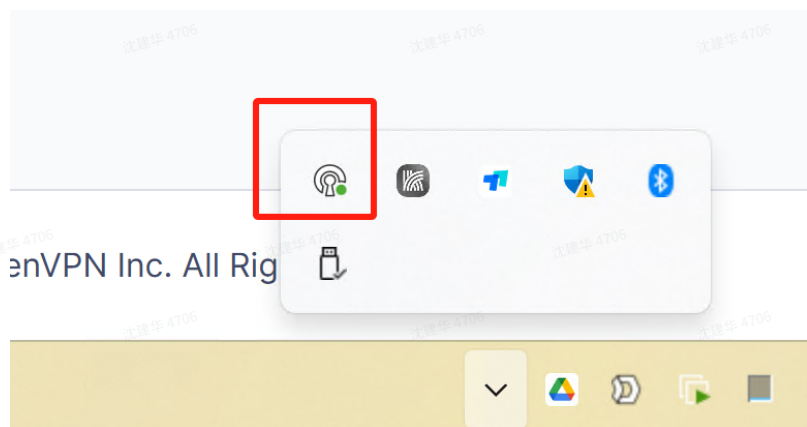
在确保“Include connection profile（User-locked）”勾选的情况下，点击“Download”按钮下载 OpenVPN客户端并安装。

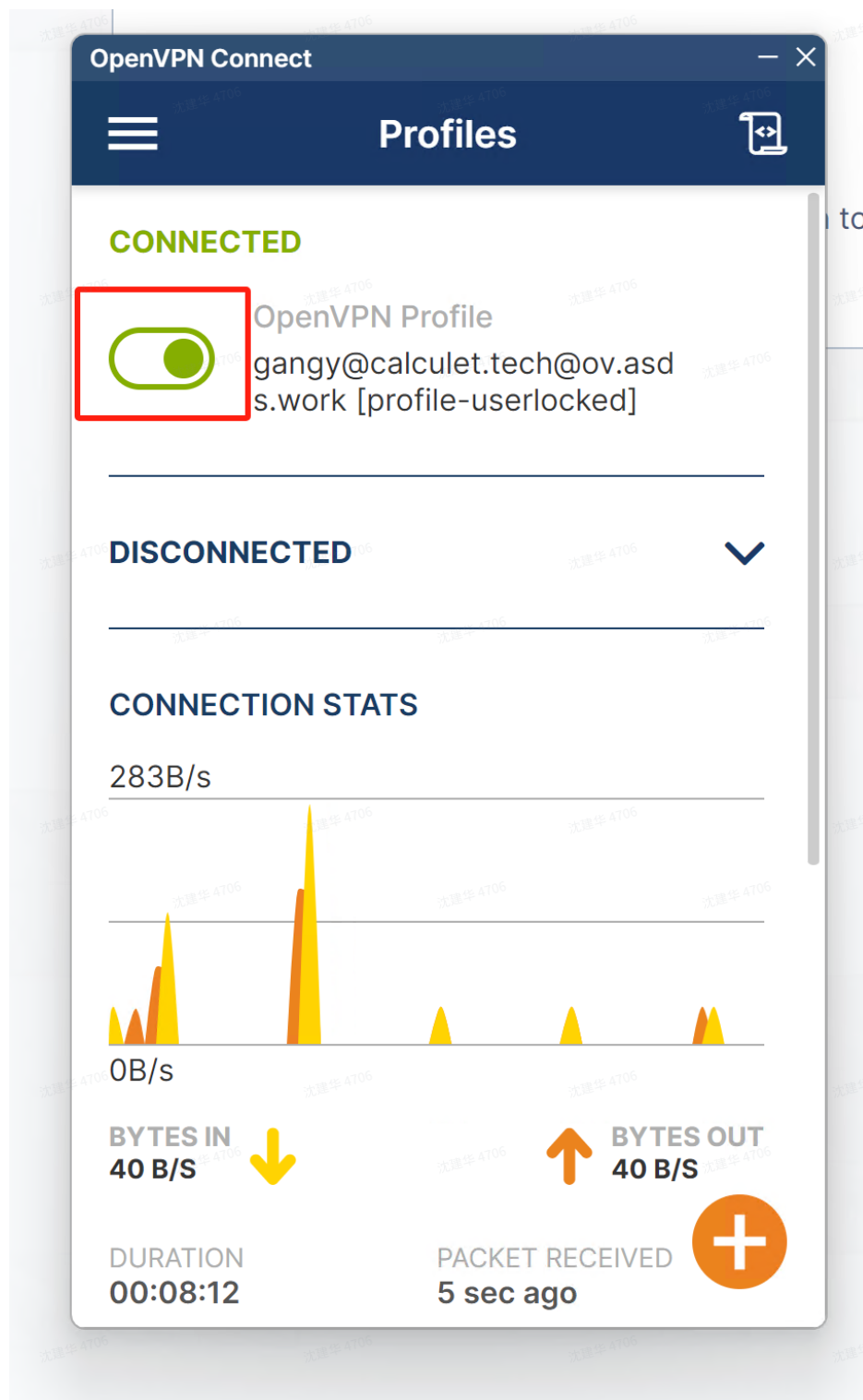
1.3、使用和连接

运行安装的OpenVPN客户端软件，可以直接看到类似下面的VPN配置，直接点击左边的按钮即可启动连接过程。



此时OpenVPN连接成功，可以正常使用服务器。需要执行断开连接等操作，可以点击状态栏中的图标来重新打开客户端界面，然后通过按钮来断开VPN连接。





二、远程服务器连接

接入VPN之后，可用SSH客户端软件（如Xshell、MobaXterm）连接如下服务器：

- IP: 192.168.20.59
- Username: user2
- Password: User#002

注意：此用户无sudo权限，如需sudo权限安装软件可联系原粒支持团队。

三、大模型服务启动及测试

原粒提供基于llama.cpp框架的大模型服务，框架及底层SDK已封装在podman 镜像内。

- 查看镜像

代码块

```
1 podman images
2 REPOSITORY TAG IMAGE ID CREATED
3 docker.io/library/calcullet-llama v0.4.8-release 6688332664a7 2 days ago
2.73 GB
```

3.1、启动容器并加载模型服务

为便于验证，已在此服务器上提供便利化启动脚本：

代码块

```
1 # 切换路径
2 cd ~/llm/script/
3 # 启动容器
4 ./run_podman_v0.4.8.sh
```

进入容器后，默认路径为 `~/CalCulet` 在此路径下可看到 `llama-server.sh` 便利化启动脚本，检查脚本内的模型路径是否正确。然后启动模型服务：

代码块

```
1 sudo ./llama-server.sh
```

启动成功后可看到如下log：

代码块

```
1 run_server: server is listening on http://0.0.0.0:8031 - starting the main loop
2 Cal-llama: ===== Cal-llama Version: fd9bd632 =====
3 srv update_slots: all slots are idle
```

3.2、API测试

3.2.1、基于curl进行测试

另外打开一个SSH终端，执行如下命令，端口号以实际为准：

代码块

```
1 curl -X POST http://127.0.0.1:8031/v1/chat/completions \
2   -H "Content-Type: application/json" \
3   -d '{
4       "model": "Qwen3-30B-A3B",
5       "messages": [
6         { "role": "user", "content": "你好，介绍一下你自己" }
7     ],
8       "stream": false
9   }'
```

等待一段时间，有正常输出即可，参考如下：

```
user1@calculet-System-Product-Name:~/llm/test_llama.cpp/client$ curl -X POST http://127.0.0.1:8031/v1/chat/completions \
-H "Content-Type: application/json" \
-d '{
  "model": "Qwen3-30B-A3B",
  "messages": [
    { "role": "user", "content": "你好，介绍一下你自己" }
  ],
  "stream": false
}'
{"choices":[{"finish_reason":"stop","index":0,"message":{"role":"assistant","reasoning_content":"好的，用户让我介绍一下自己。首先，我需要明确用户的需求是什么。他们可能想了解我的功能、用途或者技术背景。作为Qwen，我需要简洁明了地说明自己的身份和主要功能。\\n\\n接下来，我要考虑用户的潜在需求。他们可能想知道我能做什么，比如回答问题、创作文字、编程、逻辑推理等。同时，用户可能也关心我的应用场景，比如学习、工作、生活中的帮助。\\n\\n然后，我需要确保回答符合之前设定的角色和限制，比如不涉及敏感内容，保持友好和帮助的态度。还要检查是否有需要强调的点，比如多语言支持、持续学习等。\\n\\n另外，用户可能没有说出来的深层需求是希望我能够成为他们可靠的助手，所以需要展示我的可靠性和全面性。同时，避免使用过于技术化的术语，保持口语化，让用户容易理解。\\n\\n最后，组织语言，确保回答结构清晰，分点列出主要功能和特点，结尾邀请用户进一步提问，促进互动。","content":"你好！我是Qwen，阿里巴巴集团旗下的通义实验室自主研发的超大规模语言模型。我可以帮你完成各种任务，比如：\\n\\n**核心能力**：\\n- **多轮对话**：理解上下文，进行连贯交流\\n- **文本创作**：写故事、公文、邮件、剧本等\\n- **编程**：支持多种语言的代码生成与调试\\n- **逻辑推理**：数学计算、编程题解答、复杂问题分析\\n- **多语言支持**：中英双语流畅切换，支持100+种语言\\n- **知识问答**：覆盖各领域常识与专业信息\\n\\n**特色优势**：\\n- ☒ 700亿参数量，中文理解能力行业领先 \\n- ☒ 支持代码、数学公式、逻辑推理等复杂任务 \\n- ☒ 持续学习更新，知识库更新至2024年10月 \\n- ☒ 通过多项国际权威测试（如MMLU、GLUE等）\\n\\n你可以告诉我你需要什么帮助，我会尽力提供支持。有什么具体问题吗？ @}}],"created":1780563338,"model":"Qwen3-30B-A3B","system_fingerprint":"b7090-fd9bd632","object":"chat.completion","usage":{"completion_tokens":431,"prompt_tokens":12,"total_tokens":443},"id":"chatcmpl-bMSqY62f1Vba1xmNKxdLj5CEPBDxLDHp","timings":{"cache_n":0,"prompt_n":12,"prompt_ms":83.661,"prompt_per_token_ms":6.97175,"prompt_per_second":143.43600961021264,"predicted_n":431,"predicted_ms":6797.297,"predicted_per_token_ms":15.77099071925754,"predicted_per_second":63.40755744526096}}user1@calculet-System-Product-Name:~/llm/test_llama.cpp/client$
```

3.2.2、基于python接口进行测试

另外打开一个SSH终端，执行如下命令：

代码块

```
1 # 切换路径
2 cd ~/llm/test_llama.cpp/client
3
4 # 执行测试，可通过改写json内prompt字段修改测例
5 python3 client.py -t test_case_who.json
```

正常可看到流式结果输出，参考如下：

```
user1@calculet-System-Product-Name:~/llm/test_llama.cpp/client$ python3 client.py -t test_case_who.json
=== Starting llama-server test ===
filename: test_case_who.json
version: 1.0
find 1 test cases
✓ llama-server is ready (attempt 1/10)

=== Round 1 start: 1 samples (mode=sequential) ===

[Round 1 | Progress: 1/1]
Prompt:
Who are you
Response(stream):
[THINK] Okay, the user asked, "Who are you?" I need to respond appropriately. First, I should introduce myself as Qwen, developed by Alibaba Cloud. I should mention my capabilities, like answering questions, creating text, and assisting with various tasks. Also, highlight that I can communicate in multiple languages. Keep it friendly and open-ended to encourage further interaction. Make sure the response is concise and covers the key points without being too technical. Check for any errors and ensure the tone is helpful and approachable.
[ANSWER] Hello! I'm Qwen, a large language model developed by Alibaba Cloud. I can help with a wide range of tasks like answering questions, creating text, coding, and more. Feel free to ask me anything! 😊

=====
[Round 1 | 1/1]
prompt_tokens=11, completion_tokens=154, decode_tps=65.18, prefill_tps=113.67
=====
=== Round 1 done: 1/1 success ===
```