浙江大学

本科实验报告

课程名称：	计算机网络基础
实验名称：	实现一个轻量级的 WEB 服务器
姓　　名：	于倚岑
学　　院：	计算机学院
系：	计算机科学与技术
专　　业：	计算机科学与技术
学　　号：	3190105224
指导教师：	张泉方

2021 年　　12 月　　20　日

浙江大学实验报告

实验名称：实现一个轻量级的 WEB 服务器　实验类型：　编程实验

同组学生：　　实验地点：　计算机网络实验室

一、　实验目的

深入掌握 HTTP 协议规范，学习如何编写标准的互联网应用服务器。

二、　实验内容

- 服务程序能够正确解析 HTTP 协议，并传回所需的网页文件和图片文件
- 使用标准的浏览器，如 IE、Chrome 或者 Safari，输入服务程序的 URL 后，能够正常显示服务器上的网页文件和图片
- 服务端程序界面不做要求，使用命令行或最简单的窗体即可
- 功能要求如下：
  1. 服务程序运行后监听在 80 端口或者指定端口
  2. 接受浏览器的 TCP 连接（支持多个浏览器同时连接）
  3. 读取浏览器发送的数据，解析 HTTP 请求头部，找到感兴趣的部分
  4. 根据 HTTP 头部请求的文件路径，打开并读取服务器磁盘上的文件，以 HTTP 响应格式传回浏览器。要求按照文本、图片文件传送不同的 Content-Type，以便让浏览器能够正常显示。
  5. 分别使用单个纯文本、只包含文字的 HTML 文件、包含文字和图片的 HTML 文件进行测试，浏览器均能正常显示。
- 本实验可以在前一个 Socket 编程实验的基础上继续，也可以使用第三方封装好的 TCP 类进行网络数据的收发
- 本实验要求不使用任何封装 HTTP 接口的类库或组件，也不使用任何服务端脚本程序如 JSP、ASPX、PHP 等

三、　主要仪器设备

联网的 PC 机、Wireshark 软件、Visual Studio、gcc 或 Java 集成开发环境。

四、　操作方法与实验步骤

- 阅读 HTTP 协议相关标准文档，详细了解 HTTP 协议标准的细节，有必要的话使用 Wireshark 抓包，研究浏览器和 WEB 服务器之间的交互过程
- 创建一个文档目录，与服务器程序运行路径分开
- 准备一个纯文本文件，命名为 test.txt，存放在 txt 子目录下面
- 准备好一个图片文件，命名为 logo.jpg，放在 img 子目录下面
- 写一个 HTML 文件，命名为 test.html，放在 html 子目录下面，主要内容为：

```
<html>
    <head><title>Test</title></head>
    <body>
        <h1>This is a test</h1>
        <img src="img/logo.jpg">
        <form action="dopost" method="POST">
            Login:<input name="login">
            Pass:<input name="pass">
            <input type="submit" value="login">
        </form>
    </body>
</html>
```

- 将 test.html 复制为 noimg.html，并删除其中包含 img 的这一行。
- 服务端编写步骤（**需要采用多线程模式**）

a) 运行初始化，打开 Socket，监听在指定端口（**请使用学号的后 4 位作为服务器的监听端口**）

b) 主线程是一个循环，主要做的工作是等待客户端连接，如果有客户端连接成功，为该客户端创建处理子线程。该子线程的主要处理步骤是：

1. 不断读取客户端发送过来的字节，并检查其中是否连续出现了 2 个回车换行符，如果未出现，继续接收；如果出现，按照 HTTP 格式解析第 1 行，分离出方法、文件和路径名，其他头部字段根据需要读取。

✧ **如果解析出来的方法是 GET**

2. 根据解析出来的文件和路径名，读取响应的磁盘文件（该路径和服务器程序可能不在同一个目录下，需要转换成绝对路径）。如果文件不存在，第 3 步的响应消息的状态设置为 404，并且跳过第 5 步。
3. 准备好一个足够大的缓冲区，按照 HTTP 响应消息的格式先填入第 1 行（状态码=200），加上回车换行符。然后模仿 Wireshark 抓取的 HTTP 消息，填入必要的几行头部（需要哪些头部，请试验），其中不能缺少的 2 个头部是 Content-Type 和 Content-Length。Content-Type 的值要和文件类型相匹配（请通过抓包确定应该填什么），Content-Length 的值填写文件的字节大小。
4. 在头部行填完后，再填入 2 个回车换行
5. 将文件内容按顺序填入到缓冲区后面部分。

✧ **如果解析出来的方法是 POST**

6. 检查解析出来的文件和路径名，如果不是 dopost，则设置响应消息的状态为 404，然后跳到第 9 步。如果是 dopost，则设置响应消息的状态为 200，并继续下一步。
7. 读取 2 个回车换行后面的体部内容（长度根据头部的 Content-Length 字段的指示），并提取出登录名（login）和密码（pass）的值。如果登录名是你的学号，密码是学号的后 4 位，则将响应消息设置为登录成功，否则将响应消息设置为登录失败。
8. 将响应消息封装成 html 格式，如

<html><body>响应消息内容</body></html>

9. 准备好一个足够大的缓冲区，按照 HTTP 响应消息的格式先填入第 1 行（根据前面的情况设置好状态码），加上回车换行符。然后填入必要的几行头部，其中不能缺少的 2 个头部是 Content-Type 和 Content-Length。Content-Type 的值设置为 text/html，如果状态码=200，则 Content-Length 的值填写响应消息的字节大小，并将响应消息填入缓冲区的后面部分，否则填写为 0。
10. 最后一次性将缓冲区内的字节发送给客户端。
11. 发送完毕后，关闭 socket，退出子线程。

c) 主线程还负责检测退出指令（如用户按退出键或者收到退出信号），检测到后即通知并等待各子线程退出。最后关闭 Socket，主程序退出。

- 编程结束后，将服务器部署在一台机器上（本机也可以）。在服务器上分别放置纯文本文件（.txt）、只包含文字的测试 HTML 文件（将测试 HTML 文件中的包含 img 那一行去掉）、包含文字和图片的测试 HTML 文件（以及图片文件）各一个。
- 确定好各个文件的 URL 地址，然后使用浏览器访问这些 URL 地址，如 http://x.x.x.x:port/dir/a.html，其中 port 是服务器的监听端口，dir 是提供给外部访问的路径，请设置为与文件实际存放路径不同，通过服务器内部映射转换。
- 检查浏览器是否正常显示页面，如果有问题，查找原因，并修改，直至满足要求
- 使用多个浏览器同时访问这些 URL 地址，检查并发性

五、 实验数据记录和处理

请将以下内容和本实验报告一起打包成一个压缩文件上传：

- **源代码：需要说明编译环境和编译方法，如果不能编译成功，将影响评分**
- **可执行文件：可运行的.exe 文件或 Linux 可执行文件**

- 服务器的主线程循环关键代码截图（解释总体处理逻辑，省略细节部分）

```
while (1)
{
    int szClntAddr = sizeof(clntAddr);
    sClient = accept(sListen, (SOCKADDR*)&clntAddr, &szClntAddr);
    if (sClient == INVALID_SOCKET)
    {
        printf("accept() error! code:%d\n", WSAGetLastError());
        closesocket(sClient);
        return 0;
    }
    else
    {
        std::thread(solve, std::ref(sClient)).detach();
    }
}
```

使用 accept 监听客户端，当一个连接建立成功后，新建一个线程，并在这个线程里处理请求。

- 服务器的客户端处理子线程关键代码截图（解释总体处理逻辑，省略细节部分）

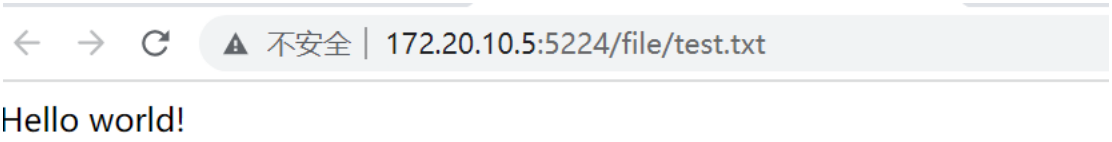
```
SOCKADDR_IN client_info = { 0 };
int addrsize = sizeof(client_info);
std::thread::id thread_id = std::this_thread::get_id();
getpeername(sClient, (struct sockaddr*)&client_info, &addrsize);
char* ip = inet_ntoa(client_info.sin_addr);
int port = client_info.sin_port;
Client now;
now.socket = sClient;now.ip = ip;now.id = count++;
now.port = port;now.thread_id = thread_id;
Client* pnow = &now;
lock.lock();
client_list.push_back(pnow);
lock.unlock();
bool finish = 0;
char buffer[MAXBUFFER] = "\0";
while (!finish)
{
    ZeroMemory(buffer, MAXBUFFER);
    int state = recv(sClient, buffer, MAXBUFFER, 0);
    std::cout << buffer << std::endl;

    if (state > 0)
    {
        message msg(buffer, sClient, now.id);
        handleMessage(msg);
    }
    else if (state == 0)
    {
        printf("Connection closed.\n");
        finish = 1;
    }
    else
    {
        printf("rev() error!\n");
        closesocket(sClient);
        finish = 1;
    }
}
closesocket(sClient);
std::vector<Client*>::iterator it;
for (it = client_list.begin();it != client_list.end();it++)
{
    if ((*it)->id == pnow->id && (*it)->ip == pnow->ip) break;
}
lock.lock();
client_list.erase(it);
lock.unlock();
return;
```

处理一个子线程时，使用一个循环来接收客户端发送的请求（注意客户端可能会发送很多请求来获得一个网页的全部资源，必须逐个处理），在与客户端建立连接后，我们把这个线程加入到线程池中，等到连接关闭后把它从线程池移除。

- 服务器运行后，用 netstat –an 显示服务器的监听端口

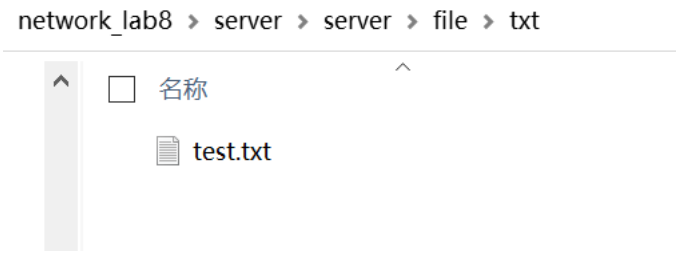
```
TCP    0.0.0.0:135            0.0.0.0:0              LISTENING
TCP    0.0.0.0:443            0.0.0.0:0              LISTENING
TCP    0.0.0.0:445            0.0.0.0:0              LISTENING
TCP    0.0.0.0:902            0.0.0.0:0              LISTENING
TCP    0.0.0.0:912            0.0.0.0:0              LISTENING
TCP    0.0.0.0:1029           0.0.0.0:0              LISTENING
TCP    0.0.0.0:3306           0.0.0.0:0              LISTENING
TCP    0.0.0.0:5040           0.0.0.0:0              LISTENING
TCP    0.0.0.0:5224           0.0.0.0:0              LISTENING
```

- 浏览器访问纯文本文件（.txt）时，浏览器的 URL 地址和显示内容截图。



Hello world!

服务器上文件实际存放的路径：



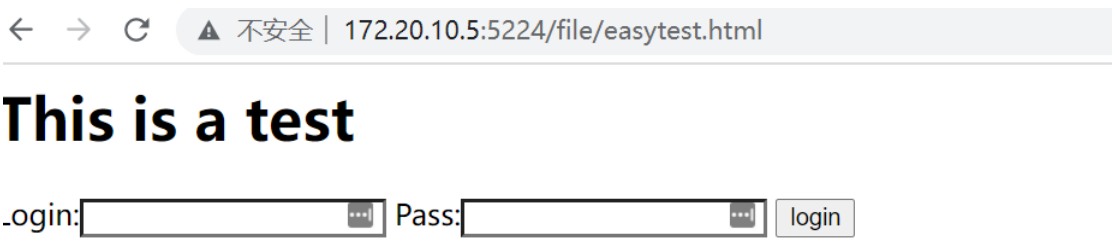
服务器的相关代码片段：

```
if (data.substr(0, 6) == "/file/")
{
    if (data.substr(6, 13) == "../img/")
    {
        int pos = data.find('.');
        std::string type = data.substr(pos + 1);
        if (type == "jpg")
        {
            std::string filepath = "file/img/" + data.substr(13);
            std::cout << filepath << std::endl;
            sendFile(filepath, s);
        }
    }
    int pos = data.find('.');
    std::string type = data.substr(pos + 1);
    if (type == "html")
    {
        std::string filepath = "file/html/" + data.substr(6);
        std::cout << filepath << std::endl;
        sendFile(filepath, s);
    }
    else if (type == "jpg")
    {
        std::string filepath = "file/img/" + data.substr(6);
        std::cout << filepath << std::endl;
        sendFile(filepath, s);
    }
    else if (type == "txt")
    {
        std::string filepath = "file/txt/" + data.substr(6);
        std::cout << filepath << std::endl;
        sendFile(filepath, s);
    }
```

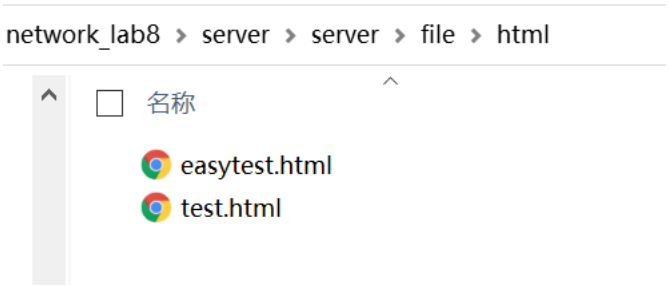
Wireshark 抓取的数据包截图（通过跟踪 TCP 流，只截取 HTTP 协议部分）：

```
    5708 66.678711    172.20.10.1     172.20.10.5     TCP      78 53141 → 5224 [SYN, ECN, CWR] Seq=
    5709 66.678893    172.20.10.5     172.20.10.1     TCP      66 5224 → 53141 [SYN, ACK] Seq=0 Ack=
    5710 66.693205    172.20.10.1     172.20.10.5     TCP      54 53141 → 5224 [ACK] Seq=1 Ack=1 Wir
    5711 66.693206    172.20.10.1     172.20.10.5     HTTP    445 GET /file/test.txt HTTP/1.1
    5712 66.702055    172.20.10.5     172.20.10.1     TCP     133 5224 → 53141 [PSH, ACK] Seq=1 Ack=
    5713 66.712041    172.20.10.1     172.20.10.5     TCP      54 53141 → 5224 [ACK] Seq=392 Ack=80
    5714 66.712107    172.20.10.5     172.20.10.1     HTTP     66 HTTP/1.1 200 OK  (text/html)
    5715 66.724038    172.20.10.1     172.20.10.5     TCP      54 53141 → 5224 [ACK] Seq=392 Ack=92
    5805 97.260461    172.20.10.1     172.20.10.5     TCP      54 53141 → 5224 [FIN, ACK] Seq=392 A

> GET /file/test.txt HTTP/1.1\r\n
  Host: 172.20.10.5:5224\r\n
  Upgrade-Insecure-Requests: 1\r\n
  Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8\r\n
  User-Agent: Mozilla/5.0 (iPhone; CPU iPhone OS 14_7_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko)
  Accept-Language: zh-cn\r\n
  Accept-Encoding: gzip, deflate\r\n
```

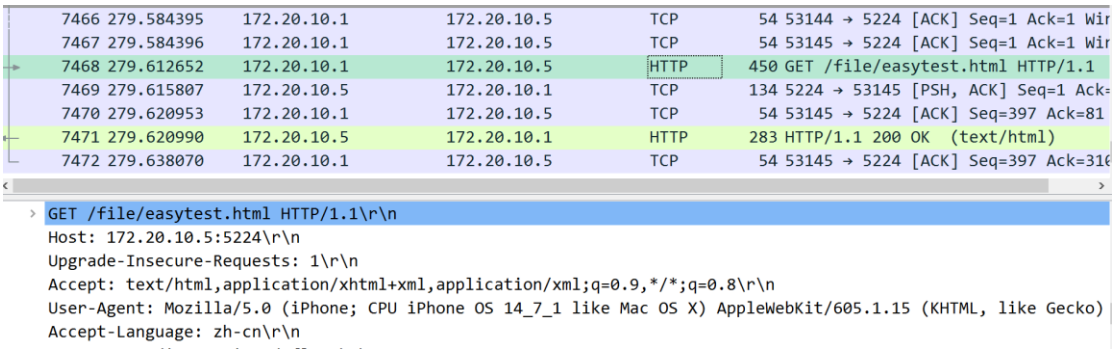
- 浏览器访问只包含文本的 HTML 文件时，浏览器的 URL 地址和显示内容截图。

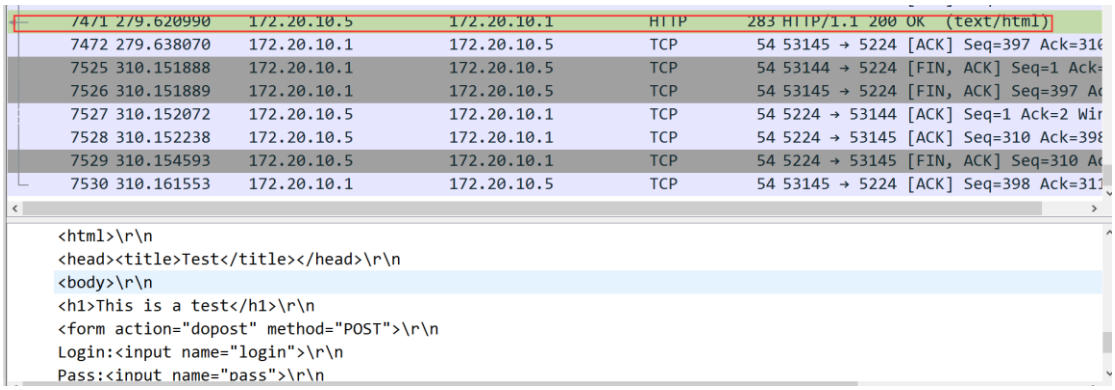


服务器文件实际存放的路径：

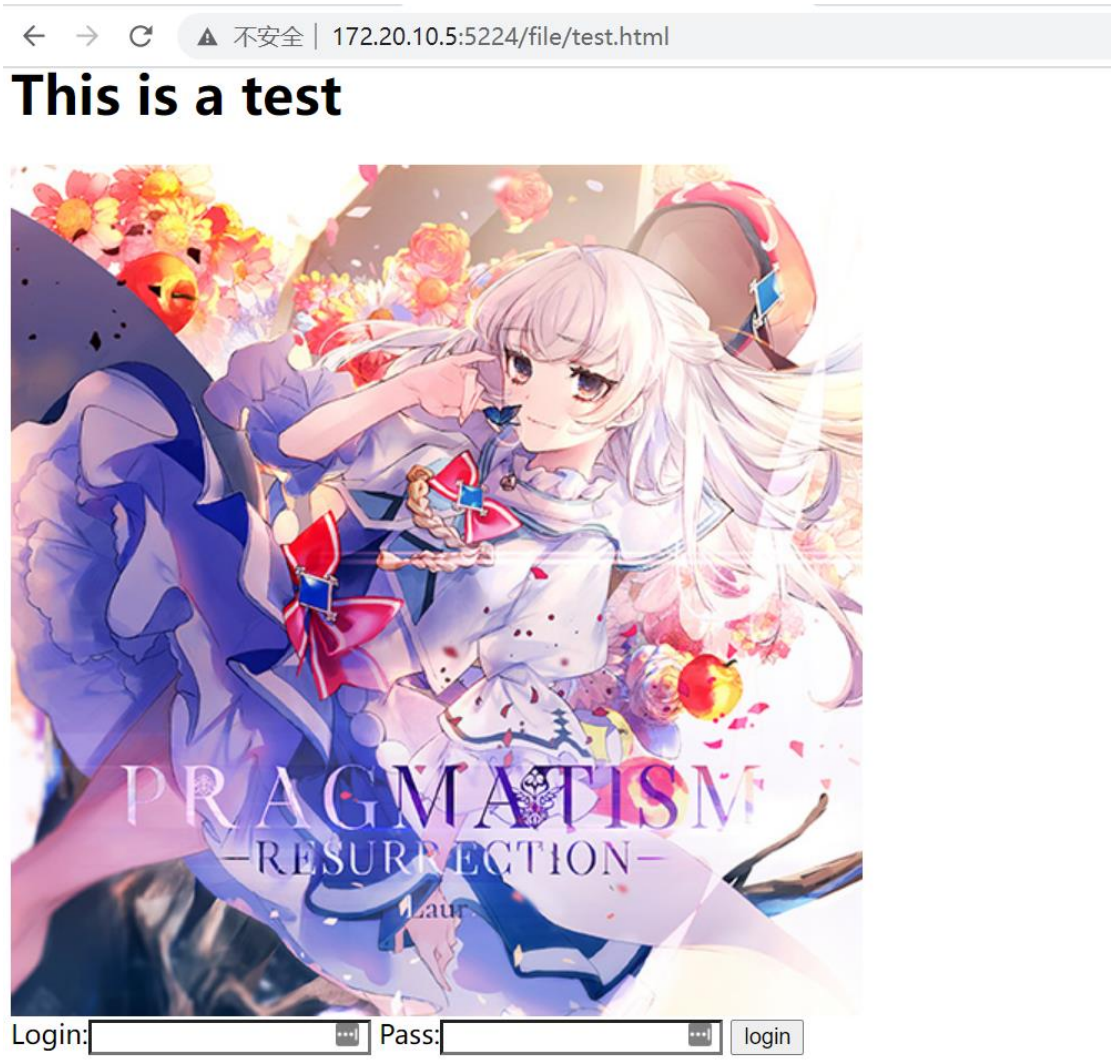


Wireshark 抓取的数据包截图（只截取 HTTP 协议部分，包括 HTML 内容）：

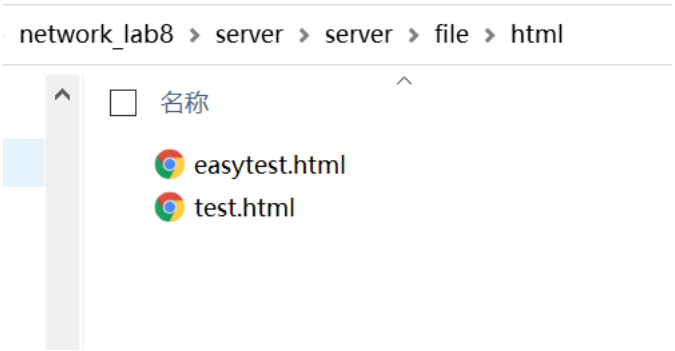




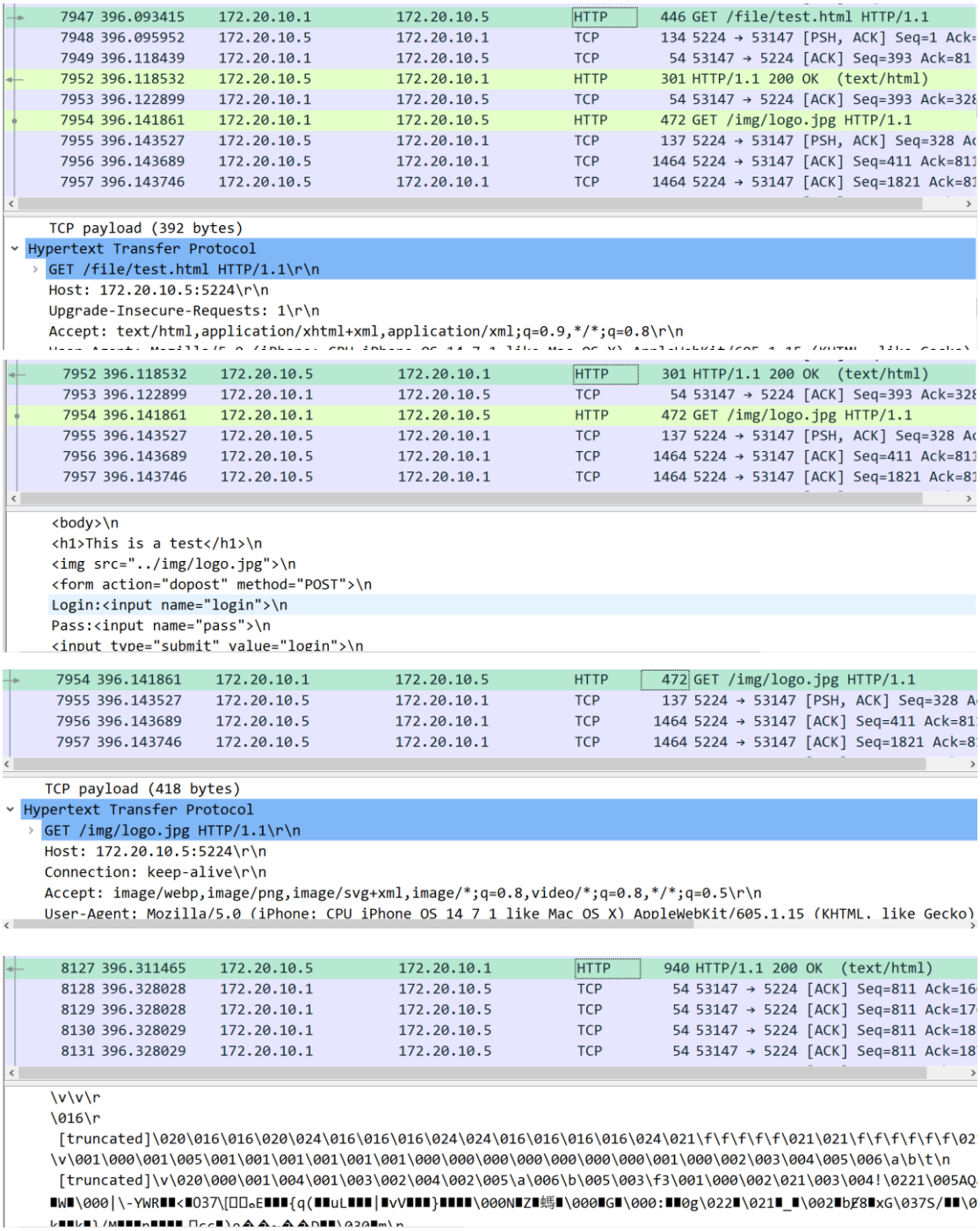
- 浏览器访问包含文本、图片的 HTML 文件时，浏览器的 URL 地址和显示内容截图。



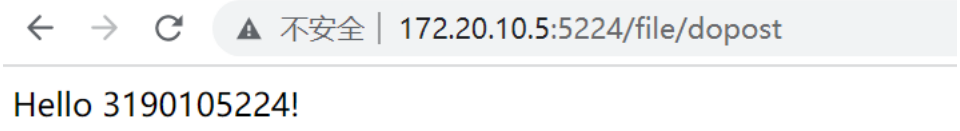
服务器上文件实际存放的路径：



Wireshark 抓取的数据包截图（只截取 HTTP 协议部分，包括 HTML、图片文件的部分内容）：



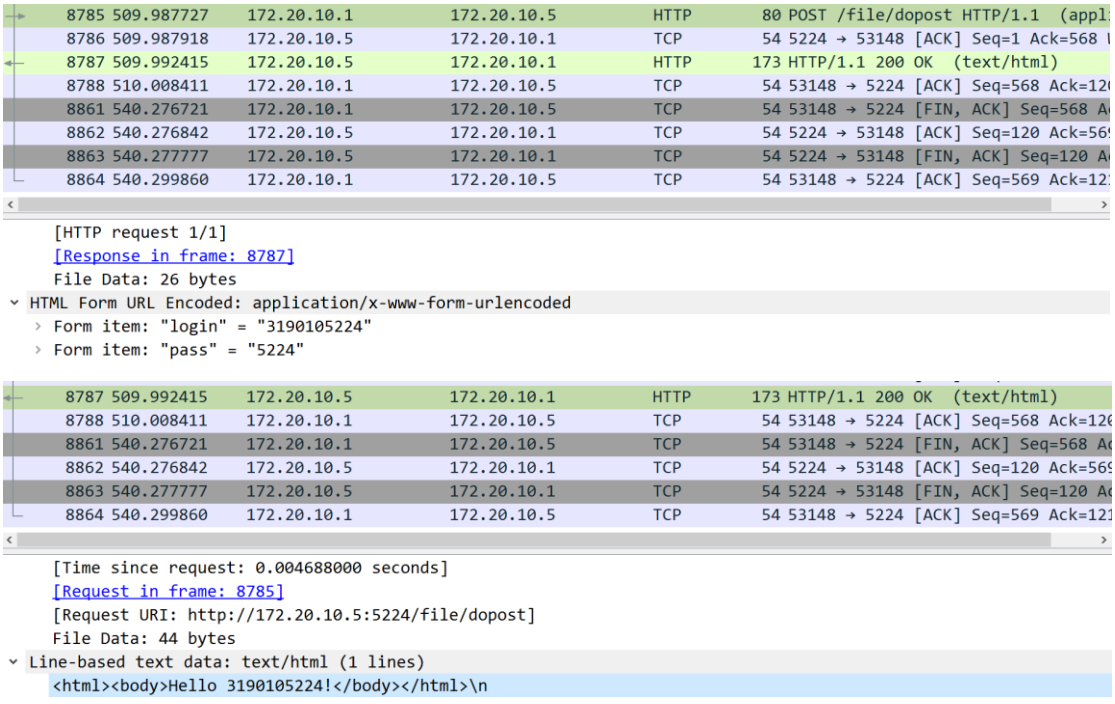
- 浏览器输入正确的登录名或密码，点击登录按钮（login）后的显示截图。



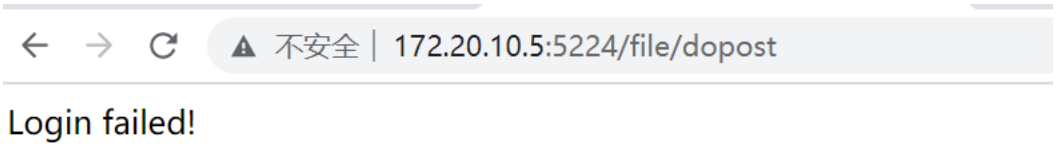
服务器相关处理代码片段：

```
std::cout << "Get a post" << std::endl;
std::string name, password,data=s.data;
data = s.data + 5;
int pos1 = data.find("login"),pos2=data.find("pass");
name = data.substr(pos1 + 6, pos2 - pos1 - 7);
password = data.substr(pos2 + 5);
std::string tmp;
//std::cout<<data<<s
std::cout << name << " " << password << std::endl;
if (name == "3190105224" && password == "5224")
{

    std::cout << "login success";
    tmp += "<html><body>Hello " + name + "!</body></html>\n";
}
else
{
    tmp += "<html><body>Login failed!</body></html>\n";
}
int len = tmp.length();
std::string res;
res += "HTTP/1.1 200 OK\n";
res += "Content-Type: text/html;charset=gb2312\nContent-Length: ";
res += std::to_string(len);
res += "\n\n";
res += tmp;
std::cout << res << std::endl;
if (send(s.sClient, res.c_str(), res.length(), 0) == SOCKET_ERROR)
{
    printf("send failed!");
}
else
{
    printf("send successfully!");
}
```

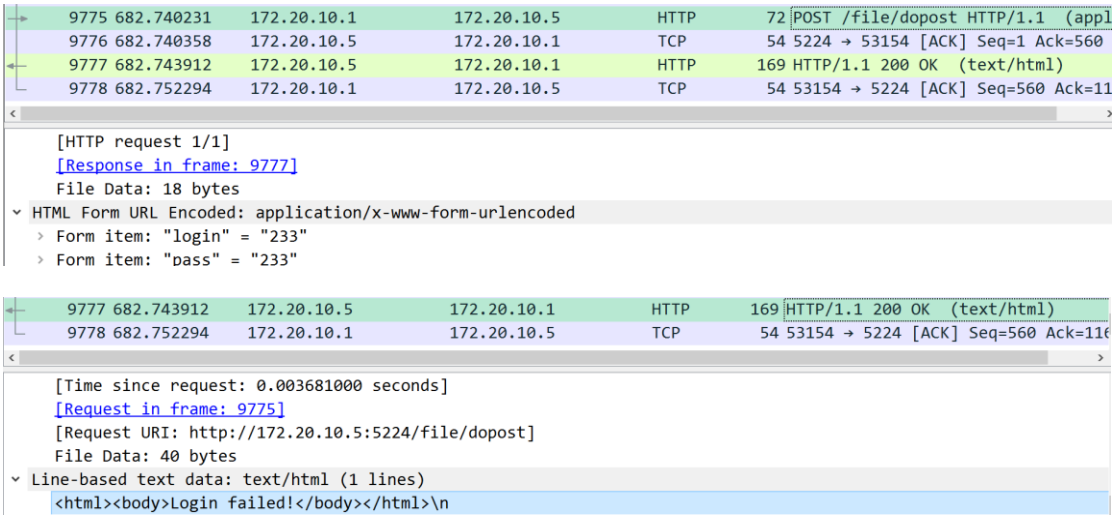
Wireshark 抓取的数据包截图（HTTP 协议部分）



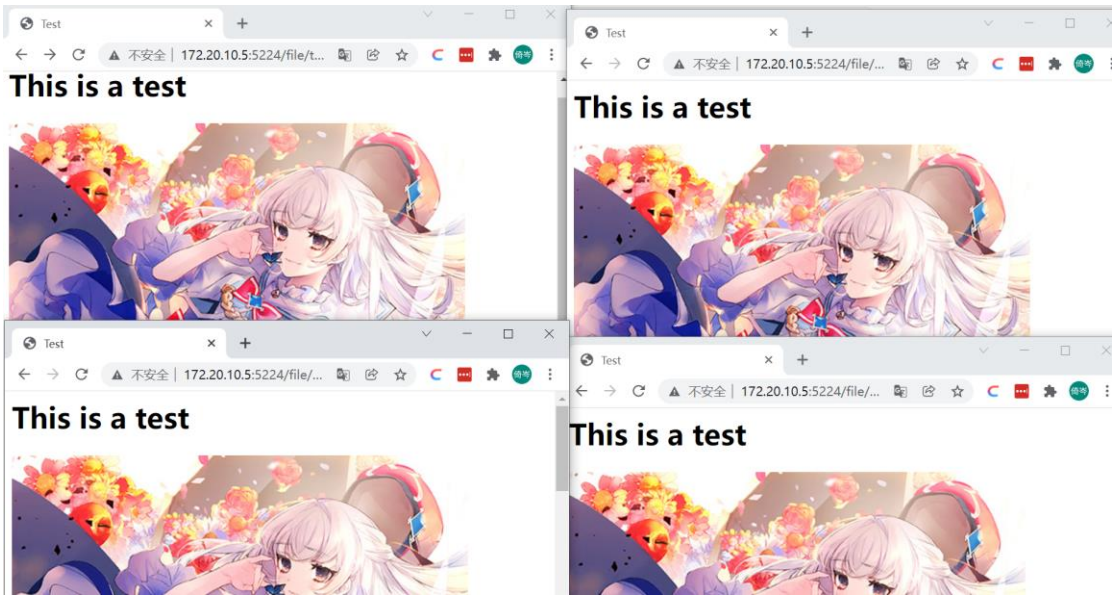
- 浏览器输入错误的登录名或密码，点击登录按钮（login）后的显示截图。



- Wireshark 抓取的数据包截图（HTTP 协议部分）



- 多个浏览器同时访问包含图片的 HTML 文件时，浏览器的显示内容截图（将浏览器窗口缩小并列）



- 多个浏览器同时访问包含图片的 HTML 文件时，使用 netstat –an 显示服务器的 TCP 连接（截取与服务器监听端口相关的）

```
TCP    172.20.10.5:5224       172.20.10.1:53142      CLOSE_WAIT
TCP    172.20.10.5:5224       172.20.10.1:53144      CLOSE_WAIT
TCP    172.20.10.5:5224       172.20.10.5:1942       CLOSE_WAIT
TCP    172.20.10.5:5224       172.20.10.5:2162       ESTABLISHED
TCP    172.20.10.5:5224       172.20.10.5:2182       ESTABLISHED
TCP    172.20.10.5:5224       172.20.10.5:2184       ESTABLISHED
TCP    172.20.10.5:5224       172.20.10.5:2185       ESTABLISHED
```

六、 实验结果与分析

- HTTP 协议是怎样对头部和体部进行分隔的？

  通过一个空行来分隔。发送连续的两个\r\n 代表体部开始，接下来是要发送的数据。

```
login successHTTP/1.1 200 OK
Content-Type: text/html;charset=gb2312
Content-Length: 44

<html><body>Hello 3190105224!</body></html>
```

- 浏览器是根据文件的扩展名还是根据头部的哪个字段判断文件类型的？

  根据头部的 Content-Type 判断类型，如 text/html 代表 html。

- HTTP 协议的头部是不是一定是文本格式？体部呢？

  头部是文本格式。体部不一定是文本格式，如传输文件是二进制格式。

- POST 方法传递的数据是放在头部还是体部？两个字段是用什么符号连接起来的？

  传递的数据放在体部。使用&连接起来的。

七、 讨论、心得

在使用这个服务器时有时会出现初次连接需要较长时间，在校园网环境下还有可能出现时而访问成功时而访问失败，或者有些设备访问成功有些设备访问失败的情况，首先应该把本机的防火墙关闭，并在弹出的提示信息允许访问网络。必要的话使用手机开启个人热点，

可能会更稳定一点。猜测校园网的玄学表现可能与校园网的路由方式和子网管理有关。

实现时必须把 rev 放到线程中处理，否则请求有图片 html 时会出现无法得到图片的情况，因为在收到 html 后连接就断开了。

在处理 http 请求时相对简单，只要分析字符串得到请求类型，然后发送对应的数据就好了。请求并不是直接访问一个文件路径，而是处理这个字符串得到正确的路径。