Web Application Protocols

1.1 HTTP

HTTP, or Hypertext Transfer Protocol, is the foundational communication protocol of the World Wide Web. The fundamental interaction involves a client sending a message (a request) to a server, which then processes it and sends a message back (a response) to the client. HTTP typically utilizes TCP (Transmission Control Protocol) connections for reliable data transfer. Crucially, HTTP is a stateless protocol, meaning that each request-response exchange is independent and self-contained; the server does not inherently remember past interactions with the client. While various requests might use different TCP connection variants, the protocol's core request/response mechanism remains autonomous.

Request. An HTTP Request is structured with several key parameters:

- *Method (or Verb)*: Specifies the action to be performed on the target resource (e.g., `GET`, `POST`, `PUT`, `DELETE`).
- *Resource*: Specifies the part of the URL identifying the target resource.
- *HTTP Protocol Version*: Indicates the version of the HTTP protocol being used (e.g., `HTTP/1.1`, `HTTP/2`, `HTTP/3`).
- *Host*: Specifies the hostname (domain name) of the server for the requested resource.
- *User-Agent*: Provides information about the client application (usually a browser) that generated the request, including its type and operating system.
- *Referer*: Indicates the URL of the page that linked to the requested resource.

```HTTP
POST /api/login HTTP/1.1
Host: www.example.com
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
Content-Type: application/json
Content-Length: 52
Referer: https://www.example.com/login


{"username": "alice123", "password": "superSecret123"}
```

Figure 1 / Example HTTP POST request for user authentication

Other important request parameters (often included as headers) include:

- *Connection*: Controls whether the underlying network connection should remain open or be closed after the current transaction finishes (e.g., `keep-alive`, `close`).
- *Cache-Control*: Used to specify caching directories for both requests and responses, influencing how intermediaries and the client should store and reuse responses.

- *Upgrade-Insecure-Requests*: Sends a signal to the server (typically with a value of `1`) expressing the client's preference for an encrypted and authenticated response, often prompting a redirect to a secure HTTPS server.
- *Sec-Fetch-**: A set of experimental headers (like `Sec-Fetch-Mode`, `Sec-Fetch-Dest`, `Sec-Fetch-Site`) designed to provide contextual information about the request (e.g., its origin and purpose) to allow the server or intermediaries to determine a priori if the request should be served, primarily for security against cross-site leaks.
- *Accept*: Provides a list of media types (e.g., `text/html`, `application/json`) that the client is capable of processing. This can be refined by related headers such as `Accept-Encoding` and `Accept-Language`.

In a standard URL structure, parameters are often passed using a query string, which begins with a question mark (`?`). This character signifies the start of the key-value pairs used to pass data to the server. The query string is used to pass parameters or data to the resource identified by the URL. The ampersand character (`&`) is used as a delimiter to combine multiple parameters within the query string. REST (Representational State Transfer) architecture provides an alternative and often cleaner way to represent resource addresses, favoring path variables over query strings for identifying a resource or a collection of resources. In RESTful URLs, resources are typically identified by hierarchical paths that align with their logical structure, promoting better clarity and meaning. Instead of `.../users?id=123`, a RESTful approach might use `.../users/123` to uniquely identify user 123.

URL Encoding defines the set of Allowed ASCII codes as ranging from `0x20` to `0x7e`. Some problematic characters (for example: `0x20` for space, `0x0a` for line feed) are encoded using the percent sign followed by the character's hexadecimal value, like `%20` or `%0a`. This process is necessary to ensure that characters that have special meaning in a URL, or those that are non-printable, are transmitted safely. In some cases, especially for characters belonging to other languages, UNICODE is used. This typically involves a two-byte representation in older URL encoding standards, as seen in the example `%u2215` → `/`. UTF-8 is the modern and most common standard, being a multibyte way to represent characters. An example of a UTF-8 encoding is `%e2%89%a0`. This encoding is widely used for data transmitted across the web, including parameters like those found in HTTP cookies. Beyond URL and character set encodings, other formats are employed for data integrity and representation. HTML encoding uses entities to represent special characters that might conflict with the document's markup (e.g., `&quot;` → "). Base64 encoding is useful to represent binary data (like cryptographic keys or images) as ASCII strings, so they can be safely included in text-based protocols like HTTP headers. Finally, Hex encoding is also useful to represent raw binary data in a readable, two-character hexadecimal format.

Response. The following parameters are typically found in the response header:

- *HTTP Protocol Version*: Indicates the version of the HTTP protocol the server used for the response.
- *Status Code*: A three-digit number that indicates the result of the request attempt. The first digit defines the class of the response: `1xx` (Informational), `2xx` (Success, e.g., 200 OK), `3xx` (Redirection, e.g., 301 Moved Permanently), `4xx` (Client Error, e.g., 404 Not Found), and `5xx` (Server Error, e.g., 500 Internal Server Error).

- *Reason Phrase*: A short, readable sentence that further explains the status code (e.g., for status code 200, the reason phrase is "OK").
- *Data*: Specifies the data and time when the response was generated by the server.
- *Accept-Ranges*: If this field is present and its value is different from `none` , it signifies that the server can accept and process partial requests for the resource.
- *Last-Modified*: The data and time of the last modification of the requested resource on the server. Clients can use this to optimize caching by sending conditional requests.
- *Access-Control-**: A group of fields (e.g., `Access-Control-Allow-Origin`) related to Cross-Origin Resource Sharing (CORS). These fields define which external domains are allowed to access the resource and how, acting as a crucial part of access control for web applications.

```
HTTP
HTTP/1.1 200 OK
Date: Mon, 21 Oct 2024 14:30:00 GMT
Server: Apache/2.4.41 (Ubuntu)
Last-Modified: Fri, 18 Oct 2024 09:15:22 GMT
Accept-Ranges: bytes
Content-Length: 1256
Content-Type: text/html; charset=UTF-8
Access-Control-Allow-Origin: *
Connection: keep-alive


<!DOCTYPE html>
<html lang="en">
<head>
<title>Alice's Profile</title> ...
```

Figure 2 / Example HTTP response for successful login.

Cookies. Cookies are tokens that a server sends to the user's web browser. Their primary purpose is to help maintain state in the otherwise stateless HTTP protocol, allowing the server to remember information about the user across multiple requests. A cookie is initially created and sent by the server via the response header `Set-Cookie` (e.g., `Set-Cookie: tracking=tI8rk7joMx44S2Uu85nSWc`); multiple cookies can be issued by sending multiple `Set-Cookie` headers in a single response. In subsequent requests to the same server, the client automatically retransmits the saved cookie data using the `Cookie` header (e.g., `Cookie: tracking=tI8rk7joMx44S2Uu85nSWc`). Cookies are typically stored as key-value pairs, though they can also be a single string without spaces.
The behavior and score of a cookie are controlled by parameters set within the initial `Set-Cookie` header:

- *Expires*: Defines a specific date and time after which the cookie will be deleted by the browser. If set, this causes the browser to save the cookie to persistent storage on the user's hard drive, allowing it to be reused in subsequent browser sessions until the expiration data is reached.
- *Domain*: Specifies the domain for which the cookie is valid. The value must be the same domain that set the cookie or a parent domain. The browser will only send the cookie to

requests made to this specified domain or its subdomains.
- *Path*: Specifies the URL path on the server for which the cookie is valid. The cookie will only be submitted for requests whose path starts with this value.
- *Secure*: A flag indicating that the cookie should only be submitted by the browser over secure channels, meaning the cookie will only be sent with HTTPS requests, preventing transmission over unencrypted HTTP.

1.2 Server/Client-Side Technologies

Server-Side. Parameters are sent to the server in multiple ways: by using the query string (starting with `?` in the URL), by employing the REST interface style (where they are embedded in the URL path), by embedding them in HTTP cookies (sent via the `Cookie` header), or by embedding them in the request body when using `POST` requests. The server processes various parts of the HTTP request, and these parameters can have a huge impact on the response. For example, a certain value of the `User-Agent` header can influence the specific page or content that is visualized by the user (e.g., serving content optimized for a mobile browser). Multiple components are used on the server's side, including: scripting languages (such as PHP and Perl), web application platforms/frameworks, web servers (to handle requests), databases and filesystems (for asset storage).

Client-Side. The user interface is essential for enabling proper communication with the server, allowing results to be presented to the user and data to be sent back to the server.
- *Hyperlinks*: These are a compact way for a user to navigate and external URL (e.g., `<a href="https://www.example.com/products">View Products</a>`). HTML Forms are extensively used to collect data from the user and send it to the server, often using `GET` or `POST` requests.
- *CSS*: CSS (Cascading Style Sheets) is used to describe the presentation of a document written in markup languages (e.g., HTML). CSS instructs the browser on how to render the contents of a resource. CSS syntax uses selectors to define a class of markup elements (e.g., all paragraphs, or elements with a specific ID) to which a given set of visual and layout attributes should be applied.
- *JavaScript*: This is a scripting language that enables the client (browser) to perform actual data processing. This can improve the application's performance by offloading part of the workload from the server to the client. It also enhances usability by allowing parts of the user interface to be dynamically updated without full page reloads. JavaScript is used to:
  (1) Validate the user's data before it gets submitted to the server, catching errors locally.
  (2) Control the browser's behavior by updating the Document Object Model (DOM). The DOM is an abstract, tree-like representation of an HTML document that can be manipulated through APIs. It allows scripts to access and manipulate individual HTML elements.
- *Ajax*: Ajax (Asynchronous JavaScript and XML) is a set of techniques that employs scripting to handle certain user actions asynchronously. By doing this, the application can exchange data with the server and update parts of a page without requiring a full page reload, significantly improving responsiveness. In Ajax applications, the client communicates their action to the server using the `XMLHttpRequest` API (or the more modern `fetch` API). The server replies with compact data, often formatted as JSON. JSON (JavaScript Object Notation) is a lightweight data interchange format used to serialize data. This compact

JSON data is then received and further processed by the client-side scripting language do dynamically update the DOM.

Sessions. The concept of a Session is crucial for maintaining a sense of continuity for a user across the stateless HTTP protocol. Once a user is authenticated, they can perform multiple actions, and the web application must be sure that each subsequent request is issued by the same user. For this reason, the server maintains a data structure that holds the user's current state and related information. This server-side data structure is called the session. Since HTTP is stateless, a unique session identifier must be constantly sent by the client and received by the server to look up and update the user's activity. There are many ways to manage and implement sessions, with the most commonly used mechanism being HTTP Cookies. A unique session ID is typically stored in a cookie on the client side; this cookie is then retransmitted with every request. The reliance on these external parameters can be dangerous if an attacker can access or hijack them, potentially leading to session hijacking and unauthorized access to the user's account.

1.3 Burp

Burp Suite is a professional, modular, Java-based software suite designed for comprehensive security testing and analysis of web applications. It operates fundamentally as an HTTP/HTTPS proxy, positioning itself as a man-in-the-middle to intercept, inspect, and modify all traffic between the client's browser and the target server. Its modular design allows for numerous extensions (plugins) and includes tools for executing various types of attacks, such as:

- Web application enumeration.
- Bruteforcing request (Intruder).
- Manual manipulation and resending of requests (Repeater).
- Automated scanning for known vulnerabilities (Scanner).

The `Target` tool serves as the core hub for a penetration testing project, providing a consolidated, high-level overview of the target application's content and functionality. Its primary components and functions are:

- *Site map*: A hierarchical tree view that records all discovered application content (hosts, folders, files, parameters). This map is populated through traffic passing through the Proxy and through Live Passive Crawling. It helps the tester visualize the application's complete attack surface.
- *Scope*: This is used to explicitly define which hosts, protocols, and URLs are in-scope for the current testing project. Setting the scope is critical because it allows the tester to filter out irrelevant traffic in the Proxy history and Site map, and configure Burp's other tools to only process traffic directed at the intended target, preventing accidental attacks on out-of-scope systems.

Burp's security auditing capabilities are split into Active and Passive checks. Passive analysis is a fundamental, non-intrusive method of vulnerability identification. The passive scanner works by only analyzing the requests and responses that naturally pass through Burp's proxy. It does not send any new, modified, or specially crafted requests to the server, ensuring the target application is not affected or alerted.

The `Proxy` tool is the fundamental and most frequently used component of Burp Suite, acting as an intercepting web proxy. The proxy is configured to sit between your browser and the web application's server. All traffic must pass through Burp, making it a powerful MITM tool. The

primary feature is its ability to intercept all incoming and outgoing HTTP/HTTPS messages. When interception is turned on, the tool holds the request or response, preventing it from continuing its journey until the tester manually forwards it. While a message is intercepted, the tester can inspect and modify any part of the raw data, including the URL, headers, and the message body. The HTTP History sub-tab logs every request and response that passes through the proxy. This creates a complete, searchable record of the application's entire traffic flow, which is essential for discovery and later analysis.

The `Repeater` tool is designed for manual, iterative testing of individual HTTP and WebSocket requests. It is the core tool for manually exploring the behavior of a specific application endpoint. It allows a tester to take a single request, modify it, and resend it repeatedly to the server without needing to interact with the browser again. Repeater is ideal for testing input-based vulnerabilities through trial and error, such as:

- *Injection Flaws* (e.g., SQL Injection, XSS) by repeatedly adjusting parameters with various payloads.
- *Access Control Flaws* (e.g., IDORs) by changing parameter values like user or product ids to access unauthorized resources.
- *State Manipulation* by sending requests in a specific sequence to test multistep processes.

Each request sent to Repeater is given its own tab, allowing the tester to work on multiple distinct test cases simultaneously. Within each tab, a full history of the modifications and corresponding server responses is maintained, making it easy to track testing progress and compare different responses.

Spidering and Parameter Analysis. Spidering is an essential reconnaissance technique used to automatically discover and map the entire content and hierarchy of a web application, including its folders, files, and links. Automatic spidering tools work by mimicking a user's navigation to systematically build a complete site map. The tool starts with an initial URL, analyzes the HTML content of the page, automatically detects all embedded links, parses forms, and follows these links to recursively discover deeper parts of the site. The `robots.txt` file, typically located in the web app's main folder `root` , contains a list of directories and files that search engines are requested not to retrieve. It is important to note that this is only a convention. A malicious spider will often check this file precisely because it may inadvertently expose hidden or sensitive areas of the application.

Automatic tools are effective but face several limitations, especially with modern applications:

- *Dynamic Content*: They often cannot automatically parse content generated dynamically on the client-side (e.g., using complex JavaScript scripts), leading to skipped resources.
- *Non-Standard Content*: Links embedded within legacy objects like Flash or Java applets are frequently skipped.
- *Multi-State Functionality*: Complex workflows, such as multistep forms or multipage checkout processes, are often not parsed or traversed correctly as the tool may fail to maintain the necessary session state or required sequence of steps.
- *Authentication and Session*: Spiders generally cannot automatically handle authentication mechanisms. The tester must manually provide session tokens or configure Burp to manage the session state.

To overcome the limitations of automatic tools and to gain context, manual spidering is performed concurrently with the automatic process. The tester should manually attempt to access the paths listed in the `robots.txt` file, as the list itself is an information disclosure vulnerability. Manually reviewing JavaScript source code can reveal URLs, functions, and endpoints that are

never linked in the visible HTML but are called dynamically. The tester must manually log into the application and then browse through all privileged or user-specific pages. This action forces the traffic to pass through the Burp Proxy, populating the Site map with authenticated-session resources.
In single-URL applications, the functionality is determined not by a new URL path, but by modifying a parameter within the request to a single resource. Spidering the base URL repeatedly will be unproductive. The key is to analyze and test the request parameters. The internal application logic is often controlled by a specific parameter, such as a hidden form field or a query string value.

```
HTTP
POST /bank.jsp HTTP/1.1
Host:  wahh-bank.com
Content-Length:  106
servlet=TransferFunds&method=confirmTransfer&fromAccount=10372918&to
        Account=3910852&amount=291.23&Submit=Ok
```

In the request above, the functionality being accessed is completely controlled by the `servlet` and `method` parameters. To enumerate other possible functionalities, a tester would use the Repeater or Intruder tools to modify these parameter values and observe the server's response. This is essential for discovering hidden or undocumented application functions.