Velum

# Introduction

Velum is an HTTP proxy management solution that ensures that the best proxy is always being used by the client. Velum itself is not an HTTP proxy service, it is only used to coordinate the usage of available proxies among clients that use it. It provides Velum clients with HTTP proxy information and the clients in turn communicate directly with the proxy.

Velum keeps metrics on the throughput of connections between proxy and endpoint. It also keeps metrics on how often an endpoint is blocking HTTP requests. It uses this information to determine if the client should switch to a different proxy in order to have the optimal connection with the endpoint.

# Definitions

**HTTP Proxy:** Often referred to in this document as the proxy. A proxy is a service running on a remote host that passes

**Endpoint:** The Web host that the client is ultimately making the request to.

**Velum Client:** Often referred to in this document as the client. The client is the program that is using the proxy to make requests to an endpoint.

**Geolocation:** The physical geographic location of a host. This is provided in latitude and longitude coordinates. This information is stored in publicly available geolocation databases. The contents of the database come from IP address registrations and so it is not always accurate.

**Domain Name System (DNS):** A system for determining the IP address of a host given the domain name.

**Authoritative Name Server (ANS):** A name server in the DNS that is most knowledgeable about the IP addresses of a given domain. ANS’s are often distributed around the world so the DNS can return a different set of IP addresses depending on the location of the host that the query is coming from.

# Overview

Velum does not maintain any state information about individual Velum clients. It is only concerned with endpoints and proxies. Clients are considered transient and many different clients may be using the same proxies.

## Web Crawling Use Case

Velum can be used by any client that wants to have optimal proxy usage. Its principle use case, however, is to support anonymous web crawling. Velum is used for anonymity because from the endpoint’s perspective all requests are coming from a proxy. This is just one piece of a truly anonymous web crawling solution.

Web crawling can incur a lot of traffic on a web host. Endpoints have various methods for detecting unwanted traffic and blocking it. The IP addresses of web crawlers that are not from well-known organizations and mainstream search engines bubble to the top of their block list due to the amount of traffic on the site. Web crawlers using Velum should do everything that they can to avoid getting into this situation, but if that should happen, Velum can be used to allow the web crawler to continue making requests from a different IP address.

Velum also has special support for web crawlers that want to download resources that are restricted by various robot exclusion standards. Web hosts may be more likely to block IP addresses that are accessing these resources too often. It is unknown how much of a factor this is in web hosts’ blocking decisions, however Velum can be used to limit the number of proxies that are blocked for this reason. It does this by maintaining a list of proxies that are only used for accessing robot excluded resources. The client usually knows which resources these are before they are requested and can use those proxies to request the resources instead. This preserves most of the other proxies for standard requests and may reduce the chances of getting blocked. Additionally, once a proxy has been used to access a robot restricted resource it will never be used to access standard resources on domains with the same second level domain (SLD) name.

If a proxy managed by Velum gets blocked by an endpoint, the client can request a new proxy for the endpoint. The next best proxy can then be used. However, the web crawler suffers because it is no longer using the optimal proxy to access the endpoint.

# Geolocation

Velum uses geolocation databases to determine the approximate geographic location of proxy hosts and endpoint hosts. This information is important to Velum because when a proxy is being requested for an endpoint there is often no latency and throughput performance history for the proxy/endpoint pair. In these cases, the geographic distance between the proxy host and the endpoint host becomes the most important factor for determining the best proxy to use. As more performance history is acquired for the pair, the geographic distance becomes less important in determining the optimal proxy.

The geolocation service provides the latitude and longitude coordinates of the host. The geographic distance between two hosts is calculated using the Haversine distance formula.

## Endpoint Geolocation

The geolocation of the endpoint is determined relative to the location of the computer running the Velum manager. This is because Velum is usually being run in the same geographic location as its clients.

Velum will query the local DNS server for the IP address of the endpoint’s domain name. The DNS server will return the nearest IP address on the network for that domain. This IP address often corresponds to the geographically nearest endpoint. DNS doesn’t report the geolocation of the IP address so the only way to determine this is to use a geolocation database. Even though the endpoint’s domain name may have many IP addresses dispersed all over the world we only connect to the ones that are closest to Velum.

## Proxy Geolocation

The geolocation of the proxy is more straightforward because the proxies do not use DNS or even if they do Velum already has the IP address that it will use for the proxy. If no geolocation information can be found for the proxy then the proxy will not be considered in the best proxy algorithm until all other proxies have been evaluated for the endpoint.

# Proxy Selection

Proxy selection for each endpoint is done based on these parameters in this order of significance:

1. The number of times the endpoint has been blocked by the endpoint (lower is better)
2. The average throughput between the proxy and the endpoint (higher is better)
3. The geographic distance between the proxy and the endpoint (lower is better)

Proxies that have previously been used to download resources that are marked as robots excluded cannot be used to download regular resources from the same endpoint.

# Actions

Velum clients can do the following three actions:

* Get the proxy currently assigned to the endpoint from Velum
* Get the next best proxy for the endpoint from Velum
* Send proxy/endpoint performance metrics to Velum

## Get Current

Given the domain name of an endpoint and robot flag return the proxy currently assigned to it. If no proxy is assigned yet assign and return the best proxy for the endpoint.

## Get New

Given the domain name of an endpoint and robot flag return the best proxy that is not the currently assigned proxy. This is called in response to an endpoint blocking the proxy too many times.

## Update Status

Each time a client makes a request it should send the throughput metric to Velum to update. Velum keeps a running average of the throughput of requests made using a proxy and between a proxy and endpoint. The client should also tell Velum if the request was made using HTTPS or HTTP so that Velum can determine if the proxy supports HTTPS.