Trường Đại Học Bách Khoa Hà Nội

Viện Công Nghệ Thông Tin & Truyền Thông

======o0o======

BÀI TẬP LỚN

MÔN: NHẬP MÔN AN TOÀN THÔNG TIN

Đề tài: Proxy – Cấu trúc và hoạt động của proxy

Giáo viên hướng dẫn : Nguyễn Linh Giang

Nhóm 20

Nguyễn Minh Đăng:	20172998
Võ Đức Quân:	20173220
Vũ Quang Huy:	20173178
Nguyễn Văn Giảng:	20173084

Ngày 06 tháng 06 năm 2020

Mục lục

I. Proxy

1. Khái niệm

Proxy là một Internet server làm nhiệm vụ chuyển tiếp thông tin và kiểm soát tạo sự an toàn cho việc truy cập Internet của các máy khách, còn gọi là khách hàng sử dụng dịch vụ Internet. Proxy hay trạm cài đặt proxy có địa chỉ IP và một cổng truy cập cố định. Ví dụ: 123.234.111.222:80.

2. Vai trò của máy chủ proxy

Các máy chủ proxy cung cấp các chức năng, bảo mật và riêng tư khác nhau phụ thuộc vào nhu cầu của bạn hoặc chính sách công ty. Tuy nhiên có thể kể đến một vài chức năng tiêu biểu sau đây của máy chủ proxy:

- Chuyển tiếp các yêu cầu: Nếu đang sử dụng máy chủ proxy, lưu lượng truy cập Internet sẽ truyền qua máy chủ proxy theo đường của nó đến địa chỉ bạn yêu cầu. Sau đó, yêu cầu này sẽ trở lại cùng một máy chủ proxy (cũng xảy ra trường hợp ngoại lệ đối với quy tắc này) và máy chủ proxy đó sẽ chuyển tiếp dữ liệu nhận được từ website đến người dùng.

- Bảo mật dữ liệu: Proxy server giống như một vệ sĩ bảo vệ khỏi những rắc rối trên Internet. Một Proxy server thường nằm bên trong tường lửa, giữa trình duyệt web và server thật, làm chức năng tạm giữ những yêu cầu Internet của các máy khách để chúng không giao tiếp trực tiếp Internet. Người dùng sẽ không truy cập được những trang web không cho phép (bị cấm).

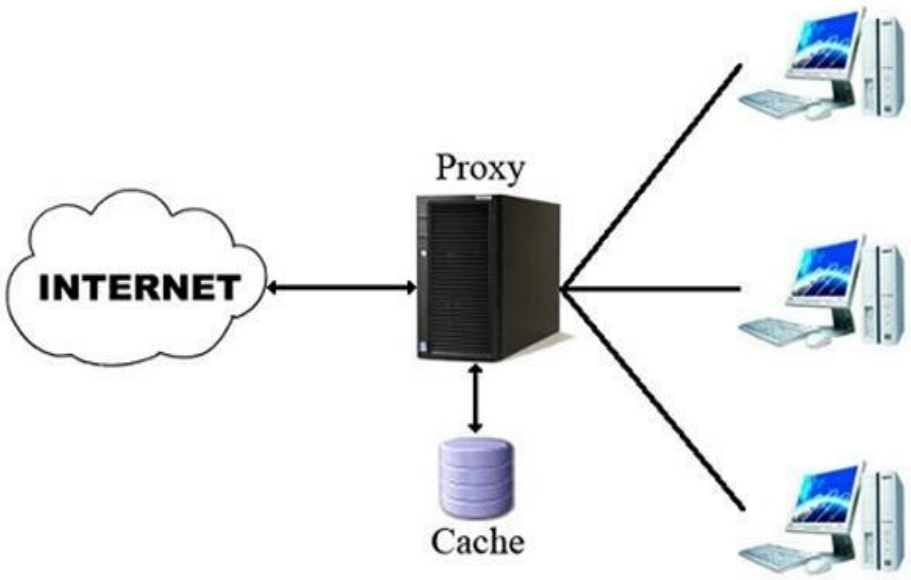
- Tăng hiệu suất mạng: Mọi yêu cầu của máy khách phải qua Proxy server, nếu địa chỉ IP có trên proxy, nghĩa là website này được lưu trữ cục bộ, trang này sẽ được truy cập mà không cần phải kết nối Internet, nếu không có trên Proxy server và trang này không bị cấm, yêu cầu sẽ được chuyển đến server thật, DNS server… và ra Internet. Proxy server lưu trữ cục bộ các trang web thường truy cập nhất trong bộ đệm để giảm chi phí kết nối, giúp tốc độ duyệt web nhanh hơn.

3. Các tính năng của máy chủ proxy (proxy server)

3.1 Tường lửa và Filtering

Proxy servers làm việc ở lớp Application, lớp thứ bảy trong mô hình tham chiếu OSI. Chúng không được phổ biến như các tường lửa thông thường mà làm việc ở mức thấp hơn và hỗ trợ lọc ứng dụng một cách độc lập. Proxy servers cũng khó khăn hơn trong việc cài đặt và duy trì so với tường lửa. Mặc dù vậy, nếu proxy server được cấu hình đúng cách sẽ cải thiện được vấn đề bảo mật và hiệu suất cho mạng. Các proxy đều có khả năng mà các tường lửa thông thường không thể cung cấp.

Một số quản trị viên mạng sử dụng cả tường lửa và proxy server để làm việc cùng nhau. Muốn thực hiện như vậy, họ phải cài đặt cả phần mềm tường lửa và phần mềm proxy server trên một server gateway.

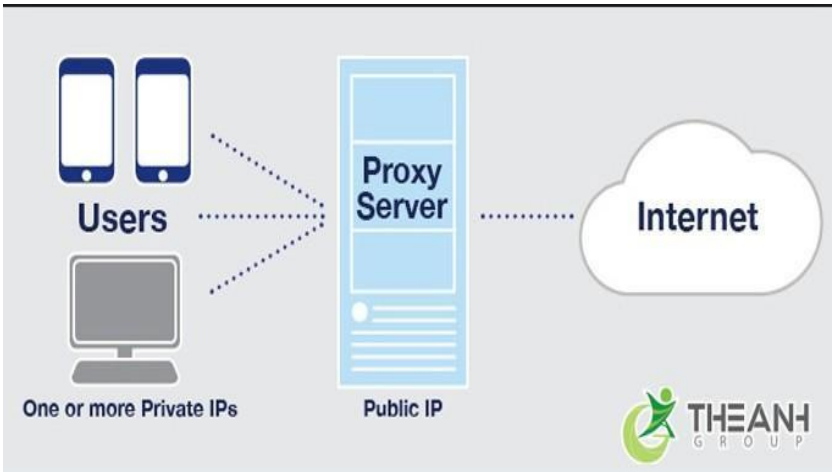


Hình 1. Quá trình truy cập từ máy tính qua proxy server

3.3 Proxy Server và Caching

Caching của các trang web có thể cải thiện chất lượng dịch vụ của một mạng theo 3 cách. Thứ nhất, nó có thể bảo tồn băng thông mạng, tăng khả năng mở rộng. Tiếp đến, có thể cải thiện khả năng đáp trả cho các máy khách. Ví dụ, với một HTTP proxy cache, Web page có thể load nhanh hơn trong trình duyệt web. Cuối cùng, các proxy server cache có thể tăng khả năng phục vụ. Các Web page hoặc các dòng khác trong cache vẫn còn khả năng truy cập thậm chí nguồn nguyên bản hoặc liên kết mạng trung gian bị offline.

4. Cách hoạt động của máy chủ proxy



Hình 2: Cách hoạt động của máy chủ proxy.

Mọi máy tính trên Internet đều phải có địa chỉ IP duy nhất. Hãy nghĩ địa chỉ IP này là giống như địa chỉ đường nhà bạn. Cũng giống như bưu điện cần biết địa chỉ đường của bạn để gửi thư, Internet cũng cần biết địa chỉ IP của máy tính để gửi dữ liệu đến đúng máy tính.

Máy chủ proxy về cơ bản là một máy tính trên Internet với địa chỉ IP của riêng nó mà máy tính của bạn biết. Khi gửi một yêu cầu web, nó sẽ đến máy chủ proxy đầu tiên. Sau đó máy chủ proxy sẽ thực hiện yêu cầu web của bạn, thu thập phản hồi từ máy chủ web và chuyển tiếp dữ liệu trang web để bạn nhìn thấy trang web trong trình duyệt.

Khi máy chủ proxy chuyển tiếp yêu cầu web của người dùng, nó có thể thay đổi dữ liệu đó mà vẫn lấy thông tin theo đúng yêu cầu. Máy chủ proxy có thể thay đổi địa chỉ IP của bạn, để máy chủ web không biết chính xác vị trí của bạn. Nó có thể mã hóa dữ liệu để không ai có thể đọc được trong quá trình vận chuyển. Và cuối cùng máy chủ proxy có thể chặn truy cập vào các trang web cụ thể dựa trên địa chỉ IP

5. Phân loại

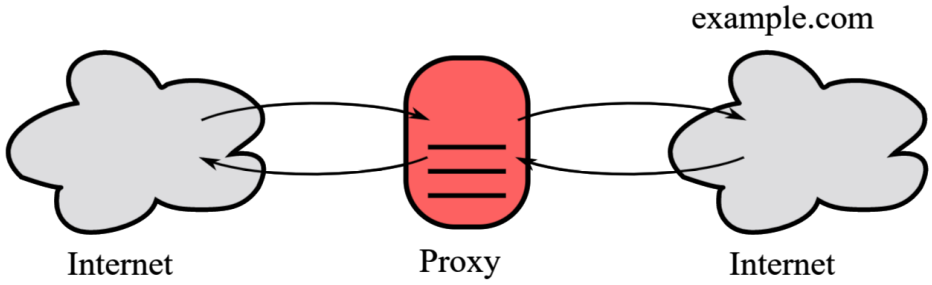
Có nhiều các phân loại proxy server khác nhau như theo chức năng: HTTP Proxy,SOCKS Proxy, CGI Proxy, FTP Proxy; theo mức độ bảo mật: Transparent Proxy, Antonymity Proxy, Distorting Proxy, High Anonymity Proxy. Tuy nhiên trong khuôn khổ bài tiểu luận này chúng em sẽ phân loại Proxy Server thành hai loại sau đây.

5.1 Forward Proxy

Forward proxy (hay còn gọi là open proxy) là một proxy server có thể được truy cập bởi bất kỳ người dùng nào. Tính đến năm 2008, Gordon Lyon (một chuyên gia bảo mật mạng máy tính) ước tính có hàng trăm ngàn open proxy đang hoạt động trên Internet. Có hai loại open proxy:

- *Anonymous proxy (proxy ẩn danh)* – Loại proxy này hoạt động như là một máy chủ nhưng không tiết lộ địa chỉ IP gốc của client. Mặc dù loại proxy này có thể bị phát hiện một cách dễ dàng, nhưng nó vẫn mang lại lợi ích khi có thể che dấu địa chỉ IP.

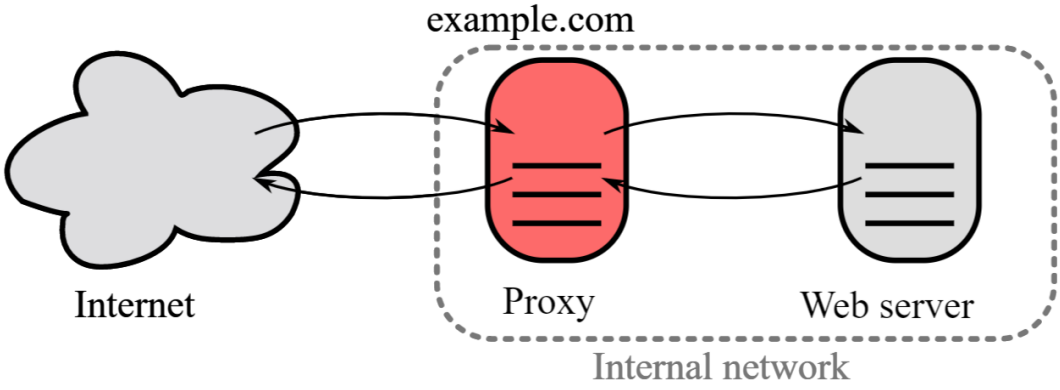
- *Transparent proxy (proxy minh bạch)* – Ngược lại với Anonymous proxy, loại proxy này hoạt động như là một proxy và chúng forward request bằng các HTTP header (ví dụ như X-Forwarded-For). Nhờ có những HTTP header này mà địa chỉ IP gốc có thể được tìm thấy. Lợi ích chính khi dùng loại proxy này là khả năng cache một website.



Hình 3: minh họa một Forward proxy chuyển tiếp request đến các server trên Internet

5.2 Reverse Proxy

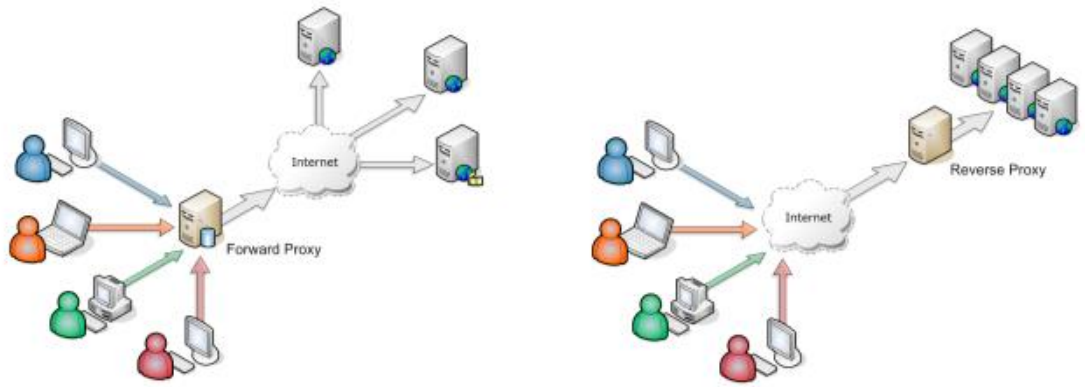
Reverse proxy là một proxy server mà khi đứng trước client, chúng hoạt động giống như những server bình thường. Reverse proxy chuyển tiếp request đến một hoặc nhiều server thật, kết quả sau đó trả về cho client như thể là chúng được trả về từ reverse proxy, khiến cho client không biết về những server thật nói trên. Reverse proxy được cài đặt trong một private network của một hoặc nhiều server, và tất cả lưu lượng truy cập đều phải đi qua proxy này.



Hình 4: minh họa một Reverse proxy nhận request từ bên ngoài và chuyển tiếp đến server trong mạng nội bộ.

Thông thường, những server sẽ sử dụng cơ chế reverse proxy này để bảo vệ các ứng dụng có khả năng xử lý HTTP yếu kém. Ví dụ như khả năng xử lý cực lớn các request, những hạn chế về xử lý sự đa dạng của các loại request (các dạng request có thể kể đến như: HTTP(S) 1.x, HTTP(S) 2.x, ...) hay khả năng chuyển đổi HTTPS thành HTTP, cache request, xử lý dữ liệu của cookies/session, chia một request thành nhiều request nhỏ hơn rồi tổng hợp lại các response, ...

5.3 Sự khác biệt giữa forward proxy và reverse proxy



Hình 5: Sự khác biệt giữa forward proxy và reverse proxy.

Khác biệt lớn nhất giữa hai loại proxy này là forward proxy được sử dụng bởi client (ví dụ như trình duyệt). Trong khi đó, reverse proxy được sử dụng bởi server (ví dụ như web server). Forward proxy có thể nằm trong một mạng nội bộ cùng với client hoặc cũng có thể công khai trên Internet.

Hay nói một cách dễ hiểu hơn, forward proxy đại diện cho client, còn reverse proxy đại diện cho server.

6. Các lợi ích khi sử dụng máy chủ Proxy

- Để kiểm soát việc sử dụng Internet của nhân viên và trẻ em: Tổ chức và phụ huynh thiết lập máy chủ proxy để kiểm soát và giám sát nhân viên hoặc trẻ em sử dụng Internet. Hầu hết các tổ chức không muốn nhân viên của họ xem các trang web cụ thể trong thời gian làm việc và họ có thể cấu hình máy chủ proxy để từ chối truy cập vào trang web cụ thể, điều hướng bạn bằng một ghi chú yêu cầu bạn không xem các trang web này trên mạng công ty. Họ có thể giám sát và ghi lại tất cả các yêu cầu web, do đó mặc dù không chặn trang web nhưng họ vẫn biết thời gian bạn dành cho những việc làm khác ngoài công việc.

- Tiết kiệm băng thông và cải thiện tốc độ: Các tổ chức cũng có thể nhận được hiệu suất mạng tổng thể tốt hơn khi sử dụng máy chủ proxy. Các máy chủ proxy có thể lưu vào bộ nhớ cache (lưu một bản sao trang web cục bộ) các trang web hay truy cập. Do đó khi yêu cầu trang www.hust.edu.vn, máy chủ proxy sẽ kiểm tra xem có bản sao mới nhất của trang web này hay không và sau đó sẽ gửi cho bạn bản sao đã lưu. Điều này có nghĩa là khi hàng trăm người truy cập vào www.hust.edu.vn cùng một thời điểm từ cùng một máy chủ proxy, máy

chủ này chỉ cần gửi một yêu cầu đến www.hust.edu.vn. Điều này giúp tiết kiệm băng thông của công ty và cải thiện hiệu suất mạng.

- Bảo mật riêng tư: Cá nhân và tổ chức cũng sử dụng máy chủ proxy để duyệt Internet riêng tư hơn. Một số máy chủ proxy sẽ thay đổi địa chỉ IP và thông tin nhận dạng khác. Điều này có nghĩa là máy chủ đích không biết ai thực sự đã thực hiện yêu cầu ban đầu, giúp giữ thông tin cá nhân và thói quen duyệt web của bạn riêng tư hơn.

- Cải thiện bảo mật: Bạn có thể cấu hình máy chủ proxy để mã hóa yêu cầu web để không ai có thể đọc được giao dịch của bạn. Ngoài ra, người dùng cũng có thể tránh các trang web độc hại thông qua máy chủ proxy. Các tổ chức có thể kết nối máy chủ proxy của họ với Mạng riêng ảo (VPN), do đó người dùng từ xa có thể truy cập Internet thông qua proxy của công ty. VPN kết nối trực tiếp đến mạng công ty để có thể kiểm soát và xác minh người dùng của họ có quyền truy cập vào các tài nguyên họ cần (email, dữ liệu nội bộ) đồng thời cũng cung cấp kết nối an toàn cho người dùng để bảo vệ dữ liệu công ty.

- Truy cập vào các tài nguyên bị chặn: Máy chủ proxy cho phép người dùng phá vỡ các hạn chế nội dung do công ty hoặc một số tổ chức áp đặt. Nếu truy cập vào trang web bị chặn, bạn có thể đăng nhập vào máy chủ proxy ở nơi khác và xem từ đó. Máy chủ proxy khiến bạn giống như ở Mỹ nhưng thực ra bạn đang ở Việt Nam.

7. Các rủi ro khi sử dụng máy chủ Proxy

Cần thận trọng khi chọn máy chủ proxy, dưới đây là một số rủi ro khi sử dụng máy chủ proxy.

- Rủi ro từ máy chủ proxy miễn phí: sử dụng máy chủ miễn phí có thể đem đến nhiều rủi ro ngay cả những dịch vụ sử dụng model doanh thu dựa trên quảng cáo. Miễn phí thường có nghĩa là họ không đầu tư nhiều vào phần cứng hoặc mã hóa phụ trợ. Bạn có thể thấy các vấn đề hiệu suất và các vấn đề bảo mật dữ liệu tiềm ẩn. Nếu bạn đang sử dụng một máy chủ proxy hoàn toàn miễn phí, hãy cẩn thận, một số trong số đó chỉ tìm cách ăn cắp số thẻ tín dụng của bạn.

- Nhật ký lịch sử duyệt web: máy chủ proxy có địa chỉ IP và thông tin yêu cầu web ban đầu của bạn do đó, hãy chắc chắn kiểm tra nhật ký máy chủ proxy và lưu dữ liệu đó.

- Không mã hóa: nếu bạn sử dụng máy chủ proxy không có mã hóa, điều đó có nghĩa là bạn đang gửi yêu cầu của mình dưới dạng văn bản thuần túy. Bất cứ ai cũng có thể thấy tên người dùng, mật khẩu và thông tin tài khoản. Đảm bảo máy chủ proxy bạn sử dụng có tính năng mã hóa.

II. Cache

Như đã trình bày ở phần trước, một trong các ứng dụng của proxy server đó chính là tăng hiệu năng của mạng bằng cách lưu trữ các nội dung được truy cập nhiều, từ đó làm giảm các request ra bên ngoài. Để hiểu thêm về chức năng này, ta sẽ tìm hiểu về cache và các thuật toán cache.

1. Cache là gì ?

Cache hay bộ nhớ đệm là phần cứng hoặc phần mềm được tích hợp sẵn với tác dụng lưu trữ dữ liệu tạm thời trong môi trường máy tính. Nội dung lưu có thể là: logo, banner, hình ảnh tĩnh, các file định dạng css, javascript, tập tin có thể tải về, tập tin media,...

2. Phân loại cache

2.1 Cache memory

Cache memory thường được gắn trực tiếp trên CPU. Nó có khả năng lưu trữ lệnh/chức năng thường được yêu cầu bởi các chương trình đang chạy, giúp bộ vi xử lý máy tính truy cập dữ liệu nhanh hơn so với RAM thông thường. Nếu xét về khả năng truy xuất thì cache memory có tốc độ rất nhanh (hơn hẳn disk cache và cả RAM cache) vì vị trí của nó gần với CPU nhất

2.2 Cache server (Proxy cache)

Thông thường, các máy chủ kết nối mạng chuyên dụng (dedicated network server) hoặc dịch vụ hoạt động như máy chủ (service acting as server) sẽ lưu trữ dữ liệu trang web và các nội dung internet một cách cục bộ. Hình thức lưu trữ này gọi là cache server hay cache proxy.

2.3 Disk cache

Disk cache ghi nhớ các nội dung đã được đọc trong thời gian gần và những dữ liệu liền kề khác có khả năng sẽ được truy cập lại. Nhiều disk cache lưu trữ dữ liệu theo tần suất đọc. Theo đó, những khối lưu trữ (storage block) truy cập thường xuyên (gọi là các khối nóng – hot block) sẽ tự động được ghi nhớ trên cache. Disk cache giúp cải thiện tốc độ đọc hoặc ghi dữ liệu lên đĩa cứng.

2.4 Flash cache

Flash cache là thiết bị lưu trữ tạm thời dữ liệu trên chip bộ nhớ NAND (thường lưu trữ dưới dạng SSD). Nó có khả năng truy xuất dữ liệu với tốc độ cao hơn so với bộ nhớ cache trên ổ đĩa truyền thống HDD.

3. Các thuật toán cache

3.1 Đặt vấn đề

Cache có giúp tăng tốc độ truy xuất dữ liệu, tuy nhiên dung lượng của nó cũng có giới hạn. Mặt khác, xu hướng truy cập vào các tài nguyên của người dùng luôn luôn thay đổi theo thời gian, do đó không thể cố định lưu trữ các nội dung trong cache được mà luôn phải cập nhật.

Nhưng làm thế nào để cập nhật nội dung lưu trong cache thế nào cho hợp lý nhất. Đó chính là lý do tại sao chúng ta lại cần các thuật toán cache.

3.2 Tổng quan các thuật toán

3.2.1 First In First Out (FIFO)

Thuật toán này hoạt động tương tự so với hàng đợi queue. Cache sẽ loại bỏ các nội dung theo thứ tự mà chúng được lưu vào cache mà không cần quan tâm đến tần suất hay số lần xuất hiện trước đó của nội dung đó.

3.2.2 Last in first out (LIFO) or First in last out (FILO)

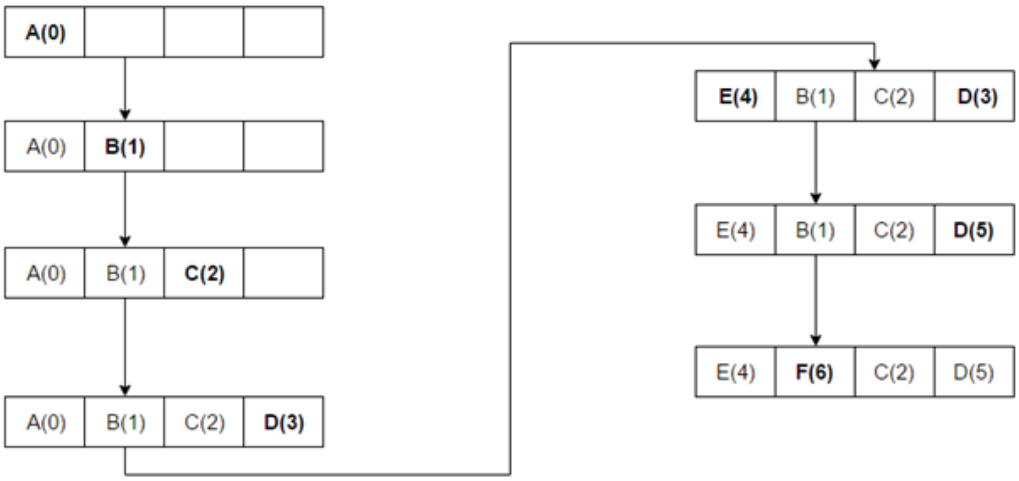
Trái ngược với thuật toán FIFO, thuật toán này có cơ chế hoạt động tương tự ngăn xếp stack. Cache sẽ loại bỏ nội dung vừa được lưu gần đây nhất mà không cần quan tâm đến tần xuất hay số lần xuất hiện trước đó của nội dung đó.

3.2.3 Least recently used (LRU)

Loại bỏ các nội dung lâu chưa được sử dụng tới nhất (có lần sử dụng cuối cùng lâu nhất tính tới thời điểm hiện tại). Chính vì thế, thuật toán này đòi hỏi chúng ta cần phải theo dõi xem những gì đã được sử dụng khi nào. Để có thể đảm bảo loại bỏ được nội dung lâu chưa được sử dụng tới nhất sẽ rất tốn thời gian (do phải duyệt qua toàn bộ nội dung trong cache).

Để triển khai thuật toán này đòi hỏi việc phải giữ các “bit tuổi” trong cache, từ đó theo dõi xem nội dung nào có lần sử dụng cuối cùng là xa nhất.

Sau đây là hình minh hoạ cho quá trình hoạt động của thuật toán LRU. Chú ý rằng thứ tự các nội dung được truy cập là A B C D E D F.



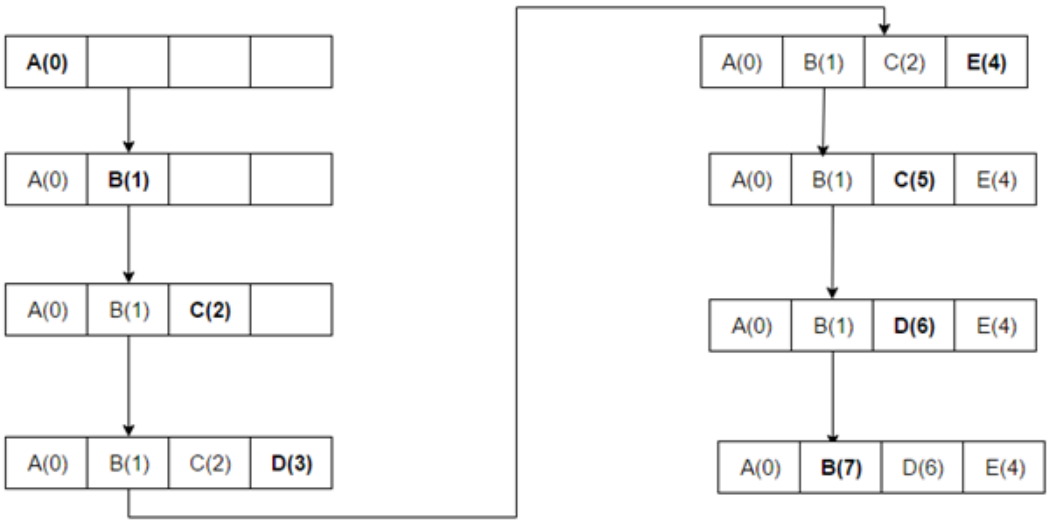
Hình 6: Quá trình thực hiện của thuật toán LRU.

Trong ví dụ này, sau khi lưu cả A, B, C, D thì cache đã đầy, sau đó khi lưu nội dung E, thuật toán sẽ tìm kiếm xem nội dung nào được sử dụng lần cuối cùng cách đây xa nhất, đó chính là A. Vì thế A sẽ được xóa khỏi cache để lưu E.

3.2.4 Most recently used (MRU)

Trái ngược lại so với LRU, thuật toán này sẽ lại bỏ nội dung được sử dụng gần đây nhất. Thuật toán sẽ có số lần truy cập vào bộ nhớ chính để lấy dữ liệu nhiều hơn so với LRU vì nó có xu hướng lưu giữ các nội dung cũ. Nó sẽ có hiệu quả trong trường hợp nội dung càng cũ thì càng có nhiều khả năng truy cập.

Trình tự truy cập cho ví dụ sau đây là A B C D E C D B.



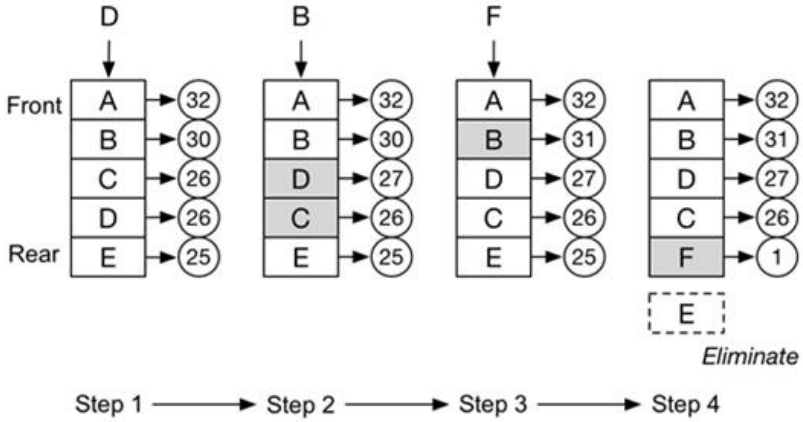
Hình 7. Quá trình thực hiện thuật toán MRU

Ở đây A, B, C, D được lưu trữ ngay vì cache vẫn còn trống. Ở lần truy cập thứ tiếp theo, ta thấy nội dung D được thay thế bởi E vì nó được sử dụng gần đây nhất. Sau đó là một lần truy cập vào nội dung C và tại lần truy cập tiếp theo, C sẽ được thay thế bởi D và cứ tiếp tục như vậy

3.2.5 Least Frequently Used (LFU)

LFU là một loại thuật toán cache được sử dụng để quản lý bộ nhớ máy tính. Đặc trưng chủ yếu của phương pháp này là theo dõi hệ thống về số lần một đối tượng được tham chiếu đến trong bộ nhớ. Khi cache đầy và yêu cầu nhiều khoảng trống hơn, hệ thống sẽ loại bỏ đối tượng có tần suất tham chiếu tới thấp nhất

Phương pháp đơn giản nhất để sử dụng thuật toán LFU là chỉ định một bộ đếm cho mỗi đối tượng được nạp vào cache. Mỗi lần tham chiếu tới đối tượng đó, bộ đếm tăng lên một. Khi bộ nhớ đầy và có đối tượng mới đang chờ để thêm vào, hệ thống sẽ tìm kiếm một khối có bộ đếm thấp nhất và loại bỏ khỏi cache.



Hình 8. Quá trình thực hiện thuật toán LFU.

3.2.5 Least Frequent Recently Used (LFRU)

Chiến lược LFRU kết hợp những điểm mạnh từ LFU và LRU. LFRU phù hợp với những ứng dụng cache 'trong mạng' như Information centric networking (ICN), Content Delivery Network (CDNs) và các mạng phân tán nói chung.

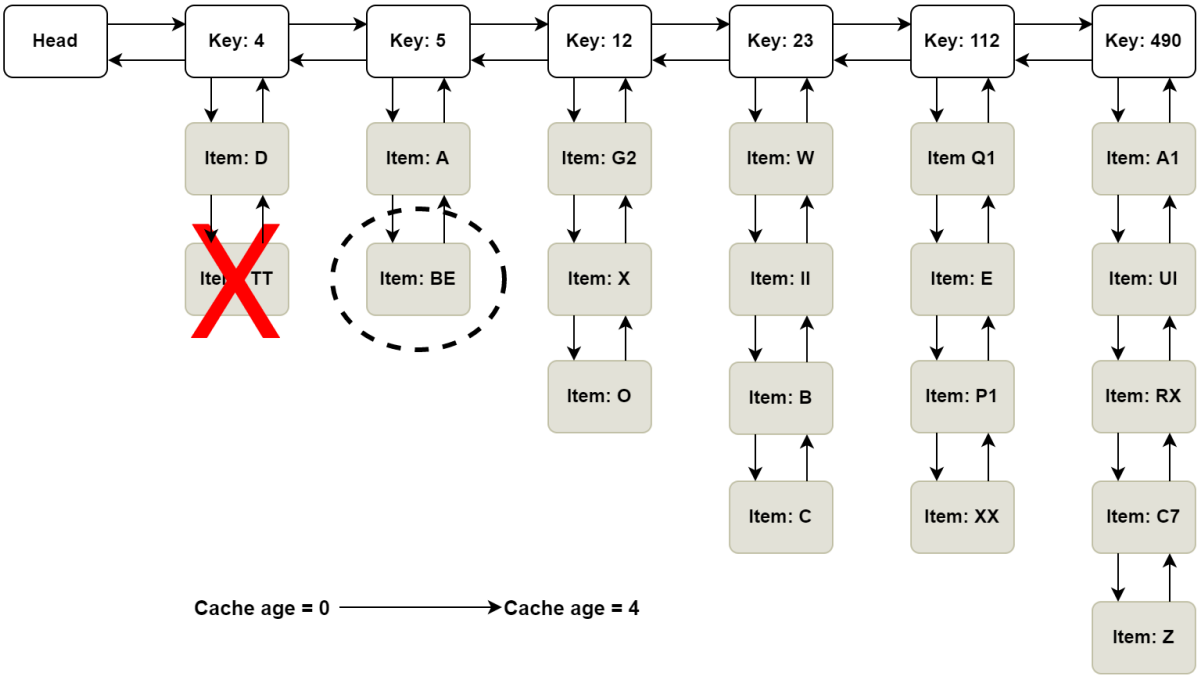
Với LFRU, cache được chia thành hai phân vùng là các phân vùng đặc quyền và không đặc quyền. Phân vùng đặc quyền có thể được định nghĩa như một phân vùng được bảo vệ. Nếu nội dung quá phổ biến, nó sẽ được đẩy vào phân vùng này. Sự thay đổi của phân vùng đặc quyền được thực hiện như sau: LFRU xóa nội dung từ phân vùng không đặc quyền và đẩy nội dung từ phân vùng đặc quyền vào, và cuối cùng là thêm đối tượng mới vào phân

vùng đặc quyền. Trong quy trình trên, LRU được dùng cho phân vùng đặc quyền và LFU tương đối (ALFU) được sử dụng cho phân vùng không đặc quyền.

3.2.7 Least Frequently Used with Dynamic Aging (LFUDA)

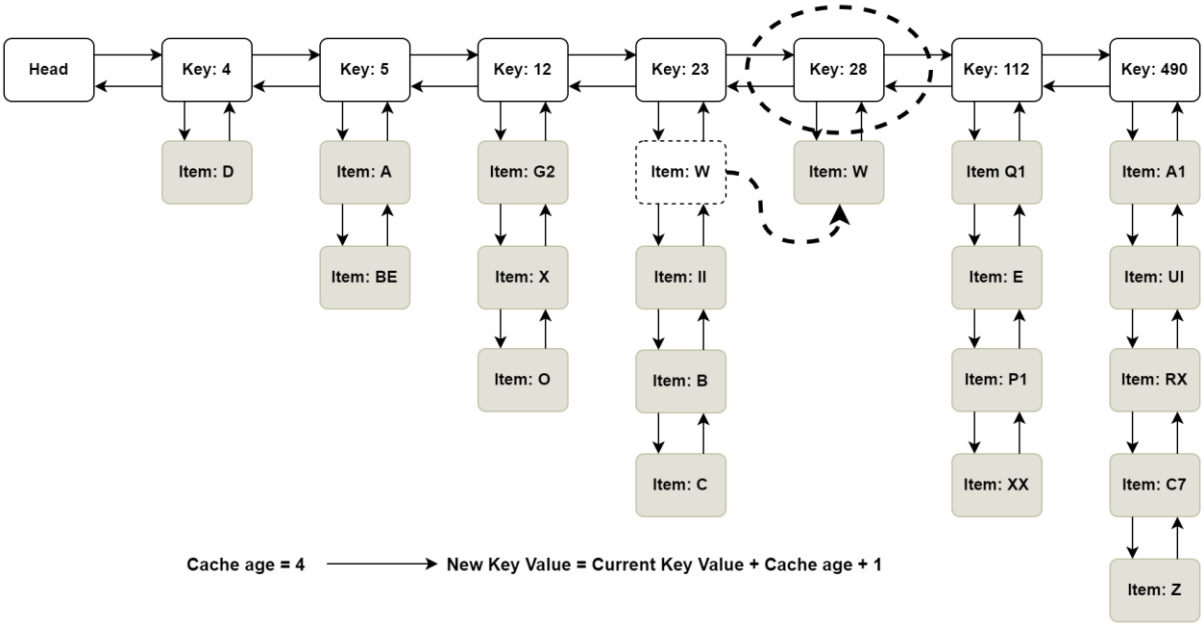
Một biến thể của LFU, được gọi là LFUDA, sử dụng dynamic aging để lưu các thay đổi trong tập các đối tượng phổ biến. Thuật toán này thêm một hệ số cache age vào bộ đếm tham chiếu. Khi một đối tượng được thêm vào cache hoặc một đối tượng đã có được tham chiếu lại, LFUDA sẽ thay đổi cache age theo giá trị khóa của đối tượng.

Ta có ví dụ sau đây: Bộ nhớ cache đã đầy, đối tượng BE được thêm mới. LFUDA tìm đối tượng có tần suất tham chiếu thấp nhất (Key: 4), TT bị loại bỏ để tạo thêm không gian. Cache age được cập nhật bằng giá trị Key của đối tượng bị loại bỏ (4). Đối tượng mới hiện tại có số lần truy cập là 1 nên Key = 1 + Cache age (4) = 5. Trong danh sách đã có Key: 5 nên đối tượng BE sẽ được thêm vào.



Hình 9: Quá trình thực hiện thuật toán LFUDA.

Chúng ta thử tham chiếu tới một đối tượng đã tồn tại, giả sử là W ở Key: 23. Số lần truy cập là 1 và Key được cập nhật = 23 + 1 + Cache age (4) = 28. Key: 28 được tạo mới và W sẽ được di chuyển qua.

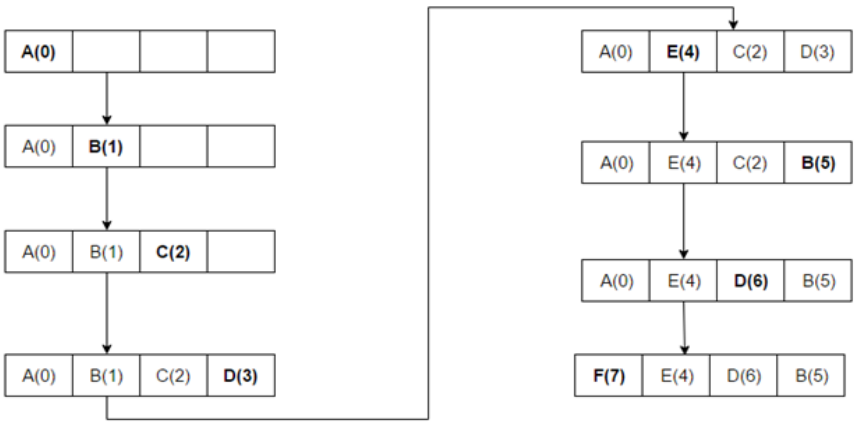


Hình 10. Quá tình thực hiện thuật toán LFUDA.

LFUDA giải quyết được một số vấn đề mà LFU gặp phải. Đó là khi một đối tượng được truy cập thường xuyên trong quá khứ và hiện tại thì không được thường xuyên nữa nhưng nó vẫn sẽ tồn tại trong cache, từ đó cản trở các đối tượng mới hoặc các đối tượng ít phổ biến hơn thay thế nó. LFUDA được ra mắt để làm giảm các đối tượng như vậy, làm chúng có thể bị thay thế. Đối với các bộ nhớ cache nhỏ thì vấn đề này rất quan trọng.

3.2.8 Random cache

Đây là thuật toán mà hệ thống sẽ lựa chọn một đối tượng ngẫu nhiên bất kỳ để loại bỏ nó ra bộ nhớ khi cần thiết. Thuật toán không cần thông tin nào về lịch sử truy cập của đối tượng. Vì sự đơn giản của nó, thuật toán này thường được sử dụng trong các chip sử dụng kiến trúc ARM. Sau đây là 1 ví dụ đơn giản về thuật toán. Thứ tự truy cập dữ liệu lần lượt là: A → B → C → D → E → B→ D → F



Hình 11. Quá trình thực hiện thuật toán Random cache.

3.2.9 Adaptive Replacement Cache (ARC)

Đây là thuật toán sử dụng trong quá trình quản lý phân trang trong bộ nhớ ảo của máy tính và nó có hiệu quả cao hơn so với thuật toán LRU. Nó sẽ đưa ra quyết định dựa vào những dữ liệu được sử dụng thường xuyên và gần đây nhất cùng lịch sử truy vấn trở lại những dữ liệu đó.

ARC cải tiến chiến lược của thuật toán LRU bằng cách chia danh sách cache thành 2 phần T1 và T2, được sắp xếp theo thứ tự truy cập gần nhất và tần suất truy cập. Sau đó mỗi phần sẽ được mở rộng thêm 1 danh sách “ma" lần lượt là B1 và B2 ngay bên dưới 2 phần. danh sách ma sẽ hoạt động như một bảng ghi điểm để ghi lại lịch sử xoá cache gần đây và thuật toán sẽ sử dụng danh sách này để thích ứng với những thay đổi gần đây trong việc sử dụng tài nguyên. Danh sách các dữ liệu lưu trữ bao gồm:

T1: các bộ nhớ đệm gần đây

T2: các bộ nhớ đệm thường xuyên dùng, được tham chiếu ít nhất 2 lần

B1: Các bộ nhớ đệm bị xoá gần đây của T1, chỉ lưu lại meta tham chiếu

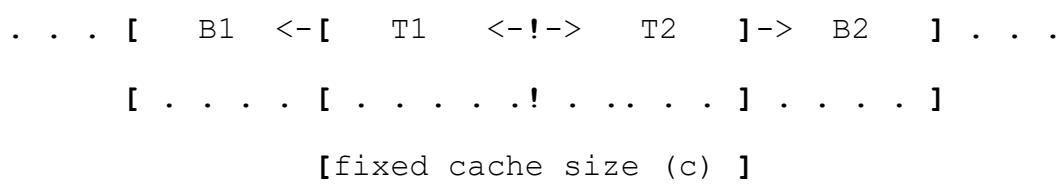
B2: Giống B1 nhưng là của T2

Phương thức hoạt động:

Các bộ nhớ đệm mới vào T1 và được đẩy dần về B1 cho đến khi bị đẩy ra ngoài

Nếu bộ nhớ đệm được truy vấn lần nữa sẽ có cơ hội vào T2 sau đó lại bị đẩy dần qua B2 cho đến khi ra ngoài.

Thay thế: Đây là một biểu diễn mô tả cho T1, T2, B1, B2



Dấu ngoặc mô tả bộ nhớ đệm thực tế có thể di chuyển tự do trong B1, B2

T1, B1 đánh dấu là L1, T2, B2 đánh dấu là L2. L1 được bắt đầu từ dấu !. Việc truy xuất lại dữ liệu dẫn đến luồng dữ liệu sẽ chạy từ ! đến dấu ^. Nếu không có ô nhớ còn trống, vị trí ^ sẽ quyết định việc T1 hay T2 sẽ xoá bỏ bớt dữ liệu

Số lượt truy cập của dữ liệu ở B1 sẽ tăng kích thước ở T1và đẩy ^ về bên phải khiến cho dữ liệu ô nhớ cuối của T2 sẽ đẩy qua B2.

Tương tự ở B2 sẽ đẩy ô nhớ cuối của T1 qua B1.

Thuật toán ARC hiện nay chủ yếu được ứng dụng bên trong bộ điều khiển lưu trữ DS6000/ DS8000 của IBM.

3.2.10 Clock with Adaptive Replacement (CAR)

Là sự kết hợp của cả ARC và Clock nên CAR có hiệu suất tương đương với ARC nhưng vượt trội hơn hẳn so với Clock và LRU. Giống với ARC, CAR hiệu chỉnh dữ liệu một cách tự động mà không cần đến một số thông số phải cài đặt. Nó sử dụng 4 danh sách liên kết nhau theo thứ tự: 2 đồng hồ T1, T2 và 2 danh sách B1, B2. T1 chứa thông tin dữ liệu dựa theo lần truy cập gần đây và khả năng truy cập ngắn hạn, còn T2 thì dựa vào tần suất truy cập và khả năng truy cập dài hạn của dữ liệu.

Còn B1, B2 thì chứa những dữ liệu vừa bị xóa ra khỏi T1, T2. Trang mới sẽ được chọn lựa và lưu vào T1, T2. Còn quy tắc để thích ứng thì được sử dụng cùng quy tắc với ARC

III. Giới thiệu Squid Proxy Server

Mục này sẽ giới thiệu về cách cài đặt, cấu hình và khởi động Squid Proxy Server trên Windows

1. Cài đặt Squid for Windows

Bước 1: Truy cập trang web https://squid.diladele.com/

Bước 2: Chọn "DOWNLOAD MSI"

Bước 3: Chạy file "squid.msi" mới download về và **Next** cho đến khi kết thúc

2. Cấu hình Squid for Windows

2.1 Lấy địa chỉ IP của máy tính

Sau khi cài đặt thành công, Squid for Windows sẽ tự động khởi động. Nhưng trước tiên, chúng ta cần ghi lại địa chỉ IP của máy và ghi nhớ nó

Bước 1: Open "Command Prompt"

Bước 2: Gõ "ipconfig" và Enter

Bước 3: Hãy ghi nhớ "IPv4 Address"

```
IPv4 Address. . . . . . . . . . . : 192.168.1.7
```

Hình 12. Địa chỉ IP của máy tính

Tất nhiên từng kết nối mạng nhé!

2.2 Cấu hình cơ bản

Bước 1: Ấn vào biểu tượng Squid for Windows ở thanh taskbar và chọn "Stop Squid Service"

Bước 2: Làm quen với file cấu hình.

- Squid for Windows -> "Open Squid Configuration"
- Tìm với từ khóa "http_port" để xem Squid chạy ở cổng nào. Các bạn cũng có thể tùy chỉnh cổng mà mình thích
- Tìm với từ khóa "cache_dir", thêm dòng

cache_dir ufs /var/cache/squid 3000 16 256

ufs	Định dạng lưu trữ của Squid, bên cạnh đó còn có **aufs, diskd**
/var/cache/squid	Thư mục lưu trữ cache mặc định của Squid
3000	Dung lượng ổ đĩa sử dụng ở thư mục trên. Mặc định là 100MB. Không nên đặt giá trị này bằng với kích thước ổ cứng. Nếu Squid sử dụng ổ đĩa này thì đặt bằng 80% kích thước ổ.
16	Số lượng thư mục con level 1 được tạo trong thư mục bên trên. Mặc định bằng 16
256	Số lượng thư mục con được tạo trong mỗi thư mục level 1 bên trên. Mặc định bằng 256

- Lưu file

Bước 3: Tạo Inbound rule cho cổng 3128

- Windows Defender Firewall -> Advanced settings
- Click vào “Inbound Rules” và chọn “New Rule…” ở cột bên tay phải.
- Tick vào “Port” rồi “Next”
- Chọn “TCP” và nhập 3128, “Next”
- Chọn “Allow the connection”, “Next”
- Chọn tất cả, “Next”
- Nhập tên “Squid Proxy Server” và “Finish”

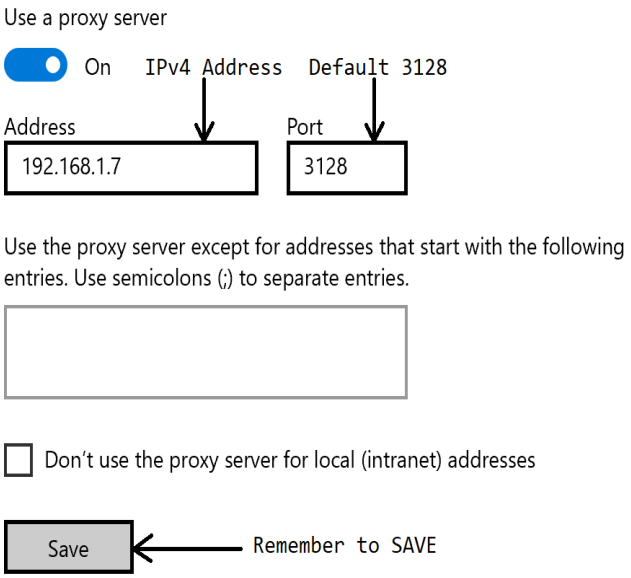
Bước 4: Bây giờ, chạy Squid Terminal với chế độ “Run as administrator”

- Vào thư mục “bin” bằng cách gõ “cd bin” rồi Enter
- Gõ “squid -z -F”

Bước 5: Khởi động Squid

- Ấn vào biểu tượng Squid for Windows và chọn “Start Squid Service”

- Settings -> Network & Internet -> Proxy
- Bật “Use proxy server”
- Điền địa chỉ IP bạn đã tìm lúc đầu và port là 3128 (mặc định) hoặc port bạn đã tùy chỉnh ở trên
- Save, nhớ lưu lại nhé



Hình 13. Cấu hình proxy server trên window 10.

3. Cấu hình chặn

Như vậy, chúng ta cơ bản đã cấu hình xong. Bây giờ là các bước cấu hình để chặn truy cập tới các tên miền, tệp tin, các trang web chứa từ khóa,...Chúng ta sẽ thêm các cài đặt ở phần **# INSERT YOUR OWN RULE(S) HERE TO ALLOW ACCESS FROM YOUR CLIENTS** trong file squid.conf.

Đầu tiên chúng ta sẽ phải tắt proxy của hệ thống cũng như dừng Squid

- Settings -> Network & Internet -> Proxy -> Tắt “Use a proxy server”
- Squid for Windows -> Stop Squid Service

Tiếp theo, chúng ta sẽ bắt đầu cấu hình trong file squid.conf. Mở file bằng cách

- Squid for Windows -> Open Squid Configuration

Tại phần **# INSERT YOUR OWN RULE(S) HERE TO ALLOW ACCESS FROM YOUR CLIENTS**, chúng ta sẽ chia thành hai đoạn

- Một đoạn để khai báo các đối tượng
- Một đoạn để đặt ràng buộc cho các đối tượng đã khai báo

3.1 Chặn tên miền

3.1.1 Cách cấu hình

Khai báo **acl blocked_sites dstdomain '/etc/squid/blocked_sites.txt'**

đối với danh sách các tên miền được lưu trong blocked_sites.txt hoặc

acl blocked_sites dstdomain .yahoo.com

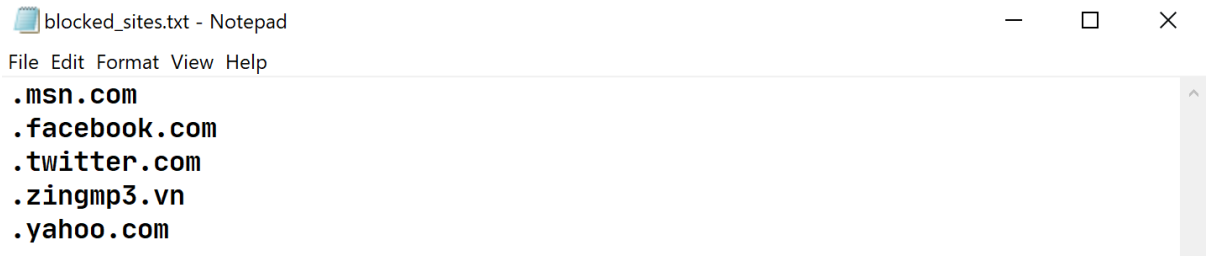
để chặn trực tiếp tên miền cụ thể

Đặt ràng buộc **http_access deny blocked_sites**

Với "blocked_sites" là tên chúng ta đặt cho đối tượng. Bạn có thể tùy biến nó.

Đối với tên miền hoặc danh sách tên miền được lưu trong file, chúng ta cần tạo file "blocked_sites.txt"

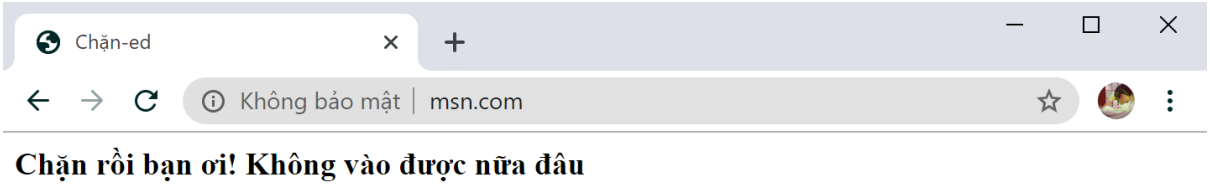
- Ấn vào biểu tượng Squid for Windows và chọn "Open Squid Folder"
- Vào thư mục etc -> squid
- Tạo file "blocked_sites.txt"
- Thêm các tên miền bạn muốn chặn



Hình 14. Danh sách các tên miền bị chặn tong file blocked_sites.txt.

- Lưu file. Khởi động Squid

3.1.2 Kết quả demo



Hình 14. Kết quả khi truy cập vào trang web bị chặn

3.2 Cấu hình chặn file

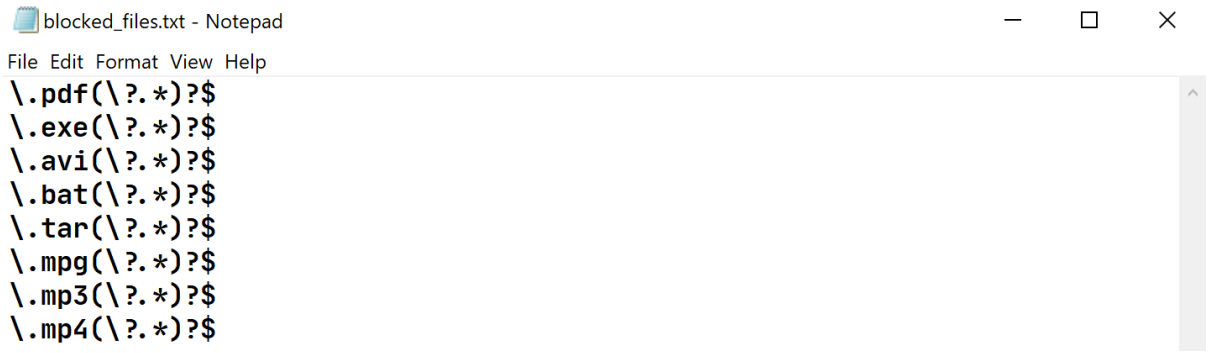
3.2.1 Cách cấu hình

Khai báo	**acl blocked_files urlpath_regex '/etc/squid/blocked_files.txt'** đối với danh sách các định dạng file được lưu trong blocked_files.txt hoặc **acl blocked_files urlpath_regex \.pdf(\?.*)?$** để chặn trực tiếp một định dạng file cụ thể
Đặt ràng buộc	**http_access deny blocked_files**

Với "blocked_files" là tên chúng ta đặt cho đối tượng. Bạn có thể tùy biến nó.

Đối với các định dạng file được lưu trong blocked_files.txt, chúng ta cần tạo file "blocked_files.txt"

- Ấn vào biểu tượng Squid for Windows, chọn "Open Squid Folder"
- Vào thư mục etc -> squid
- Tạo file "blocked_files.txt"
- Thêm các định dạng bạn muốn chặn



Hình 15. Danh sách các loại tệp bị chặn trong files blocked_files.txt

- Lưu file. Khởi động Squid

3.2.2 Kết quả demo

Hình 16. Kết quả khi truy cập vào một loại file đã bị chặn.

3.3 Cấu hình chặn từ khóa

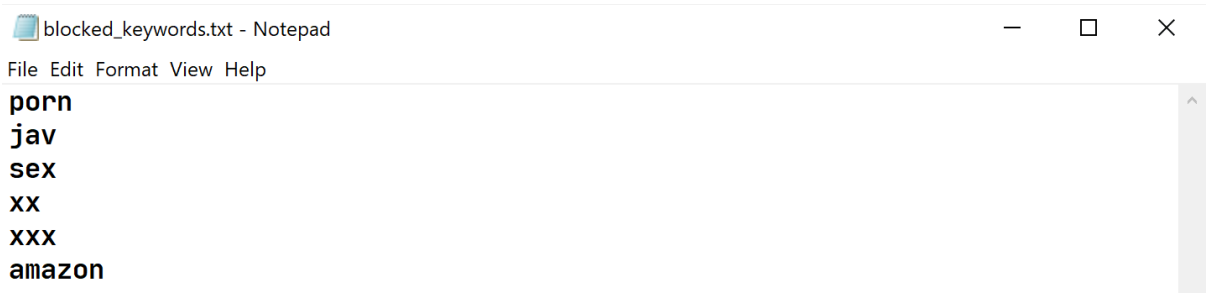
3.3.1 Cách cấu hình

Khai báo	**acl blocked_keywords url_regex '/etc/squid/blocked_keywords.txt'** đối với danh sách các từ khóa được lưu trong file "blocked_keywords.txt" hoặc **acl blocked_keywords url_regex porn** để chặn trực tiếp từ khóa cụ thể
Đặt ràng buộc	**http_access deny blocked_keywords**

Với "blocked_keywords" là tên chúng ta đặt cho đối tượng. Bạn có thể tùy biến nó.

Đó là phần cấu hình, bây giờ chúng ta sẽ tạo file "blocked_keywords.txt"

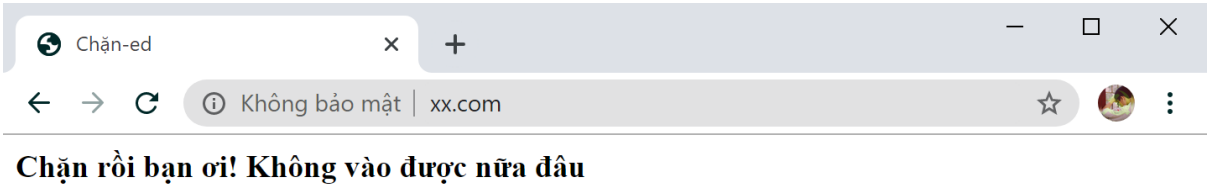
- Ấn vào biểu tượng Squid for Windows, chọn "Open Squid Folder"
- Vào thư mục etc -> squid
- Tạo file "blocked_keywords.txt"
- Thêm các từ khóa bạn muốn chặn



Hình 17. Các từ khóa bị chặn trong file block_keyworks.txt

- Lưu file. Khởi động Squid

3.3.2 Kết quả demo



Hình 18. Kết quả khi truy cập vào trang web có keywork đã bị chặn

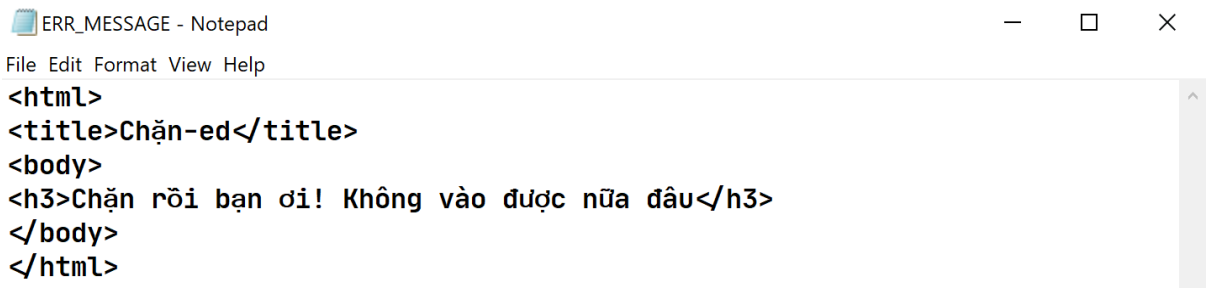
4. Tùy chỉnh thông báo

Ở trên các bạn có thể thấy trình duyệt trả về một thông báo rất thú vị. Chúng ta có thể tùy chỉnh thông báo trả về

4.1 Tạo file thông báo

- Squid for Windows -> Open Squid Folder -> usr -> share -> squid -> errors -> vi-vn

- Tạo file trống, không có định dạng. Ở đây tôi chọn tên ERR_MESSAGE
- Chèn đoạn mã HTML theo ý của bạn



Hình 18. File thông báo khi truy cập bị chặn.

- Lưu file

4.2 Cấu hình

- Squid for Windows -> Open Squid Configuration
- Chèn đoạn mã sau vào cuối file config

error_directory /usr/share/squid/errors/vi-vn

deny_info ERR_MESSAGE blocked_sites

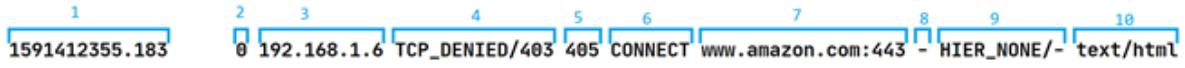
deny_info ERR_MESSAGE blocked_files

deny_info ERR_MESSAGE blocked_keywords

- Các bạn có thể tùy chỉnh thông báo nào được dùng cho đối tượng nào bằng câu lệnh **deny_info** như trên

5. Hướng dẫn đọc Squid log

- Squid for Windows -> Open Squid Folder -> var -> log -> squid
- Mở file access.log



Hình 19. Một dòng trong file access.log

Mỗi dòng trong file log có cấu trúc như hình trên. Sau đây là ý nghĩa của các trường.

1	Time	Thời điểm kết thúc, khi toàn bộ yêu cầu nhận được phản hồi. Tính theo giây từ 01/01/1970
2	Duration	Khoảng thời gian quá trình giao dịch sử dụng bộ nhớ cache. Tính theo mili giây
3	Client Address	Địa chỉ IP của máy khách thực hiện yêu cầu
4	Result Code	Cột này được tạo thành từ hai mục được phân tách bằng dấu gạch chéo thể hiện kết quả của quá trình giao dịch
5	Bytes	Kích thước của lượng dữ liệu được chuyển đến cho máy khách. Lưu ý đây không phải kích thước của một đối tượng mạng
6	Request Method	Phương thức yêu cầu tới đối tượng
7	URL	URL được yêu cầu
8	User	Xác thực người dùng đối với máy khách. Nếu không có xác thực người dùng, cột sẽ chứa một dấu gạch nối (-)
9	Hierarchy Code	
10	Type	Kiểu nội dung của đối tượng được lưu trong header của thông điệp HTTP

Phụ lục 1: danh sách Cache Result Code.

TCP_HIT	Indicates that a valid copy of the requested object was in the cache and that the proxy sent the object to the client.
TCP_MISS	Indicates that the requested object was not in the cache and that the proxy retrieved the object from the origin server or from a parent proxy and sent it to the client.
TCP_REFRESH_HIT	Indicates that the object was in the cache but was stale. Content Gateway made an if-modified-since request to the origin server and the origin server sent a 304 not-modified response. The proxy sent the cached object to the client.

TCP_REF_FAIL_HIT	Indicates that the object was in the cache but was stale. Content Gateway made an if-modified-since request to the origin server but the server did not respond. The proxy sent the cached object to the client.
TCP_REFRESH_MISS	Indicates that the object was in the cache but was stale. Content Gateway made an if-modified-since request to the origin server and the server returned a new object. The proxy served the new object to the client.
TCP_CLIENT_REFRESH	Indicates that the client issued a request with a no-cache header. The proxy obtained the requested object from the origin server and sent a copy to the client. Content Gateway deletes any previous copy of the object from the cache.
TCP_IMS_HIT	Indicates that the client issued an if-modified-since request and the object was in the cache and fresher than the IMS date, or an if-modified-since to the origin server found that the cache object was fresh. The proxy served the cached object to the client.
TCP_IMS_MISS	Indicates that the client issued an if-modified-since request and the object was either not in cache or was stale in cache. The proxy sent an if-modified-since request to the origin server and received the new object. The proxy sent the updated object to the client.
TCP_SWAPFAIL	Indicates that the object was in the cache but could not be accessed. The client did not receive the object.
ERR_CLIENT_ABORT	Indicates that the client disconnected before the complete object was sent.

ERR_CONNECT_FAIL	Indicates that Content Gateway could not reach the origin server.
ERR_DNS_FAIL	Indicates that the Domain Name Server could not resolve the origin server name, or that no Domain Name Server could be reached.
ERR_INVALID_REQ	Indicates that the client HTTP request was invalid. Content Gateway forwards requests with unknown methods to the origin server.
ERR_READ_TIMEOUT	Indicates that the origin server did not respond to the Content Gateway request within the timeout interval.
ERR_PROXY_DENIED	Indicates that client service was denied by access control configuration.
ERR_UNKNOWN	Indicates that the client connected but subsequently disconnected without sending a request.

Phụ lục 2 - Danh sách HTTP Status

Mã trạng thái	**Cụm từ chỉ lý do**	**Giải thích lỗi**
100	Continue	Yêu cầu đã được hoàn thành và phần còn lại của tiến trình có thể tiếp tục.
101	Switching Protocols	Khi yêu cầu một trang, trình duyệt có thể nhận được mã trạng thái 101, theo sau là header **"Upgrade",** cho thấy máy chủ đang thay đổi sang phiên bản HTTP khác.
102	Processing	
200	OK	Phản hồi tiêu chuẩn cho các yêu cầu HTTP thành công.

201	Created	Khi các trang mới được tạo bởi dữ liệu biểu mẫu đã đăng hoặc bởi tiến trình CGI, đây là dấu hiệu xác nhận rằng trang đó đã hoạt động.
202	Accepted	Yêu cầu của client đã được chấp nhận, nhưng chưa được xử lý.
203	Non-Authoritative Information	Thông tin chứa trong tiêu đề thực thể không phải từ trang web gốc, mà là từ máy chủ của bên thứ ba.
204	No Content	Nếu nhấp vào một liên kết không có URL mục tiêu, phản hồi này được máy chủ suy ra và không cảnh báo người dùng về bất cứ điều gì.
205	Reset Content	Điều này cho phép máy chủ reset lại bất kỳ nội dung nào được CGI trả về.
206	Partial Content	Các file được yêu cầu không được tải xuống hoàn toàn. Ví dụ, mã trạng thái này xuất hiện khi người dùng nhấn nút dừng trước khi trang được load.
207	Multi-Status	
300	Multiple Choices	Địa chỉ được yêu cầu đề cập đến nhiều hơn một file. Tùy thuộc vào cách máy chủ được cấu hình, bạn sẽ gặp lỗi hoặc được lựa chọn trang nào mong muốn.
301	Moved Permanently	Nếu máy chủ được thiết lập đúng cách, nó sẽ tự động chuyển hướng người đọc đến vị trí mới của file.
302	Found	Trang đã được di chuyển tạm thời và URL mới có sẵn. Bạn sẽ được máy chủ điều hướng đến đó.
303	See Other	Dữ liệu ở một nơi khác và phương thức GET được sử dụng để truy xuất nó.
304	Not Modified	Nếu header yêu cầu bao gồm tham số 'if modified since', mã trạng thái này sẽ được trả về, trong trường hợp file không thay đổi kể từ ngày đó.
305	Use Proxy	Người nhận dự kiến sẽ lặp lại yêu cầu thông qua proxy.

307	Temporary Redirect	
308	Permanent Redirect	
400	Bad Request	Có một lỗi cú pháp trong yêu cầu và yêu cầu bị từ chối.
401	Unauthorized	Header yêu cầu không chứa mã xác thực cần thiết và client bị từ chối truy cập.
402	Payment Required	Việc thanh toán là bắt buộc. Code này vẫn chưa hoạt động.
403	Forbidden	Client không được phép xem một file nhất định. Mã trạng thái này cũng được trả lại vào những thời điểm mà máy chủ không muốn có thêm khách truy cập.
404	Not Found	Các file được yêu cầu không có trên máy chủ. Có thể bởi vì những file này đã bị xóa, hoặc chưa từng tồn tại trước đây. Nguyên nhân thường là do lỗi chính tả trong URL.
405	Method Not Allowed	Phương pháp đang sử dụng để truy cập file không được cho phép.
406	Not Acceptable	File được yêu cầu tồn tại nhưng không thể được sử dụng, vì hệ thống client không hiểu định dạng mà file được cấu hình.
407	Proxy Authentication Required	Yêu cầu phải được cho phép trước khi diễn ra.
408	Request Time-out	Máy chủ mất quá nhiều thời gian để xử lý yêu cầu. Lỗi này thường gây ra bởi lưu lượng truy cập mạng cao.
409	Conflict	Quá nhiều yêu cầu đồng thời cho một file.
410	Gone	Các file đã được sử dụng ở vị trí này, nhưng không còn nữa.
411	Length Required	Yêu cầu thiếu header **Content-Length.**
412	Precondition Failed	Một cấu hình nhất định được yêu cầu để chuyển file này, nhưng client chưa thiết lập cấu hình đó.

413	Request Entity Too Large	Các file được yêu cầu là quá lớn để xử lý.
414	Request-URI Too Large	Địa chỉ đã nhập quá dài cho máy chủ.
415	Unsupported Media Type	Loại file của yêu cầu không được hỗ trợ.
416	Request Range Not Satisfiable	
417	Expectation Failed	
421	Misdirected Request	
422	Unprocessable Entity	
423	Locked	
424	Failed Dependency	
425	Unordered Collection	
426	Upgrade Required	
428	Precondition Required	
429	Too Many Requests	
431	Request Header Fields Too Large	
451	Unavailable For Legal Reasons	
500	Internal Server Error	Phản hồi khó chịu thường xảy ra do sự cố trong code Perl, khi chương trình CGI chạy.
501	Not Implemented	Yêu cầu không thể được máy chủ thực hiện.
502	Bad Gateway	Máy chủ cố truy cập đang gửi lại lỗi.
503	Service Unavailable	Service hoặc file đang được yêu cầu hiện không có sẵn.
504	Gateway Time-out	Cổng đã hết thời gian. Giống như 408 timeout error, nhưng lỗi này xảy ra tại cổng của máy chủ.
505	HTTP Version Not Supported	Giao thức HTTP yêu cầu không được hỗ trợ.

506	Variant Also Negotiates	
507	Insufficient Storage	
508	Loop Detected	
510	Not Extended	
511	Network Authentication Required	

Phụ lục 3 - Danh sách Hierarchy Code

https://wiki.squid-cache.org/SquidFaq/SquidLogs