Tạp chí Tin học và Điều khiển học, T.26, S.1 (2010), 17–27

PHÂN TÍCH KIẾN TRÚC NHỚ ĐỆM WEB KẾT HỢP CỦA MẠNG INTERNET

HỒ KHÁNH LÂM

Tập đoàn Bưu chính-Viễn thông, Công ty Điện toán và Truyền số liệu

Abstract. This paper uses markov chains to analyzes the performance of hybrid web caching architecture used by most of ISPs. The hybrid caching architecture has combines the avantages of both hierarchical and distributed caching, reducing the connection time as well as transmission time, helps ISPs to plan and save network resources at every level optimally. The analysis based on define web hit time at every network level, and common web hit time for ISP network with levels with hybrid caching architecture.

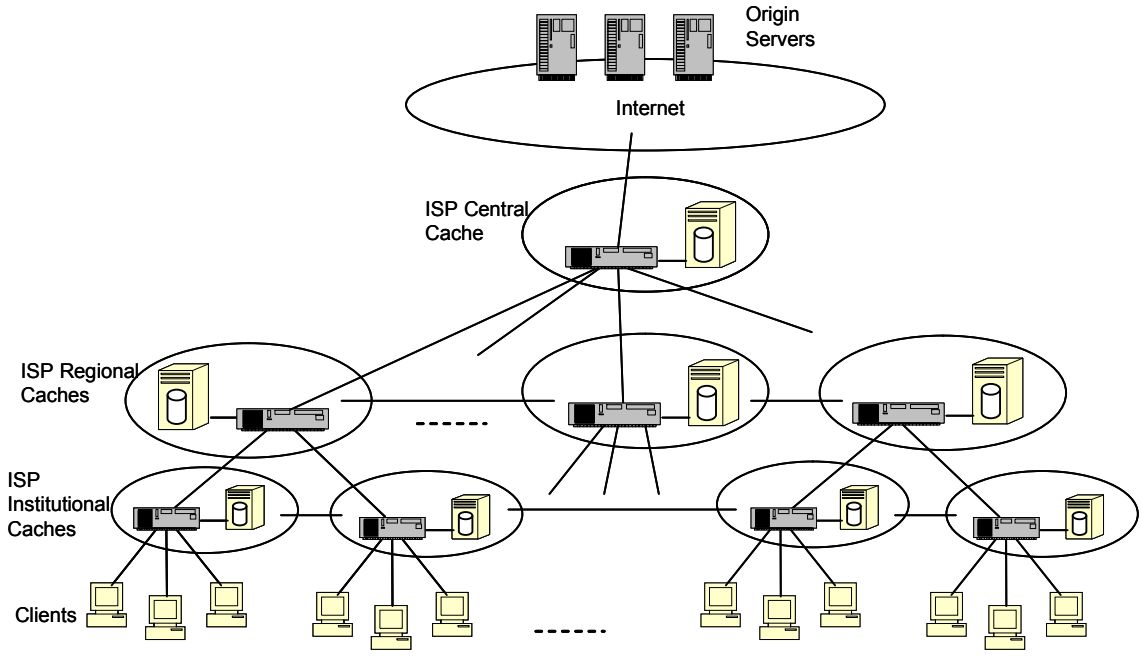
Tóm tắt. Bài báo này sử dụng các chuỗi Markov để phân tích hiệu năng của kiến trúc nhớ đệm web kết hợp được hầu hết các nhà cung cấp dịch vụ Internet sử dụng. Kiến trúc nhớ đệm web kết hợp có các ưu điểm của cả kiến trúc nhớ đệm web phân tầng và phân tán, giảm thời gian kết nối và thời gian truyền, giúp các nhà cung cấp dịch vụ Internet lập kế hoạch và tiết kiệm tài nguyên mạng ở từng cấp một cách tối ưu. Phân tích dựa trên sự xác định thời gian trúng web ở từng cấp mạng, và thời gian trúng web chung của toàn mạng của nhà cung cấp dịch vụ Internet với n cấp mạng sử dụng kiến trúc kết hợp nhớ đệm web.

1. GIỚI THIỆU

Một trong những giải pháp nâng cao hiệu năng của dịch vụ web: giảm trễ truy nhập web cho các client giảm chi phí băng thông, là có được kiến trúc web caching phù hợp. Có 3 loại kiến trúc web caching mà đa số các ISP áp dụng, đó là: kiến trúc phân cấp (hierarchical web caching architecture), kiến trúc web caching phân tán (distributed web caching) và kiến trúc web caching kết hợp (hybrid web caching). Kiến trúc web caching phân tầng cho phép các yêu cầu từ người sử dụng đầu cuối định tuyến từ mạng cấp thấp (cấp mạng truy nhập) đến các mạng cấp cao hơn: cấp địa phương (institutional network), cấp khu vực (Regional network), cấp quốc gia (national network) của mạng ISP, nếu tất cả cấp mạng của mạng ISP đều không có nội dung web mà client yêu cầu thì yêu cầu được chuyển lên Internet quốc tế. Như vậy đây là trường hợp xấu nhất, và nó cho trễ mạng lớn nhất để có được một nội dung web mà client yêu cầu. Đồng thời kiến trúc này cho tỷ lệ trúng cache không cao ở từng cấp mạng, nhưng lại yêu cầu băng thông lớn giữa các cấp mạng.

Kiến trúc web caching phân tán lại đảm bảo các hệ thống web cache liên kết ngang hàng với nhau trong từng cấp mạng, như vậy đảm bảo tỷ số trung web cao ở từng cấp mạng, và như vậy sẽ tiết kiệm băng thông giữa các cấp mạng. Tuy nhiên kiến trúc này lại yêu cầu đầu

tư chi phí lớn cho các hệ thống web cache ở từng cấp mạng: băng thông và các web server.



Hình 1. Kiến trúc web caching phân tầng của Internet

Hệ thống web caching kết hợp là một giải pháp mà các ISP thường sử dụng, Nó kết hợp hai loại kiến trúc phân tầng và kết hợp. Tại từng cấp mạng thực hiện kiến trúc web caching phân tán, song không phải ở tất cả các nút có các hệ thống web cache, vì lý do đảm bảo tiết kiêm chi phí, mà chỉ ở những nút có yêu cầu băng thông cao do có số đông dân cư sử dụng mạng Internet. Và tất cả các hệ thống web cache như vậy liên kết ngang hàng trong một tầng mạng (kết nối P2P) với nhau dể tăng mức độ sử dụng của hệ thống web caching ở từng cấp mạng. Kiến trúc web cache phân tầng kết hợp ở đây đảm bảo các hệ thống web cache của các tầng liên kết với nhau, như vậy đảm bảo tỷ lệ trúng web cao khi yêu cầu của các client chuyển lên tầng mạng trên và cũng làm tiết kiện băng thông giữa các tầng mạng. Hình 1 là sơ đồ của kiến trúc web caching kết hợp của các tầng mạng liên kết với có 4 cấp chính. Cấp cao nhất của toàn kiến trúc web caching kết hợp là hệ thống web caching trung tâm (cấp 1) của mạng trục Internet quốc gia, ISP CC (Central Cache). Tại các mạng khu vực của mạng ISP (ISP Regional Network) có các hệ thống Web cache khu vực (cấp 2), RC (Regional Cache). Cấp tiếp theo (cấp 3) là các hệ thống web cache của các mạng địa phương, IC (Institutional Cache). Những người sử dụng đầu cuối (client) ở cấp mạng 4. Chúng kết nối với các mạng truy nhập địa phương. Các viễn thông tỉnh, thành phố có các nút POP (Network Point of Presence) là các nút truy nhập địa phương của mạng Internet. Tại các POP đặt các hệ thống web cache, IC (Institutional Cache). Các client có thể là một máy tính PC, một điện thoại di động, trực tiếp hay thông qua mạng LAN, kết nối với Internet qua các POP bằng các mạng truy nhập như Dial-up, ADSL, mạng di động. Các POP có các liên kết

truyền dẫn tốc độ cao với các mạng khu vực.

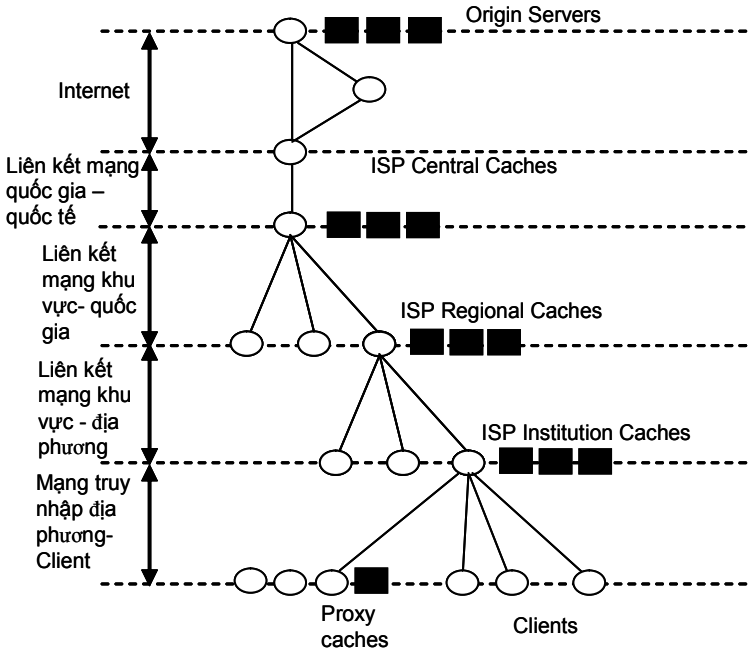
Các client bằng trình duyệt web có thể trực tiếp hoặc gián tiếp qua Proxy server cục bộ của LAN gửi yêu cầu về trạng web qua mạng truy nhập đến mạng địa phương. Trong trường hợp Proxy server cục bộ không có nội dung của trang web yêu cầu, thì Proxy server chuyển yêu cầu của client đến hệ thống IC ở các POP địa phương. Nếu hệ thống IC có nội dung mà client yêu cầu (trúng IC) thì IC chuyển nội dung yêu cầu về cho client (đồng thời Proxy server cục bộ cũng lưu nội dung trang web này). Trường hợp trúng IC (IC hit) thời gian đáp ứng yêu cầu của client (hay thời gian trễ đáp ứng) là nhỏ nhất. Khi trượt IC (IC miss) nghĩa là nội dung trang web mà client yêu cầu không có tại hệ thống IC, thì từ hệ thống IC yêu cầu được chuyển lên mạng cấp trên- mạng khu vực. Tại mạng khu vực, nếu hệ thống RC có nội dung yêu cầu (RC hit) thì nó chuyển nội dung về hệ thống IC, từ hệ thống IC chuyển tiếp về Proxy server và client. Nếu trượt RC (RC miss), yêu cầu của client được chuyển lên hệ thống CC của mạng quốc gia. Nếu hệ thống CC có nội dung thì nội dung (CC hit) được chuyển về hệ thống RC, rồi hệ thống IC, và đến Proxy server, đến client. Nếu trượt CC (CC miss), yêu cầu của client được chuyển đến Internet quốc tế, đến web server gốc.

2. CÁC NGHIÊN CỨU VỀ HIỆU NĂNG CỦA CÁC KIẾN TRÚC WEB CACHINH TRÊN INTERNET

Các tác giả Pablo Rodriguez, Christian Spanner,... [1] đã đưa ra mô hình của các kiến trúc web caching phân tầng và phân tán, và phân tích hiệu năng của các kiến trúc này với thời gian kết nối, thời gian truyền, trễ, và tốc độ trúng cache dựa theo lý thuyết xếp hàng với mô hình markov $M/D/1$. Tuy nhiên trong nghiên cứu này các tác giả cho rằng thời gian kết nối trung bình ở các tầng mạng là giống nhau và chỉ phụ thuộc vào số tầng mạng. Điều này là không chính xác vì thực tế ở mỗi tầng mạng băng thông khác nhau, tốc độ truyền dữ liệu khác nhau, trễ khác nhau và các hệ thống web cache công suất khác nhau, mức độ kết hợp (hợp tác) cũng khác nhau, và ở tầng mạng thấp nhất - tầng mạng truy nhập cũng phụ thuộc rất nhiều vào các công nghệ mạng truy nhập, mức độ tập trung dân cư ở các nút địa phương và các mạng đầu cuối của người dùng. Điều này cũng đã được các tác giả Guangwei Bai, Carey Williamson [2] phân tích các đặc tính tải trong kiến trúc web caching phân tầng, hoặc của các tác giả Abdullah Balamash and Marwan Krunz [3] khi phân tích hệ thống caching cho lưu lượng Web. Các nghiên cứu đánh giá hiệu năng của các hệ thống web caching của các tác giả ở các bài báo $[4-8]$ đều cho thấy ở từng cấp mạng phải có những nghiên cứu đánh giá riêng do sự khác nhau về lưu lượng. Công cụ chuỗi Markov được sử dụng phổ biến để đánh giá phân tích hiệu năng của các kiến trúc web caching, của web Proxy server. Công cụ chuỗi markov và các mô hình hàng đợi phù hợp cho các đánh giá phân tích hiệu năng của các hệ thống viễn thông. Và nội dung bài báo này cũng sử dụng chuỗi Markov để phân tích đánh giá hiệu năng của kiến trúc web caching kết hợp.

3. GIẢI PHÁP

Dựa theo kết quả nghiên cứu của các tác giả của bài báo [1], ở đây đề xuất sơ đồ mô hình cây của kiến trúc web caching kết hợp (Hình 2). Các hình chữ nhật bôi đen ở từng tầng mạng thể hiện sự kết hợp ngang hàng (P2P) của các hệ thống web cache.

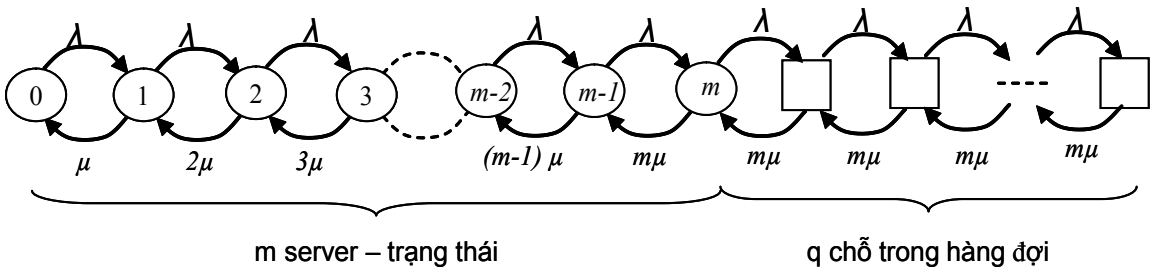


Hình 2. Mô hình cây của kiến trúc web caching phân tầng

Dễ nhận thấy kiến trúc web caching kết hợp làm cho trễ truy nhập các nội dung web của các client được giảm đi nếu tỷ lệ trúng web (web hit) cao ở từng tầng mạng. Nhưng nếu trượt web (web miss) xẩy ra ở nhiều tầng mạng thì trễ truy nhập web càng lớn.

Để phân tích hiệu năng của hệ thống web caching kết hợp, khác với các phương pháp phân tích của các nghiên cứu trước đây, bài báo này đưa ra mô hình xếp hàng với chuỗi Markov thời gian liên tục (CTMC) cho hệ thống web caching. Vì ở mỗi tầng mạng các web server về nguyên tắc có cấu hình, công suất giống nhau, chúng liên kết ngang hàng nhau, và các truy nhập đến chúng đều có thể tới bất kỳ, do vậy mỗi hệ thống web caching ở từng cấp mạng có thể được coi gồm nhiều server kết nối song song và được biểu diễn bằng một mô hình hệ thống M/M/m/q với cả mất mát và trễ. Cho rằng các yêu cầu HTTP từ hệ thống web caching của các client đến độc lập với nhau, và số lượng các client có yêu cầu phát sinh HTTP không bị giới hạn. Thời gian giữa các yêu cầu HTTP đến hệ thống và thời gian phục vụ của các hệ thống có phân bố mũ. Hệ thống web caching có m server giống nhau, $m = 1, 2, ...$ Chúng có dung lượng nhớ đệm (cache) bị giới hạn được mô hình như các hàng đợi nhận các yêu cầu HTTP với độ dài là q. Vì độ trễ truyền dẫn do môi trường mạng ở từng cấp mạng khác nhau nên ta cho rằng tốc độ đến trung bình của các yêu cầu HTTP ở

từng cấp là $\lambda_i,\ i = 0, 1, 2, ..., n$ trong đó n là cấp mạng, và tốc độ phục vụ trung bình của các hệ thống web caching (server) ở từng cấp mạng là $\mu_i,\ i = 0, 1, 2, ..., n$. Mức độ sử dụng của từng web server là $U = \lambda_i/\mu_i$. Tổng quát nếu, nếu hệ thống web caching ở từng tầng mạng gồm có m server kết nối song song và có q chỗ trong hàng đợi thì ta có độ thị trạng thái của CTMC cho ở Hình 3.



Hình 3. Đồ thị trạng thái CTMC của hệ thống web cache $M/M/m/q$

Đối với chuỗi CTMC này các đẳng thức cân bằng luồng được thỏa mãn như sau:

+ Khi $0 \leq k \leq m$:

$$\begin{cases} \lambda_i p_{i0} = \mu_i p_{i1},\ p_{i1} = U_i p_{i0} \\ \lambda_i p_{i1} = 2\mu_i p_{i2},\ p_{i2} = \dfrac{U_i^2}{2} p_{i0} \\ \lambda_i p_{i2} = 3\mu_i p_{i3},\ p_{i3} = \dfrac{U_i^2}{6} p_{i0} \\ \lambda_i p_{ik-1} = k\mu_i p_{ik},\ p_{ik} = \dfrac{U_i^k}{k!} p_{i0} \\ \lambda_i p_{im-1} = m\mu_i p_{im},\ p_{im} = \dfrac{U_i^m}{m!} p_{i0} \end{cases} \Rightarrow p_{ik} = \frac{U_i^k}{k!} p_{i0},\ \forall k = 1, 2, 3, ..., m. \quad (1)$$

trong đó, p_{i0} là xác suất mà yêu cầu HTTP của client đến khi trúng web ở cấp mạng i và hệ thống web cache đang ở trạng thái rỗi.

p_{ik} là xác suất mà yêu cầu HTTP của client đến khi trúng web và hệ thống web caching của cấp mạng i đang ở trạng thái k (đang phục vụ k yêu cầu HTTP).

+ Khi $m \leq k$ nhưng không gian trong hàng đợi có từ 1 đến q:

$$\begin{cases} \lambda_i p_{im} = m\mu_i p_{i(m+1)},\ p_{i(m+1)} = \dfrac{\lambda_i}{m\mu_i} p_{im} = \dfrac{U_i^m}{m!}(\dfrac{U_i}{m}) p_{i0} \\ \lambda_i p_{i(m+1)} = m\mu_i p_{i(m+2)},\ p_{i(m+2)} = \dfrac{\lambda_i}{m\mu_i} p_{i(m+1)} = \dfrac{U_i^m}{m!}(\dfrac{U_i}{m})^2 p_{i0} \\ \lambda_i p_{i(m+k-1)} = m\mu_i p_{i(m+k)},\ p_{i(m+k)} = \dfrac{\lambda_i}{m\mu_i} p_{i(m+k-1)} = \dfrac{U_i^m}{m!}(\dfrac{U_i}{m})^k p_{i0} \end{cases}$$

$\Rightarrow p_{i(m+k)} = \dfrac{U_i^m}{m!}(\dfrac{U_i}{m})^k p_{i0},\ \forall k = 1, 2, 3, ..., q.$

+ Đối với mô hình $M/M/m/q$ này điều kiện bình thường hóa phải được thỏa mãn, đó là tổng các xác suất phải bằng 1:

$$1=\sum_{k=0}^{m+q}p_{ik}=p_{i0}\Big(1+U_i+\frac{U_i^2}{2!}+\frac{U_i^3}{3!}+...+\frac{U_i^{m-1}}{(m-1)!}+\frac{U_i^m}{m!}+\frac{U_i^m}{m!}(\frac{U_i^m}{m})+...+\frac{U_i^m}{m!}(\frac{U_i}{m})^q\Big).$$

Ta có:

$$1=p_{i0}\Big(\sum_{k=0}^{m}\frac{U_i^k}{k!}+\frac{U_i^m}{m!}\sum_{k=1}^{q}(\frac{U_i}{m})^k\Big)=p_{i0}S,$$

$$S=\Big(\sum_{k=0}^{m}\frac{U_i^k}{k!}+\frac{U_i^m}{m!}\sum_{k=1}^{q}(\frac{U_i}{m})^k\Big).$$

Suy ra:

$$p_{i0}=\frac{1}{S}=\Big(\sum_{k=0}^{m}\frac{U_i^k}{k!}+\frac{U_i^m}{m!}\sum_{k=1}^{q}(\frac{U_i}{m})^k\Big)^{-1}. \tag{2}$$

+ Khi toàn bộ m server và cả q chỗ trong hàng đợi của hệ thống web caching của từng cấp mạng đều bận thì yêu cầu HTTP mới đến hệ thống sẽ bị khóa (không được đưa vào hàng đợi). Đây chính là trường hợp nghẽn tại tầng mạng i khi toàn bộ hệ thống web caching với m server đã quá tải. Trạng thái này được xác định bằng xác suất khóa hay xác suất mất mát tại cấp mạng i và bằng $B_i=p_{i(m+q)}$,

$$B_i=p_{m+q}=\frac{U_i^m}{m!}(\frac{U_i}{m})^q p_{i0}. \tag{3}$$

Như thế xác suất khách hàng mới đến phải chờ đợi chính là xác suất mà m server bận và hàng đợi còn có chỗ trống là $p_{i(m+k)}$ với $1\le k\le q$.

Theo định luật Zipf và Internet [9], ta có thể xác định số lượng truy nhập trên số đông dân cư tại các địa phương. Phải có thống kê dự báo dân cư truy nhập ở từng khu vực, và dựa vào kết quả này xây dựng các hệ thống web caching với các server công suất tối ưu (CPU và dung lượng nhớ) đảm bảo cho giá trị q chứa được tối đa số lượng các yêu cầu HTTP.

Với mô hình $M/M/m/q$ này ta có thể tính các thông số hiệu năng cho hệ thống web caching cho từng cấp mạng i như sau:

1) Số lượng các yêu cầu HTTP có trong hàng đợi của hệ thống web caching của tầng mạng i, $E[N_{iq}]$:

$$E[N_{iQ}]=\sum_{k=1}^{q}kp_{i(m+k)}=p_{i0}\frac{U_i^m}{m!}\Big((\frac{U_i}{m})+2(\frac{U_i}{m})^2+3(\frac{U_i}{m})^3+...+q(\frac{U_i}{m})^q\Big)=p_{i0}\frac{U_i^m}{m!}\sum_{k=1}^{q}k(\frac{U_i}{m})^k. \tag{4}$$

2) Thời gian chờ đợi trung bình của một yêu cầu HTTP ở hệ thống web caching của tầng mạng i để được phục vụ $E[W_{iQ}]$ được xác định theo luật Little:

$$E[W_{iQ}] = \frac{E[N_{iQ}]}{\lambda_i} = \frac{1}{\lambda_i}\Big(p_{i0}\frac{U_i^m}{m!}\sum_{k=1}^{q} k(\frac{U_i}{m})^k\Big). \quad (5)$$

3) Thời gian đáp ứng trung bình $E[C_i]$ của hệ thống web caching ở từng cấp mạng i:

Đây là thời gian trung bình mà một yêu cầu HTTP của client (nội dung web) được xử lý trong hệ thống web cache (gồm thời gian chờ ở hàng đợi và thời gian được phục vụ (nội dung web được tìm ra):

$$E[C_i] = E[W_{iQ}] + E[S_i] = \frac{E[N_{iQ}]}{\lambda_i} + \frac{1}{\mu_i}. \quad (6)$$

Tổng quát, nếu kiến trúc web caching của một mạng ISP có n cấp mạng, yêu cầu HTTP của client trượt web ở cấp mạng thứ n, trúng web ở hệ thống web caching ở cấp mạng thứ i mà $n > i$ thì đáp ứng của hệ thống web caching ở cấp mạng thứ i cho yêu cầu HTTP của client sẽ bằng:

$$\begin{aligned} E[R_i] =&(D_{nM} + D_{nREQ}) + (D_{n-1M} + D_{n-1REQ}) + ... + (D_{i+1M} + D_{i+1REQ}) + E[C_i] + \\ & D_{i+1} + ... + D_{n-1} + D_n, \end{aligned} \quad (7)$$

trong đó:

D_{nM}- trễ do trượt web ở cấp mạng thứ n.

D_{nREQ} - trễ phụ thuộc băng thông kênh truyền dẫn mà yêu cầu HTTP của client (hay từ proxy server cục bộ) chuyển từ cấp mạng thứ n đến cấp mạng thứ $n-1$.

D_n - trễ trả về nội dung web yêu cầu cho client phụ thuộc băng thông kênh truyền dẫn từ cấp mạng thứ $n-1$ đến cấp mạng thứ n, và phụ thuộc kích thước của nội dung web.

Diễn giải tương tự cho các tầng mạng của cấu trúc mạng Internet của hầu hết các nhà cung cấp dịch vụ (ISP), ta có một đồ thị biểu diễn thời gian trễ của giao dịch HTTP ở Hình 4. Trong đồ thị này ta có các thời gian như sau:

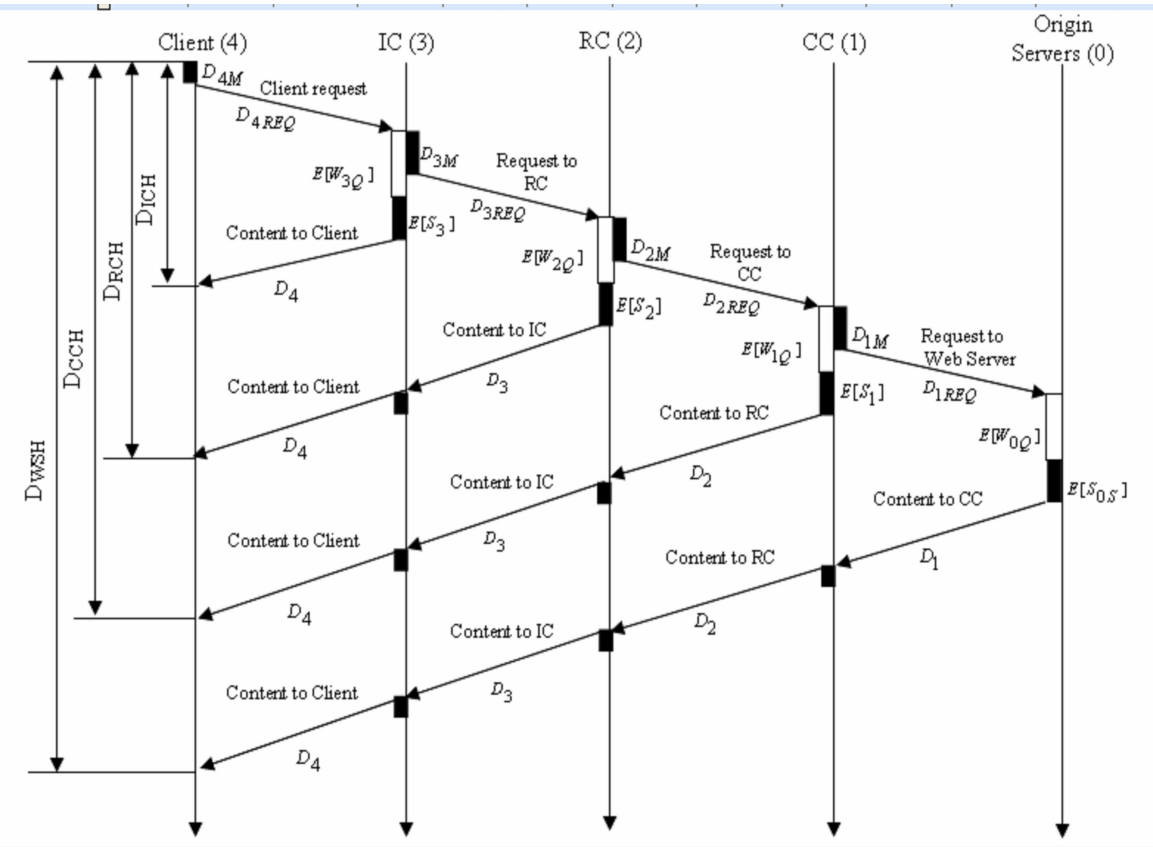
D_{4M} - thời gian trượt web cục bộ tại mạng của client, thời gian này phụ thuộc vào tốc độ của LAN (trong đó có proxy server) của client. Nếu client là một máy tính đơn lẻ không qua LAN thì thời gian này có thể bỏ qua.

D_{4REQ} thời gian mà yêu cầu HTTP của client được gửi đến mạng địa phương/cơ sở (đến POP hoặc đến Router cơ sở). Thời gian này phụ thuộc vào tốc độ đường truyền kết nối client qua mạng truy nhập (dial-up, ADSL, vô tuyến, di động, đường truyền trực tiếp) tới các nút mạng địa phương/cơ sở , và phụ thuộc kích thước gói yêu cầu, trễ qua các nút mạng truy nhập trung gian của từng khu vực như quận, huyện đến các POP.

D_4 - trễ trả về client khi trúng web ở hệ thống IC (trúng web cấp 3), phụ thuộc vào băng thông kênh truyền dẫn và kích thước nội dung web tr về từ cấp mạng địa phương đến client ở mạng cấp 4.

D_{ICH} - thời gian đáp ứng trung bình của hệ thống web caching IC khi trúng web ở IC (IC hit). Thời gian này bao gồm: thời gian tổng thời gian ($D_{4M} + D_{4REQ}$), thời gian đáp

ứng trung bình của hệ thống web caching ở cấp mạng 3, $E(C_3)$ và thời gian trả nội dung web D_4.



Hình 4. Đồ thị thời gian trễ của giao dịch HTTP của client trên mạng Internet với kiến trúc Web caching kết hợp

$$\begin{aligned} D_{ICH} =& (D_{4M} + D_{4REQ}) + E[C_3] + D_4 = (D_{4M} + D_{4REQ}) + E[W_{3Q}] + E[S_3] + D_4 \\ &= (D_{4M} + D_{4REQ}) + \frac{E[N_{3Q}]}{\lambda_3} + \frac{1}{\mu_3} + D_4. \end{aligned} \tag{9}$$

D_{RCH} - thời gian đáp ứng trung bình của hệ thống web caching ở mạng khu vực khi trúng RC (RC hit):

$$\begin{aligned} D_{ICH} =& (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + E[C_2] + D_3 + D_4 \\ &= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + E[W_{2Q}] + E[S_2] + D_3 + D_4 \\ &= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + \frac{E[N_{2Q}]}{\lambda_2} + \frac{1}{\mu_2} + D_3 + D_4. \end{aligned} \tag{10}$$

D_{CCH} - thời gian đáp ứng trung bình của hệ thống web caching ở mạng quốc gia khi trúc CC (CC hit):

$$D_{CCH} = (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + E[C_1] + D_2 + D_3 + D_4$$
$$= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + E[W_{1Q}] + E[S_1] + D_2 + D_3 + D_4$$
$$= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + \frac{E[N_{2Q}]}{\lambda_2} + \frac{1}{\mu_2} + D_2 + D_3 + D_4. \quad (11)$$

D_{WSH} - thời gian đáp ứng trung bình của Internet quốc tế và web server nguồn:

$$D_{WSH} = (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + (D_{1M} + D_{1REQ}) +$$
$$E[C_1] + D_2 + D_3 + D_4$$
$$= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + E[W_{1Q}] + E[S_1] + D_2 + D_3 + D_4$$
$$= (D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + (D_{2M} + D_{2REQ}) + \frac{E[N_{1Q}]}{\lambda_1} + \frac{1}{\mu_1} + D_2 + D_3 + D_4. \quad (12)$$

Như vậy, trường hợp xấu nhất là không trúng web yêu cầu ở tất cả các cấp mạng của ISP trong quốc gia và chỉ trúng web trên cấp mạng Internet quốc tế ở web server nguồn. Công thức (12) là công thức tổng quát để tính trễ truy nhập web cho trường hợp xấu nhất này. Công thức này cho thấy sự phụ thuộc vào từng hệ thống web caching ở từng cấp mạng: các liên kết ngang hàng, số lượng nút hệ thống web caching, giao thức thay thế web cache, các liên kết của các hệ thống web caching của các tầng mạng với nhau, và số lượng tầng mạng.

Tính toán minh họa một số thông số của mô hình *M/M/m/q*

Cấp mạng 4: Cho rằng, hệ thống web caching có $m = 5$ server; hàng đợi (số các yêu cầu HTTP) $q = 10$. Để đạt được mức độ sử dụng của hệ thống web cahing ở cấp mạng 4, $U_4 = \frac{\lambda_4}{\mu_4} = 1$ thì tốc độ đến trung bình của các yêu cầu HTTP phải bằng tốc độ phục vụ trung bình của hệ thống web caching ở cấp mạng 4. Như vậy, từ biểu thức (2) ta suy ra $p_{40} = 1$. Xác suất khóa khi toàn bộ 5 server của hệ thống web caching ở tầng 4 sẽ là vô cùng nhỏ, nghĩa là ít có khả năng nghẽn ở tầng mạng 4, và bằng:

$$B_4 = \frac{1}{5!}(\frac{1}{5})^{10}(1) \approx (0,008)(1,024)10^{-7}.$$

Nếu chỉ tăng m - số lượng server kết nối của hệ thống web caching của tầng mạng (giữ nguyên giá trị q), hoặc tăng cả q và m nguyên thì xác suất khóa sẽ tiến nhanh đến 0. Như vậy theo các công thức (4) và (5) số lượng trung bình các yêu cầu HTTP ở trong hàng đợi và thời gian chờ đợi sẽ bằng 0. Vậy các công thức (2), (3) và (4) giúp các nhà thiết kế và qui hoạch mạng xác định số lượng các server và dung lượng bộ nhớ tương ứng của hệ thống web caching của tầng mạng. Theo công thức (6) ta nhận thấy trong trường hợp xác suất khóa bằng 0 thời gian đáp ứng trung bình của hệ thống web caching của tầng bằng chính thời gian phục vụ trung bình của tất cả các server trong hệ thống.

Các cấp mạng còn lại: Cũng với phân tích tương tự ta có thể tính được các thông số hiệu năng theo các công thức (2),(3),(4),(5),(6) cho từng tầng mạng với các hệ thống web caching tương ứng. Và tổng quát theo công thức (7) ta tính được đáp ứng trung bình của hệ thống web caching của tầng i.

Kết luận. Bài báo này đề xuất áp dụng mô hình $M/M/m/q$ cho kiến trúc web caching kết hợp của Internet với các kết quả của mô hình này là các công thức nhận được (1), (2), (3), (4), (5), (6), (7) giúp cho các nhà thiết kế qui hoạch mạng một công cụ để tính toán hiệu năng của các hệ thống web caching cần xây dựng nếu có các giá trị về các kênh truyền dẫn, số nút mạng, thống kê dự báo số người dùng ở từng địa phương, khu vực. Các kết quả đưa ra không giống với các nguyên cứu tương tự trước đây [1] ở chỗ nó chi tiết đến từng cấp mạng với các thông số về thời gian trễ của truyền dẫn, của từng hệ thống web caching. Và nó dễ dàng ở chỗ chỉ cần áp các giá trị cụ thể của mạng Internet định thiết kế hay đang khai thác có thể xác định các giá trị hiệu năng cần thiết. Những kết quả này chưa được công bố ở bất cứ tạp chí khoa học nào.

Hướng nghiên cứu tiếp sẽ là mô hình chuỗi markov $M/M/m/q$ được áp dụng để tính toán cụ thể cho từng hệ thống web caching riêng biệt cho một số dịch vụ băng thông rộng trên Internet như Game-online, IPTV, web conferencing, Internet telephony,...

TÀI LIỆU THAM KHẢO

[1] Pablo Rodriguez, Christian Spanner, Ernst W.Biersack, Web caching architectures: Hierarchical and distributed caching, http://workshop99.ircache.net (4^{th} International WWW Caching Workshop), Institut EUROCOM, france, 1999.

[2] Guangwei Bai, Carey Williamson, Workload characterization in Web caching hierarchies, *10^{th} IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems,* (MASCOTS02), 2002.

[3] Abdullah Balamash, Marwan Krunz, and Philippe Nain, Performance analysis of a client-side caching/prefetching system for Web traffic, *Computer Networks* **51** (13) (12 September 2007) 3673–3692.

[4] Carey Williamson, Mudashiru Busari, "Simulation Evaluation of Web Caching Architectures", M.Sc. Thesis, June 2000, Department of Computer science, University of Saskatchewan, http://www.cs.usask.ca/faculty/carey/.

[5] Haohuan Fu, Pui-On Au, Weijia Jia, Performance evaluation of replacement algorithms in hierarchical Web caching, *Book series Lecture notes in computer science*, **3129** (2004) 1611–3349 (Springer Berlin/Heidelberg", ISSN 0302-9743 (online)).

[6] A. Rousskov, On performance of caching proxies, *ACM SIGMETRICS*, Madison, USA, september 1998.

[7] C. Maltzahn, J.Richardson, "Performance Issues of Enterprise Level Web Proxies", 1998.

[8] M. Deshpande, G. Karypis, Selective Markov models for pedicting Web page access, *ACM Transactions on Internert Technology* (may 2004)

[9] Lada A.Adamic, Bernardo A.Huberman, "Zipfs Law and Internet", HP Laboratories, Glottometrics 3, 2002, 143–150.

[10] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, P. Owezarski, A flow- based model for internet backbone traffic, internet measurement conference, *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment, Year of Publication,* 2002 (ISBN:1-58113-603-X).

[11] Công ty VDC, "Mạng Internet của VDC", Tài liệu lưu hành nội bộ, 2008.

Nhận bài ngày 21 - 12 - 2009
Nhận lại sau sửa ngày 15 - 4 -2010