1A-5

独立アクセスモデルに基づくCDNとアクセス解析

石井　充†　　加藤崇†　　服部　進実†

金沢工業大学工学部情報工学科†

1.序論

ADSL, FTTH, CATV などのブロードバンドネットワークが急速に普及するにつれて、ネットワーク上でやり取りされるコンテンツの量も増加している。コンテンツそのもののサイズの増加と、アクセス数の増加という効果により、今後ネットワークへの負荷がますます大きくなるものと予想される。こういった背景の下で、コンテンツを効率よく配信するため、CDN（Content Distribution/Delivery Network）やマルチキャストなどの技術が注目され研究される[1]と同時に商業サービスとしての試みもされている。

我々の以前の研究においても、CDN を対象として、効率的なコンテンツの配置を行うための多段階キャッシュシステムを調べた[2],[3]。

しかしながら、それらの研究は、さまざまなコンテンツ配置方針のもとでシミュレーションを行い、その結果を元に、いわば事後的に、どういったやり方が適しているかを調べるものであり、何らかの数学的モデルに基づいて、合理的に最適な配置方法を見出すという方法ではなかった。

例えば、複数のキャッシュサーバー間で連携しながら、アクセス頻度の多いコンテンツを、末端のユーザーに近いところに配置するといったことはしばしば行われるが、本来はアクセス頻度が多いコンテンツでも、たまたま特定の期間のアクセス数が少なかったがために、末端のユーザーから遠いところに配置されてしまうという場合も考えられる。こういったことがどの程度の確率で生じるのか、また、こういったことを避けるためにはどうすればよいのかということを数学的に解析しておく必要がある。

そこで、本稿では、アクセスが独立に生じるものとして、多段階キャッシュサーバーにおいて、上記のような不都合が生じる可能性を検討し、更にその解決方法を提示する。

2.基本的枠組み

最も簡単な多段階キャッシュサーバーとして、図 1(a)に挙げたものを考える。一般的には、図1(b)のように下流のキャッシュサーバーが複数になっているであろうが、本章で以下に述べる枠組みはどちらにも適用可能である。

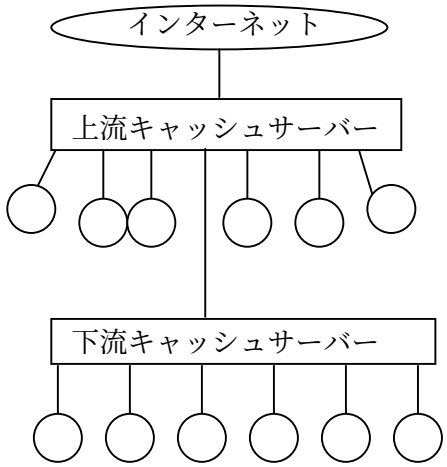

図 1(a)　本稿で取り上げる多段キャッシュ。円はキャッシュサーバーを利用するエンドユーザーのクライアントを表す

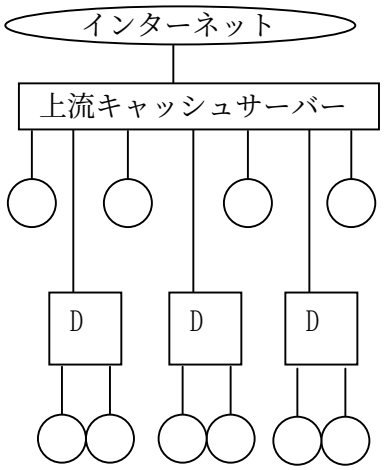

図１(b) 一般的には、上流キャッシュサーバーの下に複数のキャッシュサーバーがある。Dは下流キャッシュサーバーを表す。

下流キャッシュサーバーが空である状態から始めるものとする。下流キャッシュサーバー下にあるエンドユーザーのクライアントからのアクセスがそれぞれのコンテンツに対して独立に生じるものとすると、最終的に上流キャッシュサーバーでの、個々のコンテンツへの単位時間内のアクセス回数が n である確率 $p(n)$は、単位時間あたりの平均アクセス回数 q を用いて、ポアソン分布

$$p(n) = \frac{q^n}{n!} e^{-q}$$

で与えられる。

このコンテンツが単位時間内に m-1 回以下のアクセス数である確率は、

$$\sum_{k=0}^{m-1} p(k) = \frac{\Gamma(m,q)}{\Gamma(m)}$$

で与えられる。ここで$\Gamma(m)$ はガンマ関数であり、$\Gamma(m, q)$は

$$\Gamma(m,q) = \int_q^{\infty} t^{m-1} e^{-t} dt$$

により与えられる不完全ガンマ関数である。

特に、平均アクセス回数である q よりも小さい回数しかアクセスされない確率は、

$$\frac{\Gamma(q,q)}{\Gamma(q)}$$

となる。また、アクセスされる確率が rである m の値は

$$\frac{\Gamma(m,q)}{\Gamma(m)} = r$$

を解くことにより求められる。

3.具体例とアクセス解析結果

単位時間あたり 100 回のアクセスがあるコンテンツを、上流キャッシュサーバーから下流キャッシュサーバーに移動させたいものとする。あるコンテンツは、単位時間あたり 100 回のアクセスが期待されるものとすると、特定の単位時間あたりにアクセスが 100 回未満である確率は、

$$\frac{\Gamma(100,100)}{\Gamma(100)} = 0.487$$

で与えられる。このことは、単純に、単位時間内に 100 回のアクセスがあったものを下流のキャッシュサーバーにコピーするという方法では、実際に単位時間あたり 100 回のアクセスが期待されるコンテンツのうち、約半分を取りこぼしていることを意味する。100 回のアクセスが期待されるコンテンツでも、実際の特定の単位時間内に 100 回以上アクセスされることもあれば 100 回未満のこともあり、そのどちらになるかがほぼ等確率であると理解することができ、妥当な結果であるといえる。

以上のような解釈を踏まえて、実際のアクセスログを理論式に照らし合わせて解析した。商業インターネットプロバイダーのキャッシュサーバーのログを解析し、単位時間に 1 日をとって、1 日あたり 100 回のアクセスがあるコンテンツを選び出すため、5 日間にわたるログから、1 日あたりの平均のアクセス数を調べた。1 日あたりの平均でちょうど 100 回になるコンテンツの数が少なく、統計的に有意の結果が得られなかったので、1 日あたりの平均のアクセス数が 95 から 115 までの間のコンテンツ 52 個を選び出して、特定の 1 日の間のアクセス数を調べた。これを上記理論で q=100，105，110 としたものと比較したのが表 1 である。

m	実データ	q=100	q=105	q=110
70	100%	100%	100%	100%
80	94%	98%	99%	100%
90	88%	85%	93%	97%
95	75%	70%	85%	93%
100	62%	51%	75%	84%
105	50%	32%	51%	70%
110	38%	17%	32%	51%
115	25%	8%	18%	33%
120	13%	3%	8%	18%
130	2%	0%	1%	3%

表 1　商業プロバイダーのキャッシュサーバーから得られたデータと理論値との比較。左端の m はアクセス数を表し、m 以上のアクセス数を持つコンテンツの割合が、第 2 列以降に示してある。m<100 では q=100 の場合がおおむね妥当であるが、100<m<105 では q=100 と q=105 の間の値になっており、m>110 では q=105 と q=110 の間の値になっている。

表 1 から、標本数が十分に多くなく統計誤差があることや、1 日あたりの平均のアクセス数を得るために用いた期間が 5 日と短いことを考えると、理論式が現実のデータをおおむね説明できていると言える。

なお、本研究は部分的に情報処理推進機構の 2004 年度未踏ソフトウェアー創造事業の支援を受けて行われたものである。

参考文献

[1] R.Brussee et. al. “Content distribution network state of the art,” Telematica Instituut, June 2001

[2] Y.Ikeda et. al. “Construction and its verification of policy selection type CDN platform,” IEICE Trans. Commun., Vol.J86-B, No3, pp400-409, March 2003

[3] T.C.Hu et. al. “Total cost-aware proxy caching with cooperative removal policy,” IEICE Trans. Commun., Vol.E86-B, No.10, pp3035-3062, Oct. 2003

CDN and Access Analysis Based on Random Access Assumption

† Division of Information and Computer Science, Kanazawa Institute of Technology