分散処理システムにおける最適負荷分担

宮原秀夫（大阪大学　基礎工学部）

1. まえがき

ある処理装置が現在それに入流している負荷以上のものを処理する能力があるとき、その装置は"軽負荷"(lightly loaded)といい、またそれとは逆に、処理能力以上の入流負荷があるとき"過負荷"(over loaded)状態にあると言う。前者の状態にある処理装置から後者の状態にある装置へと負荷を転送することを負荷分担(load sharing)と呼ぶ。本稿では、複数個の処理点(ノード)がネットワーク状に結合された分散処理システムを対象とし、そこにおける負荷分担の問題を取扱う。分散処理システムにおいて負荷分担を行うには、ある処理装置への負荷の一部を他のどの装置へも回わして処理できるという条件、即ち処理装置の均質性(homogeneity)が要求される。従ってこの負荷分担を考える上で対象とするシステムは機能分散というよりむしろhomogeneousな系を対象とすることになる。あるいは一般の分散処理システムにおけるhomogeneousな部分だけを抽出して負荷分担を行うものとみなしてもよい。この負荷分担は、あるノードへ系外から到着するジョブ流(job stream)を幾つかの流れに分岐させ、その流れの1つをローカルな処理点へ、その他の流れを他の処理点へ転送することによってなされる。即ち入力流のどれだけの割合をどこへ転送するかを決定する問題である。負荷分担に関し2つのタイプがある。1つは固定方式で、流入するトラヒックのかなり長い時間にわたる確率的特性に基づいて分割割合を決定するものである。従ってこの方式は、瞬時的なシステム状態には独立である。他の1つはアダプティブな方式であり、瞬時的な負荷の変動に速応して適宜入流量に対する分割割合を変えて行くものである。両者はちょうどパケット交換網における固定ルーティング方式とアダプティブルーティング方式とにそれぞれ対応する。もちろんシステムの実際の稼動状況下において、瞬時的なトラヒックの変動を吸収するにはアダプティブな方式によらなければならないが、固定方式を採用した場合、ネットワークにおける定常的なフローパターンをあらかじめ知ることができる。即ちどの種のどれだけの量のジョブがどこで処理されるかを知り得るため、その種のジョブを処理するのに必要なデータ等の設置が自ら決定される。従って分散型データベースを保有する計算機ネットワーク等の設計に際し、この固定方式は有用な情報を提供するものと考えられる。固定方式に関しては P.V. McGregor and R.R. Boorstyn が文献[1]で、アダプティブな方式については J.T. Fitzgerald[2]が扱っている。[1]における方式はN個のノードより成るネットワークにおいて、ネットワーク全体での重み付け待ち時間の平均 W ($W=(1/\lambda)\sum_{i=1}^{N}\lambda_i W_i$, W_i：ノードiに入流したジョブがシステムに到着してから系外へ去るまでの時間の平均値, λ_i：ノードiへのジョブの到着率)を最小とするアルゴリズムを与えている。この評価基準においては、各ノードでの待時間を到着レイトで重み付けしているためトラヒックの少ないノードより投入(Submit)されたジョブは時として多大の遅延を受けることになりかねない。

そこで本稿では、いずれのノードへサブミットされたジョブも、ほぼ同様のターンアラウンド時間(TAT)を経験するような負荷分担法について考える。即ち、

各ノードにおける平均TATの差の最大を最小とするジョブフローパターンを求めそれに基いて入流負荷の分割を行うものである。

2. モデル及び定式化

ネットワークはN個のノード $V=\{V_i\}(i=1,2,\cdots,N)$ と、M個のブランチ $B=\{b_j\}(j=1,\cdots,M)$ から成るグラフGで表わされている。ノード V_i は処理装置（コンピュータ）K_i とネットワークコーディネイター NC_i とから構成されている。系外からノード V_i へ到着したジョブは、まず NC_i に入り NC_i より処理点の指定を受ける。従ってあるものは K_i で処理を受け、他のものは V_i 以外のノードへ転送され処理を受ける。他へ転送されたジョブは処理完了後再びノード i へ返送され系外へ出力される。ノード V_i へのジョブの到着はレイト λ_i のポアソン過程であるとし、$\lambda=\sum_{i=1}^{N}\lambda_i$ とする。ジョブは平均 l_p（処理単位）の指数分布に従う処理量を有し、その処理前、処理後の長さはそれぞれ平均 l_b, l_b'（ビット）のともに指数分布に従うものとする。問題の定式化に当りさらに次の諸量を定める。

k_i：ノード V_i のコンピュータ K_i の単位時間当りの処理能力
C_{ij}：ノード V_i, V_j 間の回線の伝送速度（bps）
$f(i,j)$：ノード V_i から V_j へのジョブフロー量（到着率）

ここで解析を簡単にするため V_i において入出力のジョブを蓄積しておくバッファ容量には制限がないものとし、ノード間を転送されるジョブの長さに関して、先に Kleinrock が述べている独立仮定はやはり成立しているものとする。このようにすると、K_i での処理及び伝送回線でのジョブの転送過程はいずれも M/M/1 Queue のモデルと見なすことができる。

ノード V_i に来たユーザーはジョブをサブミットしてからその応答を得るまである時間待たなければならない。その時間をターンアラウンド時間（TAT）と呼ぶと、これはジョブの処理時間、処理待ち時間から成り、他のノードへ転送される場合はさらに往復の伝送時間および伝送待ち時間が加わることになる。各ノードへのジョブの到着率及びコンピュータの処理能力が異なるとき、適切な負荷分担を行わなければ、当然ノード間にTATの差が生じる。そこで本稿では、ノード間におけるTATの差をできるだけ小さくする様な負荷分担について考える。即ちノード間のTATの差の最大を最小にするものである。

平均到着率 α、平均処理率 β の M/M/1 Queue における平均システム時間 W_s は、$W_s=1/(\beta-\alpha)$ と簡単な式で表わし得る。この式を用いると V_i へサブミットされたジョブの平均TAT W_i は次の様になる。

$$W_i=\left\{1-\frac{\sum_{j=1}^{N}f(i,j)}{\lambda_i}\right\}\frac{1}{\sigma_i-\left[\lambda_i+\sum_{j=1}^{N}\{f(j,i)-f(i,j)\}\right]}+\sum_{j=1}^{N}\frac{f(i,j)}{\lambda_i}\left[\frac{2}{u(i,j)-f(i,j)}\right.$$

$$\left.+\frac{1}{\sigma_j-\left[\lambda_j+\sum_{k=1}^{N}\{f(k,j)-f(j,k)\}\right]}\right] \qquad (1)$$

ただし　$\theta_i \equiv k_i/\ell_p$,　$\mu_{ij} \equiv C_{ij}/\ell_b \ (\equiv C_{ji}/\ell_b')$

(1)式オ1項は、ノード V_i へサブミットされたジョブのうち V_i で処理されるもののTATの平均で、オ2項は V_i 以外のノードへ転送されそこで処理を受け再び V_i へ返送されてくるジョブの平均TATである。従って本稿で考えている負荷分担の問題は次の様に表現できる。

「制約条件

$$0 \le f(i,j) < \mu(i,j) \quad (i,j=1,2,\cdots,N) \tag{2}$$

$$0 \le \lambda_i + \sum_{j=1}^{N}\{f(j,i) - f(i,j)\} < \theta_i \quad (i,j=1,2,\cdots,N) \tag{3}$$

のもとで　評価関数

$$\text{minimize}\{\max_{i,j} |w_i - w_j|\} \tag{4}$$

を満す最適フローパターン $F^* = \{f(i,j)\}\ (i,j=1,2,\cdots,N)$ を見い出すことである。」

一方ユークリッドノルムの定義より $\|X\|_\infty = \max_i |x_i|$ であるが、ある値以上の p に対して

$$\|X\|_p = \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}} \simeq = \max_i |x_i| \tag{5}$$

が成立するものと仮定して上の問題を次の様に変形する。

$$\text{minimize} \sum_{i,j=1}^{N} (w_i - w_j)^p \tag{6}$$

$$\text{subject to} \quad -f(i,j) \le 0 \tag{7}$$

$$f(i,j) - \mu(i,j) < 0 \tag{8}$$

$$-\lambda_i - \sum_{j=1}^{N}\{f(j,i) - f(i,j)\} \le 0 \tag{9}$$

$$-\lambda_i + \sum_{j=1}^{N}\{f(j,i) - f(i,j) - \theta_i \le 0 \tag{10}$$

制約条件式(2)または(7),(8)は、$i \to j$ への回線を流れるジョブフローの平均が回線の容量を越えないようにするためのもので、条件(3)または(9)(10)はノード V_i で処理されなければならないジョブの平均到着率を K_i の処理率以下に保つためのものであり、いずれもネットワークを定常状態下に置くための条件である。ここで制約条件式の数は $2N^2$ で、変数の数は高々 $N(N-1)$ である。本稿ではこの非線型問題を解くに当り　不等号制約条件式を含む場合に有効な方法として知られるHestenes [3] あるいは Powell [4] の乗数法 (multiplier method) を用いる。

3. 数値例

実際の数値計算において(6)式のφの値はφ=20とし、計算の実行はFACOM S.S.L.のDAVIDDを用い京大大型計算機センタFACOM M-190において行った。

まずはじめに、図2に示す様な2つのノードの場合において $\lambda_1=0.99$, $\lambda_2=1.0$ $\delta_1=1.0$, $\delta_2=2.0$ とした時を考える。この場合 V_1 における K_1 の利用率は0.99, V_2 におけるそれは0.5と極端にアンバランスな状態にある。即ち負荷分担をしなければ V_1 におけるTATは $W_1=100.0$, V_2 においては $W_2=1.0$ となる。このモデルに対し、種々の回線容量($\mu=2.0\sim100.0$)に対して負荷分担を行った場合の W_1 及び W_2 が表1に示してある。これによるとおよそ V_1 へ到着するジョブの1/2を V_2 へ回わして処理することにより W_1 は以前の1/50と改善され、W_2 は以前のわずか2倍程度に留まっている。図3は回線容量の増加に対する V_1 と V_2 におけるTATの差(W_1-W_2)を示している。これによると回線容量のわずかの増大で W_1 と W_2 の差を急激に減少させ得ることが判る。

次に図4に示す様な $N=3$ の場合についてみる。図6は $\lambda_1=0.99$, $\lambda_2=1.0$, $\lambda_3=1.0$ $\delta_1=1.0$, $\delta_2=2.0$, $\delta_3=2.0$, $\mu_{ij}=100.0 (i,j=1,2,3)$ としたとき負荷分担を行わないときと、最適な分担を行ったときの W_i ($i=1,2,3$) を示している。その時の最適フローパターンは図5が示すように $f(1,2)=0.45$, $f(1,3)=0.45$ である。

4. むすび

定常的な入力トラヒック特性をもとに負荷分担を行う方法について述べたがこれによって各ノードで処理されなければならないトラヒック量についての情報が得られるため、それらを処理するに必要なプログラム、データ等の設置条件も同時に得られる。瞬時的なトラヒック変動に対処するにはアダプティブな方法によらなければならないため、実システムにおいては固定方式で得られたデータをベースとしてアダプティブな方式を採用するのが効果的と考えられる。しかしながら、このアダプティブな方式については、筆者が知る限りにおいては文献[2]以外あまり報告されていない。むろんパケット交換網におけるルーティングの問題に関連するものについては種々のアダプティブな方式が提唱されてはいるがこれらを直ちにここで述べた様な負荷分担の問題するには多少難があるように思われる。従って今後このアダプティブな方式に関する考察が望まれる。

〔参考文献〕

[1] P.V. McGregor and R.R. Boorstyn, "Optimal Load Sharing in a Computer Network", Conference Record of the International Conference on Communications (ICC-76), 1976.

[2] J.T. Fitzgerald, "Load Regulation and Dispatching in a Network of Computers." Master Thesis of Dept. of Computer Science, Univ. of Illinois, 1972.

[3] M.R. Hestenes, "Multiplier and Gradient Methods," Journal of Optimization Theory and Applications, Vol. 4, No. 5, pp 303-320 (1969)

[4] M.J.D. Powell, "A Method for Nonlinear Constraints in Minimization Problems," in "Optimization" edited by R. Fletcher, Academic Press, NY, pp. 283~298 (1969)

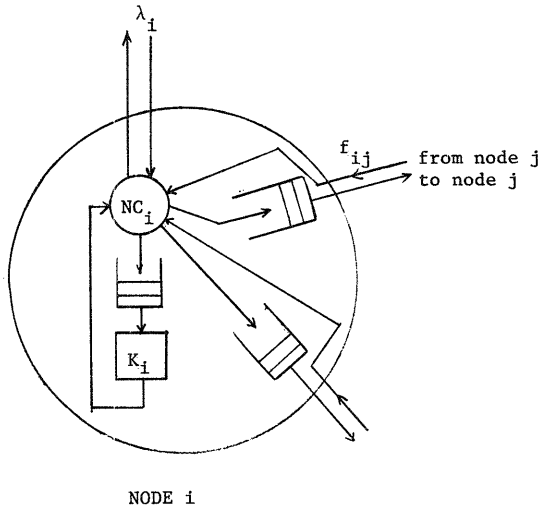

図1. ノードモデル

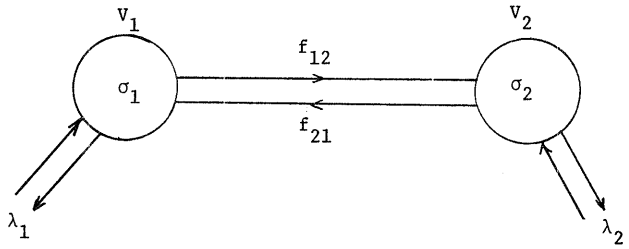

図2. 2ノードネットワーク

表1

チャネル容量に対するW1, W2およびフローレイト

($\mu=\mu_{12}=\mu_{21}$)

μ	2.0	4.0	7.0	10.0	30.0	50.0	80.0	100.0
W_1	3.8	2.3	2.2	2.0	2.0	2.0	2.0	2.0
W_2	2.9	2.4	2.2	2.0	2.0	2.0	2.0	1.9
f/λ_1	0.65	0.59	0.57	0.55	0.52	0.51	0.50	0.48

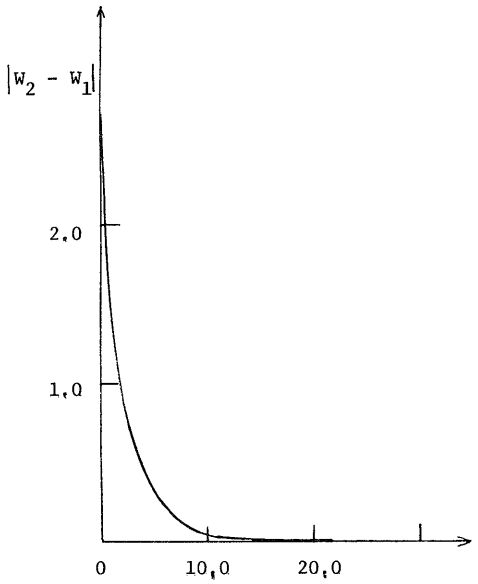

図3　チャネル容量に対する2ノード間におけるTATの差

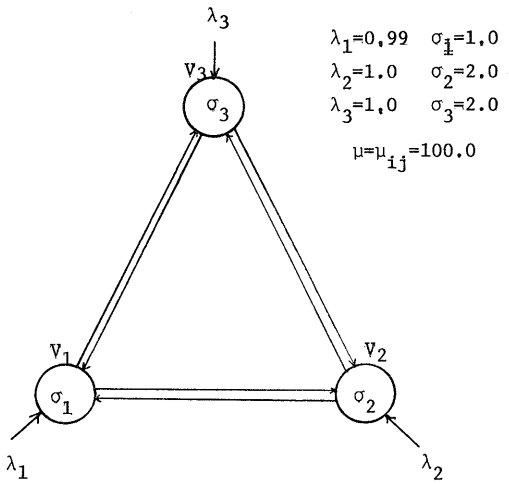

図4　3ノードネットワーク

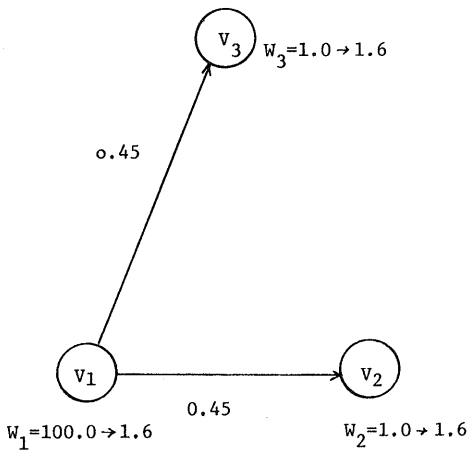

図5　最適フローパターン

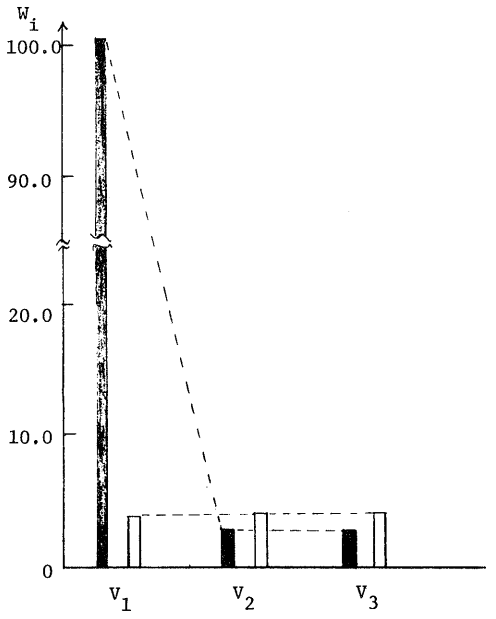

図6　各ノードにおける負荷分担前、後のTATの変化