負荷均一化に対するカクタス上の効率的なメッセージ伝搬

北陸先端科学技術大学院大学
知識科学研究科
林 幸雄 (yhayashi@jaist.ac.jp)

キーワード ：
拡散方程式, 2 次計画問題, 分散アルゴリズム, Bethe 近似.

—

1. 分散コンピューティングにおける基本課題

分散システムは並列計算機とは異なる特徴を持つ.

- ゆるやかな結合, 基本的に非同期な処理,
- 物理的, 論理的なプロセスの独立性 （負荷の分割性）,
- 異質性 （結合構造, サーバ性能, データ転送速度）.

メッセージ複雑度を最小にする負荷均一化

- 局所情報に基づく非集中制御,
- 無駄な通信や負荷移動を避ける （移動量の計算が先決）,
- 均一化ための通信や処理コストより, 実質負荷が大
  （遊んでるマシンを含む個別処理より効率化できること）.

代表的な負荷均一化法

C. Xu, and F.C.M. Lau, '97 など.

- Dimension Exchange Method (超並列計算機向き)
- Diffusion Method (非同期, マルチポート向き)

局所的な拡散によるフロー計算後, 負荷移動 (計算資源の配分)

⇒ 木上のメッセージ伝搬 + カクタス上のバイパスの活用 ！
最低限の連結性を保ちつつ，負荷移動量に応じてバイパス化

「できるだけ少ない通信回線で, できるだけ負荷移動量を抑えて, できるだけ早く/少ない手順で負荷を均一にする.」

2. グラフ上の拡散法による負荷均一化

多重辺や各頂点の自己ループがない無向グラフ (V, E) に対する, 辺の伝達効率などの重み付き離散 Laplacian を考える.

$$L\mathbf{f} = \begin{bmatrix} \cdots & \cdots & -w_e & \cdots & \cdots \\ \vdots & \ddots & 0 & \vdots & \vdots \\ -w_e & 0 & \sum w_e & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{pmatrix} f(1) \\ \vdots \\ f(u) \\ \vdots \\ f(m) \end{pmatrix},$$

$$Lf(u) = -\sum_{v \sim u} w_e(f(v) - f(u)).$$

拡散方程式 $\frac{\partial f}{\partial t} = -Lf$ の差分版, サーバの負荷量 $f(u)$ の均一化,

$$\mathbf{f}^k = (I - \Delta t L)\mathbf{f}^{k-1} = \underbrace{F \times \ldots \times F}_{k}\,\mathbf{f}^0.$$

従来の多項式スキーム R. Diekmann et al., '99

$p_0(t) = 1,\ p_1(t) = \frac{1}{\gamma_1}[(\alpha_1 - t)p_0(t)],$

$$p_k(t) = \frac{1}{\gamma_k}\left[(\alpha_k - t)p_{k-1}(t) - \beta_k p_{k-2}(t)\right], \quad (k \geq 2),$$

$$\mathbf{f}^k = p_k(F)\mathbf{f}^0 = \frac{1}{\gamma_k}\left[\alpha_k \mathbf{f}^{k-1} - F\mathbf{f}^{k-1} - \beta_k \mathbf{f}^{k-2}\right].$$

FOS, SOS, Chebyshev スキーム, OPS, OPT（有限収束）など.

⇒ ネットワーク上の最近接頂点との局所処理で実行可能だが,

- L の固有値の計算が予め必要,
- 例え, フロー量が求まっても, 負荷移動の順番をどうするか？

が問題（並列：○, 分散：×）.

3. 分散システム上の負荷移動

グラフ上の拡散方程式の差分版: $\mathbf{f}^k = (I - \Delta t L)\mathbf{f}^{k-1}$

$\Downarrow$

2 次計画問題と等価 （反復計算が必要） Y.F. Hu, and R.J. Blake, '99

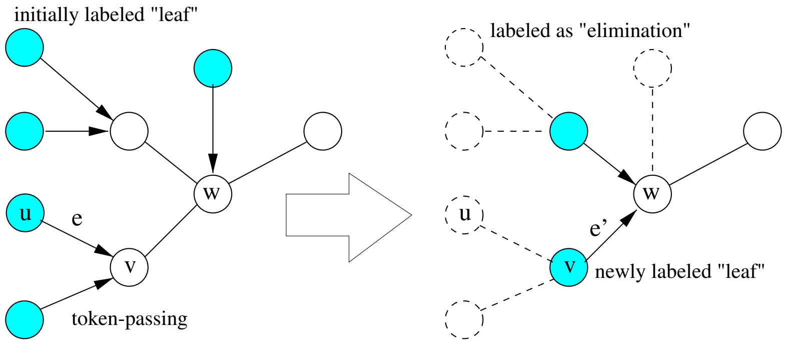
$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T W^{-1}\mathbf{z}, \\ s.t. & B\mathbf{z} = \mathbf{f}^0 - \bar{\mathbf{f}}, \end{array}$$

B: 接続行列, z_e: 負荷移動フロー, $\mathbf{f}^0$: 初期負荷配分, $\bar{\mathbf{f}}$: 均衡解.

$\surd$ 均衡は可解でしかなく, 無駄な巡回フローを含む.

この問題を逆手にとって, 巡回路を持たない木構造を考える ！
（局所既知 ：サーバ ID, 隣接頂点と辺, 葉や間接点）

最小全域木上の効率的なメッセージ伝搬



1. 葉から根への累積による, 総負荷量$\sum_{u\in V} f(u)$の計算.
2. 根から葉への, 均一解$\bar{f}$のブロードキャスト.
3. 葉から根への直接フロー計算 :$z_{e'} = f(v) - \bar{f} + \sum_e z_e$.
4. 葉から根への, 非同期な負荷移動 (または負荷要求) .

⇒ 根は動的に決定, 計算量や通信量も他頂点と同程度 !

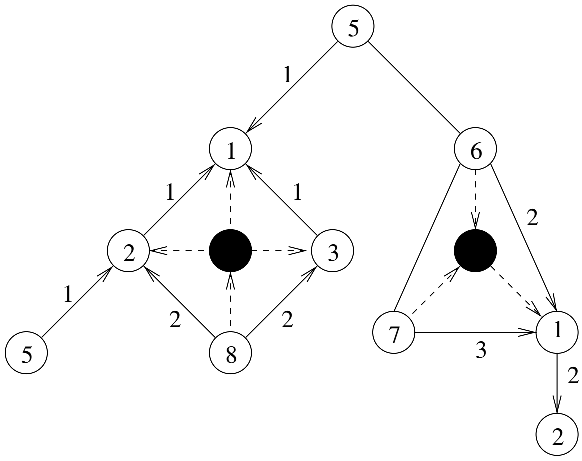
4. クラスタ化された木の拡張: カクタスの活用

カクタス固定 ： ghost を考えて仮想木上のフローを割り当て.
フロー量に適応的でない, 葉から根への伝搬の待時間必要.

カクタス生成 ： いくつかの基準
- フロー量に応じたバイパス選択, $|z_e|^2/w_e$大を優先.
- 辺の伝達効率 w_eに基づく (辺の向きを除いて同上).
- バイパス数を出来るだけ多くする ？
  カクタス制約による組合せ問題, 適応的でない.

⇒ 一般のグラフでは閉路が非独立で互いに影響し合い, バイパス量が局所的に求まらない (結局, 2 次計画問題に逆戻り).

固定したカクタスの場合



独立なクラスタを持つカクタス上の効率的なメッセージ伝搬

⇒ 統計物理（高次相関），誤り訂正符号, 因果関係の確率推定：
Loopy Belief Propagation 等における近似計算と共通の考え方.

バイパス化によるボトルネック辺の改善

2 次計画問題に再度着目して，

変分/摂動的な総通信コストの削減:　$\min \ \frac{1}{2}\mathbf{z}^T W^{-1}\mathbf{z}$

バイパス化なので負荷均衡は不変:　$s.t. \ \ Bz = \mathbf{f}^0 - \bar{\mathbf{f}}$

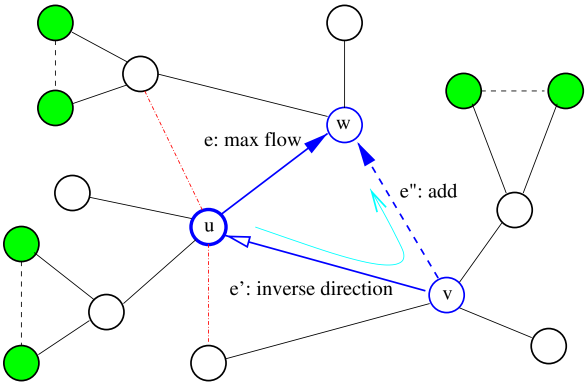
$\Downarrow$ 基本閉路/e''ごと独立な計算

頂点 u に接続する最大辺コストとその逆向きフローに対して,

$$\delta C(\Delta z) \stackrel{\mathrm{def}}{=} \frac{(z_e - \Delta z)^2}{w_e} + \frac{(z_{e'} - \Delta z)^2}{w_{e'}} + \frac{\Delta z^2}{w_{e''}} - \left(\frac{z_e^2}{w_e} + \frac{z_{e'}^2}{w_{e'}} \right),$$

最大コスト削減: $\frac{\partial(\delta C)}{\partial(\Delta z)} = 0, \ \Delta z_{opt} = \frac{w_{e'} w_{e''} z_e + w_e w_{e''} z_{e'}}{w_{e'} w_{e''} + w_e w_{e''} + w_e w_{e'}} > 0$ の時,
$\delta C(\Delta z_{opt}) < 0$. 結局, フロー Variance が最小化 (流量均衡化).

負荷均一化の分散アルゴリズム



1. 最小全域木の生成
2. 木上のフロー計算
3. 各頂点における非同期なバイパスの決定: Greedy
4. カクタス　(木とバイパス) 辺上の非同期な負荷移動

Ternary(三角閉路) の妥当性

バイパス決定の際

- 調整相手の数少 (影響範囲が限定), 最近接との直接通信,
- 交互の組合せに限定 (最適への全順序性や非再現性はない),
- 単なる仲介転送のみの長いパスを回避,
- 地理的にも近い相手の可能性大.

⇒ 最適 (困難) とは限らないが, 実際的に妥当なトポロジー !

閉路が独立でΔz_{opt}が一意↔ カクタス性による辺の相互排除,
完全グラフに近いほどフローは分散するが通信回線数が多,
拡散伝搬の速度は L の固有値/ネット構造に依存.

5. まとめ

分散ネットワーク上のサーバの負荷均一化問題を考え, 重み付き離散 Laplacian による拡散法に対して,

1. 負荷の均衡のみならず, 無駄な巡回フローを抑えることが本質的に重要であることを $\boxed{\text{等価な 2 次計画問題}}$ から指摘.
2. 巡回路を持たない最小全域木に着目 (反復処理が不要) : メッセージ伝搬による高速計算と, 非同期な負荷移動.

   $\boxed{\text{カクタス上のバイパス化}}$ ⇓ ボトルネック辺の改善

3. フロー量に応じて変分/摂動的に総通信コストを最小化する, 非同期, 分散アルゴリズムを提案.