

## テクニカルノート

# 余剰計算資源共有を行うための 通信遅延を考慮したネットワークの構築

中村 貴<sup>†1</sup> 菅谷 至寛<sup>†1</sup> 大町 真一郎<sup>†1</sup>

本研究はネットワーク上に存在する遊休計算資源の共有を行うための P2P ネットワークの構築を目的としている。ネットワーク上で資源共有を行うため接続する計算機間の通信遅延情報を P2P 通信により保持・利用を行う。通信遅延の情報を保持する手法として、各端末間の通信遅延を座標系におけるユークリッド距離として扱うことを可能とする Vivaldi を用いる。しかしながら、Vivaldi そのものでは本研究の目的のひとつである極近距離での詳細な遅延情報を表現することが困難であるため、目的に沿うような拡張を行う。

## Constructing Network in light of Communication delay for Sharing Idle Computational Resources

TAKASHI NAKAMURA,<sup>†1</sup> YOSHIHIRO SUGAYA<sup>†1</sup>  
and SHINICHIROU OMACHI<sup>†1</sup>

Our study aims at constructing network in light of communication delay for sharing idle computational resources. For maintaining information of communication delay, we use Vivaldi which enables us to deal with the delay as Euclidian distance in coordinate system. However, it is difficult to express the delay information of very near computers by using Vivaldi as is. Therefore, we will expand Vivaldi.

<sup>†1</sup> 東北大学大学院工学研究科

Graduate School of Engineering, Tohoku University

### 1. はじめに

日本における一般世帯の Personal Computer(以下 PC) 普及率は内閣府消費動向調査<sup>1)</sup>によると平成 21 年 3 月時点 73.2%と 70%を越え、大部分の家庭で PC が利用されているといっても過言ではない状況になっている。またこれらの PC の性能の向上も目覚しく、最新の PC では数年前のスーパーコンピュータを凌駕する性能を持つものも存在する。しかしながらブラウザソフトなどの利用を行っている際に PC が計算処理を行っている時間はごくわずかで、『遊んでいる』PC が多いのが現状である。一方、PC だけでなくネットワークの高速化・広帯域化も著しく、数年前には ADSL が主流であったのが最近では光回線 (FTTH) へと多くの利用者が移行している。以上の 2 点から、ネットワーク上には多くの遊休資源が存在し、有効に利用する手法を考案することは非常に魅力的である。

遊休資源の利用を行っている先例として、ボランティアコンピューティング (以下 VC) と呼ばれるものが存在する。VC は資源の利用を望む者が自身の資源利用目的を公開し、その理念などへの賛同者がボランティアとして資源を無償で提供し、提案者が一方的にタスクの送信を行うことで大規模な並列分散環境を実現するものである。例として、宇宙から受信した信号を解析し地球外生命体の存在を見出そうとする SETI@home<sup>2)</sup>、タンパク質の折り畳みや折り畳みに関する疾病を解明する Folding@home<sup>3)</sup> などが挙げられるが、資源の提供者は提供者、利用者は利用者として厳密に区別される。

そこで、参加者の誰もが資源の提供・利用ができ、一般家庭に存在する PC の計算資源を参加者の誰もが共有できるシステムを考える。適切に並列分散処理を行うシステムの構築を行うことができれば、手軽かつ安価に強力な計算環境を実現することが可能である。想定される利用法としては、動画エンコードの高速化、空調を最適化するための室内の気流のシミュレーションなどが挙げられ、将来的には一般家庭においても多大な計算能力に対する一時的な需要があると考えられる。

そのような要求に応えるための一つとして、本研究では一般的に使われている PC の遊休資源の共有を行い、参加者の誰もが高い計算能力を獲得することを可能とするシステムの構築を目的としている。

参加者の同等性、システムの耐故障性などを考慮し、システムの構築は P2P ネットワークの構築によって行う。参加者が自身のタスクを他者の PC に送信し結果を受けとることでシステムが成り立つことを考えると、通信に要する時間や各 PC の計算性能がシステムの性能に与える影響は大きい。したがって、P2P ネットワークは以下の要件を満たす必要がある

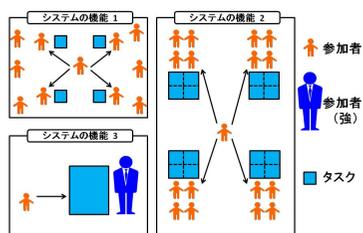


図1 システムの利用概念図  
Fig.1 Utilization model of the system

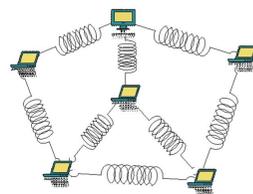


図2 Vivaldi モデル図  
Fig.2 Model of Vivaldi

と考えられる。

- (1) ネットワークにおいて近いPCを検索可能
- (2) ネットワーク内のまとまったPCの集団を検索可能
- (3) ネットワーク内の高性能PCを検索可能

以上のシステムの利用概念を図1に示す。強力な計算資源を保有する参加者を参加者(強)とし、タスクは並列計算のために分割された処理内容を表す。システムの機能2においてはタスクの送信者が送ったものをさらに分割可能であるが、通信を必要とする並列計算などが該当する。

また巨大な計算資源を獲得しようとする、数多くのシステムへの参加者が必要となる。多くの参加者を集めようと考え、必然と参加者は地理的に散在し、このことは一般的にネットワーク上でも参加者が散在することにつながる。さらに一般家庭に存在する遊休資源を利用することを考えると、参加者が自身のPCの利用を行う際には資源の提供は期待できず、さらに参加者ごとのPCの計算性能もばらばらだと考えられる。したがって参加者は以下の性質を持つと想定される。

- 地理的に散在
- 保有するPCの能力は千差万別
- 永久的な参加は保証しない(突然参加・離脱する)

次章では各参加者間の通信遅延情報を保持する手法である Vivaldi の概要を述べ、本研究の目的に即した拡張を提案する。

## 2. ネットワーク座標系

### 2.1 概要

提案するシステムはネットワーク上に並列分散処理環境を構築する。よって参加者間の通信遅延は処理時間に少なからず影響をあたえるため、システムへの参加者間の通信遅延情報を効率的に保持する必要がある。本研究では参加者同士の通信遅延をユークリッド距離として表現可能な座標系を作成することにより、情報の保持・共有を行うことを提案する。

既存手法<sup>5)</sup>として、実際に通信遅延を計測しながらネットワーク上で近い参加者間でクラスタを形成する手法が存在する。この手法にはクラスタの維持・更新に要する通信の量が多いことや、ネットワーク上で近い参加者が別々のクラスタに振り分けられる可能性があるといった問題が存在する。これらの問題は座標系により通信遅延の管理を行うことで解消されることが期待できる。

本システムの座標系が完全に端末間の通信遅延を表現可能であると仮定すると、各参加者の座標データベースを P2P ネットワークによって分散管理することにより通信遅延の情報管理が可能である。座標系により管理を行う利点としてはシステムの参加者が全体における位置関係を把握できることや、座標によって簡単に通信遅延の判断が可能であるので、参加者間の通信遅延を厳密にデータベースとして保持していなくても簡単に推定できることが挙げられる。また各参加者間の通信遅延がユークリッド距離で表現されることにより、直接通信を行わなくても通信遅延の推定を行うことが可能であり、このことはネットワーク上のトラフィックの軽減につながるという副次的な利点も存在する。

上記は座標系が正確であることを前提としているが、実際にはネットワークへの参加・離脱が発生することや通信の回線性能差、ネットワークの形状などから生じるゆがみによって正確な座標系を構築することは難しい。当然ではあるが、座標系の次元数を増やせば増やすほど精密な座標系を形成することが可能である。しかし、前提条件であるシステムの利用者の参加・離脱が容易に起こることや、高次元の座標系を有効に利用するためには参加者間の通信遅延の計測が大量に必要であることを考えると、低次元の座標系においていかに座標系を精密に形成するかを考える必要がある。そこで逐次的に座標を修正することで座標系を正確に近づける手法である Vivaldi<sup>4)</sup>を用いる。しかしながら Vivaldi には本研究の目的とするシステムが求める性能が足りない部分もあるので、より適合するよう拡張を行う。

### 2.2 Vivaldi

Vivaldi はネットワークに存在する参加者間の通信遅延を座標系に表現する手法で、参加

者間を物理的なばねで結ぶモデルによって座標修正を行う。この手法の特徴はアプリケーションレベルの通信に便乗して参加者間の通信遅延を計測することにより、ネットワークを維持するための余情通信が発生しないことである。この計測によって実際の通信遅延と座標系のユークリッド距離に誤差が検知された際には座標の修正を行い誤差を小さくする。誤差はばねモデルにおける引力・斥力に対応する。修正を行う参加者が通信を行った相手以外にも通信経験を持つ場合にはその相手との間にもばねが存在し、力のつりあいがとれる位置に近付くように修正される。Vivaldi のモデル図を図 2 に示す。

Vivaldi が性能を発揮し、実通信遅延と座標系から得られるユークリッド距離の誤差を小さくするためにはネットワーク上で近傍の参加者との通信が多い必要があるとされているが、遠方の参加者ともある程度通信を行うことも必要とされている。本システムでは単純に近傍の参加者と通信を行うことが多いと考えられるが、通信遅延を問わず性能が良い参加者と通信を行うことや参加者群を検索して通信を行うことから Vivaldi が性能を発揮する条件に合致し、有効に働くことが期待される。

### 2.3 Vivaldi の問題点

提唱者によると、Vivaldi は参加者間の実際の通信遅延と座標系から得られるユークリッド距離の誤差を 10 %程度に抑えることが可能だとされている。実ネットワークの回線性能差などから生じる歪みを考慮すれば、十分正確な座標系が構築されていると考えられる。しかし、本研究が必要とする局所的な精密さや近傍ノードの精密な順位を表現する性能に関しては不十分だと考えられる。

座標系として 2 次元、もしくは高さの概念を導入した 2+h 次元の利用が考えられているが、Vivaldi がばねモデルであることから、ネットワーク上で近い参加者間が互いに遠ざけあうことで誤差が生じることや、実際のネットワークのトポロジーに多大な影響を受けることは容易に想像ができる。ルーター間の距離が比較的大きく、同一のルーターに接続を行っている参加者が少ないようなネットワークでは参加者のまとまりを区別しやすく、誤差も小さい上にネットワーク上で近い参加者も識別しやすいと考えられる。逆に同一のルーターに接続する参加者が多く、ルーター間の通信遅延が小さいようなネットワークでは実際の通信遅延が小さいため誤差が少ない可能性はあるが、座標系上では別々のルーターに接続を行っている参加者の座標が重なることも容易に起こりうる。2+h 次元の座標系を用いることによって別のルーターに接続を行っている参加者同士が同一座標になることが避けられる可能性もあるが、同一のルーターに接続を行っている参加者間の通信遅延を高さの 1 次元のみで表現することになり、極近傍の参加者間の通信遅延が不正確になるといった問題が存在する。

表 1 座標系から得られた上位 30%近傍の推定精度  
 Table 1 Accuracy of estimate top near 30%

circle	star	mix
30%	31%	30%

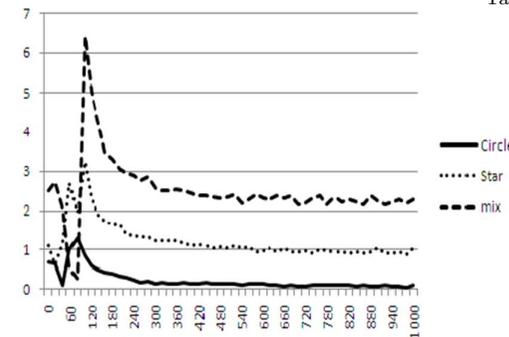


図 3 Vivaldi のネットワーク毎の誤差  
 縦軸：誤差平均/通信遅延平均  
 横軸：時間 [ms]

Fig. 3 Error in network.  
 vertical axis: error average/delay average  
 horizontal axis: time[ms]

1000 ノードによる circle, star, および二つの mix の三種類のネットワークにおいてシミュレーションを行った結果を図 3、表 1 に示す。2 次元を用いた Vivaldi の評価ではあるが、誤差の収束はネットワークによりまちまちである。また circle において誤差はほぼ 0 に収束しているものの推定精度の向上につながっていないことがわかる。このシミュレーションは mix を除いては局所的なネットワークを想定して行っているが、いずれも期待した性能を発揮できていない。

ネットワークの性能が向上するにつれ参加者間の通信遅延が小さくなることも考えられ、多少離れた参加者同士が互いの距離を誤認識することがシステムの問題とならない事態も想定される。しかし通信を行った参加者同士（あるいは通信経験があるもの同士）の通信遅延情報を利用して座標系を修正することを考えると、座標系におけるわずかな誤差がひとつの参加者の座標を大きく狂わす可能性もあるため、局所的な精密さはやはり必要である。

### 2.4 Vivaldi の拡張

ネットワークの通信遅延を座標系に表す際に障害となるのは、ルーターに接続している参加者群 (A) のルーターとの通信遅延を同程度とした際に、他ルーターに接続している

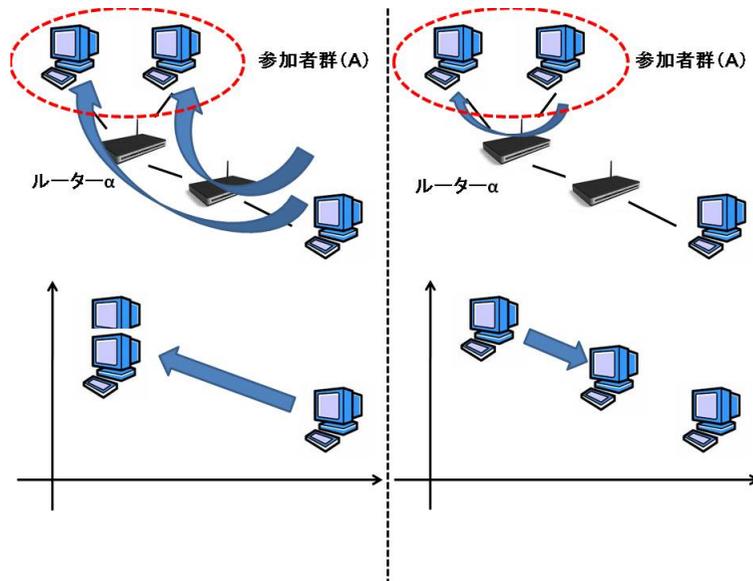


図 4 座標系に表現する際の問題点  
Fig. 4 The problem in expressing coordinate system

他参加者群から見た参加者群 (A) は同じ位置に存在するべきではあるものの、参加者群 (A) 内では確実に通信遅延が存在することにある (図 4)。同一群内で修正を行った際、他の参加者群との距離が変わらないような修正を行う場合は良いが、ばねをつないでいる参加者の位置関係によっては他の参加者群との位置関係を大幅に狂わせる修正を行ってしまう可能性がある。実運用ではネットワーク上で近い参加者同士が通信を行うことが多いことを考えると、参加者群 (A) 内での通信は頻繁に起こることが想定される。したがって近傍の参加者間の通信によって参加者群内の位置関係が正しく表現されてはいるが、近傍の参加者群の位置関係を表現するためある程度の面積が必要となるため、その参加者群が存在するエリアの外縁では他の参加者群が近くに存在する可能性がある。このことにより外縁部周辺で他の参加者群を近傍であると誤認識し、検索を精密にできなくなる事態は容易に起こり得る。

この問題を解決するために同一参加者群で別次元の座標系を形成することが考えられ、2次元座標系に高さの概念 (1次元) を追加したものが Vivaldi の提唱者によって提案された 2+h次元である。しかしながら、局所的なネットワークのトポロジーによっては、1次元で

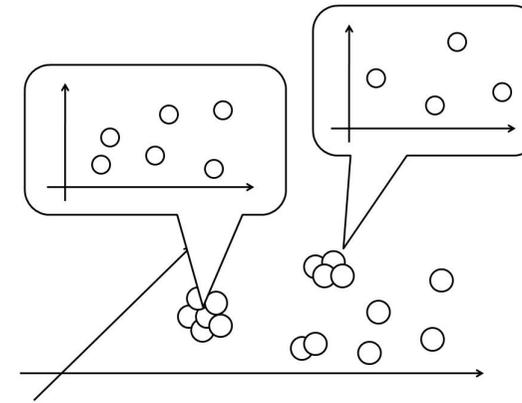


図 5 2+2次元の座標系概念図  
Fig. 5 The image of 2+2D coordinate system

精密な通信遅延を表現する座標系を構築するのは難しいと考えられる。

そこで 2次元の座標系によって全体の距離関係を保持しながら近傍参加者間で別途新しく 2次元の座標系を構築し、局所的に正確な座標系を構築することを提案する。この座標系においては、近傍の参加者群は元の座標系においては同一の座標に存在すると見なされ、他の参加者との通信遅延はこの座標とのユークリッド距離によって表現されるものとする。また近傍の参加者間の通信遅延情報は新たに形成した座標系におけるユークリッド距離として表現される。2+2次元の座標系のモデルを図 5 に示す。元となる座標系において密集していると考えられる部分において、その部分に存在する参加者のみで新しく座標系を構築することで、元の座標系では判別しづらい近傍の参加者との通信遅延が推定しやすくなることが期待できる。拡張を行わない Vivaldi で用いられる座標系を大局座標系、拡張により新しく作成される座標系を局所座標系と呼ぶこととする。

### 3. P2P ネットワークによる座標情報の管理

構築した座標系の情報は、前述のとおり P2P ネットワークによって保持・共有する。また、将来的なさらなる拡張として、新たに作成した 2次元座標系の中でさらに新しい座標系を再帰的に作成することも考えられる。システムの利用によりネットワーク上の局所的な範囲で通信が頻繁に起こることを考えれば、座標情報の問い合わせも狭い範囲内で収まること

が望ましい。以上の点から、座標系ごとに内部の参加者の座標を管理するスーパーノードが存在する木構造のネットワークが適していると考えられる。木構造のネットワークを利用して地理情報を管理する手法として Globase.KOM<sup>(6)</sup> が存在するが、この手法を参考にできると考えている。スーパーノードをある程度計算能力が大きなノードが担うとすれば、計算能力の大きなノードの検索を行う際にも有用に働くことが期待される。

#### 4. 実験

2+2次元に拡張した Vivaldi の性能をシミュレーションによって評価をした。本シミュレーションは、1000 ノードを 1 台の PC 上で仮想的に動作させることで行った。シミュレーションを行う際にネットワークの形状として circle, star, mix1, mix2 の 4 種類を用いた。circle は完全なサークル形であり、各通信路の通信遅延は 1~2 に設定し、ノード間の通信遅延は最小で 1、最大で 750 程度となるように設定を行った。star は 1 ノードを中心として残りの 999 ノードが接続を行う完全なスター形であり、中心ノードとの通信路の通信遅延を 20%のノードが 1~20、60%のノードが 21~50、残りの 20%のノードが 51~100 と設定し、ノード間の通信遅延は最小で 1、最大で 200 程度となるように設定をおこなった。mix1, mix2 はともにサークル形ネットワークを構成する各ノードを中心として、いくつかのノードがスター形に接続を行う形である。サークルにおける通信路の通信遅延は 10~20、スターにおける通信路の通信遅延は 1~5 に設定を行ったが、mix1 はサークルを構成するノードを 100 ノード、mix2 はサークルを構成するノードを 10 ノードとした。形成する座標系は 2 次元を用い、提案手法においてはさらにいくつかの 2 次元座標を用いた。

##### 4.1 シミュレーション手順

###### 4.1.1 座標の初期値設定

まず準備段階として 1000 ノードをランダムな順番で逐次的にネットワークに参加させた。一番最初のノードは座標系の中心、二番目の座標は一番目のノードとの通信遅延分垂直方向、もしくはは並行方向へ正負ランダム移動したものとした。三番目以降は既に参加済みのノードからランダムに 2 つを選び、そのノード間の通信遅延と座標系におけるユークリッド距離の誤差を最小とする座標とし、座標の各成分が 0 以下の場合は 0、最大値以上の場合は最大値となるように修正をした。

###### 4.1.2 Vivaldi による座標修正

Vivaldi による座標修正を行うにあたって、1000 ノードが通信を行うノードを決定し、そのノードとの通信遅延と座標系から得られるユークリッド距離の誤差を 0 に修正するとい

う動作を行った。つまり物理モデルに例えるなら、ノードをつなぐばねは修正を行うと即消滅するもので、修正の際考慮するのは一本のばねについてのみである。通信ノードの選びかたは Vivaldi の性質を考慮し、以下のように行う。

- (1) 各ノードが過去 5 回の通信遅延に関する履歴を保持する。
- (2) 他の 999 ノードからランダムに選ばれたノードとの通信遅延が履歴の平均を下回る場合に通信を行う。
- (3) 通信遅延が平均以上であった場合は、通信ノードを再びランダムに選び履歴平均との比較を行う。
- (4) 以上の手順を 10 回繰り返し通信候補との通信遅延が履歴平均を下回らなかった場合、遠方ノードとの通信の必要性も考慮し 10 回目のノードと通信を行う。

以上の手順で 1000 ノードが通信相手を探し自座標を修正する動作を、実際に設定した通信遅延と座標系から得られるユークリッド距離との誤差が収束するまで行った。

##### 4.2 局所座標系生成

提案手法においては 1000 ノードが座標の修正を行う動作をした後、さらに以下の動作を行った。

- (1) 各ノードについて、自身が所属する局所座標系が存在する場合、大局座標系において局所座標系が管理するエリアに所属するか判別を行い、エリア外に出ている場合は局所座標系から離脱する。
- (2) 局所座標系に所属していないノードからランダムにノードを選び出し自身を中心とした 20[ms] 四方に存在するノードが 20 以上ある場合自身がエリアの管理者としてそのエリア内のノードを対象に局所座標系を生成する。この動作を 100 回行う。

提案手法においてはノード間の通信遅延推定を以下の二つの状態に分けて行う

- 同一の局所座標系に所属しているノード同士:その座標系でのユークリッド距離
- その他:大局座標系でのユークリッド距離

##### 4.3 評価方法

本研究が目指すシステムが近傍ノードの探索に重点を置く検索を必要とするため、誤差が収束した際に各ノードが自身に近いノードを座標系から探索できる精度の評価を行った。具体的には、各ノードが座標系から上位 30%の近傍ノードを探索し、そのノードが設定した通信遅延から得られる上位 30%の近傍ノードとの合致精度を評価した。

##### 4.4 実験結果と考察

実験結果を図 6 に示す。

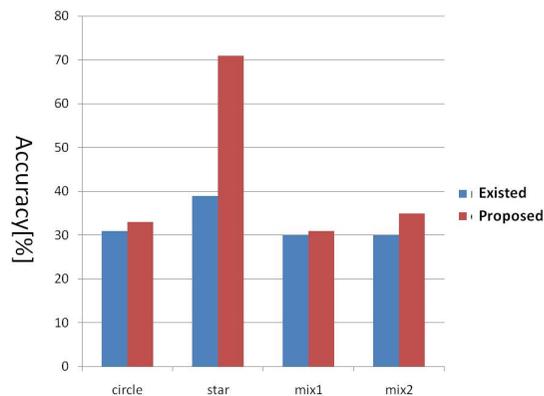


図 6 座標系から得られた上位 30%近傍の推定精度

Fig. 6 Accuracy of top-30% nearest neighbors estimated by coordinate system

star 形に対しては良好な結果が得られたものの、その他に対しては微小な精度の増加しか見られなかった。この原因としては、ネットワークに円環構造が加わることによる座標の不安定さが考えられる。すなわち、円環構造においては輪の右端と左端部分が上端と下端から見ると同程度の距離に位置すると思われるが、実際に右端と左端は最大に離れている部分であり簡単に遠方の座標にいるべきノードが近傍座標にいるべきだという誤認識が必然的に起こる。また円上のあるひとつのノードからその逆側へと時計回りにたどった際に形成されるべき座標系と逆時計回りにたどった際に形成されるべき座標系とで円の逆側において決定的な誤差が生じるといった問題も起こる。

推定精度が最もよかった star 形については、70%の精度と決して高い数値だとは言えないものの、近傍ノードとのばねの存在時間を長くすることや実際にネットワーク上で運用する際には近傍ノードとの通信が多く起こるといったことなどから精度の向上が期待できる。

mix1 と mix2 を比較した際に mix2 の方が若干よい結果が得られているが、新座標系を導入したことによる近傍ノードの推定精度の向上だと考えられる。しかしながら精度の向上は満足できるものではなく、さらなる精度向上のための手法が必要である。

## 5. 課題と改善策

今回得られた課題として以下の 2 点を挙げる。

- 円環構造のネットワークへの対処
- 局所座標系の構築場所や広さの考慮

前者には各ノードに対して複数の新しい座標系に所属を許可することで改善が期待できる。局所座標系は近傍のノードの距離を把握しやすくするために構築するものであり、ある程度座標系の上限を小さく設けることにより、その座標系内で決定的に近傍にはないノードを弾き出すことが可能であると考えられる。よって拡張された座標系内に存在するノードはそのノード同士で位置情報を保持しやすく、これは大局座標系で遠くに弾かれたとしても新座標系内に留まることができれば実際は近いノードだと判断でき、近傍ノードの探索に役に立つ。実際にネットワークを構築する際に座標情報を木構造のネットワークで保持することを考えると、高さのレベルがひとつ上のノードに対しては複数リンクを張ることで実現できる。

後者に関しては前者の導入によりさらに考慮が必要であり、すぐには答えが出せないのが現状である。実際のネットワーク構築を行う際には通信の集中が起こるのが新しい座標系を構築するノードであるのでいまいち条件が厳しくなることが予想される。ノードの密集度と推定精度の考察から密集度をパラメータとしてエリア構築を行う方法を現在考慮中である。また、ネットワークがある程度ツリー状である前提ではあるが、同一ルーターに接続を行ってネットワークに接続を行っているのであれば、ネットワークの構造推定により近傍ノードを探索し、局所座標系の構成ノードとする手法も考えられる。IP アドレスの利用により近傍ノードを推定するといった手法も考えられ、今後さらなる考察が必要である。

いずれの場合も情報をより精密に保とうと考えれば、各ノードが保持する情報量や通信量も増加し、実運用を考えた時には考慮が必要である。今後の課題としたい。

## 6. ま と め

一般家庭で強力な計算環境を実現する方法として遊休資源を共有する手法を目的とし、各参加者間の通信遅延情報を保持・共有する手段として座標系の導入し、座標系を構築する手法である Vivaldi の拡張手法を提案した。拡張手法として新たに局所的な座標系を導入することで、局所的に精密な通信遅延の推定を可能とする効果を得た。シミュレーションによって star 形に関して近傍ノードの推定精度の向上が得られたことを確認した。円環構造を持ったネットワークに対するさらなる考慮や、一つのノードが複数の局所座標に存在可能とすることによる改善、ネットワーク構造の推定や IP アドレスの利用による改善は今後の課題としたい。

## 参 考 文 献

- 1) 内閣府消費動向調査結果: <http://www.esri.cao.go.jp/jp/stat/shouhi/shouhi.html>
  - 2) SETI@home: <http://setiathome.berkeley.edu/>
  - 3) Folding@home: <http://folding.stanford.edu/Japanese/main>
  - 4) Frans Kaashoek, Frank Debek, Russ Cox and Robert Morris. Vivaldi: A decentralized network coordinate system *Proceeding of the ACM SIGCOMM '04 Conference*, pp.146-160, Portland, Oregon, August 2004.
  - 5) 氏家武志, 菅谷至寛, 阿曾弘具. 動的負荷分散システムのための自律的オーバーレイネットワーク. HPCS2008, Vol.2008, No,2, p.65, January 2008.
  - 6) Aleksandra Kovačević, Nicolas Liebau and Ralf Steinmetz. Globase.KOM - A P2P Overlay for Fully Retrievable Location-based Search. Seventh IEEE International Conference on Peer-toPeer Computing, pp.87-94.
-