WIDE Technical-Report in 2010

分散ハッシュテーブルの研究動向 (2008-2009)

wide-tr-ideon-dht-survey2010-00.pdf

WIDE Project : http://www.wide.ad.jp/

If you have any comments on WIDE documents, please contact to board@wide.ad.jp

Title: 分散ハッシュテーブルの研究動向 (2008-2009)
Author(s): 斉藤 賢爾
Date: 2010-12-04

分散ハッシュテーブルの研究動向 (2008-2009)

斉藤 賢爾 <ks91@sfc.wide.ad.jp>

1 概要

分散ハッシュテーブル (DHT: Distributed Hash Table) は、ネットワーク上に < キー , 値 > ペアを格納し、キーから値を検索するサービスを提供する分散データ構造およびアルゴリズムの一種である。DHT は、ノードの識別子 (IP アドレス等) とキーをハッシュ関数 (典型的には暗号学的ハッシュ関数) を用いて同一の固定長ビット空間に配置し、当該空間上でキーに近接するノードに < キー , 値 > ペアを格納するべく P2P オーバレイネットワークを構成する手法であると一般化できる。

この手法では、ノードとキーが空間上に均一に配置されると期待できることから、均質な負荷分散を実現可能である。また、冗長化やデータの輸送によりチャーン (churn; ノードの頻繁な出入り) に耐性を持てるように設計されるという特徴を持つ。ノード数 N の増加に対して、一般に検索性能を $O(\log N)$ にできることから、規模拡張性を持つ分散ストレージとしての応用が期待され、実際に数々の応用例がある。

第 1 世代の DHT としては、リング構造のオーバレイネットワークを利用した Chord[29], プラクストン・メッシュ (Plaxton Mesh) 構造 [24] を利用した Tapestry[36] および Pastry[27], 高次元トーラスを利用した CAN[25] (以上 2001 年), ハッシュ値間の XOR を距離の尺度とした Kademlia[19] (2002 年) 等がある。

これらの DHT は、特に Chord, Pastry といったものを中心に、その後の分散アルゴリズムの研究の基盤として用いられている。Kademlia は、BitTorrent[4] でトラッカー (インデクスサーバ) の代わりに用いられ、また、eMule[6] で採用されるといったように、特に実用上の応用例が多い。

その後に発表された DHT の例としては、バタフライグラフ (Butterfly Graph) に基づく Viceroy[16] (2002 年) や de Bruijn グラフに基づく Koorde[9] (2003 年) 等がある。

本稿では、最近の DHT の研究動向について、特に 2008~2009 年に行われた研究を中心にまとめる。

本稿は次のように構成される。第 2節では、検索機能の向上に向けたアプローチについて述べる。第 3節では、検索性能の向上に向けたアプローチについて述べる。第 4節では、資源の効率的利用に向けたアプローチについて述べる。第 5節では、カウツグラフ (Kautz Graph) 等を用いた DHT の新しい世代について述べる。最後に第 6節で最近の DHT の研究動向について総括する。

2　検索機能の向上に向けたアプローチ

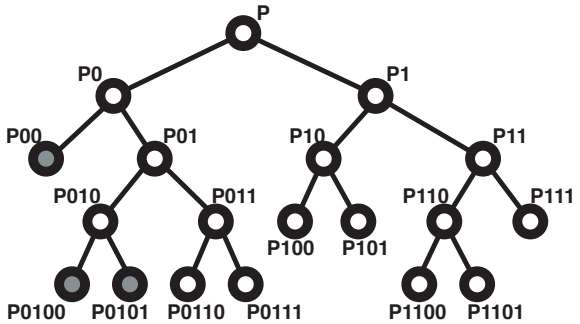
DHT ではハッシュ値を用いるため、一般に、格納されているデータの配置はキーの順序を反映していない。そのため、例えば地理上の特定の緯度経度内の地点といったように、ある範囲に含まれるキーを持つ値を検索するといった応用に向かないという難点がある。これに対し、ハッシュ関数を用いない分散化により範囲検索を可能にしている分散データ構造およびアルゴリズムとして、2003 年に発表された SkipGraph[1] がある。SkipGraph はスキップリストに基づき、可塑性を組み込むことにより、分散システムにおいて平衡木の機能を提供する。

この節では、これと異なり、DHT 上に範囲検索を実現する手法について解説する。

2.1　プレフィックスハッシュ木を用いた範囲検索

2005 年には、DHT を下位構造とし、下位の DHT に変更を加えずに、範囲検索を含む高度な検索機能を上位モジュールとして実現した例として、PlaceLab[5] にて使用された手法 [3] が発表された。PlaceLab は、無線通信基地局の ID を用いた測位サービスである。PlaceLab では、OpenDHT[26] (2005 年) 上にプレフィックスハッシュ木 (PHT: Prefix Hash Trees) と呼ばれる多元範囲検索のためのデータ構造を実現することによりサービスを実装した。PHT は、2 進符号化されたキーのトライ木である。範囲検索を行う場合、範囲の最小値と最大値に対する最大の共通プレフィクスを頂点とするサブツリーに含まれるノードを並列に検索することで、当該範囲に含まれるすべてのキーに対応する値を取得できる。この手法では、空間充填曲線 (この場合 Z 曲線) を利用し、多次元に分布するキーを 1 次元に配置することにより多元 (緯度および経度) の範囲検索に対応している。

図 1に、簡単な PHT の例を示す。PHT ノードは、特定のキーの範囲を代表し、そのラベルは、当該範囲に含まれるキーのプレフィックスとなっている。



PHT ノードのラベル	Z 曲線上のキー (2 進) の例
P00	000010
P0100	010000
	010011
P0101	010101
	010110
	010111

図 1: 簡単な PHT の例

2.2 カウツグラフを用いた範囲検索

2006 年には、DHT に基づく範囲検索手法として Armada[14] が提案された。Armada は多元の範囲検索をサポートし、N 個のノードが参加するオーバレイネットワークで $2\log_2 N$ 以内のホップ数で結果を返すことを保証する、遅延制限付き (delay-bounded) の手法である。Armada はカウツグラフに基づく DHT である FissionE[15](2005 年) 上で動作する。

カウツグラフは有向グラフであり、度数 d のカウツグラフは、各々 d 個の外向きリンクと d 個の内向きリンクを持つノードから成る。各ノードは、隣り合う数字がすべて異なる $d+1$ 進数の番号 (カウツ文字列) により区別され、その外向きのリンクは、自己の番号を左にシフトし、空いた桁を利用可能な数字で埋めた番号を持つノードに接続される。内向きのリンクに対してはその逆の計算を行う。例えば、度数が 3 のカウツグラフにおけるノード 132 は、外向きに番号 $320, 321, 323$ のノードと繋がり、内向きに番号 $013, 213, 313$ のノードから繋げられる。

このことにより、例えば、度数が 3 でカウツ文字列長が 6 のカウツグラフでは、972 個の全ノードの中のいかなる 2 個のノードも、6 ホップ以内で繋がることが保証され、また、3 つの完全に異なる (1 つとして同じ中継ノードを持たない) 経路を持つ。DHT における最悪ケースのホップ数を「直径」と呼ぶことがあるが、カウツグラフでは直径はカウツ文字列長に一致する。また、カウツグラフに基づく DHT では、度数は経路表のサイズを表す。

図 2に、度数 2、カウツ文字列長 3 の場合のカウツグラフの例を示す。カウツグラフ上

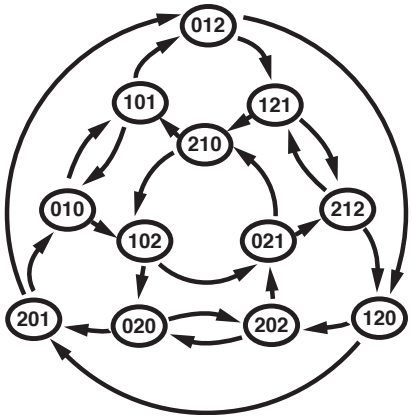


図 2: カウツグラフの例 (度数 2, カウツ文字列長 3)

の基本的なルーティングは、送信元のカウツ文字列を左にシフトしながら送信先のカウツ文字列に近づけることで行える。例えば図 2の例では、102 から 012 に向かう経路は、$102 \rightarrow 020 \rightarrow 201 \rightarrow 012$ となる。

FissionE ネットワークは度数 4、直径は $2\log_2 N$ であり、平均ホップ数は $\log_2 N$ 以下となる。

Armada では、実数の範囲を木構造に分割するハッシュアルゴリズムを用い、キーの順序を保存したままカウツ名前空間にマップする。その上で、 FissionE ノードから構成され、外向きのリンクに順序関係を意味づける FRT (Forward Routing Tree) を用いて、カウツ名前空間内の範囲を検索することができる。

2009 年に新たに提案された ERQ (Efficient scheme for delay-bounded Range Query)[32] は、カウツグラフに基づく DHT である DK(第 5節参照) の上で並列検索と刈り込みを行うタイプの遅延制限付き手法である。ERQ では、 DK の上で PHT をエミュレートする。ERQ はノード数 N と度数 d の下位 DHT において、$\log_d N(2\log_d \log_d N + 1)$ ホップ以内で検索を終了することを保証する。ERQ でも空間充填曲線 (Z 曲線) を利用し、多次元に分布するキーを 1 次元に配置する。

ERQ は、度数が 4 のとき、Armada より高性能であり、処理コストも低いことが示されている。

3　高速化に向けたアプローチ

DHT では、検索にあたり、オーバレイネットワーク上のホップを必要とする。複数のノードに跨る処理が行われること自体、オーバヘッドが高いことに加え、下位層のトポロジに無関係にオーバレイのトポロジが作られる場合、下位層での無駄な通信を引き起こしやすい。

この節では、経路表のサイズを大きくすることによってホップ数を減らす試みや、ノードの非均質性を考慮して DHT を階層化することにより性能の向上を図る試みについて解説する。

3.1　1 ホップ DHT

DHT は、チャーンに対する耐性や、規模拡張性を重視するために、基本的に経路表を小さく抑えるという発想で設計されている。経路表が小さいということは、ノードの新たな参加や離脱が起こった場合、テーブルを書き換えなければならないノードの数が低く抑えられていることを意味するからである。

しかし、DHT の設計における前提ともいえるこの条件を見直す動きも出ている。すなわち、隣接するノードからその隣接するノードへのポインタを取得し、経路表を成長させる「先読み」の手法がいくつか考案されている。

現在は、ハードウェアの進歩により、ノードのメモリや利用可能な帯域に対する制限が緩くなっており、数万ノードといった中規模の DHT では、実際に各ノードが他のほぼすべてのノードに対するポインタを維持することも現実的な意味を帯び始めている。

これは、クラウドコンピューティングのように、計算のための資源をデータセンタに集約でき、チャーンは起きないがノードの管理コストを低く抑えたい場合には、特に有用と考えられる手法である。

この 1 ホップ DHT の手法においては、一貫性のあるポインタ情報を如何に高速かつ低コストで全ノードに配信するかが設計上の鍵となっている。

2005 年に提案された D1HT[20] は EDRA (Event Detection and Dissemination/Reporting Algorithm) により DHT のメンバシップに関わるイベントを共有する。EDRA は、基本的にはスキップリストを用いたフラッディングである。

同じく 2005 年に提案された 1h-Calot[30] は同様にスキップリストを用いてメンバシップに関わるイベントの拡散的な配信を行う。1 ホップ DHT を実現する 1h-Calot では数千

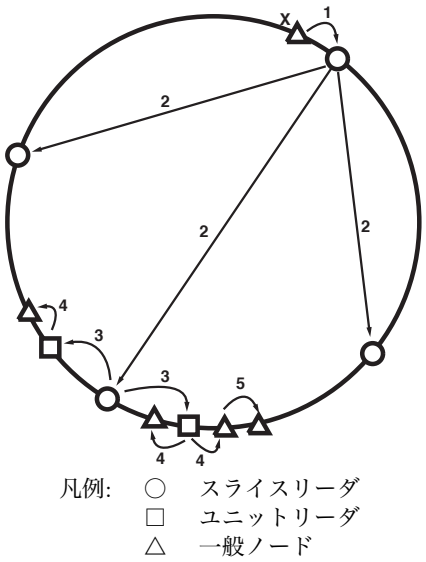


図 3: OneHop におけるメンバシップ情報の拡散手法

ノード、2 ホップ DHT を実現する 2h-Calot では数百万ノードを維持できるとされるが、高いチャーン耐性を目指して設計されてはいない。

OneHop[7](2004-2009 年) は、初めて 1 ホップ DHT の概念を実現したものであり、Chord のプロトコルを応用して先読みを行う手法である。OneHop では、各ノードにてメンバシップの完全な情報を維持し、当該情報が最新のものであれば 1 ホップにて目的のノードに到達し、そうでなければ少数のホップ数で到達する。最新の論文では、99%の検索において 1 ホップで目的のノードに到達できることが示されている。OneHop では、高速かつ低帯域でメンバシップ情報をシステム全体に拡散させるために、ハッシュ値の空間を k 個のスライスに分割している。各スライスは、その中間点の値に対応するノードをスライスリーダとして選出する。また、各スライスは更に u 個のユニットに分割され、それぞれの中間点の値に対応するノードをユニットリーダとして選出する。メンバシップの変更 (ノードの参加あるいは離脱) を検出したノードは、自己が属するユニットのスライスリーダにメッセージを送る (図 3)。スライスリーダは、単位時間内にスライス内で生じたメンバシップの変更の通知を集約し、他のスライスリーダに通知する。各スライスリーダは、単位時間待った後に、受け取った通知を集約してスライス内のユニットリーダに送信する。ユニットリーダは、受け取った情報を通常の Chord の保守メッセージに載せて近隣のノードに拡散させる。

OneHop は、現在、Chord の拡張ルーティングの選択肢のひとつとして実装・提供されており、Chord の応用として作られているアプリケーションは、未修正で OneHop の恩恵を受けることができる。

2009 年には、OneHop, D1HT, および 1h-Calot の性能の比較 [21] が行われている。1,000 万ノードまでに規模を拡大させた場合の 3 者の振る舞いの検証と、データセンタ環境における D1HT と 1h-Calot の比較を行った結果、著者ら (D1HT の開発者でもある) は、D1HT が最もオーバヘッドが低く、また、データセンタにおける高性能コンピューティン

グに最も向いていると結論づけている。

3.2 循環ルーティング

先読みは、DHT の安定した運用に向けた応用も可能である。

DHT を実用的に応用していく上では、NAT やファイヤウォール等の存在を前提とすると、IP での外部からの到達性がない環境を想定する必要がある。また、悪意のあるノードや、過負荷により事実上停止しているノード等も考慮する必要がある。

循環ルーティング (CR: Cyclic Routing)[13] (2009 年) は、そのような環境において一般的に利用できる、先読みを用いた既存の DHT の拡張ルーティング手法である。CR により拡張された Chord では、5~10%のノードが悪意を持つ場合、2 倍低い検索の失敗率を、40~50%のノードが悪意を持つ場合は、2 倍高い検索の成功率を観測できたとしている。

CR では、ノード s から d にメッセージを送り、ノード d から s に返信が返るまでの経路に含まれるノードのリストをサイクルと定義する。サイクルは、通常の探索メッセージにノードの情報を載せていくことにより取得できる。サイクルを知ることにより、s は経路表を先読みしてメッセージを送信できる。これにより、メッセージングの性能が向上する他、不都合なノードを迂回することが可能となる。

3.3 グループ化/階層化 DHT

実際にオーバレイネットワークを構成するノードは、一般に非均質である。CPU 性能、ストレージ資源、利用できる帯域等に余裕があり、また長時間ネットワークに参加しているノードと、そうでないノードが混在していることを考えると、前者の持つ資源を有効に活用するために DHT を階層化したいと考えることは自然な発想である。

そのような手法の先駆けとして、例えば Tapestry の研究グループでは、スーパーノードを設けるランドマークルーティングを行う Brocade[35] (2002 年) を提案した。

また、 2004 年に提案された Diminished Chord[11] は、Chord にグループ化を採り入れたものである。Diminished Chord のリングに参加するノードは、どのサブセットであっても、全体のリングにおけるルーティング機構を利用することで、独自のリングを形成せずにサービスを提供できる。独自のリングを形成する場合、サイズが k であるサブグループは $O(k \log k)$ のストレージ資源を消費するが、Diminished Chord では $O(k)$ しか消費しない。ただし、サブグループに属さないノードにも、当該サブグループに関する情報が格納される。

2008 年に提案された G-TAP (Grouped Tapestry)[31] は、Tapestry において、参加ノードの非均質性に基づいてオーバレイネットワークをグループに分割し、柔軟なルーティングを実現する手法である。従来の DHT におけるルーティングに加えて、グループ内のノードのみを経由し、グループ内のノードに到達する PC (Path-Constrained) ルーティング (Diminished Chord はこれを実現できない) と、最終的にグループ内のノードに到達することのみを保証する DS (Destination-Specified) ルーティングを利用できる。

これらの拡張ルーティングを利用することにより、性能の高いノードのみを用いた計算を行ったり、安定したノードのみによりサービスを提供したり、悪意のあるノードを排除した通信、またはグループ内にプライベートな通信が可能となる。

G-TAP では、DHT 内のそれぞれのグループに対し、サブ DHT のためのルーティング構造 (グループ用の経路表) と、グループの発見のためのグループメンバシップ木 (GMR tree: Group Membership Rendezvous) を備える。

G-TAP のグループ自体は階層構造を持たないが、G-TAP を階層構造向けに最適化した H-TAP も提案されている。H-TAP は、経路局所性 (path locality) および経路集約性 (path convergence) を実現する。経路局所性は、2 つのノードを繋ぐ経路が、双方を含む最小のドメインを出ないことを保証する。経路集約性は、あるキーに関わるメッセージに対して、ドメインを出る経路が 1 つのオーバレイルータを必ず通ることを保証する。

同じく 2008 年に提案された階層化 DHT によるマルチメディア配信サービス [18] は、IETF にて検討されている P2PSIP[2] に基づく手法である。これはスーパーノードを用いるものであり、スーパーノードのみが参加するオーバレイネットワークを形成して、下位の DHT を相互に接続する。ノードの識別には、プレフィックス ID とサフィックス ID から成る階層化 ID を用いる。性能に関しては、階層化された Kademlia によるシミュレーション評価が行われている。

2009 年に提案された GTPP (General Truncated Pyramid Peer-to-Peer)[22] は、トランケートされたピラミッド型アーキテクチャであり、複数段階の階層をサポートする DHT の階層化の例である。各階層は独自のオーバレイネットワークを形成し、各ノードは下位のネットワークに対するスーパーノードとして動作する。

4 資源共有の効率化に向けたアプローチ

高速化とも関連があるが、P2P はそもそも、ネットワークに接続されたコンピュータの余剰資源を効率的に利用する目的で発想されたものであり、下位層のトポロジを考慮した上で、CPU、ストレージ、ネットワーク帯域といった資源を効率的に利用できるように設計されることが望ましい。

この節では、資源の評価や近傍性を考慮した負荷分散の手法について解説する。

4.1 順位付けと評判

第 3節で示したようなグループ化/階層化 DHT を利用して、実際にサービスの質を安定させるためには、参加ノードやそれらが提供する資源を各自が評価でき、不適切なノードへの転送を避けたり、必要なレベルの資源を持つノードを要求先として採用できる必要がある。

このような評価は、DHT の分散性を考えると、特定の権威に依るのではなく、それ自体が分散化されたアルゴリズムで行える必要がある。

そのような分散化した評価システム、すなわち評判システムとしては、各々のノードによる評価を、そのノードの評判により重み付けし、再帰的に計算する、(Secure) EigenTrust[10] (2003 年) 等がある。

2006 年に発表された論文 [17] では、P2P システムにおける評判システムを分類し、詳細な要求分析を行っている。この論文では、評判システムの機能を表 1に示すように 3 つに分割する。論文では、関連用語を定義した上でこれらの機能の実現上の概念を整理し、

表 1: 評判システムの機能

情報収集	自己同一性の識別
	情報源
	情報の集約
	新規参入者に対するポリシー
採点と順序づけ	善い vs. 悪い振る舞い
	量 vs. 質
	時間に対する依存
	選択の閾値
	ピアの選択
応答	インセンティヴ
	罰

要求を明らかにしている。

2009 年に提案された局所的平衡モデル [12] は、ノードの持つ資源のランク付けの計算のための数学的モデルである。ノードが自分で持つべき資源は何か、他ノードに求めるべき資源は何で、どのノードを経由して取得すべきか、自ノードが他ノードに提供すべき資源や、自ノードを経由して他ノードに転送すべき資源・サービスの品質はどの程度であるべきか、といった判断が自動的に行えることを目的としている。

計算は反復的に行われ、反復的にノード間の資源共有に用いられる。このモデルでは、各ノードが融通する資源の価値のバランスが取れるような調整が行われる。

4.2 近傍性を考慮した負荷分散

ネットワーク全体の負荷を考えると、オーバレイネットワークで近接するノードが下位のネットワークでは離れていることにより、オーバレイのホップの度に下位層で大きなオーバヘッドがかかるような事態は避けたい。

下位ネットワークにおけるノードの近傍性を考慮した近接ノード選択 (neighbor selection) は、DHT のみならず、分散システム全体における大きな課題である。

2008 年に提案された近接クラスタリング [28] は、近隣のノードから成るクラスタを形成するものである。単にスーパーノードをルータとするのではなく、物理的に近傍なスーパーノードとのオーバレイネットワークを形成する p クラスタ (物理クラスタ) と、論理的に (ハッシュ値の近い) 近傍なスーパーノードとの接続を持つ v クラスタ (論理クラスタ) の両方を検討し、それぞれに適した応用を分析している。

同じく 2008 年に提案された P3ON (Proximity Based Peer-to-Peer Overlay Networks)[23] は、AS 番号とノードのハッシュ値を連結したものをノード ID とすることにより、AS 毎にノード ID が近接するようにした DHT である。トポロジとしては、 AS 毎のリングを持つ 2 段階の階層構造を持つ。

5　DHT の新しい世代

この節では、 DHT の新しい世代として、有向グラフ構造に基づき、一定の度数、すなわち経路表のサイズを持つ、定度数 DHT の動向について解説する。

5.1　定度数 DHT の動向

2008 年に、分散線グラフ (DLG: Distributed Line Graph) に基づいて任意の定度数を持つ DHT を生成する手法 [33] が提案された。この手法により生成される、DLG に基づき、N ノードが参加する、外向けの度数 d の DHT では、内向きの度数は $1 \sim 2d$ であり、直径が $2(\log_d N - \log_d N_0 + D_0 + 1)$ 未満であることが保証される。ここで D_0 と N_0 はそれぞれ初期ネットワークの直径およびノード数である。この手法において、ネットワークの維持コストは $O(\log_d N)$ となる。第 2節にて紹介した ERQ は、この手法を用いて作られた、カウツグラフに基づく DHT である DK 上の手法である。この手法はバタフライグラフ等、他のグラフにも適用できると考えられている。

更にこの手法を受け、2009 年には、任意の度数を持つ分散カウツグラフを用いた DHT である SKY[34] が提案された。SKY は、カウツグラフを用いる DHT としては初めて実用的なレベルを目指したものである。ハッシュアルゴリズムとしては、任意のキーを度数 2 のカウツ空間に均一にマップするカウツハッシュ (KautzHash) を、任意の度数に適用できるように拡張して用いている。

カウツグラフを実用的に用いる際の最大の問題点の 1 つは、度数 d とカウツ文字列長 D が決まると、カウツグラフの最大のノード数が $d^D + d^{D-1}$ に決まってしまう点にある。カウツ文字列長が固定であるシステムの場合、ノード数がこの値を超える際には、全ノードの番号を付け替えるといった非現実的な対応を迫られることになる。SKY では、カウツグラフを近似する分散カウツグラフを採用し、カウツ文字列長を可変にすることでこの問題を回避している。

2008 年に提案された BAKE[8] は、DLG を用いる手法とは異なる方法で生成された、均衡カウツ木 (balanced Kautz tree) を用いた DHT である。BAKE ネットワークの直径は $\log_d N$ である。著者らは、この手法は de Bruijn グラフ等にも適用可能としている。

6　まとめ

本稿では、DHT の最近の研究動向について、2008~2009 年に発表された論文を中心に、検索機能の向上、検索性能の向上、資源共有の効率化、およびそれらを踏まえた新世代の DHT の研究に着目して調査・整理した。

カウツグラフを用いた DHT は定度数と制限可能な直径を持ち (即ち最大ホップ数の制限を設けることが可能であり)、範囲検索にも応用可能であることが示されている。これらの DHT では経路表のサイズは基本的に固定である。

一方、先読みにより経路表を成長させ、ほとんどの検索を事実上 1 ホップで終了可能にする試みが、クラウドコンピューティングのように、資源を集約させることができる環境の中で現実的に動作可能であることが示されている。

これら 2 つの方向は真逆とも言えるが、互いに独立しており、今後、これらを組み合わせた、高機能で高速なサービスが実現される可能性も見えてきたと言える。

また、ノードの非均質性に注目したグループ化の手法にも、信頼できる (仕様を満たす、悪意のない) ノードが選別できると仮定した上で、それらのみを用いたオーバレイの経路が利用できるといったような新たな展開が見られる。ノードの信頼性を評価するための評判システムの研究も続けられている。

第 1 世代の DHT の研究から約 10 年が経過するが、この間、DHT は分散環境の現実に揉まれる中で成長し、現実的価値を増してきていると考えられる。

参考文献

[1] James Aspnes and Gauri Shah. Skip graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, 2003.

[2] David A. Bryan, Philip Matthews, Eunsoo Shim, Dean Willis, and Spencer Dawkins. *Concepts and Terminology for Peer to Peer SIP*, July 2008. Internet-Draft.

[3] Yatin Chawathe, Sriram Ramabhadran, Sylvia Ratnasamy, Anthony LaMarca, Scott Shenker, and Joseph Hellerstein. A case study in building layered DHT applications. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '05)*, 2005.

[4] Bram Cohen. Incentives build robustness in BitTorrent. In *Proceedings of the First Workshop on Economics of Peer-to-Peer Systems*, May 2003.

[5] Intel Corporation. Place Lab, 2006. Available electronically at http://www.placelab.org/.

[6] eMule Team. emule-project.net, 2002. Available electronically at http://www.www.emule-project.net/.

[7] Pedro Fonseca, Rodrigo Rodrigues, Anjali Gupta, and Barbara Liskov. Full-information lookups for peer-to-peer overlays. *IEEE Transactions on Parallel and Distributed Systems*, 20(9), 2009.

[8] Deke Guo, Yunhao Liu, and Xiang-Yang Li. BAKE: A balanced Kautz tree structure for peer-to-peer networks. In *INFOCOM*, 2008.

[9] Frans Kaashoek and David R. Karger. Koorde: A simple degree-optimal distributed hash table. In *2nd International Workshop on Peer-to-Peer Systems*, February 2003.

[10] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The Eigen-Trust algorithm for reputation management in P2P networks. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[11] David R. Karger and Matthias Ruhl. Diminished Chord: A protocol for heterogeneous subgroup formation in peer-to-peer networks. In *IEEE IPTPS*, 2004.

[12] Dmitry Korzun and Andrei Gurtov. A local equilibrium model for P2P resource ranking. *ACM SIGMETRICS Performance Evaluation Review*, 37(2), 2009.

[13] Dmitry Korzun, Boris Nechaev, and Andrei Gurtov. Cyclic routing: Generalizing look-ahead in peer-to-peer networks. In *ACS/IEEE International Conference on Computer Systems and Applications*, 2009.

[14] Dongsheng Li, Xicheng Lu, Baosheng Wang, Jinshu Su, Jiannong Cao, Keith C. C. Chan, and Hong va Leong. Delay-bounded range queries in DHT-based peer-to-peer systems. In *Proceedings of the 26th IEEE International Conference on Distributed Computing Systems (ICDCS '06)*, 2006.

[15] Dongsheng Li, Xicheng Lu, and Jie Wu. FISSIONE: a scalable constant degree and low congestion DHT scheme based on Kautz graphs. In *INFOCOM*, 2005.

[16] Dahlia Malkhi, Moni Naor, and David Ratajczak. Viceroy: A scalable and dynamic emulation of the butterfly. In *21st Annual Symposium on Principles of distributed computing*, July 2002.

[17] Sergio Marti and Hector Garcia-Molina. Taxonomy of trust: categorizing P2P reputation systems. *Computer Network*, 50(4), 2006.

[18] Isaias Martinez-Yelmo, Alex Bikfalvi, Carmen Guerrero, Ruben Cuevas, and An dreas Mauthe. Enabling global multimedia distributed services based on hierarchical DHT overlay networks. In *Proceedings of the 2008 The Second International Conference on Next Generation Mobile Applications, Services, and Technologies*, 2008.

[19] Petar Maymounkov and David Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In *Proceedings of IPTPS02 (Springer LNCS 2429)*, March 2002.

[20] Luiz Rodolpho Monnerat and Cláudio L. Amorim. D1HT: A distributed one hop hash table. In *The 20th IEEE Intl. Parallel & Distributed Processing Symposium*, 2005.

[21] Luiz Rodolpho Monnerat and Cláudio L. Amorim. Peer-to-peer single hop distributed hash tables. In *GLOBECOM*, 2009.

[22] Zhonghong Ou, Erkki Harjula, Timo Koskela, and Mika Ylianttila. General truncated pyramid peer-to-peer architecture over structured DHT networks. *Mobile Networks and Applications*, 15(5), 2009.

[23] Kunwoo Park, Sangheon Pack, and Taekyoung Kwon. Proximity based peer-to-peer overlay networks (P3ON) with load distribution. In *ICOIN 2007 Revised Selected Papers*, 2008.

[24] C. Greg Plaxton, Rajmohan Rajaraman, and Andrea W. Richa. Accessing nearby copies of replicated objects in a distributed environment. In *Proceedings of ACM SPAA*, June 1997.

[25] Sylvia Ratnasamy, Paul Francis, Richard Karp Mark Handley, and Scott Shenker. A scalable content-addressable network. In *Proceedings ACM SIGCOMM, San Diego, CA*, August 2001.

[26] Sean Rhea, Brighten Godfrey, Brad Karp, John Kubiatowicz, Sylvia Ratnasamy, Scott Shenker, Ion Stoica, and Harlan Yu. OpenDHT: a public DHT service and its uses. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '05)*, 2005.

[27] Antony Rowstron and Peter Druschel. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In *Proceedings of the 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001)*, November 2001.

[28] Haiying Shen and Cheng-Zhong Xu. Hash-based proximity clustering for efficient load balancing in heterogeneous DHT networks. *Journal of Parallel and Distributed Computing*, 68(5), 2008.

[29] Ion Stoica, Robert Morris, M. Frans Kaashoek David Karger, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of ACM SIGCOMM*, August 2001.

[30] Chunqiang Tang, Melissa J. Buco, Rong N. Chang, Sandhya Dwarkadas, Laura Z. Luan, Edward So, and Christopher Ward. Low traffic overlay networks with large routing tables. *ACM SIGMETRICS Performance Evaluation Review*, 33(1), 2005.

[31] Yiming Zhang, Dongsheng Li, Lei Chen, and Xicheng Lu. Flexible routing in grouped DHTs. In *Proceedings of the 2008 Eighth International Conference on Peer-to-Peer Computing*, 2008.

[32] Yiming Zhang, Ling Liu, Dongsheng Li, Feng Liu, and Xicheng Lu. DHT-based range query processing for web service discovery. In *International Conference on Web Services*, 2009.

[33] Yiming Zhang, Ling Liu, Dongsheng Li, and Xicheng Lu. Distributed line graphs: A universal framework for building DHTs based on arbitrary constant-degree graphs. In *The 28th International Conference on Distributed Computing Systems*, 2008.

[34] Yiming Zhang, Xicheng Lu, and Dongsheng Li. SKY: efficient peer-to-peer networks based on distributed Kautz graphs. *Science in China Series F: Information Sciences*, 52(4), 2009.

[35] Ben Zhao, Yitao Duan, Ling Huang, Anthony Joseph, and John Kubiatowicz. Brocade: Landmark routing on overlay networks. In *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS 2002)*, March 2002.

[36] Ben Y. Zhao, John D. Kubiatowicz, and Anthony D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB//CSD-01-1141, U.C.Berkeley, April 2001.