DHT 기반 P2P 네트워크에서 네트워크 토폴로지를 고려한 메시지 라우팅 매커니즘

손영성* · 김희정* · 김경석**

*한국전자통신연구원 · **부산대학교

Message Routing Mechanism using Network Topology Information in DHT based P2P Network

Young-Sung Son* · Hee-Jeong Kim* · Gyung-Seok Kim**

*Electronics and Telecommunication Research Institute · **Pusan National University

E-mail : ysson@etri.re.kr, heejkim@etri.re.kr, gimgs@asadal.cs.pusan.ac.kr

요 약

인터넷 상의 불특정 다수의 컴퓨팅 자원을 연계하여 오버레이 네트워크를 구성하여 새로운 컴퓨팅 인프라를 구성하려는 분산 해쉬 테이블 (Distributed Hash Table) 방식의 Peer-to-Peer(P2P) 네트워크 관련 연구가 진행되고 있다. DHT 방식의 P2P 네트워크는 자원의 복제 및 공유하여 컴퓨팅 시스템 전반에 걸친 신뢰성과 결함 감내 능력을 향상시키는 장점이 있는 반면에 하부 네트워크 정보를 무시하기 때문에 전체 시스템 구성 및 메시지 전송에 있어 실제적인 성능상의 문제점을 드러낸다. 이 논문에서는 하부 네트워크의 구성 정보를 이용하여 DHT 기반의 P2P 네트워크의 메시지 전송 방식의 효율을 높이는 방법을 소개한다.

키워드

P2P Network, Distributed Hash Table, Routing Algorithm

Ⅰ. 서 론

하루가 다르게 사용자 PC의 성능이 향상되고, 네트워크 속도와 대역폭이 발전되어 방대한 규모의 인터넷이 구성되는 현실에서, 전통적인 서버 집중식 모델은 현재의 인터넷 자원을 충분히 활용하지 못하고 있다. 그러한 현실 인식을 바탕으로 떠오르고 있는 Peer-to-Peer (P2P) 기술은 분산컴퓨팅 기술의 출현은 부분적인 기술의 변형으로 성공적인 상업화로 진행되는 것으로 보인다. P2P 기술은 불특정 다수가 참여하는 대규모의 분산시스템의 핵심 기술로 정보검색과 정보전송, 대규모 연산 응용 등의 분야에서 성능향상과 함께 컴퓨팅 시스템 전반적에 걸친 신뢰성과 결함 감내 능력을 향상시키고 있다. 기존의 P2P 네트워크 관련 연구는 대규모 불특정 다수의 자원을 효과적으로 관리하기 위한 방법으로 분산해쉬테이블 (DIstributed Hash Table) 방식을 선택하였다. 이 방식은 자원 정보를 해쉬함수를 이용하여 임의적인 분산을 지원하였고 이를 통해서 네트워크의 부하 상황에 따른 전체 네트워크 붕괴의 위협을 해결하고 저비용의 관리 부담으로 순수 분산 환경에서 대규모의 자원 관리를 하기에는 매우 적합하여 전체적인 시스템 효율을 높일 수 있었다. 그러나, 실제 DHT 방식의 P2P 네트워크 기술을 인터넷에 적용하기에는 실제적인 성능 문제로 한계를 보이고 있다.

이에 본 논문에서는 기존의 DHT 기반의 P2P 네트워크 기술에 네트워크 토폴로지 정보를 추가할 수 있도록 확장하여 보다 효과적인 메시지 전송을 지원할 수 있는 방법을 제안하고 이에 따른 성능의 안정성을 설명한다.

II. 관련연구

2.1 Magic Square [1]

Magic Square Protocol은 시스템에 참여하는 불특정 다수의 노드들의 접속과 탈락에 영향을 덜 받는 *P2P* 메시지 라우팅을 지원하기 위해서 고안되었다. 각 노드는 해쉬 함수를 이용해서 m비트의 자신의 노드 아이디를 소유한다. 고정 ip 를 가진 노드는 ip 를 이용해서 노드 아이디를 만들고 가변 ip 를 가진 노드는 사용자 지정 정보를 이용해서 노드 아이디를 생성한다. 각 노드는 두 종류의 연결 테이블을 이용해 피어를 유지한다. 두 테이블은 각 노드가 시스템에 접속시에 네트워크에서 임의의 노드를 선택하는 전역 테이블(global table) 과 자신의 노드 아이디를 중심으로 순차적으로 연결성을 갖는 라우팅 테이블(routing table) 이다. 전체 시스템 구성은 그림 1 과 같다. 지역 테이블은 양방향 스킵 리스트 (bi-directed skip list) 를 구성하기 위해서 사용된다. 스킵 리스트[2]는 검색 효율을 위해서 자동 조정(self-balanced) 기능을 가진 자료 구조이다. 각 노드는 자신의 네트워크 요구 처리 능력에 따라서 지역 테이블의 크기를 정한다. 그림 1은 스킵 리스트로 지역 테이블을 구성한 경우를 나타낸다. 노드 A, E, J 는 네트워크 대역폭과 네트워크 처리 능력이 뛰어나고 노드 G 의 경우는 전화 모뎀으로 접속되었다고 해석할 수 있다.

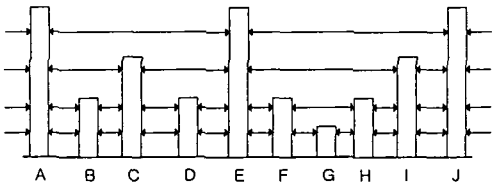

그림 1. Magic Square Routing 방식에 따른 노드 연결성 예제

2.2 Chord [3]

Chord 시스템은 P2P 응용을 위한 분산처리 자원탐색 프로토콜로, 분산 탐색을 지원하며 해쉬 함수를 이용하여 데이터의 삽입과 탐색을 수행한다. Chord는 2^k 크기의 원형 식별자 공간을 사용하며 각 노드는 IP주소를 SHA-1과 같은 해쉬 함수로 해쉬하여 nodeID를 구한 다음 원형 식별자 공간의 해당 nodeID에 위치한다. 데이터의 위치 정보는 (key, value) 쌍으로 표현되며 데이터가 저장될 노드의 위치는 key를 해쉬한 값에 의해 정해진다. 원형 식별자 공간에서 각 노드는 *successor, predecessor*의 정보를 유지하여 링을 형성하고 노드가 fail되였을 때 시스템을 복원하기 위해 successor list의 정보를 유지한다.

Chord에서 key를 해쉬한 값을 식별자 공간에 대응시킬 때 원형 식별자 공간에서 해쉬된 키 값과 같은 nodeID를 가지는 노드에 저장하거나 같은 nodeID를 가진 노드에 저장하거나, 같은 *nodeID*를 가진 노드가 없을 때는 바로 뒤의 노드에 저장한다. 이 노드를 successor 노드라 부른다. 각 노드는 전체 네트워크 상의 노드에 대한 정보를 분산된 동적 환경에서 효과적으로 만족시키기 위해 finger table을 유지한다. Finger table은 데이터를 삽입하거나 데이터를 관리하는 노드를 찾기 위해 lookup 메시지를 해당 노드에게 전달하기 위해 이용된다.

2.3 Grapes [4]

Grapes는 Topology를 고려한 P2P시스템이다. Grapes는 leader들로 구성되는 super-network와 물리적으로 가까운 노드들로 구성된 sub-network으로 이루진 시스템이다.

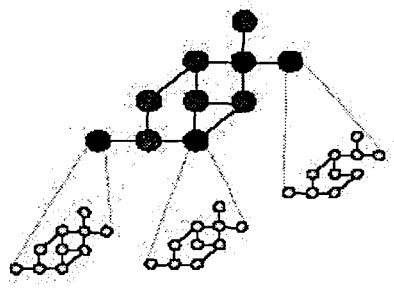

그림 2. Grapes 시스템 구조

조인을 원하는 노드는 bootstrap node 와 물리적 거리를 측정하여 그 값이 threshold 값보다 작으면 bootstrap node에 속하는 sub-network에 조인하고 threshold 값보다 크면 super-network의 leader와 물리적인 거리를 측정하여 threshold 값보다 작은 leader의 sub-network에 조인된다. 만약 모든 leader와 물리적 거리를 측정한 값이 Threshold 값보다 크면 새로운 노드는 super-network에 조인되고 sub-network의 leader가 된다. Sub-network의 모든 노드는 다른 sub-network의 데이터를 탐색하기를 원할 때 자신의 sub-network에 속해 있는 *leader*를 통해서만 가능하기 때문에 leader에 대한 부담 크면 leader가 fail될 때 leader의 교체하는 비용 크다. 또한 각sub-network의 크기가 동일하지 않으므로 node의 수가 적은 sub-network의 노드는 데이터를 탐색하기 위해 다른 sub-network에서 데이터를 찾는 확률이 높아 탐색의 효율성이 떨어진다. 우리가 제안하는 논문은 subnet의 사이즈를 고정하여 subnet에 노드의 수를 균일하게 만들기 때문에 탐색의 효율성을 향상시킨다.

III. 문제점

다음 그림3은 DHT기반의 P2P네트워크의 예를 보인다. 노드 A에서 J까지 그림과 같이 상호 연결성을 유지한다. 메시지가 노드A에서 노드G로 전송될 경우 노드A->노드E->노드F->노드G 순서로 전달된다. 그러나, 실제 네트워크 구조가 그림과 같이 구성되어 있다면 3번의 글로벌 네트워킹을 요구하게 된다. 실제 글로벌 네트워킹은 로컬 네트워킹의 수십-수백배의 전송시간을 보인다.

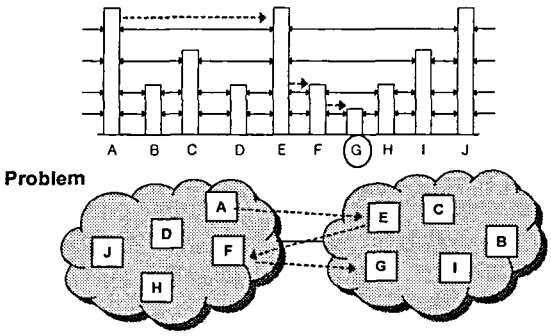

그림 3. DHT 기반 P2P 네트워크에서의 비효율적인 메시지 전송

위의 예제에서 최적의 메시지 전송은 다음 그림 예제와 같이 동일 네트워크에서 최대한 목적지에 가깝게 로컬 네트워킹을 통해서 메시지 전송을 수행한 뒤 글로벌 네트워킹 메시지 전송을 수행해야 한다.

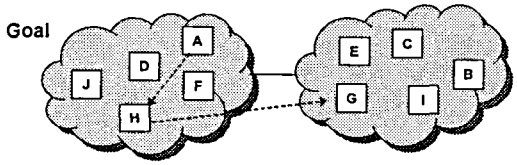

그림 4. 최적의 메시지 전송 예제

IV. 네트워크 토폴로지를 이용한 메시지 라우팅 메커니즘

DHT 기반의 P2P 네트워크에서는 물리 네트워크 구성 (Network Topology) 정보를 추상화한 오버레이 네트워크를 구성하기 때문에 본 논문에서 재기한 전송 효율문제를 해결할 수 없다. 따라서 본 문제점을 해결하기 위해서는 네트워크 토폴로지에 대한 정보를 DHT 기반 라우팅 테이블에 포함시키야 한다. 이를 위해서 본 논문에서는 기존 Magic Square 프로토콜의 라우팅 정보를 위한 글로벌 테이블을 로컬 네트워크 정보를 위한 새로컬 테이블 필드로 사용하여 노드가 P2P 네트워크에 참여할 때 이 정보를 저장하고 이를 라우팅에 활용한다. 로컬 테이블은 각 노드가 Magic Square 네트워크에 참여하고 해제하는 절차에 따라서 정보가 수정된다. 각 노드가 Magic Square 네트워크에 참여할 때 자신의 정보를 IP 멀티캐스트를 통해서 로컬네트워크에 전송한다. 이를 받은 각 노드들은 자신이 속해있는 로컬 네트워크에 새로운 노드가 추가되었음을 파악하고 로컬 테이블을 업데이트한다. 각 노드가 Magic Square 네트워크에서 탈퇴할 때에도 자신의 정보를 IP 멀티캐스트를 통해서 로컬네트워크에 전송한다. 이를 통해서 각 노드의 로컬 테이블을 업데이트한다.

Node A routing table

left	right
	E
	C
	B
	B

Local table

D
F
H
J

그림 5. 노드 A 의 라우팅 테이블

각 노드가 메시지를 전송하기 위해서 메시지의 목적지 주소를 기존의 라우팅 테이블과 로컬 테이블 필드를 참조하여 메시지 전송을 로컬 네트워킹 패스를 통할 것인지 글로벌 네트워킹 패스를 통할 것인지를 결정하여 메시지 전송의 효율을 높인다. 메시지를 수신한 노드는 메시지의 목적지 주소(Destination Address)를 기준으로 자신의 라우팅 테이블과 로컬 테이블에 들어있는 각 노드의 주소를 비교하여 상대적으로 가장 가까운 노드로 메시지를 재전송 한다. 이를 통해서 전체적인 메시지 전송 효율을 높일 수 있다.

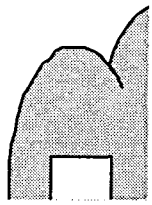

그림 6. 네트워크 토폴로지를 고려한 메시지 라우팅 예제

그림 6 은 메시지 전송 예제를 표현하였다. 기존의 DHT 기반의 메시지 전송방식을 붉은색으로 표시하고 본 논문에서 제안하는 네트워크 토폴로

지를 기반으로한 메시지 전송 방식을 푸른색으로 표시하였다.

V. 성능분석

Magic Square 프로토콜 및 개선된 프로토콜의 성능을 평가하기 위해서 전체 시스템을 모델링하는 파라메터를 정하고 이를 이용하여 메시지 전송시의 전체 전송 지연을 계산하였다. 본 논문에서 사용하는 파라메터는 다음과 같다.

1. 전체 노드 수는 N 개이다.
2. 로컬 네트워크 수는 M 개이다. 로컬 네트워크는 IP 멀티캐스트로 전달되는 범위로 한다.
3. 로컬 네트워크에 속한 평균 노드 수는 LN = $\frac{N}{M}$ 이다.
4. Magic Square 프로토콜에서 메시지 전송시 출발 노드 (Source Node)에서 목적지 노드 (Destination Node) 까지 전달되는 메시지 재전송 횟수는 $O(logN)$번이다.
5. 로컬 네트워크 메시징 비용은 D_L 이다.
6. 글로벌 네트워크 메시징 비용은 D_G 이다.
7. 글로벌 네트워크 메시징 비용과 로컬 네트워크 메시징 비용은 간편하게 상수(C) 배라고 가정했다. $D_G = C \cdot D_L$
8. 각 노드의 로컬 테이블 크기는 L 개이다.
9. 각 노드의 라우팅 테이블 크기는 R 개이다.

기존의 Magic Square 프로토콜의 메시지 전송 비용은 다음과 같다.

$$\left(\frac{R}{N}\cdot D_L+\frac{N-R}{N}\cdot D_G\right)^{O(logN)}=\left(\frac{R}{N}D_L+\frac{N-R}{N}C\cdot D_L\right)^{O(logN)}$$
$$=\left(\frac{R+(N-R)\cdot C}{N}\cdot D_L\right)^{O(logN)}=\left(\frac{R+CN-CR}{N}\cdot D_L\right)^{O(logN)}$$
$$=\left(\frac{CN+(1-C)\cdot R}{N}\cdot D_L\right)^{O(logN)}\fallingdotseq (C\cdot D_L)^{O(logN)}$$

본 논문에서 제안하는 네트워크 토폴로지를 고려한 메시지 전송 비용은 다음과 같다.

$$\left(\frac{L}{L+LN}\cdot D_G+\frac{LN}{L+LN}\cdot D_L\right)^{O(logN)}=\left(\frac{L}{L+\frac{N}{M}}\cdot D_G+\frac{\frac{N}{M}}{L+\frac{N}{M}}\cdot D_L\right)^{O(logN)}$$
$$=\left(\frac{L}{\frac{LM+N}{M}}\cdot D_G+\frac{\frac{N}{M}}{\frac{LM+N}{M}}\cdot D_L\right)^{O(logN)}=\left(\frac{LM}{LM+N}\cdot D_G+\frac{N}{LM+N}\cdot D_L\right)^{O(logN)}$$
$$=\left(\frac{(C\cdot LM+N)}{LM+N}\cdot D_L\right)^{O(logN)}\fallingdotseq (D_L)^{O(logN)}$$

VI. 결 론

본 논문은 IP 멀티캐스트 방식을 활용하여 로컬네트워크의 정보를 수집하여 각 노드의 글로벌 테이블을 구성하고 이를 통해 메시지 전송에 있어 로컬 메시징의 비율을 높임으로서 전체적인 전송 효율을 높이는 방법을 제안한다. 간단한 성능 비교를 통해서 기존 메시지 전송방식과의 성능 비교를 통하여 성능향상의 근거를 보인다.

참고문헌

[1] Sun-Mi Park, "Magic Square: Scalable peer-to-peer lookup protocol considering peer's characteristics", IASTED CIC, 2003.
[2] William Pugh, "Skip Lists: a probabilistic alternative to balanced Trees", communications of the ACM, Vol.33 No.6 pp.668-676, June 1990

[3] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek and H.Balakrkshnan, "Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications," In Proceedings of SIGCOMM (August 2001)
[4] K. Shin, S. Lee, G.. Lim, H. Yoon, and J. S. Ma, "Grapes: Topology-based Hierarchical Virtual Network for Peer-to-peer Lookup Services," In Proceedings of the International Conference on Parallel Processing Workshops (ICPPW' 02), 2002.