지연과 신뢰성을 고려한 오버레이 멀티캐스트 제공 방안

이상옥[0] 김상하
충남대학교 컴퓨터과학과 네트워크 연구실
{solee, shkim}@cclab.cnu.ac.kr

Reliable Data Delivery in Delay Bounded Overlay Multicast

Sang-Ok Lee[0] Sang-Ha Kim
Department of Computer Science, ChungNam National University

요 약

오버레이 멀티캐스트는 인터넷에서 확장성 있는 일-대-다, 다-대-다 데이터 전송을 제공하기 위한 메커니즘으로 제안되었다. 하지만, 데이터 전송이 각 멤버들의 패킷 전송에 의존하게 되므로 한 멤버의 고장은 하위 멤버들이 데이터를 받을 수 없게 만든다. 더욱이, 보다 높은 상위 계층의 멤버가 고장 날수록 더 많은 하위 멤버들이 데이터를 받을 수 없게 된다. 본 논문에서는 오버레이 멀티캐스트에서 신뢰성 있는 데이터 전송을 위한 메커니즘을 제안한다. 제안된 메커니즘은 단-대-단 지연을 일정 값 이하로 유지시킬 수 있는 동시에 노드의 고장 확률에 기반하는 오버레이 데이터 전송 트리(DDT)를 구성한다.

1. 서 론

중간 라우터의 상태 정보 유지등과 같은 IP 멀티캐스트 적용상의 문제점을 해결하기 위하여 오버레이 멀티캐스트가 제안 되었다. 오버레이 멀티캐스트는 라우터에서 수행하던 멀티캐스트 포워딩을 각 그룹 멤버가 수행하게 된다. 따라서, 각 그룹멤버는 받은 데이터 패킷들을 복사하여 다른 멤버들에게 전달해야 한다. 지금까지 오버레이 멀티캐스트 메커니즘들은 효율적인 오버레이 데이터 전송 트리를 구성하는데 초점이 맞추어져 있었다. 따라서, 다른 오버레이 멀티캐스트의 다른 관점은 지적되고 있지 않다.

본 논문은 오버레이 멀티캐스트의 신뢰성 있는 데이터 전송 특성에 초점을 맞추고자 한다. 이 신뢰성은 데이터 전송 트리 참가자의 고장에 대한 강건함과 또한 이러한 고장의 영향 정도에 영향을 받게 된다. 첫번째로, 라우터가 데이터 전송 트리를 구성하고 있는 순수 멀티캐스트와는 달리 오버레이 멀티캐스트에서는 각 그룹 멤버들과 함께 데이터 전송 트리를 구성하고 있다. 일반적으로 호스트들은 라우터보다 고장에 더욱 취약하기 때문에 오버레이 데이터 전송 트리는 IP 멀티캐스트보다 더 높은 고장 가망성이 있다. 또 한편으로는, 순수 멀티캐스트에서 어떤 그룹의 멤버가 고장 나더라도 다른 그룹 멤버의 패킷 전송률에는 전혀 영향이 없다. 하지만, 오버레이멀티캐스트의 데이터 전송 트리는 하나의 멤버가 고장나면, 그 멤버의 하위 멤버들은 전부 데이터를 받을 수 없다. 게다가, 고장난 노드의 위치가 상위 레벨에 있을 경우에는, 더 많은 히위 멤버들이 데이터를 받을 수 없게 된다. 이런 결과로, 멤버들의 빈번한 고장이 순수 멀티캐스트보다 신뢰성있는 데이터 전송을 방해하는 동시에 다른 그룹 멤버의 전송률에 또한 영향을 준다. 그래서, 중대한 성능감소가 일어난다.

이러한 문제점을 해결하기 위하여 몇 가지 메커니즘이최근에 제안되었다. [1]은 사전결정 방식 또는 요구 기반 방식으로 적은 지연 범위와 높은 데이터 전송률을 독자적으로 보장할 수 있는 PRM(Probabilistic Resilient Multicast)를 제안했다. 사전 결정 방식으로서, 무작위로 추출한 포워딩 방식은 각각 다른 그룹 멤버에게 무작위로 데이터를 포워딩한다. 다른 한편으로는, 트리거드 NAK은 네트워크 혼잡과 링크 에러 때문에 발생한 데이터손실을 처리하기 위해 소개 되어졌다. [2]는 이와 유사한 RJ(Random Jump)개념을 소개했다. RJ를 기반으로 하나의 노드는 자신의 부모 노드 뿐만 아니라 트리 안에 있는 다른 노드로부터 데이터를 전송 받는다. [3]은 단-대-단까지의 높은 회복성을 위해 디자인된 새로운 오버레이 멀티캐스트 프로토콜(Nemo)를 묘사하고 있다. 시뮬레이션 결과를 통하여 어떻게 Nemo가 높은 긴장 상태하에서 재전송 패킷과 추가적인 지연에 관해서 상당히 적은 비용으로 높은 연결성과 전송률을 제공하는 것을 보여주고 있다. ROMA는 [4]에 의해 제안되었다. 이것은 다중연결을 설립하는 그룹 멤버에서의 TCP수행성능감소에 대한 적용 이슈를 언급하고 있다.

앞에서 설명한 것처럼, 이전 개념들은 중복된 포워딩을 사용해서 신뢰성있는 데이터 전송을 제공한다. 이러한 다중 전송을 통하여 서비스 중단을 최소화할 수 있다. 그러나, 이런 중복전송은 불필요한 자원낭비뿐만 아니라 각 그룹 멤버의 유지를 위한 오버헤드가 매우 커진다. 게다가, 임의로 선택된 추가적인 에지들은 여러 상위 노드 허용함으로써 메쉬 구조로 변화시킨다. 이와 같이, 데이터 패킷의 루프를 예방하는 추가적인 고려사항이 있어야 한다. 이와 비슷한 이유로서, 임의로 추가된 트리 에지가 부모보다 상위 노드로 설정된다면, 신뢰성 증가에 아무런 효과가 없다.

본 논문에서는 오버레이 멀티캐스트에서 지연과 신뢰성을 고려한 데이터 전송을 위한 기본구조를 제안한다. 이것을 ROM(Reliable Overlay Multicast)라고 부른다. ROM의 목적은 잃어버린 패킷을 재전송하거나 빠르게 복구하는 것이 아니라 데이터 전송율을 높일 수 있는 오버레이 데이터 전송 트리를 구성하는 것이다. 이는 오버레이 데이터 전송 트리 구성시 더욱 신뢰성이 있는 그룹 멤버는 트리에서 더 높은 상위 레벨에 위치시키는 반면에 적은 신뢰성을 가진 그룹 멤버는 낮은 하위 레벨에 위치시키게 된다. 또한, 우리가 이러한 오버레이 데이터 전송 트리를 구성할 때, 기존의 메커니즘들과 마찬가지로 소스노드로부터 각 그룹 멤버들까지 단-대-단 지연을 미리 명시된 임계치 값으로 제한되어야만 한다. 일단, 오버레이 데이터 전송 트리가 구성되면, 각 노드는 자신의 고장으로 인하여 전체 시스템에 영향을 미치는 정도를 나타내는 가중치와 새로운 데이터 구조를 유지 관리해야 한다. 자신의 노드의 가중치가 크다면, 하위노드는 각각 다중 부모 노드를 설정해야 한다. 루프문제는 동적인 그룹-식별자를 지정하여 해결한다.

이 논문은 다음과 같이 구성된다. 2장에서는 ROM의 기본구조, 그룹참가와 탈퇴를 설명한다. 마지막으로, 3장에서는 결론 및 향후 연구방향에 대하여 설명한다.

2. 제안 메커니즘

ROM에서, 오버레이 데이터 전송 트리는 2가지 특성을 갖는다. 첫번째로, 오버레이 데이터 전송 트리는 응용계층 멀티캐스트간의 계층적인 클러스터를 기본으로 한다. 클러스팅은 동일한 네트워크에 속하는 그룹 멤버들로 구성된다는 것을 의미한다. 동일한 네트워크는 지리적 인접성을 의미하는 것이 아니고 Autonomous System (AS)와 같은 동일한 관리자에 의해 관리된다. 본 논문에서는, 각 네트워크를 식별하기위해 라우터의 AS 넘버를 사용한다. 각 멤버에서는 AS 넘버를 호스트에게 광고하기 위해서 ICMP 라우터 광고를 변경한다. 둘째로, 클러스트 리더들 사이(CL_DDT)와 쿨러스터 안에서 클러스터 리더와 그룹 멤버들 사이(GM_DDT)의 오버레이 데이터 전송 트리는 최소-힙 구조를 가지게 된다. 최소-힙은 부모의 값이 자식들의 값 보다 더 크지 않도록 만든 구조이다. 이 때, 각 그룹 멤버의 값은 노드 고장 가능성이다. 최소 힙 구조를 토대로, 한 노드가 그룹 참가를 원한다면, 최소 힙에서 노드 삽입방법처럼 비슷한 알고리즘에 의해서 처리된다. 이와 비슷하게, 최대 힙으로부터 한 노드가 삭제되는 알고리즘이 그룹탈퇴를 하려는 그룹 멤버가 있을 때 적용된다. 그러나, 일반적인 최소 힙 구조와 다른점은 각 그룹 멤버의에서 가질 수 있는 자식의 노드 수가 2로 제한되지 않는 다는 점이다.

2.1 자료구조

네트워크 상에 있는 각 그룹 멤버는 다음과 같은 변수들과 자료 구조로 유지되어야 한다.

- 부모 노드와 자식노드 : 오버레이에 데이터 전송 트리에 있는 모든 부모 노드와 자식 노드
- 순서 번호 : 각 TIME_WINDOW 마다 순서 번호가 매겨진 패킷을 받을 수 있는 것을 의미한다.
- *nff* : *nff* 는 정규화된 고장 분포를 나타낸다. 이것은 얼마나 많은 패킷이 손실되었는지를 나타낸다. 각 멤버들은 NFF_TIME_WINDOW동안에 *nff* 의 값을 유지해야 한다. 본 논문에서는 *nff* 를 노드 고장 확률로 정의한다.

$$nff_{cur} = \frac{\#\ of\ lost\ packets}{TIME\_WINDOW \times (last\ sequence - first\ sequence)}$$

$$nff = \alpha * nff_{cur} + (1-\alpha) * nff$$

여기서 α 는 0에서 1까지의 범위로서 이전 *nff* 값의 정도를 나타낸다. 초기에 0으로 설정된다.

- 클러스터 ID와 리더 : 각 그룹 멤버는 클러스터 ID와 리더 주소의 기록을 유지한다.
- 그룹 ID: 패킷 전송의 순번을 나타낸다.
- 지연 : 소스로부터 해당 노드까지의 단-대-단 지연을 의미한다.
- 고장 영향 요소 : 고장 노드에 의한 영향을 의미한다. 이 값은 고장 노드확률만큼 모든 자식 멤버의 총수에 의해서 영향을 받는다.

2.2 프로토콜 설명

이 장에서는 그룹 참가와 탈퇴를 설명한다. ROM에서는 기본적으로 멀티캐스트 그룹 주소와 포트 넘버는 초기에 URL와 같이 온라인 또는 오프라인을 통해서 모든 멤버들에 알려져 있다고 가정한다. 또한, 모든 그룹 멤버들은 핵심 포인트(CP)라고 불리는 특정 호스트를 알고 있다고 가정한다. CP는 모든 그룹 멤버들을 관리하지 않지만, CL(cluster leader)의 목록을 유지한다.

2.3 그룹 참가

그룹 멤버가 멀티캐스트 그룹에 참가하려고 할 때, 현재 그룹멤버를 지정하고있는 클러스터 리더 (CL)를 먼저 알아야 한다. 클러스터 리더의 정보를 얻기 위해 자신에게 허용된 지연과 자신이 속한 네트워크의 AS넘버를 포함한 JOIN_REQUEST_TO_CP 패킷으로 CP로 연결을 시도한다. CP가 JOIN_REQUEST_TO_CP 패킷을 받은 경우, AS넘버에 해당되는 CL을 찾는다. 그러나, CL이 존재하지 않은 경우에는 현재 그룹 참가 요청 노드를 CL로 지정하고 이 노드를 포함한 CL_DDT를 만들려고 시도한다. 일단 데이터 전송 트리가 만들어지면, CP는 논리적으로 인접한 부모와 자식 CL을 가진 JOIN_REPLY를 보낸다. 요구하는 노드가 CL을 포함한 JOIN_REPLY를 받았을 때, 해당 노드는 지연경계를 가진 JIN_REQUEST_TO_CL 패킷을 보낸다. CP에 참가 요구를 하는 비슷한 경우, 측정된 노드 고장 가능성이 없을 때, CL은 단지 지연 경계를 가진 GM_DDT가 일시적으로 만들어진다. 지연 경계 CL_DDT와 GM_DDT는 다음단계처럼 만들어진다.

2.4 그룹 탈퇴

그룹 멤버중에 하나가 그룹 통신을 중단하고, 멀티캐스트

그룹을 탈퇴할 때, 더욱 복잡한 절차가 요구된다. 그 절차는 그룹멤버가 CL인지 아닌지와, 그룹멤버가 오버레이 DTT에서 자식노드를 가지고 있는지 아닌지에 의존한다. 모든 경우에, 그룹을 탈퇴하려는 그룹 멤버는 새로운 데이터 전송 트리가 구성되기 전에 이미 전송되어진 패킷 손실을 줄이기 위해 일정기간 패킷 포워딩을 지속한다.

표 1. 그룹 탈퇴의 경우

	자식 노드	비자식노드
클러스터 리더	경우 1	경우 2
비클러스터 리더	경우 3	경우 4

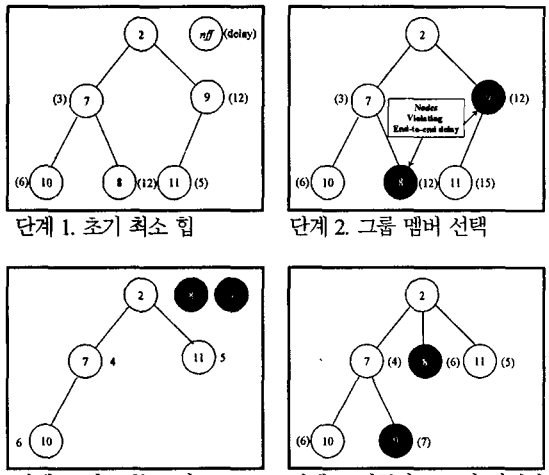
- 경우 1 : CL은 클러스터 안에 그룹멤버들의 정보와 관련된 전송과 새로운 CL를 지정한 후에 그룹 통신을 완전히 끝낼 수 있다. 고장 가능성이 가 장 적은 자식 노드를 선택하고 이를 새로운 CL로 지정한다. 그 다음에 탈퇴하는 CL은 새로운 CL에게 새로운 CL_DDT를 구성하기 위해서 CP를 알린 다. 새로운 이웃 클러스터 리더는 CP에 의해서 알려진다. 또한, 탈퇴하는 CL은 부모 노드를 교체하기 위해서 자식 노드들에게 알려야 한다.
- 경우 2 : 이 경우는, 단지 그룹 멤버가 출발점을 CP에게 알려준다.
- 경우 3 : 한 그룹 멤버는 자식 노드 중에 고장가능성이 적은 자식 노드를 선택한 다음 선택된 그룹멤버로 교체한다. 탈퇴하는 부모 노드로부터 지정된 후, 자식 노드는 탈퇴한 부모의 노드의 상위 노드의 멤버인 할아버지 노드에게로 데이터 포워딩을 요구한다. 그런 다음에, 새로운 부모 노드는 동일 멤버들에게 자신의 주소를 알린다.
- 경우 4 : 이것은 가장 단순한 절차를 보여준다. 이런 상황은, 한 그룹 멤버가 단지 포워딩 테이블의 엔트리를 삭제하는 것을 부모 노드에게 요구한다.

2.5 오버레이 데이터 전송 트리 유지

한 노드가 처음에 그룹 멤버가 되는 경우에 CL_DDT와 GM_DDT를 구성할 경우, 둘 다 측정된 노드 고장 가능성이 없기 때문에 단-대-단 지연 경계로 새롭게 재구성된다. 노드의 고장 가능성을 반영하기 위해서, 각 그룹 멤버는 주기적으로 CL과 교신한다. 한편으로는, CL은 힙이 재구성될 때까지 순서 없는 상태로 유지된다. 지연 경계를 가진 오버레이 데이터 전송 트리를 만들기 위해서, 우리는 발견적인 휴리스틱한 중앙 집중화 알고리즘을 제안한다. 이 알고리즘은 다음과 같다.

① 완전 이진 트리 형태로 최소-힙을 구성한다.
② 각 그룹을 위한, 소스에서 그룹 멤버까지 단-대-단 지연이 지연경계보다 크다면, 이 그룹멤버는 최소 힙으로부터 제거된다. 중간 그룹 멤버가 제거될 때, 현재 최소 힙은 그룹 탈퇴 절차를 적용하는 동일한 알고리즘에 의해 조정된다.
③ 단-대-단 지연을 어기는 그룹 멤버는 명시된 한계 내에서 단-대-단 지연이 제한될 때까지 알고리즘1에 의해 최소 힙에서 적당한 위치를 찾기 시작한다. 전체 최소 힙이 완전히 최적화됨에도 불구하고 어떤 그룹 멤버도 지정되어 있지 않다면, 한 그룹 멤버는 단-대-단지연을 최소화할 수 있는 그룹 멤버에게 연결을 요구한다.

[그림 1]은 오버레이 데이터 전송 트리 유지단계를 보여준다. 단계 1. 에서, 최소 힙은 초기에 구성되고 지연이 측정된다. 괄호 안은 측정된 지연을 나타낸다. 측정된 지연을 통해, 단-대-단 지연을 쉽게 알 수 있다. 만약에 단-대-단 지연경계를 10으로 설정하면, 8과 9 같이 노드 고장 가능성을 가지고 제한된 지연을 초과한 노드를 선택할 수 있다. 왜냐하면 그것들은 단-대-단 지연이 제한보다 크기 때문이다. 이와 같이, 최소 힙으로 부터 제거된 다음에 새로운 최소 힙은 단계 3. 에서 보여 주는 것처럼 구성된다. 지연은 역시 업데이트된다. 그 결과로, 신뢰성있는 트리로 제한된 지연이 단계 4.처럼 완전히 조정된다.



단계 1. 초기 최소 힙 단계 2. 그룹 멤버 선택

단계 3. 최소 힙 조정 단계 4. 어긋난 노드의 재배치

[그림 1] 알고리즘에 따른 오버레이 데이터 전송 트리의 예

3. 결 론

본 논문에서는 IP 멀티캐스트 개념의 대안으로서 최근 연구되고 있는 오버레이 멀티캐스트 개념의 신뢰성 이슈를 언급하였다. 오버레이 멀티캐스트의 신뢰성 있는 데이터 전송을 위해 본 논문에서는 각 그룹 멤버의 노드 고장 확률과 단-대-단 지연을 고려한 데이터 전송 트리를 구성하는 방안을 제안하였다.

[참고문헌]

[1] S. Banerjee et al., "Resilient Multicast using Overlays," in Proceedings of ACM SIGMETRICS, San Diego, CA, Jun. 2003, pp. 102 – 113.
[2] S. Birrer et al., "Resilient Overlay Multicast from the Ground Up," Tech. Report NWU-CS-03-22, Department of Computer Science, Northwestern University, 2003
[3] L. Xie et al., "An Approach to Reliability Enhancement of Overlay Multicast Trees," in Proceedings of PostGraduate Networking Conference, Jun. 2003.
[4] G. I. Kwon et al., "ROMA: Reliable Overlay Multicast with Loosely Coupled TCP Connections," To appear in Proceedings of IEEE INFOCOM, Hong Kong, Mar. 2004.