RTE:
신뢰적 멀티캐스트를 위한 라우팅 트리 추정 기법

윤원용, 이동만
한국정보통신대학원대학교

RTE: Routing Tree Estimation Scheme for Reliable Multicast

Wonyong Yoon (wyyoon@icu.ac.kr), Dongman Lee (dlee@icu.ac.kr)
Information and Communications University

요 약

트리 기반 신뢰적 멀티캐스트에서 효율적이고 확장성 있는 손실 복구를 하기 위한 RTE(Routing Tree Estimation) 기법을 제안한다. 에러 비트맵(error bitmap) 정보를 통하여 멀티캐스트 라우팅 트리와 유사한 논리적 트리(logical tree)를 구성함으로써 멀티캐스트 라우팅 트리에서 상위에 위치하는 수신자들이 재전송을 요청한 수신자의 신뢰성을 책임지도록 보장한다. 논리적 트리는 세션 멤버쉽이나 멀티캐스트 경로의 변화에 따라 적응적으로 재구성되는데 이는 멀티캐스트 세션 진행 동안 논리적 트리와 멀티캐스트 라우팅 트리 사이에 불일치를 최소화함으로써 멤버쉽과 경로가 변하는 상황에서도 implosion과 exposure를 감소시키는 장점을 지닌다. 제안한 기법과 정적 트리기반의 신뢰적 멀티캐스트 프로토콜의 시뮬레이션 결과는 세션의 크기가 커짐에 따라 제안한 적응형 트리 기반의 복구방식의 효율성을 입증한다

1 서 론

신뢰적인 멀티캐스트는 소스로부터의 각각의 패킷이 모든 수신자들에게 에러 없이 전달되는 것을 보장한다. 사용자 수와 지리적 범위 측면에서 통신의 규모가 점차 커짐에 따라서 신뢰적인 멀티캐스트의 두 가지 내재적 문제인 implosion과 exposure의 해결은 더욱 중요한 사안이 된다 [1, 2]. 확장성(scalability)을 높이기 위한 한가지 방법이 분산 복구(distributed recovery)이다 [3]. 분산 복구는 손실 복구를 소스뿐만 아니라 다른 수신자들도 수행하도록 하여 복구의 짐을 분산시키며 implosion을 줄이는데 도움을 준다. 또 다른 방법으로 지역적으로 손실을 복구함으로써 재전송을 위한 대역폭과 exposure를 감소시키는 지역 복구(local recovery)가 있다 [3,4,5]

트리 기반의 프로토콜은 트리구조 상으로 분산복구와 지역적 복구를 자연스럽게 결합시킨다 [6,7,8]. 트리 기반 복구에서 멀티캐스트 라우팅 트리의 상부에 위치하는 수신자들이 재전송을 요청하게 될 수신자들의 부모가 되도록 논리적 트리를 구성하는 것이 중요하다. 그러나 세션 멤버가 동적으로 가입/탈퇴하거나 하부 멀티캐스트 라우팅 트리가 변함에 따라, 그림 1 (a)과 같이 물리적 부모-자식 관계가 논리적 트리에 반영되지 못하는 전위(inversion) 문제가 발생할 수 있다 그림 1 (a)의 논리적 트리에서 논리적 자식 C는 피드백을 논리적 부모 P에게 피드백을 보낸다 그러나 P는 실제 멀티캐스트 라우팅 트리에서는 물리적 자식이므로 C의 신뢰성(reliability)를 책임질 수가 없다

본 논문에서는 트리 기반 신뢰적 멀티캐스트에서 효율적이고 확장성 있는 손실 복구를 하기 위한 RTE(Routing Tree Estimation) 기법을 제안한다.

2 RTE (Routing Tree Estimation)

전위 현상을 피하기 위해 RTE 기법은 IP 멀티캐스팅 [9]의 fate sharing property를 이용함으로써 라우팅 트리의 부모-자식 관계를 최대한 반영하는 논리적 트리를 구성한다. 멀티캐스트 세션의 각 수신자들은 에러 비트맵 정보를 유지하고 논리적 트리 상의 부모에게 피드백한다. 에러 비트맵은 번호 S와 비트맵으로 되어 있다. 비트맵의 각 비트는 해당 패킷을 수신하였으면 1로, (처음의 전송이) 실패하였으면 0으로 설정한다 가령 S=5, 비트맵 11010은 이 수신자가 순서번호 5,6,8의 패킷들을 성공적으로 수신하였음을 나타낸다. Fate sharing property란 만일 어떤 노드가 1이면 라우팅 트리 상의 부모 노드가 1일 가능성이 높은 것을 말하는데, 이러한 에러정보로써 라우팅 트리를 추정할 수 있다. 에러 비트맵의 취지는 손실 상관관계(loss correlation)가 유지되는 부모-자식 관계를 구성하자는 것이다.

논리적 토리 구성 기법의 제시에 앞서 연산자 ⊂와 ⊃를 다음과 같이 정의한다 노드 N의 에러 비트맵에 있는 모든 비트가 M의 그것보다 같거나 작을 경우에 N⊂M 관계를 가진다고 한다. 또한 N⊂M 이면 N⊃M 이고 그 역도 성립한다

신뢰적 멀티캐스트 세션에 들어온 새로운 멤버는 일단 소스에게 자신의 IP 주소를 알림으로써 소스의 임시 자식(tentative child)이 된다. 새로운 멤버는 멀티캐스트 그룹의 주소와 소스의 IP 주소를 안

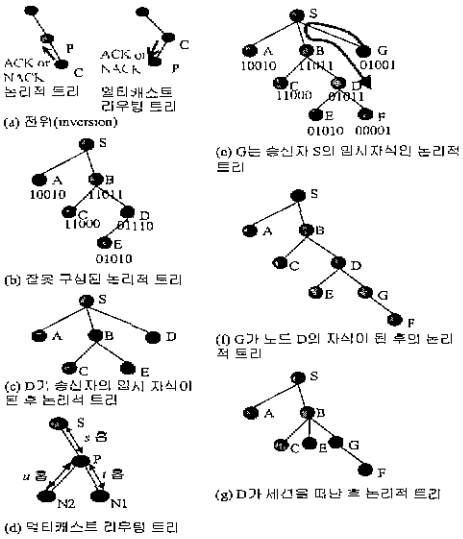


그림 1 RTE 기법

다고 가정한다. 소스는 새 멤버의 에러 비트맵 정보를 받으면 그것을 다른 자식들과 비교하여 새 멤버 N이 다른 자식들과 어떠한 관계를 가지는 지 검사한다. 첫째 N이 어떤 자식과도 관계를 가지지 않는다면 소스는 N을 정식 자식(regular child)으로 받아들인다. 둘째 N이 다른 자식 C와 N⊂C 관계를 가지면 소스는 N을 자신의 자식으로 채택하고 C는 N의 자식이 된다. 셋째 N이 C 와 N⊃C 관계를 가지면 소스는 N의 에러 비트맵 정보와 IP 주소를 C에게 알려준다. 그러면 노드 C에서 똑같은 과정이 반복하는데 적절한 N의 부모를 찾을 때까지 계속된다. 넷째 N과 C과 동일한 에러 비트맵을 가지면 N은 C의 자식으로 들어간다

세션을 탈퇴하려는 참여자는 그 사실을 부모에 알리면 부모는 탈퇴하려는 참여자의 직속 자식들을 자신의 자식으로 받아들이고 그 자식들에게 부모정보를 자기로 변경하라고 통보한다 실패(failure)로 인해 미처 알리지 못하고 트리에서 나가는 경우가 가능하다 이때 그 자식들은 트리에서 분리되게 된다. 자식들은 이 사실을 알게 되면 소스의 임시 자식으로 다시 트리에 붙는다. 그 다음 가입 알고리즘과 동일하게 적절한 부모를 찾아갈 수 있다

위의 같은 방법으로 논리적 트리와 라우팅 트리의 괴리가 발생하는 것은 드물지만 완전히 배제할 수는 없다. 또한 네트워크 경로가 변하면 그에 따라 두 트리 간에 괴리가 생기고 전위현상으로 인한 피해를 보게 된다. 이 때문에 논리적 트리를 재구성할 필요가 있는데 주기적으로 피드백 되는 에러 비트맵 정보를 이용한다 참여자 T가 자신의 자식 C와 그관계가 더 이상 성립하지 않음을 알게 되면 T는 자신의 부모 P에게 C의 새로운 위치를 찾아달라고 요청한다 C와 그관계가 성립하는 최초의 조상 N을 발견할 때까지 논리적 트리를 따라 상위로 같은 과정이 반복된다. 그 순간 N에서부터 트리 구성 시 가입 알고리즘과 똑같은 과정이 반복되어 C의 정확한 위치를 찾을 수 있다

그림 1 (e) (f) (g)는 RTE 기법의 논리적 트리 구성 예를, (b) (c)는 논리적 트리 재구성 예를 묘사하고 있다.

3 성능평가

3.1 최적의 에러 비트맵 크기

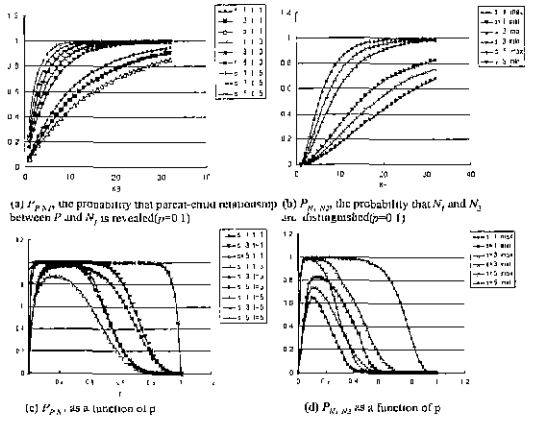


그림 2 파라미터 값 선택 : K_B

본 절에서는 '최적의' 논리적 트리 구성을 위해 에러 비트맵의 비트맵 크기(비트 수) K_B가 얼마로 설정되어야 하는지 알아본다. K_B는 그림 1 (d)의 N_1과 N_2가 서로 부모-자식 관계를 맺지 않도록 충분히 커야 한다. 또한 K_B 크기의 에러 비트맵으로 P와 N1이 부모-자식 관계를 맺어야 한다는 사실을 판단할 수 있어야 한다

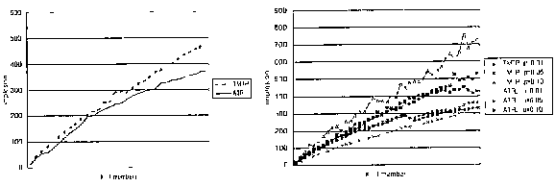
$P_{P,N1}$을 P와 N_1이 각각 부모와 자식임으로 판명될 확률을 나타내기로 하자. 이 확률은 1에서 부모-자식으로 판명되지 않을 확률을 뺀 값이다. 부모-자식으로 판명되지 않을 확률은 패킷이 P에 까지 전달되지 않을 확률과 패킷이 P와 N_1 모두에게 손실 없이 도착될 확률의 합이다. 즉 패킷이 둘 모두에 대해 공히 손실되거나 아니면 공히 전달될 확률이다. p를 링크의 패킷 손실 확률이라 할 때, $P_{P,N1}$은 아래와 같이 주어진다.

$$
\begin{aligned}
&P_{P,N1}\\
&= 1 - \{1 - (1-p)^s - (1-p)^t(1-p)^{t'}\}^{K_B}\\
&= 1 - \{1 - S - ST\}^{K_B}\\
&\text{where}\quad S = (1-p)^s, T = (1-p)^t
\end{aligned}
\tag{1}
$$

이제 N_1과 N_2이 부모-자식 관계에 있지 않은 것으로 판명될 확률을 $P_{N1,N2}$이라 하면, 이 값은 1 - P_{mis} 로 나타낼 수 있다. 여기서 P_{mis}, 은 부모-자식 관계에 있는 것으로 잘못 판명될 확률을 말한다. 이러한 오판(misjudgment)은 상관관계가 없는 손실(unrelated losses)로 인해 발생한다. P_{mis} 은 N_1이 N_2와 그관계를 가지는 것으로 잘못 판명될 확률과 N_2가 N_1와 그관계를 가지는 것으로 잘못 판명될 확률을 합한 값에서 N_1, N_2 가 동일한 수신 상태를 가질 확률을 빼면 된다. 이를 계산하면 아래와 같다.

$$
\begin{aligned}
&P_{N1,N2} = 1 - P_{mis}\\
&= 1 - [(1 - S(1-T)U)^{K_B} + (1 - S(1-U)T)^{K_B}\\
&- (1 - S + S(1-T)(1-U) + STU)^{K_B}]\\
&\text{where}\quad S = (1-p)^s, T = (1-p)^t, U = (1-p)^u
\end{aligned}
\tag{2}
$$

그림 2 (a)에서 $P_{P,N1}$은 K_B, s, t의 함수로, (b)에서 $P_{N1,N2}$은 K_B 및 s 의 함수로 나타내져 있다 링크간 홉 수 t 와 u, 는 1에서 5 사이의 범위에 있다. 식별하기 쉽게 하기 위해 확률의 최대값과 최소값만을 명시하였다. 그림 2의 (a)와 (b)에서 K_B 가 증가함에 따라 확률이 급격한 비율로 1에 가까워진다는 사실을 알 수 있다. 이 결과에 따라 K_B 는 32로 설정되면 논리적 트리 구성에는 별 무리가 없는 것으로 결론 맺을 수 있다. 주기적인 에러 비트맵 교환이 진행될수록 정보량



(a) average implosion (b) implosion at loss probability p

그림 3 Average feedback implosion

은 32, 64로 점검 증가하고 이에 따라 확률 값들도 1에 가까워지게 된다. 확률이 1에 가깝다는 것은 논리적 트리와 멀티캐스트 라우팅 트리가 거의 유사함을 의미한다

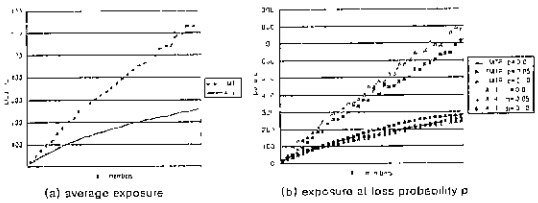
링크 손실 확률 p의 영향을 알아보기 위해 $P_{P,N1}$ 와 $P_{N1,N2}$를 p의 함수로 나타낸 것이 그림 2의 (c) 와 (d) 이다. 여기서 K_B 는 32로 정해진다. p가 0을 넘어서기 시작하면 두 확률은 급격히 증가를 보이면서 1에 근접하고 어느 정도 깅점에 머무르다 일정 구간을 지나면 점차적으로 감소하면서 결국 0에 이름을 볼 수 있다. 이 결과는 RTE 기법이 현실적인 손실 확률의 범위에서 잘 동작하는 것을 함의하고 있다. 특히 그림 2의 산술적 결과는 t 와 u 값이 클수록 두 확률이 큰데 이는 통신의 범위가 광대역일수록 즉 수신자들이 지리적으로 넓게 분포되이(sparsely distributed) 있을수록 RTE 기법이 잘 동작함을 의미한다.

3.2 시뮬레이션

RTE의 유용성을 점증하기 위해 RTE에 기반하는 적응형 트리 복구 기법인 ATRL(Adaptive Tree-based Recovery with Local groups)을 개발하였다 [11]. ATRL은 RTE로 구성된 적응형 트리 구조 속에 지역 그룹(local group)을 조직함으로써 트리 구조의 단점인 반복 지연(cascaded delay)를 줄일 수 있게 한다. ATRL은 대규모 일대다 멀티캐스트 세션의 신뢰성을 효과적으로 전송해 준다.

제안한 ATRL와 정적 트리 기반 멀티캐스트 프로토콜인 TMTP [7]을 시뮬레이션을 통해 비교한다. GT-ITM [10]을 이용하여 400개의 노드를 가진 샘플 네트웍을 만들었다. 주어진 멀티캐스트 세션에서 임의의 한 노드가 송신자로 선택되고 그 송신자는 일정한 비율로 패킷을 그룹으로 멀티캐스트하도록 프로그램 하였다. 제안한 ATRL과 TMTP에 대해 수신자가 동적으로 세션에 참가하도록 하면서 시뮬레이션 하였다. 수신자의 최대값은 300으로 하였다. 이 시뮬레이션에서 링크의 지연값은 20ms에서 100ms사이의 임의의 값을 갖도록 하였다. 링크의 손실확률은 각각의 링크에 동일하게 적용하였는데, 그 값은 0.01에서 0.1까지의 값이다. 손실은 임의로 일어난다고 가정하였다. 호스트에서 읽어서 이것을 네트웍에 쓸 때까지의 지연값은 10ms로 정하는데 이 값은 실제로 실험한 측정값이다. 단순하게 하기 위해서, NACK를 보내기 위한 타이머의 값은 요청자와 재전송자 간의 왕복지연시간으로 정하였다. 억제(suppression)타이머의 값은 0에서 왕복시간의 반값의 사이에서 선택되었다.

시뮬레이션 결과는 그림 3과 4에 나타나 있다. 그림 3 (a)는 acknowledgment당 링크들이 implosion된 평균 회수를 나타낸다. 피드백은 ACK와 NACK를 포함한다 멀티캐스트세션의 크기가 커질수록 ATRL과 TMTP 모두 피드백 implosion의 선형적 증가를 보여준다 하지만, TMTP의 implosion 오버헤드는 ATRL보다 점점 더 커진다. 특히, NACK 의 양은 현저하게 차이가 나게 된다. 그 차이는 전위 현상의 증가와 TMTP에서 NACK를 제한된 범위의 멀티캐스트(limited scope multicast)의 범위에서 비롯된다. 그림 3 (b)에서는 다양한 손실 확률(0.01-0.1)에 따른 implosion을 보여준다. 예상대로 손실 확률의 증가에 따라서 implosion도 증가한다



(a) average exposure (b) exposure at loss probability p

그림 4 Average data exposure

그림 4 (a)는 손실된 패킷을 복구하기 위한 재전송 패킷이 평균 몇 개의 링크를 지났는지 그 회수를 보여준다 결과는 그림 21과 비슷하다 그러나, ATRL이 TMTP보다 더 낫다는 것을 알 수 있다 데이터 exposure의 총합에서 ATRL과 TMTP가 차이를 보이는 것은 implosion과 같은 이유에서 비롯된다. 그림 4 (b)는 손실 확률의 변화에 따른 exposure의 변화를 보여준다. 둘 다 손실 확률에 비례하고, 흥미롭게도 exposure에 미친 손실 확률의 영향이 ATRL에서보다 TMTP에서 더 컸다.

4 결론

본 논문은 신뢰성 있는 멀티캐스트 프로토콜이 사용자의 수에 증가에 따른 성능 지하를 최소화하기 위한 손실 복구 방법을 지원하기 위하여 멀티캐스트 라우팅 트리와 유사한 복구 트리를 효율적으로 구성하기 위한 기법을 제안하였다.

참고문헌

[1] B Rajagopalan, "Reliability and Scaling Issues in Multicast Communication," *ACM SIGCOMM 92*, August 1992

[2] C. Papadopoulos, G Parulkar, and G. Varghese, "An Error Control Scheme for Large-Scale Multicast Applications," *IEEE INFOCOM 98*, March 1998

[3] S Floyd, V. Jacobson, C. Liu, S McCanne, and L Zhang, A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing, *ACM SIGCOMM 95*, pp. 342-356, August 1995

[4] M Hoffman, "A Generic Concept for Large-scale Multicast," *Int. Zurich Seminar on digital communications 96*, February 1996

[5] S K Kasera, J F Kurose, and D. F Towsley, "Scalable Reliable Multicast Using Multiple Multicast Groups," *ACM SIGMETRICS 97*, June 1997

[6] J. C. Lin and S. Paul, "RMTP· A Reliable Multicast Transport Protocol," *IEEE INFOCOM 96*, March 1996

[7] R Yavatkar, J Griffioen, and M Sudan, "A Reliable Dissemination Protocol for Interactive Collaborative Applications," *ACM Multimedia 95*, 1995

[8] B. N. Levine and J J Garcia-Luna Aceves, "A Comparison of Reliable Multicast Protocols," *ACM Multimedia Systems*, 1998

[9] S. Deering, Host Extensions for IP Multicasting, *Request For Comments 1112*, August 1989

[10] K Calvert and E Zegura, GT-ITM· Georgia Tech Internetwork Topology Models

[11] W Yoon and D Lee, "Adaptive Tree-based Recovery for Scalable Reliable Multicast," *to appear in IEEE ICCCN 99*, October 1999