이질적인 분산 시스템에서의 개선된 브로드캐스트 알고리즘 (Improved Broadcast Algorithm in Distributed Heterogeneous Systems)

박 재 현, 김 성 천
(Jae-Hyun Park) (Sung-Chun Kim)
서강대학교 공과대학 컴퓨터학과

요 약

이질적인 분산 컴퓨팅 환경에서의 브로드캐스트, 멀티캐스트 등과 같은 효율적인 그룹 통신은 매우 중요하다. 기존의 휴리스틱 알고리즘들은 각 단계에서의 최적의 해를 선택하기 때문에 지역적 최적(locally optimum)에 빠질 수 있는 단점이 있다. 본 논문에서는 보다 합리적이고, 유용성 있는 edge 선택 기준을 제시하여, 효율적인 브로드캐스트 트리를 구성해주는 개선된 휴리스틱 알고리즘을 제안한다.

1. 서 론

일반적으로, 이질적인 노드와 네트워크를 지닌 환경에서의 멀티캐스트, 브로드캐스트와 같은 그룹 통신을 효율적으로 수행하는 것은 대단히 중요하다. 이질적 분산 시스템에서의 최적의 브로드캐스트 스케쥴을 찾는 문제는 NP-complete하기 때문에, FEF(Fastest Edge First), ECEF(Earliest Completing Edge First), look-ahead와 같은 여러 가지 휴리스틱 알고리즘들이 제안되었다[3]. 이와 같은 휴리스틱 알고리즘은 노드와 네트워크 링크의 이질성을 고려한 통신 구조를 기반으로 하고 있다.

본 논문에서는, 이질적 컴퓨팅 환경에서의 브로드캐스트를 위해 유용성 있는 edge 선택 기준을 제시한, 개선된 알고리즘을 제안한다. 새로이 제안한 휴리스틱 알고리즘은 지역적 최적화(local optimum)에 빠지는 문제를 해결하여 성능 향상을 꾀하였다. 본 논문의 목표는 모든 메시지가 전달된 시간 즉, 완료 시간(completion time)을 최소화하는 것이다.

2. 이질적 분산 시스템을 위한 통신 모델

N개의 노드를 지닌 이질적인 분산 시스템을 생각해보자. G 내에서의 edge (v_i, v_j)는 노드 P_i와 노드 P_j사이의 경로(path)를 나타낸다. edge (v_i, v_j)의 값 C_{ij} $(0 \le i,j < N)$는 P_i로부터 P_j까지의 브로드캐스트 메시지를 보내는 시간을 나타낸다. 이러한 정보는 식 (1)과 같은 통신 행렬(communication matrix) C 로 나타낼 수 있다.

$$\mathbf{C} = \begin{bmatrix} 0 & 146 & 325 & 39 \\ 146 & 0 & 163 & 115 \\ 325 & 163 & 0 & 257 \\ 39 & 115 & 257 & 0 \end{bmatrix} \text{식 (1)}$$

3. 기존의 휴리스틱 알고리즘

3.1 FEF(Fastest Edge First)

FEF 휴리스틱은 각각의 단계마다 집합 A에 속한 노드 P_i와 집합 B에 속한 노드 P_j에 대한 가장 작은 값을 지닌 edge (i,j)를 선택한다. 이와 같은 edge의 선택은 매 단계상의 송신 노드와 수신 노드를 결정하게 된다. 각각의 통신 단계에서 C_{ij} 만큼의 시간이 걸리게 된다. 브로드캐스트 알고리즘에서는 N-1 단계가 걸리게 되며, 전체 수행 시간은 $O(N^2 \log N)$이 된다.

3.2 ECEF(Earliest Completing Edge Frist)

ECEF 휴리스틱 알고리즘의 구조는 FEF 휴리스틱과 유사하다. 단지, ECEF 휴리스틱에서의 edge의 선택은 각 edge의 값과 송신 노드의 준비 시간(ready time)에 의해 결정된다. 이 휴리스틱 역시 $O(N^2 \log N)$의 수행 시간이 걸린다.

3.3 Look-ahead 알고리즘

Look-ahead 알고리즘은 ECEF 휴리스틱에서 향상된 알고리즘이다. 알고리즘의 매 단계에서, 집합 B에 속한 각각의 노드 P_j를 위해 look-ahead 값인 L_j가 계산된다. 이 값은 집합 B에서 집합 A로 옮겨지면 보다 "**유용할**" 노드 P_j를 결정할 수 있도록 해준다. 매 단계에서, 알고리즘은 먼저 집합 B에 속한 모든 노드에 대해 L_j 값을 계산한다. ECEF 휴리스틱 이나, FEF 휴리스틱 에서와 같이, edge는 그 후에 선택되게 된다. 이러한 look-ahead 함수는 $O(N^2)$의 계산 복잡도를 가지며, 따라서 전체 수행 시간은 $O(N^4)$이 된다.

4. 새로운 휴리스틱 알고리즘

집합 A에 속한 각각의 송신 노드들에 대해서, 집합 B에 속한 수신 노드들로의 edge 값들을 비교한다. 이때 가장 작은 값을 가지는 $C_{ij(\min)}$ 값과 그 다음으로 작은 $C_{ij(2nd\min)}$을 찾아서 이것들의 차를 구한다. 이 때의 '차'는 edge를 선택하기 위한 **유용성**이 된다. 이를 기준으로 edge를 선택한다.

$$| C_{ij(\min)} - C_{ij(2nd\min)} | \quad \text{식(2)}$$

개선된 휴리스틱 알고리즘의 과정은 다음과 같다.

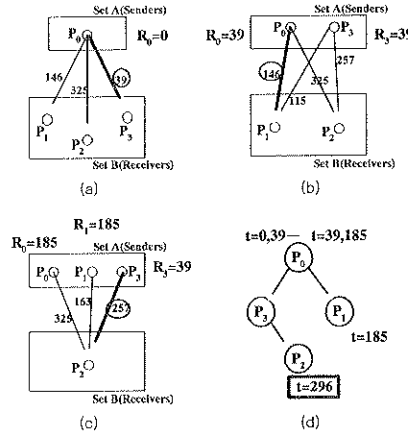
[단계 1] 근원지 노드에서 목적지 노드로의 최소 값을 지니는 edge를 선택한다. Look-ahead 방식을 사용하여도 좋다.

[단계 2] N-3번 반복한다.

{각 단계마다 식(2)에 의해 edge를 결정한다.}

[단계 N-1] edge와 R_i의 합의 최소 값을 지니는 edge를 선택한다.

[그림 1]는 4 개의 노드를 가진 시스템에서의 브로드캐스트를 위한 개선된 휴리스틱 각각의 단계를 보여주며, 식(1)의 비용 행렬을 사용하였다.



[그림 1] 4개 노드를 위한 개선된 휴리스틱 알고리즘을 적용한 통신 스케쥴

개선된 알고리즘 역시 처음에 각 노드로부터 나가는 edge들을 오름차순으로 정렬한다. 따라서 이 단계에서 $O(N^2 \log N)$의 시간이 걸리게 된다. 나머지 시간도 look-ahead를 제외한 다른 알고리즘과 유사한 과정을 지닌다. 따라서 개선된 새로운 휴리스틱 알고리즘의 전체 수행 시간은 $O(N^2 \log N)$이 된다.

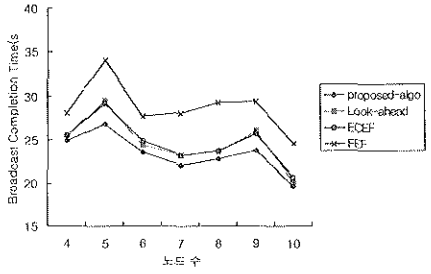
5. 성능 평가

5.1 성능 평가 요소 및 가정

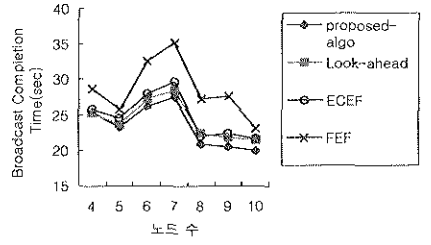
기존의 휴리스틱 알고리즘인 FEF, ECEF, look-ahead 알고리즘과 새로이 제안한 개선된 휴리스틱 알고리즘을 이질적 분산 시스템의 통신 모델을 기반으로 생성한 대칭적(symmetric) 비용 행렬과 비대칭적(asymmetric) 비용 행렬에 대해 각각 적용하였다. 통신 행렬을 위해 쌍방향 네트워크 지연시간(pair wise network latencies)은 10 μsec에서 1msec의 범위 내에서 임의로 선택하였고, 대역폭은 10kB/s에서 200MB/s의 범위 내에서 임의로 선택함을 가정한다.

5.2 성능 평가 결과

ECEF 알고리즘과 look-ahead 알고리즘의 완료 시간이 각각의 노드 수에 대해 거의 비슷하게 나타나고, 모든 경우에서 FEF가 가장 나쁜 결과를 보여주고 있다.



[그림 2] 대칭적 비용 행렬에서의 완료시간/노드 수(1MB)



[그림 3] 비대칭적 비용 행렬에서의 완료시간/노드 수(1MB)

[그림 2]에서 새로운 알고리즘은 기존의 ECEF나 look-ahead에 비해 좋은 결과를 보여 주고 있다. 이 때, 새로운 휴리스틱 알고리즘이 기존의 look-ahead나 ECEF에 비해 평균 6%의 성능 향상을 보이고 있다. [그림 3]에서 새로운 휴리스틱 알고리즘이 기존의 look-ahead 알고리즘에 비해 4%의 성능향상을 보이며, ECEF 알고리즘에 비해 평균 6%의 성능 향상을 보이고 있다.

5. 결론

새로운 휴리스틱 알고리즘은 look-ahead 알고리즘에 비해 평균 4~6%의 완료 시간에서의 성능 향상을 보였으며, 계산 복잡도를 $O(N^3)$에서 $O(N^2 \log N)$로 낮추었다. 하지만 제안한 알고리즘 역시 휴리스틱한 알고리즘이기 때문에, 항상 최적의 값을 보일 수는 없으며, 특정 통신 비용에 대해서는 기존 기법에 비해 높은 완료 시간을 보일 수 있다. 이러한 점은 각각의 시스템의 특성을 파악하여, edge 선택 시에 필요한 변수들에 적절한 가중치를 부여함으로써 보완될 수 있을 것이다.

6. 참고 문헌

[1] M.Banikazemi, V.Moorthy, and D.K.Panda, "Efficient Collective communication on heterogeneous networks of workstations", *In Proc. Intl.Conf. Parallel Processing*, pp.460-467, 1998.

[2] Pangfeng Liu, Da-Wei Wang, "Reduction Optimization in Heterogeneous Cluster Environments", *Proceedings of the 14th International Parallel & Distributed Processing Symposium*, pp.477-482, May 2000.

[3] Prashanth B. Bhat, C. S. Raghavendra, Viktor K. Prasanna, "Efficient collective communication in distributed heterogeneous systems", *Proceedings of the 19th IEEE International Conference on Distributed Computing Systems*, pp.15-24, May 1999.