다중 NIC를 위한 효율적인 데이터 분배 알고리즘

차 윤 준, 김 양 섭, 이 진 영, 김 영 찬
중앙대학교 컴퓨터공학과

Data Distribution Algorithm in Multiple NIC

YounJoon Cha, YangSeub Kim, JinYoung Lee, YoungChan Kim
Dept of Computer Science and Engineering, Chung-Ang Univ.

요약

하드웨어 기술의 발전으로 서버 시스템의 연산능력은 발전을 거듭하고 있다. 또한 인터넷 사용의 광범위한 발전으로 인한 웹에 대한 폭발적인 사용 증가는 네트워크 서버의 연산 능력에 대한 요구와 더불어 향상된 네트워크 대역폭을 요구하게 되었다. 네트워크 장비의 발전도 진일보하고 있지만, 10Mbps, 100Mbps, 기가비트 이더넷등을 거치는 표준의 변화와 함께 기존의 장비에 대한 전면적인 교체 등으로 성능향상을 위해 많은 비용의 소요를 감수할 수밖에 없는 상황에 처해 있다. 클러스터의 한 예인 Beowulf 프로젝트와 같은 경우에, 기존의 네트워크 인터페이스를 병렬적으로 사용함으로써 큰 비용 없이 더 큰 네트워크 대역폭을 얻기 위한 목적으로 이더채널(Channel-bonding)과 같은 기술이 개발되어 사용되기도 하였으나, 어디까지나 클러스터링을 위한 부수적인 기술로써 다소의 성능 향상에 만족하였고 심도있는 연구와 개발은 이루어지지 못하였다. 본 논문에서는 강력한 컴퓨팅 파워를 요구하는 특별한 서버 시스템이 아닌, 일반적으로 큰 네트워크 대역폭만을 요구하는 네트워크 서버에서 기존의 네트워크 인터페이스를 병렬적으로 이용하고, 채널-본딩에 비해 개선된 data distribution algorithm을 제안함으로써 성능의 향상을 꾀하고, 더불어 이러한 기술을 IEEE에서 제정중에 있는 802.3ad Link Aggregation 표준에 적용시키기 위한 기초를 마련하고자 한다.

1. 서론

인터넷의 급속한 발전으로 대용량 네트워크를 필요로 하는 멀티미디어 서비스가 요구되는 시점이 되었다. 또한, 연구소 등의 전유물이던 고속 통신망이 가정마다 저렴한 가격으로 보급되고 있다. 과학계산등의 목적을 위해 엄청난 연산능력을 지닌 슈퍼컴퓨터가 필요했던 것처럼, 일반 사용자를 위해 평범해진 멀티미디어와 같은 서비스를 제공하기 위해 대용량의 입출력을 감당할 수 있는 네트워크 대역폭이 필요해진 시점이 온 것이다. 물론, 연산장치의 고속화, 기억장치의 대규모화와 함께 네트워크 장비 또한 발전을 거듭해 예전과 비교할 수 없는 수준에 이르렀지만, 기하급수적으로 증대될 대역폭 증설 요구를 수용하기 위해서는 시스템에 장착되는 NIC(Network Interface Card)로부터 망 장비에 이르기까지 전면적인 교체가 불가피하다. 그러나, 서버 시스템의 서비스 다변화 등으로 매우 큰 대역폭이 필요할 경우, NIC를 추가적으로 설치해 병렬적으로 활용할 수 있다면, 이는 최소의 비용으로 장치 이용률을 극대화할 수 있는 방안임에 분명하다. 많은 연구가 이루어 진 바 있는 클러스터링에 대한 연구의 일부로써 기존의 NIC를 병렬적으로 사용하는 채널 본딩과 같은 기술은 이미 연구된 바 있다. 그러나, 부수적인 기술로 연구됨으로써 그 연구가 제한적인 수준에서 이루어졌고, 대역폭을 획기적으로 증설할 목적으로 도입하기에는 부족한 것이 사실이다.

본 연구에서는 기존의 이더넷 채널 본딩 기술을 분석하고 문제점을 제시하며, 다중 NIC를 더 효율적으로 활용할 수 있는 방안의 일환으로 새로운 data distribution algorithm을 제시함으로써 이러한 기술을 네트워크 서버의 대역폭을 최소한의 비용으로 증설할 수 있는 방안으로 발전시키고자 한다.

2장에서 기존의 채널 본딩의 기술적인 배경과 원리에 대해 기술하고, IEEE에서 제정중인 802.3ad Link Aggregation 표준에 대해 설명한다. 3장에서는 기존의 여러 가지 방안이 내포하고 있는 장단점을 설명하고 그 문제점을 제시하며 이를 극복할 수 있는 새로운 알고리즘을 제시한다. 4장에서는 기존의 방법과 새로 제시된 알고리즘의 차이를 비교·설명하여 우월성을 보이고자 한다.

2. 기반 연구

2.1 리눅스 이더넷 채널 본딩

Beowulf 프로젝트는 USRA(Universities Space Research Association)과 NASA에 의해 운영되는 CESDIS(Center of Excellence in Space Data and Information Sciences)에서 1994년 ESS(Earth and space sciences project)를 위해 16개의 노드를 가진 클러스터링 서버를 개발하기 위한 목적으로 시작되었다. 최초의 Beowulf는 DX4 프로세서들과 10Mbit/s 이더넷으로 구성되었으나 이 프로세서는 당시의 이더넷 장비를 하나만 활용하기에 여유가 있었으므로 시스템의 균형을 위해 새로운 드라이버를 제작해 트래픽을 둘 이상의 이더넷 장비로 분산시킬 수 있도록 리눅스상에서 초기의 Channel-bonded 이더넷이 구축되었다. 이후, 100 Mbit/s 이더넷이 충분한 가격대 성능비를 가지게 되어 채널 본딩의 필요성이 줄어들었으나, 네트워크 속도가 클러스터의 전체의 성능을 좌우하므로 이는 여전히 유용한 개념이다.

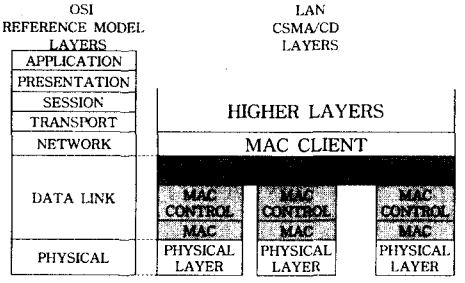
이 기술은 다중 NIC의 사용을 위해 최대한 간단하고 효율적인 방법론을 취한다. 프로토콜계층과 드라이버의 중간 계층에 가상의 NIC를 위한 본딩 드라이버가 존재해 각 NIC에 대한 data distribution을 행함으로써, 본래의 NIC와 이를 위한 디바이스 드라이버를 그대로 사용하면서도, Multiple NIC를 이용한 병렬적인 입출력을 수행할 수 있다. 즉, 기존의 프로토콜과 하드웨어간의 인터페이스에 대한 변화 없이도 다수의 장치를 통합해 성능의 향상을 꾀하는 것이다.

네트워크 장치의 고속화로, 저렴한 가격으로 패스트 이더넷을 채택할 수 있게 되었고, 이미 표준이 정립된 기가비트 이더넷이나 또는, 10 기가비트 이더넷 등의 장비도 폭발하는 수요에 의해 사용될 것이다. 그러나 증가하는 수요를 큰 비용 없이 증설할 수 있다

는 점에서, 채널 본딩과 같은 개념은 유용하며, 이에 대한 체계적이고 표준적인 접근 방법이 필요한 것은 당연하다. 이러한 필요에 의해, IEEE 802.3ad 워킹 그룹이 결성되었고 Link Aggregation을 위한 프로토콜(LACP: Link Aggregation Control Protocol)을 제정중에 있다.

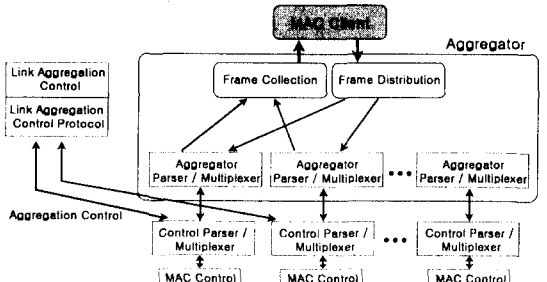
2.2 IEEE 802.3ad Link Aggregation Standard

Link Aggregation이란, 다수의 물리적인 링크를 하나의 논리적인 링크로 결합시킴으로써, 기존의 프로토콜과 NIC를 이용해 더 큰 대역폭을 얻는 것을 말한다. 결국 N개의 기가비트 링크가 결합되어, 이론상으로 최대 N Gbps의 대역폭을 제공한다는 의미이다. 현재 표준이 제정중이며, 아직 이에 대한 산업표준은 없지만, 프로토콜이나 전반적인 네트워킹 장비의 교체 없이 대역폭을 원하는 만큼 증설하기 위한 방안의 일환으로 이미 여러 업체에서 조금씩 다른 접근방법을 통해 실용화하고 있는 단계이며 이러한 움직임이 통일된 표준의 제정을 더욱 가속화시킬 수 있으리라고 본다.



[그림 1] Architectural position of Link Aggregation Sublayer

IEEE 표준은 확립된 단계는 아니며, 세부적인 방법론을 정의하고 있지 않으나, 골격이 되는 sublayer를 통해 발전 방향을 제시하고 있다. [그림 1]의 Link Aggregation의 architectural position은 기본 접근방법에 있어, 기존의 채널 본딩과 크게 다르지 않다. 결국, 제정중인 표준과 기존의 기술이 추구하는 목적도 동일하다. 기존 체제의 큰 변화 없이, 소프트웨어적인 방법만으로 다중 링크를 논리적인 하나의 링크로 사용함으로써 본래의 네트워크가 제한하는 것보다 더 큰 대역폭을 얻고자 하는 것이며, 가장 큰 목적은 BOD(Bandwidth On Demand)를 통한 비용의 절감이다.



[그림 2] Link Aggregation sublayer block diagram

IEEE 802.3ad working group에서 제안하고 있는 Link Aggregation의 가장 주된 부분은, [그림 2]에서 볼 수 있듯이, 하나의 논리적 링크와 하부에 존재하는 다중 링크간의 Frame Distribution과 Collection의 기능이다. 이 부분은, 이더넷 프레임을 분산시킴으로써 트래픽을 효율적으로 분배하고 대역폭을 극대화할 것인가를 결정해 전반적인 sublayer의 성능을 결정한다. 제정중인 표준은 이에 대한 획일적인 방법을 제시하지 않고 있고, 이와 유사한 시도를 하고 있는 업체들은 조금씩 다른 방법을 사용하고 있으며, 현재 제시되고 있는 방법으로는 크게 다음의 세 가지 방법이 알려져 있다.

1) Static distribution - PDU(Protocol Data Unit)의 특정한 factor를 근거로 해쉬 함수에 의해 PDU를 각각의 링크로 전송한다.

2) Round-Robin distribution - 전송하는 PDU를 각 링크에 교대로 한 번씩 전송한다

3) Adaptive distribution - 각 링크의 트래픽에 따라 PDU를 가장 트래픽이 적은 링크를 통해 전송한다.

1)의 방법은 해쉬 함수의 키값이 되는 요소의 설정에 따라 다르지만, 다양한 키값을 가지지 못하면 공정한 프레임의 분배를 기대하기 어렵고, 반면 2)의 방법은 그러한 단점을 가지지 않는다. 그러나, 1) 2) 모두 간단한 구조로 인해 유동적으로 변화하는 상황에 대해 능동적으로 대처하지 못한다는 단점을 안고 있다. 3)의 방안은 다중 링크의 각 부하에 따라 동적으로 분배가 행해지므로 가장 공정한 트래픽의 분배가 이루어질 것이나, PDU 전송시 각 링크의 트래픽을 측정한다는 것은 상당한 오버헤드를 유발하게 되므로, 전자의 두 방법보다 효과적이라고 보기는 어렵다.

현재 적용되고 있는 방법들을 살펴볼 때, 앞에서 예로 든 리눅스 이더넷 채널 본딩은 2)의 Round-Robin 방법을 사용하고 있으며, 내부 구조가 공개되지 않은 다른 업체들의 구현방법의 경우에도 이와 유사한 방법을 사용했다고 설명하고 있다. 이는 구현이 간단하며, 오버헤드가 적고 비교적 공정한 트래픽의 분배를 기대할 수 있다. 또, IEEE 표준안은 기존의 network layer에 대한 backward compatibility와 MAC 프로토콜 상에서의 향후 LACP의 적용을 고려해 단일 conversation내 frame ordering을 유지할 것을 권고하고 있으나, 결과적으로 1)의 방법만이 이를 충족한다.

이러한 장단점, 향후 제정될 표준안의 적용 등을 고려하여, 본 논문에서는 다음의 조건을 만족하는 새로운 data distribution algorithm을 제시하고자 한다.

첫째로, a)각 물리적인 링크에 부과되는 실질적인 트래픽을 고려하는 부하 균등화의 능력이다. 이는 NIC를 병렬적으로 사용하는 단순한 범주에서 벗어나, 사용자의 필요에 따라 현재의 표준이 제한하는 범주를 능가하는 대용량의 대역폭을 얻는 확장성을 고려한 자연스러운 요건이다. 선행조건으로 트래픽의 측정에 대한 오버헤드를 최소화한다는 전제가 필요한 것은 물론이다. 둘째로, b)최소한 단일 conversation내에서 MAC frame을 주고받는 데 있어 MAC frame들의 order가 유지되는 방향으로 운영되어야 한다. - conversation이란, 하나의 end host로부터 통신 상대방에게로 MAC frame이 전송되는 데 있어서 전송될 때에 MAC frame간의 순서가 정해져 있는 단위 communication을 말한다 - 이는 IEEE 802.3ad 표준에서 정의하고 있는 권고안이며, 정의된 sublayer 내에서 LACP등의 세부적인 방법론을 구현하기 위한 조건이 된다. 다중 NIC에 대해 PDU를 분배하는 방법으로써 제시되고 있는 세 가지 방법 중 유동적인 트래픽의 균등화를 지향하는 방안은 3)의 Adaptive distribution이며, frame ordering을 보장하는 방안은 1)의 Static distribution 방법뿐이다. 그러나, 전자는 트래픽 측정으로 인한 오버헤드를 피하기 어렵고, 후자는 distribution factor에 따라 공정한 distribution이 이루어지지 않을 수도 있다. 이것이, Linux의 이더넷 채널 본딩이나 이와 유사한 다른 업체들의 상업용 제품들이 2)의 Round-robin방법을 채택하고 있는 이유가 된다.

3. Hybrid distribution algorithm

본 논문에서는 아래의 세 가지 조건을 만족하면서도 기존의 접근방법이 가지는 단점을 해결하는 새로운 hybrid algorithm을 제시하고자 한다.

ㄱ) 단일 conversation내 frame ordering을 최대한 보장.
ㄴ) 개별 링크의 트래픽에 대한 부하 균등화
ㄷ) 트래픽 측정에 의한 오버헤드를 방지

새로운 데이터 분배 알고리즘에서는 이더넷 프레임의 ordering 유지를 전제 조건으로 Static distribution의 방법을 채택하되, 적정한 수준에서의 부하 균등화를 수행하는 방법을 취하였다. 트래픽에 따라 PDU를 분배하는 접근으로써는 ㄱ)의 조건을 충족시킬 수 없다. PDU를 분배하는 것이 아닌 conversation을 분배하는 접근이 시도되어야 한다. Conversation의 단위는 리눅스 운영체제의 경우에 모든 네트워크 응용 프로그램의 기본적인 인터페이스가 되는 BSD 소켓 인터페이스의 socket connection을 단위로 볼 수 있다. 이러한 정보는 물론 이더넷 PDU에는 포함되어 있지 않으나 NIC를 통해 frame을 전송하기까지 유지되는 정보로써, PDU를 각 링크에 분배하기 위해 사용하는 것이 가능하다.

Conversation이 발생하면, 이를 위해 가장 적은 트래픽을 갖는 링크를 결정하고 이후로 conversation의 모든 PDU는 이 링크를 통해서만 전송된다. 시스템 내에 서비스를 위해 하나의 conversation만이 발생되어 오랜 시간 유지된다면, 이는 올바른 부하 균등화를 기대할 수 없지만, 이런 종류의 네트워크 서비스는 찾아볼 수 없다. 수많은 단발적인 conversation이 형성되는 상황을 고려해 볼 때, 위의 접근 방법은 필요로 하는 모든 요건을 만족시킨다.

Strategy
- Conversation이 시작될 때 트래픽이 최소인 링크 선택
- 이후, 해당 conversation의 모든 PDU를 선택된 NIC로 전송

Procedure

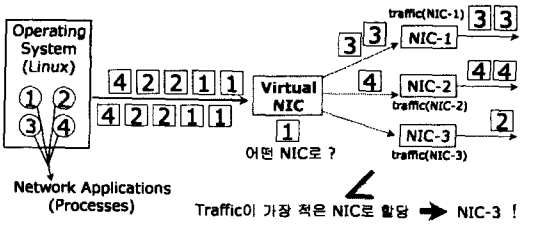
```
( Virtual NIC의 Transmit function에서 구현되어야 한다. )
NICforPDU : Target NIC ( NIC[1] ~ NIC[n] )
minTrafficNIC : NIC[1] ~ NIC[n] 중 가장 트래픽이 적은 NIC

if PDU = NewConversationPDU
    // PDU starting conversation
    NICforPDU = minTrafficNIC
    SendPDU( PDU, NICforPDU )
else
    // PDU in the middle of conversation
    SendPDU( PDU, NICforPDU )
```

[그림 3] New Hybrid distribution Algorithm pseudo-code

[그림 3]는 새로운 hybrid distribution algorithm의 세부적인 절차를 pseudo-code로써 설명하고 있다. 이러한 알고리즘의 실제 구현은 Link Aggregation sublayer중 Frame Distribution의 부분에서 구현되어야 한다. 앞의 예에서 보인 이더넷 채널 본딩의 경우라면, 본딩 드라이버의 transmit function에서 구현되어야 한다.

[그림 4]는 본 논문에서 제안한 알고리즘에 따라 이루어지는 behaviour를 묘사하고 있다. 각 링크에 부과된 트래픽을 측정하는 것은 새로운 conversation이 형성되는 시점에서 한 번씩만 이루어지며, 해당 conversation이 종료될 때까지 이 링크로만 전송된다.



[그림 4] Hybrid distribution Algorithm behaviour

4. 평가 및 결론.

소스 코드와 함께 상세한 구현 원리가 공개되어 있는 리눅스의 이더넷 채널 본딩의 경우, 간단한 Round-robin의 방식을 채택하고 있다. 다중 NIC의 수가 적을수록, 이 방법은 간단하고도 효율적으로 전송되는 이더넷 프레임을 실제 NIC로 분배한다. 그러나, 이미 앞에 기술한 바 대로, IEEE 802.3 ad 표준에서 권고하고 있는 "frame ordering"의 문제를 그대로 방치해 차후 이 표준안을 수용해 발전시키기 어렵다는 문제점을 않고 있다. 또한, 통합되는 링크의 수가 많아질수록, 단순한 구조로 인해 각 링크에 대한 효율적인 트래픽의 분배가 이루어지기 어렵게 되며 동적인 상황에 대한 대비가 이루어지지 않는다. 새롭게 제안된 알고리즘에서는 보다 공정하고 능동적인 트래픽의 분배를 실현함과 동시에 향후 확립될 표준안의 요구 조건을 수용하였다.

본 연구는 발전해 가는 하드웨어 기술의 한편에서 다중 NIC를 한 시스템 내에서 보다 효율적으로 통합해 사용함으로써 네트워크 인터페이스의 교체에 의한 성능의 향상보다도, 필요한 만큼 추가적인 성능의 향상을 보다 저렴한 비용으로 얻는 데 필요한 기반 기술을 연구하는 데 그 목적이 있다. 기존의 방법에 대해 건설적인 발전 방향을 제시함으로써 새롭게 정해지고 있는 표준을 수용하는 한 방안을 제공하는 것이다. IEEE 802.3ad 표준에서 제안하고 있는 전체 sublayer중 한 부분을 발전적으로 구체화하였지만, 앞으로, 전반적인 Link Aggregation Sublayer를 구체화하고 이를 실제 이용 가능하도록 적용해 보아야 할 것이다.

5. 참고 문헌

[1] IEEE Draft P802.3ad/D1.9CB, "Supplement to CSMA/CD Access Method & Physical Layer Specifications : Link Aggregation" June 24, 1999

[2] Salvatore Salamone, "Load Balancing's Inexpensive Performance Boost", http://www.byte.com/art/9409/sec4/art3.htm

[3] Intel, "Solving Server Bottlenecks with Intel Server Adapters"

[4] Khalil El-Khatib and Carl Tropper, "On Metrics for the Dynamic Load Balancing of Optimistic Simulation", Proceedings of the 32nd Hawaii International Conference on System Software, 1999

[5] F.Silla and J. Duato, "Improving the Efficiency of Adaptive Routing in Networks with Irregular Topology", Universidad Politeenica de Valencia