다중 NIC에서 효율적인 결함 허용 메카니즘

이진영° , 김양섭 , 차윤준 , 김영찬
중앙대학교 컴퓨터 공학과

The Effective Fault Tolerace Mechanism of Multiple NIC

Jin-Yeong Lee° , Yang-Sub Kim , Youn-Joon Cha , Young-Chan Kim
Dept. of Computer Science and Engineering, Chung-Ang Univ.

요약

최근 인터넷의 초고속 성장과 멀티미디어 데이터의 급격한 증가로 인해서, 고속의 전송매체와 이를 최적으로 이용하기 위한 인터페이스 시스템이 요구되고 있다. 이에 따라, 이더넷이 기가비트 이더넷으로 발전되어 LAN 인터페이스 시스템의 고속화를 이루고 있다. 그러나, 폭발적으로 증가하는 인터넷 환경에서 기가비트 이상의 고속 네트워크 대역폭을 지원하는 NIC(Network Interface Card)가 요구되고 있다. 이를 해결하고자, 기가비트 이상의 고속의 네트워크 대역폭을 지원하는 다중(Multiple) NIC의 연구가 진행되고 있다. 그러나, 고속의 네트워크 대역폭를 지원하는 다중 NIC를 운영할 때, 단일 NIC 결함으로 인해 시스템 운영이 중단되는 현상이 발생할 수 있다. 따라서, 효율적인 결함허용 기법을 적용하여 신뢰성 있는 시스템 운영을 지원할 필요성이 대두되고 있다. 본 논문에서는 기존의 하드웨어 결함 허용기법인 TMR, Primary-Standby Approach, Watchdog Timer 기법에서 발생되는 자원에 대한 가용성과 내구성의 비효율적인 부분을 고려하여, 동적으로 검출주기를 변환하여 다운타임(Downtime)을 최소화 할 수 있는 효율적인 결함 허용 메카니즘을 설계하여 제안하고자 한다.

1. 서론

최근 인터넷의 초고속 성장과 멀티미디어 데이터의 급격한 증가는 고속의 전송매체와 이를 최적으로 이용하기 위한 인터페이스 시스템을 요구하고 있다. 현재 여러 NIC(Network Interface Card)들이 각각의 네트워크 프로토콜을 이용하여 개발되고 사용되고 있다. 그 중에서 가장 많이 이용하고 있는 NIC로는 이더넷(Ethernet) NIC로서 1Gbps급 이더넷 NIC가 개발된 상태이다[1].

그러나, 이보다 더 많은 대역폭을 요구하는 시스템에서 1Gbps급 이상의 단일 대용량 NIC를 개발하여 사용하게 된다면 전에 구축되어 있는 네트워크망과 망에 속해 있는 시스템들의 전체적인 교환이 필요하게 되므로 상당한 비용이 소요되고, 개발과 교환 기간 동안에 요구되는 대역폭을 지원하지 못하기 때문에 많은 손실이 발생할 수 있다. 그러므로, 본 논문에서 고속의 네트워크 대역폭을 지원하기 위한 NIC로서 다중(Multiple) NIC를 기술하고 있다. 또한, 다중 NIC를 사용함으로써 기존 네트워크 환경의 큰 변화없이 고속의 LAN 환경을 구축할 수 있으므로 현 시스템과의 호환성(backward compatiblility)을 유지할 수 있고 오버헤드를 줄일 수가 있다. 아울러, 요구하는 네트워크 대역폭을 지원함으로써 고성능 저비용의 효과를 얻을 수 있다.

대용량 다중 NIC의 SPOF(Singe Point Of Failure)인 결함으로 인해서 시스템 중단이 생기면, 대용량의 멀티미디어 데이터를 서비스하는 시스템인 만큼 엄청난 손실을 가지고 오게 된다. 따라서, 본 논문에서는 결함으로 오는 손실을 방지하기 위해 결함 허용 기법을 사용하여 "결함 허용 다중 NIC" 에 대해 기술하고 있다. 즉, 기존의 하드웨어 결함 허용기법인 TMR, Primary-Standby Approach, Watchdog Timer 기법에서 발생되는 자원에 대한 가용성과 내구성의 비효율적인 부분을 고려하여, 동적으로 검출주기를 변환하여 다운타임을 최소화 할 수 있는 효율적인 결함 허용 메카니즘을 설계하여 제안하고 있다.

본 논문의 구성은 2장에서 다중 NIC에서 하드웨어 결함허용 기법에을 기술하고, 3장에서 결함 발생시 다운타임을 최소화 할 수 있는 효율적인 결함 허용 메카니즘에 대해 기술하며, 마지막으로 평가 및 결론을 내린다.

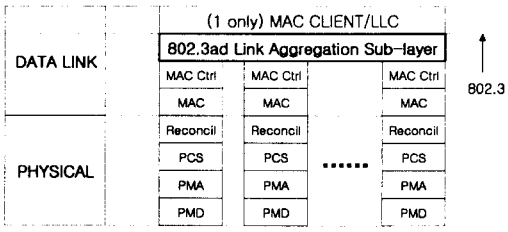
2. 기반 연구

2.1 다중 네트워크 인터페이스 시스템(다중 NIC)

이더넷이 처음 등장한 이후로 계속해서 처리속도 및 성능을 높여오면서 100Mbps Fast 이더넷까지 성능을 향상시켰고, 인터넷의 고속화 광역화 요구의 증대로 이더넷 대역폭을 향상시키기 위해서 기가비트 이더넷이 개발되었다. 기가비트 이더넷이 개발되기 전에도 Fast 이더넷보다 광역의 대역폭이 요구되었으며, Fast Ether Channel이란 메카니즘이 개발되어 다중 NIC가 시스템화 되기 시작했다. 이와 같은 요구는 기가비트 이상의 대역폭으로 발전하였고, 이를 지원하고자 다중 기가비트 NIC가 개발되고 있다. 이들은 대부분 CISCO 이더 채널을 기반으로 하고 있으며, Sun에서 Trunking , Intel에서 Server Adapter , BeoWulf에서 Linux 기반의 Channel Bonding 등이 있다[2][3]. 그러나, 아직 공통된 표준이 없

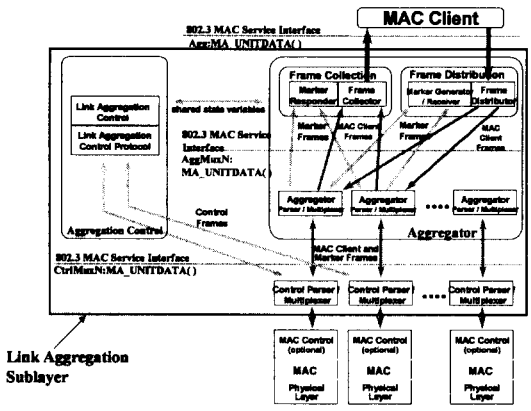
어서 이를 제정하기 위해 IEEE에서 802.3ad이라는 워킹 그룹이 결성되어, LACP(Link Aggregation Control Protocol)에 관한 표준을 제정하고 있다.

LA(Link Aggregation)는 다수의 물리적인 링크(MAC)를 하나의 논리적인 링크(Link Aggregation Sublayer)로 결합시킴으로써, 상위 계층에 있는 MAC Client에게 하나의 Virtual MAC으로 인식시킨다. 그러므로, 네트워 트래픽 전송시 여러 링크로 트래픽을 분산시킬 수 있으며, 트래픽 수신시 여러 링크로 받아들일 수 있다. 즉, N개의 링크가 결합되어 있는 경우, 최대 N x NIC의 대역폭을 제공받을 수 있다. 또한, 각 링크가 full-duplex를 지원할 경우에는 2N x NIC의 대역폭을 제공받을 수 있다.



[그림 1] Link Aggregation

LA 링크 계층 상위에 LAS(Link Aggregation Sublayer)를 추가함으로 써 논리적으로 하나의 Vitual MAC으로 인식하게 된다. LAS는 여러 모듈로 설계되어 있다. Aggregation Control은 LACP를 기반으로 LACPDU(Link Aggregation Control Protocol Data Unit)를 Frame Collection, Frame Distribution, control paser/Multiplexer의 여러 모듈과 통신하여, 다중 링크들을 하나의 Virtual MAC을 만들어줌으로써 인해서 N배의 대역폭을 제공한다 [4].



[그림 2] Link Aggregation Sublayer

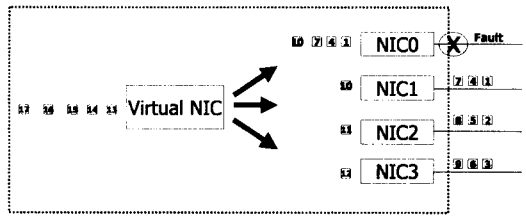
Linux 기반의 Channel Bonding 시스템을 제외한 대부분의 다중 NIC들은 LA 권고안을 고려하여 다중 NIC를 설계하였지만, 아직 표준으로 제정되어 있지 않기 때문에, 호환성에서 문제를 유발하고 있다. 그러나, BeoWulf의 Channel Bonding 시스템은 모든 Ethernet NIC를 지원하고, 시스템이 공개되어 있기 때문에, 본 논문에서 설계한 결함허용 메카니즘은 Channel Bonding 시스템 기반하에 구현하였다.

2.2 다중 NIC에서 하드웨어 결함 허용

일반적인 시스템에 대한 결함 허용(Fault Tolerance) 기법은 결함의 특징에 따라서, 하드웨어 결함 기법, 소프트웨어 결함 허용 기법, 정보 결함 허용 기법 등으로 분류된다. 본 논문에서 제안하고 있는 다중 NIC는 특성상 lan cable 결함, NIC port 결함, 허브(스위치 허브) 결함으로 시스템 운영이 중단되므로 소프트웨어나 정보에 대한 결함보다 하드웨어에 대한 결함이 발생할 수 있기 때문에 하드웨어 결함 허용 기법을 적용하고 있다.

기존의 하드웨어에 관한 fault tolerance 기법으로는 TMR(Triple Modular Redundancy), Primary-Standby Approach, Watchdog Timer로 구분할 수 있다[5]. 이러한 모든 기법들은 자원에 대한 중복(redundancy)을 원칙으로 제안되어 있다. TMR은 세 개의 중복된 모듈을 두어, 각각의 모듈의 출력을 비교해서 다른 출력이 나오는 하나의 모듈을 결함(Fault)으로 인식하고 나머지 모듈에게 운영을 인수하게 된다. Primary-Standby Approach는 모듈을 Primary 모듈과 Standby 모듈로 구분한다. Primary 모듈은 운영하고 있는 모듈이고 Standby 모듈은 결함이 발생했을 때, Primary 모듈의 운영을 인수할 수 있는 대기중인 모듈이다. 그러나 TMR과 Primary-Standby 결함허용 기법은 운영을 담당하고 있지 않은 모듈을 중복으로 대기시켜야 하기 때문에 자원의 낭비를 초래한다.

자원을 모두 활용할 수 있는 기법으로 Watchdog Timer 결함 허용 기법이 있다. Watchdog Timer는 모든 모듈이 운영되고 있는상태에서, 반복적으로 Watchdog packet을 각 모듈로 전송하여 응답이 없는 모듈을 결함으로 인식하여 시스템에서 제거하고, 결함 모듈에서 다시 응답이 오면 시스템에 추가하여 복구시에도 동적으로 시스템이 재구성된다. 그러나, 결함이 발생한 후에도 결함모듈을 반복적으로 검사하므로 응답에 해당하는 Timeout 시간동안 다운타임이 증가하게 되어서 시스템 운영 서비스의 손실이 증대된다. 본 논문에서는 이러한 문제점들을 해결하기 위한 효율적인 하드웨어 결함 허용 기법을 설계하고 제안하고 있다.



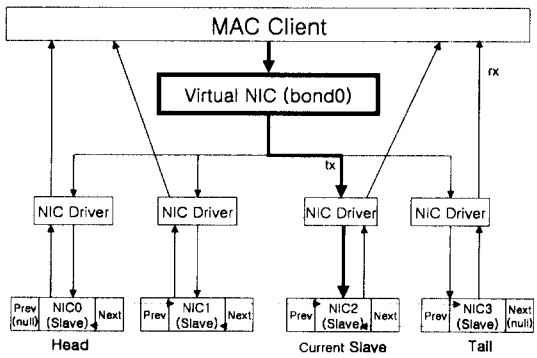
[그림 3] Watchdog Timer

3. 다중 Channel Bonding에서 효율적인 하드웨어 결함 허용

3.1 다중 Channel Bonding 시스템 구성도

Ethernet 기반의 LAN에서 시스템 간에 서로 패킷을 전달하기 위해서 각 NIC의 MAC Address를 참조하여 전송한다. 그러므로, 다중으로 NIC가 병렬화되어 있는 시스템으로 패킷을 전달하고자 할 때 IP에 할당된 하나의 MAC Address 만을 참조하게 되므로, 하나의 NIC만이 활성화되는 현상이 발생한다. 따라서, Channel Bonding에서는 이러한 점을 고려하여, 다중 NIC에 있는 각각의 NIC를 Slave NIC로 설정하여 Slave 큐에 삽입시키고, 이들을 하나

의 논리적인 Virtual NIC로 인식할 수 있도록 Master NIC인 Bond0를 설정한다. Master NIC인 bond0의 MAC Address는 처음으로 Slave 큐에 삽입되는 Slave NIC의 MAC address를 사용하고, Slave 큐에 있는 나머지 Slave NIC는 Master NIC와 같은 MAC Address가 할당된다. 결과적으로, Bonding된 NIC들은 하나의 MAC Address를 할당 받아서, 다중 NIC를 포함한 시스템에 하나의 IP와 하나의 MAC Address로 이루어진 Channel Bonding 시스템이 구성된다. [그림 4]에서 보는 바와같이 channel bonding 시스템은 패킷 전송할 때와 수신할 때 다른 모습을 보여 주고 있다. 전송(tx)할 때는 bond0에 의해서 RR(Round Robin) 방식으로 Slave NIC 가 지정되어 전송을 수행한다. 이 때, 패킷 전송을 수행하고 있는 Slave가 Currenet Slave인데, 시스템이 초기화될 때 Slave 큐의 Head가 Current Slave로 지정되고, 다음 패킷 전송시에는 Next 포인터 값을 참조하여 RR 방식으로 Current Slave가 이동한다. 그러나, 수신(rx)할 때는 상대 시스템의 bond0에 의해서 제어되므로 지정된 경로를 통해서 직접 상위 계층인 MAC Client로 이동한다. 따라서, Bond0는 전송부분만을 제어하게 된다[6].



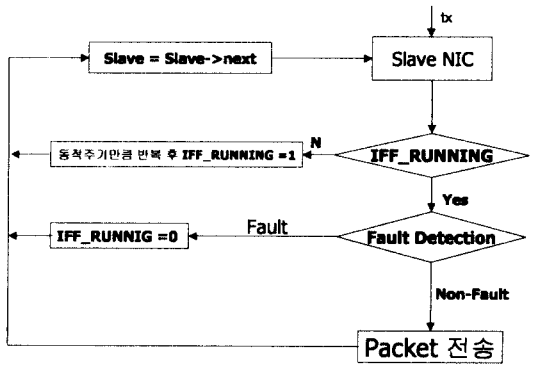
[그림 4] Channel Bonding 시스템 구성도

3.2 다중 NIC에서 효율적인 하드웨어 결함 허용 메카니즘

Channel Bonding 되어 있는 다중 NIC에서 하드웨어 결함이 발생할 수 있는 원인으로는 단일 NIC 자체의 결함, NIC에 연결되어 있는 케이블 결함, Switch(HUB) port의 결함등이 있다. 이러한 하드웨어 결함이 발생하면 공통적으로, 시스템 외부로 패킷을 전송할 수 없다는 에러를 시스템 내부에 저장하게 되므로, 결함을 검출해 내기 위한 Fault Detection 모듈에서는 패킷을 전송할 때마다 발생하는 전송 에러정보를 추출해 내게 된다.

이 때, 결함이 발생하면 다음 Slave로 포인터를 옮김으로써 전송하려는 패킷을 안정적인 Slave NIC로 넘기게 되는데, 이렇게 하면 Watchdog Timer처럼 결함을 감지하기 위해서 매번 Timeout 시간 동안 기다려야 하는 오버헤드가 발생한다. 따라서, 하드웨어 결함이 발생하면 즉시 복구될 수 없으므로 시간이 소요되며, 이시간을 고려한 결함 허용 메카니즘을 필요로 한다.

본 논문에서는 이를 고려하여 결함이 발생하면 동적으로 검출주기를 변환하여, 다운타임을 최소화 할 수 있는 메카니즘을 제시하고 있다. [그림 5]에서 알 수 있듯이, 본 논문에서 제시하고 있는 결함허용 메카니즘은 결함이 발생하면 해당하는 Slave의 IFF_RUNNING 플래그값을 비활성화 시켜서, 결함이 발생한 Slave가 Current Slave가 되면 Fault Detection 모듈을 거치지 않고, Next Slave를 Current Slave로 지정하여, Fault Detection 모듈에서 생기는 Timeout 시간을 소비하지 않고 시스템 운영 서비스를 유지할 수 있다. 결함이 발생한 Slave가 복구 되면 다시 시스템에서 운영을 수행해야 하므로, 결함이 발생한 Slave가 동적인 주기만큼 Current Slave로 지정되면 IFF_RUNNING 값을 다시 활성화시키고 Fault Detection 모듈로 이동시켜 전송 에러를 검출하게 된다. 결과적으로 결함이 발생하면 매번 Fault Detection에서 소요되는 Timeout 시간을 감소시켜서, 시스템 전반적인 다운타임을 최소화 시킬 수 있다.



[그림 5] 동적 주기 변화를 적용한 결함허용 메카니즘

4. 결론 및 향후 연구 방향

고속의 네트워크 대역폭을 지원하는 다중 NIC에서 결함이 발생하면, 자원중복으로 인한 가용성(availability)감소와 Detection 모듈에서 주기적으로 생기는 다운타임이 증가하는 문제가 발생한다. 이런한 문제를 해결하기 위해, 동적으로 검출주기를 변환하여 다운타임을 최소화 할 수 있는 "다중 NIC에서 효율적인 결함 허용 메카니즘"을 제시하였다.

향후에는 시스템 내부적으로 결함 허용 메카니즘이 수행되는 정보를 모니터링하여 GUI(Graphic User Interface)를 통해 사용자에게 제공함로써, 보다 효율적으로 결함정보를 관리하고자 한다. 또한, 본 연구에서 기반으로 한 NIC보다 고속의 NIC에서도 결함 허용을 지원할 수 있는 다중 NIC에 대한 연구에 주력할 것이다.

참고문헌

[1] "Gigabit Ethernet Comes of Age" 3COM, June.1999 Norman
[2] Finn "Port Aggregation Protocol" CISCO Systems May 1, 1998
[3] Ariel Hendel "Link Aggregation Trunking "IEEE 802 - Tutorial Session 11-November-1997
[4] Tony Jeffree ,Rich Seifert "P802.3ad/D2.0Link Aggregation ",IEEE Draft, July 17, 1999
[5] Walter L. Heimerdinger, Charles B. Weinstock "A Conceptual Framework for System Fault Tolerace" CMU/SEI-92-TR-33
[6] Jacek Radajewski, Douglas Eadline "Beowulf HOWTO" 22 November 1998