김해중
KBS보도기술국

실무 네트워크 Design - 9 : 네트워크 가상화 1

가상화란 물리적인 1대의 장비를 논리적으로 여러 대의 장비로 나누거나, 물리적인 여러 대의 장비를 1대의 논리적인 장비로 통합하는 기술입니다. 예를 들면 VLAN은 1대의 물리적인 switch를, 논리적으로 여러 대의 switch로 나누는 최초의 네트워크 가상화 기술이며, FHRP(HSRP, VRRP)는 여러 대의 router를 1대의 가상의 router로 통합하는 네트워크 가상화 기술입니다.

현재 사용되는 네트워크 가상화 기술을 정리하면 [표 1]과 같습니다.

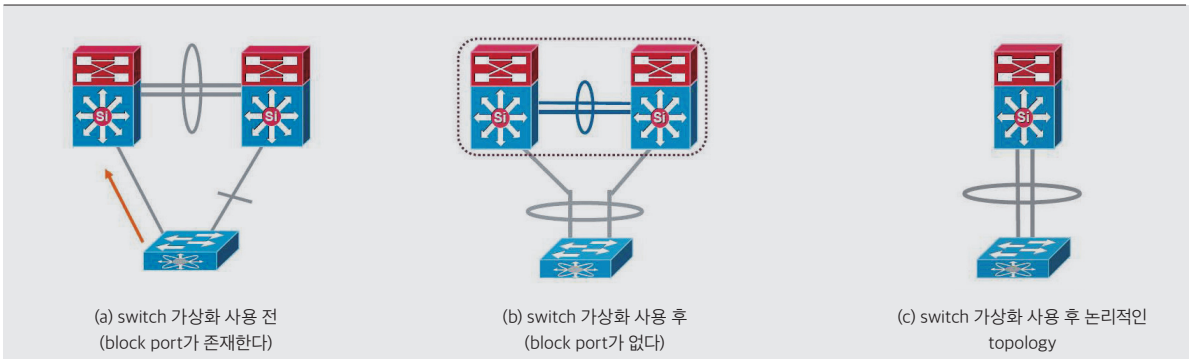
종 류		제 품
통 합	stacking	cisco catalyst 2960/3750
	Ethernet-channel	대부분의 switch
	VSS(virtual switch system)	cisco catalyst 6800, 6500, 4500X
	VPC(virtual port channel)	cisco nexus(5K, 6K, 7K, 9K)
	VC(virtual-chassis)	Juniper EX-series
	TRILL(transparent interconnection of lots of link)	HP, Huawei
	Fabricpath	cisco Nexus(5K, 7K)
	NIC teaming	Intel PROset, BACS/BASP
	Fabric extender	cisco Nexus(5k, 7K)
	bridge Port extension(802.1BR)	Juniper MX series(Fusion topology)
	FCoE(Fiber channel over Ethernet)	Emultx, Qlogic, cisco
	NPIV(N_Port_ID_virtulalization	Mcdata, IBM. Emulex, Cisco
분 산	VLAN(Virtual LAN)	대부분의 switch에서 지원
	VRF(Virtual routing and forwarding)	대부분의 router에서 지원
	VDC(virtual device context)	cisco nexus (7K)
	VIC(Virtual interface card)	cisco(VIC 1240, 1280)

표 1. 네트워크 가상화 기술 종류

이번 시간에는 [표 1]에 언급된 기술 중 access, distribution layer에서 많이 사용되는 2가지 기술만을 설명하겠습니다. 첫 번째는 switch 가상화 기술 중 여러 대의 물리적인 switch를 1대의 논리적인 switch로 통합하는 기술로, 현재 현업에 많이 사용되는 대표적인 기술인 cisco의 stackwise, juniper의 virtual chassis에 대해서 자세히 알아보겠습니다. 두 번째는 여러 개의 물리적인 NIC을 1개의 논리적인 NIC으로 바라보는 NIC teaming에 알아보겠습니다.

switch 가상화(통합)

등장배경



(a) switch 가상화 사용 전
(block port가 존재한다)

(b) switch 가상화 사용 후
(block port가 없다)

(c) switch 가상화 사용 후 논리적인
topology

그림 1

지난 시간에 자세히 설명을 했듯이, [그림 1(a)]와 같이 distribution layer switch 간의 link를 L2 trunk로 하면, loop 방지를 위해 STP가 동작을 해 block port가 존재합니다. 하지만 [그림 1(b)]와 같이 2대의 distribution switch를 1대의 논리적인 switch로 보는 가상화 기술을 사용하면, 물리적인 topology는 [그림 1(b)]지만, 논리적인 topology는 [그림 1(c)]와 같이 됩니다. 이제는 loop가 존재하지 않아서 STP가 필요 없는 디자인이 되며, block port 없이 모든 link를 사용할 수 있게 됩니다.

또 하나의 특징으로는, 기존에는 2대의 switch를 각각 관리해야 된다는 불편함이 존재했다면, 이제는 1대의 switch에 접속을 해서 2대의 switch를 통합관리할 수 있다는 장점도 제공합니다. 그리고 만약 switch 1대에서 장애가 발생했을 때, 가상화 기술을 적용하기 전보다, 짧은 복구시간을 가집니다.

Stackwise(cisco)

cisco의 3750 series switch에서 사용하는 가상화를 stackwise라고 합니다. [그림 2]와 같이 stack cable을 통해서 최대 9대의 장비까지 연결할 수 있으며, 물리적인 cabling은 ring topology로 합니다.

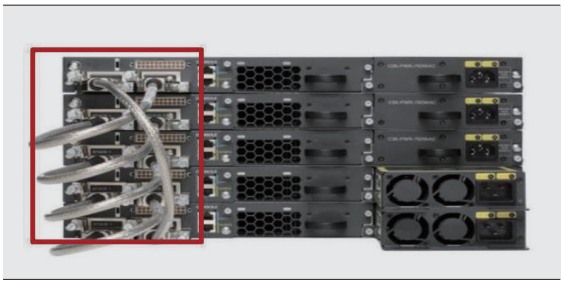


그림 2. stack cable을 통해 최대 9대까지 연결 가능하다

stackwise 내부의 신호 흐름을 보면, [그림 3]처럼 다른 방향으로 회전하는 2개의 ring이 존재하며, 동작방식은 예전에 많이 사용되던 FDDI(fiber distributed data interface)와 유사합니다.

실제 stack cable의 내부 구조는 [그림 4(a)]와 같습니다. 각 TX(송신), RCV(수신)은 각 8pair를 가지고 있으며, 각 pair는 2.5Gbps의 속도로 동작합니다. 각 pair는 10bit를 전송하며, 실제 그중에 8bit만 data 전송용도로 사용하는 8B/10B encoding을 사용해서, 1 pair당 실제 속도는 [2.5G × (8/10) = 2Gbps]이며, 총 8pair를 가지고 있어서, 총 속도는 16Gbps입니다. 각 TX, RCV pair는 다른 signaling 방식을 사용하는 2개의 trace를 가지고 있어서, cable 내부는 총 32개의 trace를 가지고 있습니다.

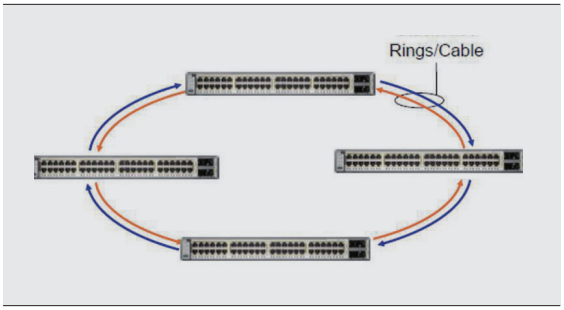


그림 3. stackwise는 2중 ring 구조를 통해 고가용성을 보장한다

김해중
KBS보도기술국

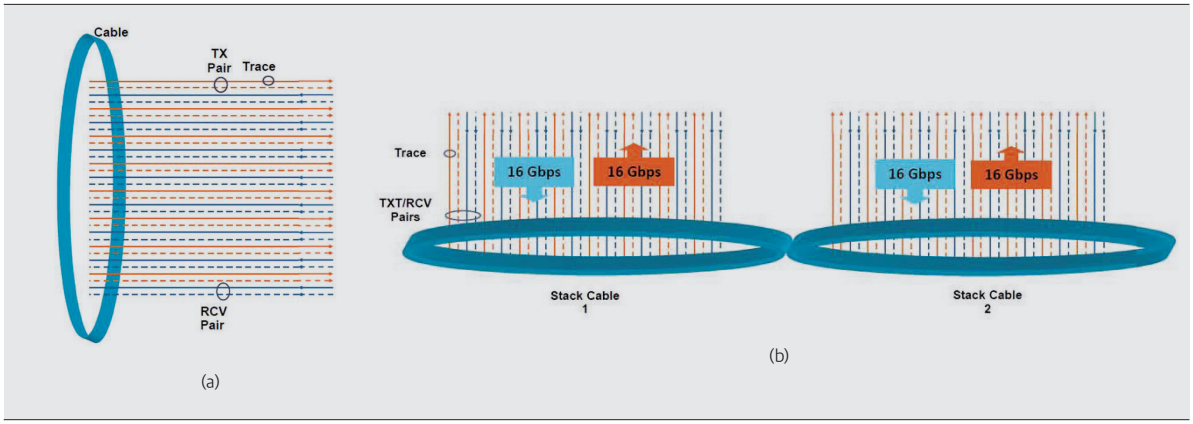
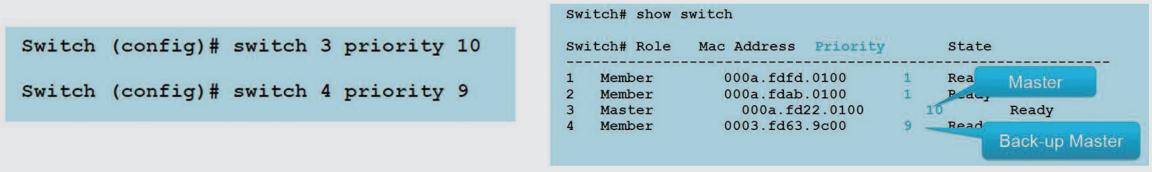


그림 4. stack cable 내부 구조

여러 대의 switch 중에서 priority가 가장 높은 switch를 master switch로 선정을 하고, 나머지는 backup switch로 동작하는 1:N 구조입니다. priority의 기본값은 10이며, 0이면 "stack에 참여 안 한다"는 의미이며, 1에서 15까지의 값을 세팅할 수 있습니다. 만약 priority가 같을 때는 높은 hardware, software 버전을 사용하는 switch가 master switch 역할을 합니다. master switch 장애를 대비해, 그 다음으로 master switch으로 될 member switch의 priority를 높여 두는 대책이 필요합니다.



(a) 각 switch별로 priority를 세팅을 한다.

(b) priority가 가장 높은 장비가 master switch 역할을 한다.

그림 5

stack에서 이더채널과 STP를 사용할 때, switch의 대표 mac 주소는 master switch의 mac 주소를 사용합니다. 문제는 master switch의 장애가 발생했을 때입니다. master switch에서 장애가 발생하면, mac 주소의 변화가 발생해 순간적인 통신의 단절이 발생합니다. 이것을 방지하기 위해 "stack-mac persist timer 0" 명령어를 사용하면, master switch가 down 되어도 해당 mac 주소를 계속 사용할 수 있습니다.

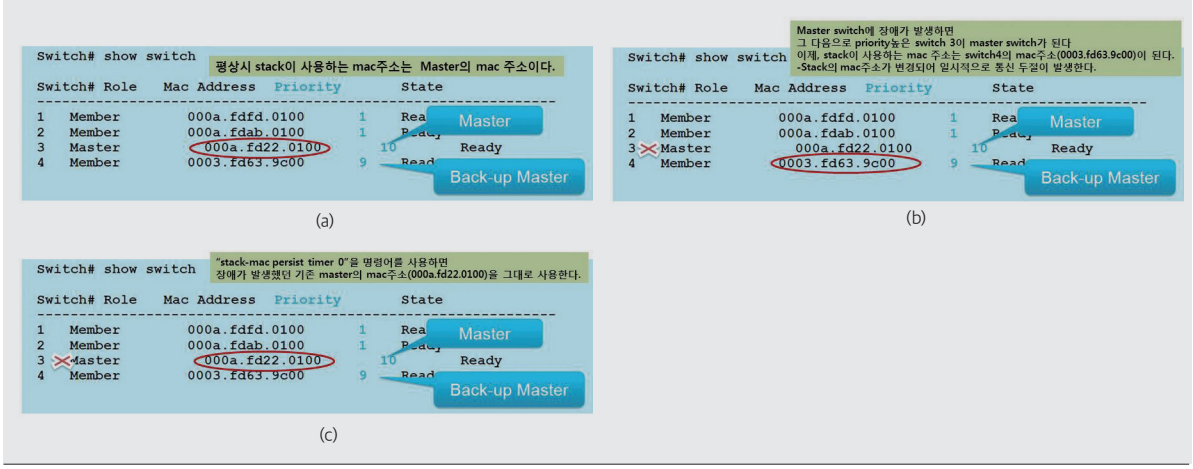


그림 6. "stack-mac persist timer 0" 명령을 사용해, 장애 발생시도 기존 master switch의 mac 주소를 그대로 사용할 수 있다

[그림 7]과 같이 관리자가 master switch에 세팅을 하면, 모든 member switch에 세팅이 자동으로 전달됩니다. 중앙 집중적 관리를 위해 member switch로 console 접속을 할지라도, master switch로 넘어가도록 설계되어 있습니다. 또한 telnet, ssh로 관리할 때도, 1개의 IP로 stack에 참여한 모든 switch를 관리할 수 있습니다. 단 mac-address table 관리와 STP는 각 장비에서 분산 처리된 후에, 장비 간에 update를 통해 sync를 맞춥니다.

그림 7. master switch에 세팅하면, 모든 member switch에 자동으로 전달된다

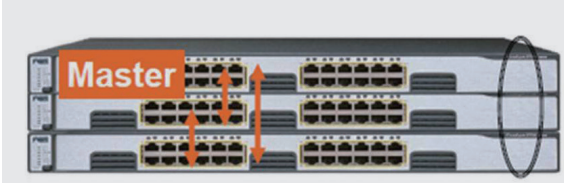


그림 8. mac table 관리와 STP는 각 장비에서 분산 처리됩니다

3750의 stacking은 토큰링/FDDI처럼 source strip 방식을 사용합니다. 즉 source switch로 패킷이 유입되면, ring을 통해 destination switch로 가게 됩니다. destination switch에 도착한 original 패킷은, 다시 ring을 타고 source switch로 도착하면 해당 패킷이 폐기되는데, 이것을 source strip이라고 합니다. 패킷이 ring을 1바퀴 돌다 보니 대역폭 낭비가 발생하는 문제점을 가지고 있습니다.

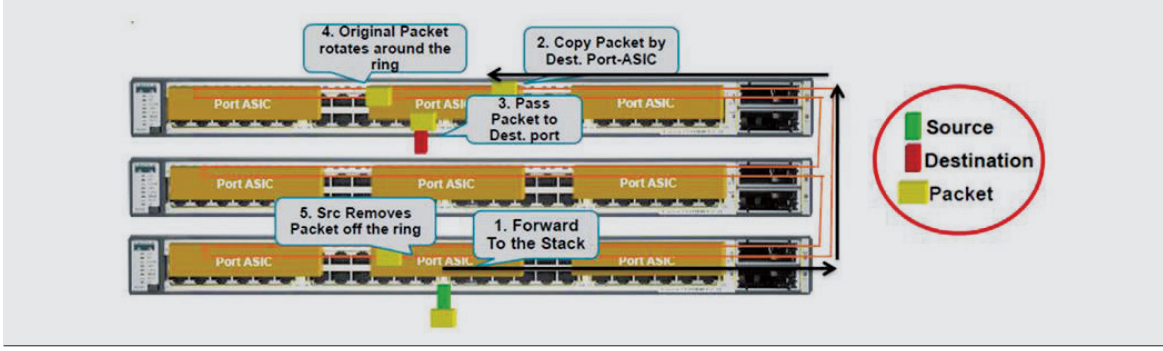


그림 9. source strip

하지만 최근에 출시된 3750x는 RPR(Resilient packet ring)처럼 destination strip 방식을 사용합니다. source switch에서 보낸 패킷이 destination switch에 도착하면, destination switch에서 패킷은 제거되고, source switch에게 패킷을 잘 받았다는 16bit ack 메시지만 보냅니다. 이것을 spatial reuse(공간 재사용)이라고 하며, 대역폭 재사용으로 인해 대역폭 증가효과를 가져오게 됩니다.

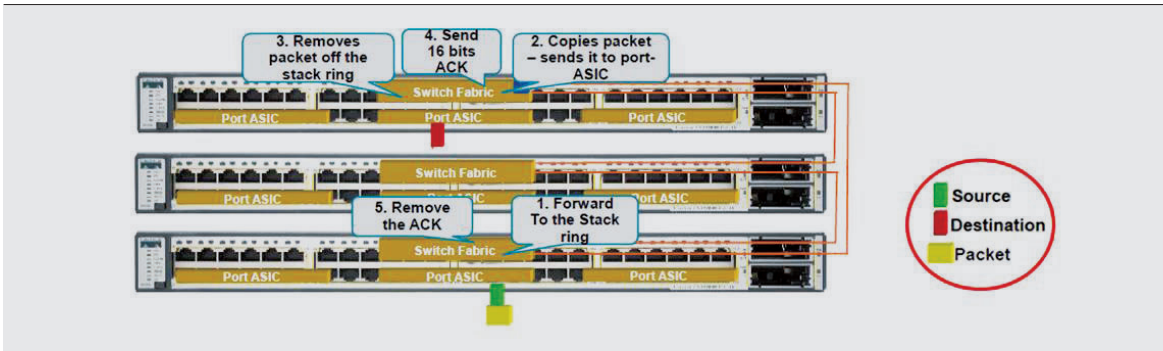


그림 10. catalyst 3750x에서 제공되는 destination strip

그리고 3750x는 destination strip 기능을 확장하여, [그림 11]과 같이, source switch와 destination switch가 동일한 경우에는, local switching 기능을 제공합니다. 이 기능을 제공하지 않았던, 예전 버전의 3750 series는 패킷이 ring을 1바퀴 도는 문제점이 존재 했습니다.

김해중
KBS보도기술국

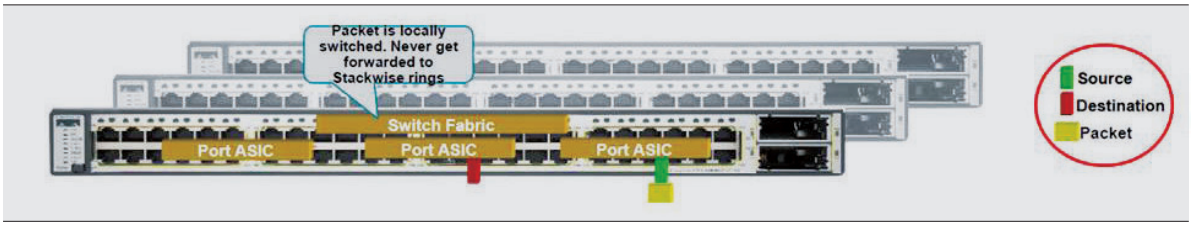


그림 11. 3750x에서 제공되는 local switching

3750에서는 data 신호에 대한 stacking 뿐만 아니라, 최대 4대까지 power에 대한 stacking 기능도 제공합니다. 장비의 power 장애 발생시, 옆 장비에서 power를 공급해주는 기능입니다.

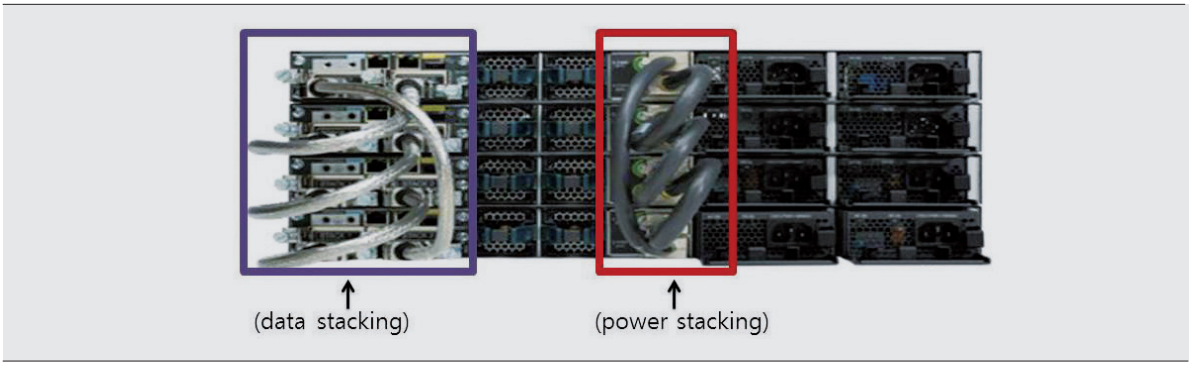


그림 12. data stacking 뿐만 아니라, power stacking도 가능하다

최근에 cisco의 catalyst 2960-S는 flex stack이라는 이름으로, 일종의 stack 기능을 제공하고 있습니다. 단 ring topology가 아니라, hop-by-hop topology를 사용하며, 최대 4개의 member switch를 가질 수 있습니다. stakcwise와 flex stack에 대한 비교는 [표 2]를 참고하시면 됩니다.

	catalyst 3750	catalyst 2960
stack 이름	stackwise	flex-stack
최대 member 수	9대	4대
topology	Ring	Hop by bop
속 도	32Gbps	20Gbps

표 2. catalyst 3750, 2960 stack 비교

Juniper의 VC(Virtual chassis)

실제 현업에서는 cisco switch 뿐만 아니라, juniper switch도 많이 볼 수 있습니다. juniper의 VC(virtual chassis)는 cisco의 stack와 유사한 특징을 가지고 있으며, [표 3]은 juniper의 EX series에서 최대로 연결할 수 있는 장비수를 보여줍니다.

Device	최대 연결 장비 수
EX 2200	4
EX 3300	10
EX 4200, 4300, 4500, 4550, 4600	10
EX 8200	4

표 3

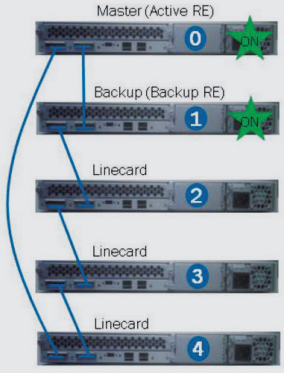
[표 4]는 cisco의 stackwise와 juniper의 virtual chassis를 간단히 비교했습니다.

	Cisco(stack wise)	Juniper(VC)
bandwidth	32Gbps	128Gbps
Role	master, member로 구성	master, backup, member로 구성
master 장애 발생시	member 중에 다시 master 선출	backup이 master 역할을 대체함
master 식별방법	master switch의 LED에 점등	LCD text에 role(master, backup) 표시
최대 연결 장비	9대	10대
연결 cable 이름	stack cable	VCP(virtual chassis port) cable
논리적 topology	ring	ring

표 4. cisco stackwise와 juniper virtual chassis 비교

juniper의 VC에 참여하는 각 switch에는 고유한 ID가 부여되며, 이것을 통해 각 switch는 구분됩니다. VC에 참여한 여러 대의 switch 중에 priority가 가장 높은 1대를 master로 선출하고, 그 다음으로 priority가 높은 장비를 backup으로 선출하며, 나머지 장비는 LCC(line card) 역할을 합니다. priority는 1~255 사이의 값을 가질 수 있으며, 기본값은 128입니다. master를 만들 switch의 priority를 255로 할당하는 것을 권장하며, backup으로 만들 switch의 priority는 254로 할당하는 것을 권장합니다.

선출된 master와 backup 장비 간에는 keepalive를 주고받으며, 이것을 통해 master 장애 발생 시, 짧은 fail-over time을 가지게 됩니다. virtual chassis에 대한 세팅에 관한 권장 순서는 [그림 13]과 같습니다.



Step 1. master switch 설치

- master로 만들 switch를, 가장 먼저 booting 해서 member ID 0을 얻게 한 후에, priority를 255로 할당합니다.

Step 2. backup switch 설치

- VCP cable로 master switch와 연결한 후에, 전원을 공급하면 member ID 1을 얻게 됩니다. backup으로 만들기 위해 priority를 254로 할당합니다.

Step 3. line card switch 설치

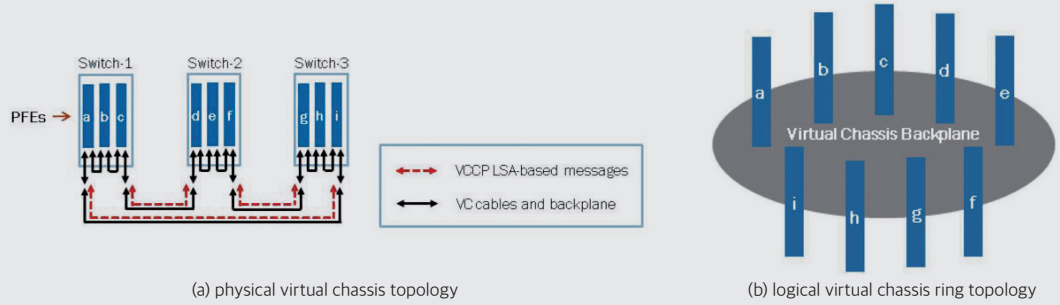
- master, backup switch와 VCP cable로 연결한 후에, 전원을 공급하면 member ID 2를 얻게 됩니다.

Step 4. 추가 line card switch 설치

- 앞의 step 3을 반복합니다.

그림 13. Virtual chassis 권장 세팅 순서

VC는 link state protocol 특징을 가지는 VCCP(virtual chassis control protocol)를 사용해서, 각 member switch를 찾은 후, 서로 간에 VC topology를 [그림 14]와 같이 공유합니다.



(a) physical virtual chassis topology

(b) logical virtual chassis ring topology

그림 14. VCCP Protocol을 이용해, 서로 간에 discovery 후 topology를 만듭니다

김해중
KBS보도기술국

만들어진 topology를 이용해, [그림 15]와 같이 SPF(shortest path first) tree를 만든 후, member switch 간에 shortest path switching을 하게 됩니다. 단 hop count가 동일할 때는 load-balancing이 되지 않으며, 낮은 member ID가 우선합니다. 참고로 VCCP는 IS-IS protocol을 변형해서 만들었습니다.

SPF가 어떻게 동작하는지 [그림 16]을 통해 확인해보겠습니다. 만약 member switch0(ge0/0/10)에서 member switch3(ge3/0/14)로 갈 때는 가장 짧은 거리인 switch 4를 경유하며, 그림에 빨간색으로 표현되어 있습니다, 만약 member switch0(ge0/0/28)에서 member switch 2(ge2/0/47)로 갈 때는 가장 짧은 거리인 switch 1을 경유하며, 그림에는 파란색으로 표현되어 있습니다.

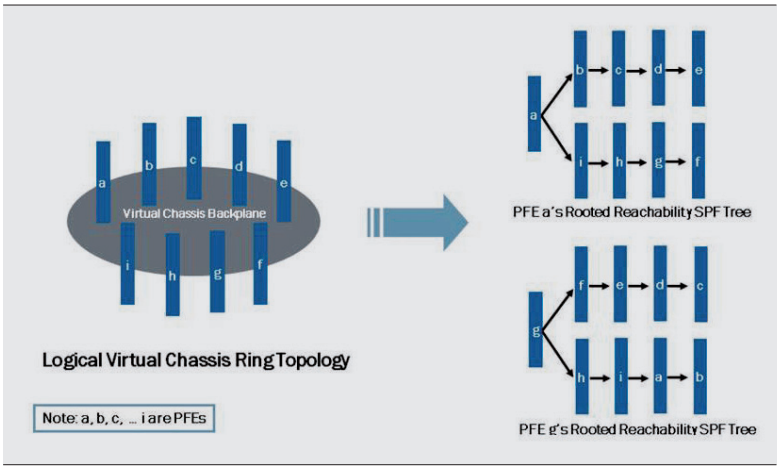


그림 15. VCCP로 인해 만들어진 SPF Tree

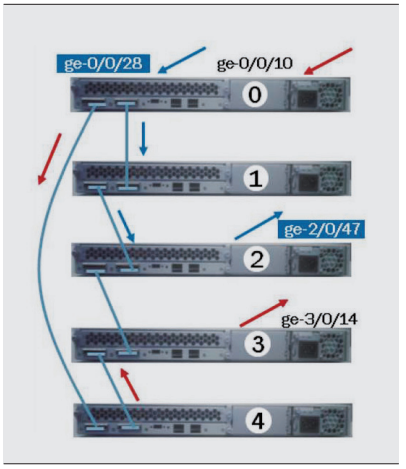


그림 16. VC는 shortest path switching을 한다

NIC teaming(bonding)

[그림 17(a)]와 같이 서버에서 1개의 NIC 만을 사용할 때는, uplink switch에서 장애가 발생하거나, 서버 NIC에 장애가 발생하면, 모든 통신이 두절되는 문제점이 발생합니다. 이것을 해결하기 위해, [그림 17(b)]와 같이 여러 개의 물리적인 NIC을 1개의 논리적인 NIC으로 묶어서 사용하는 NIC teaming(bonding)을 사용합니다.

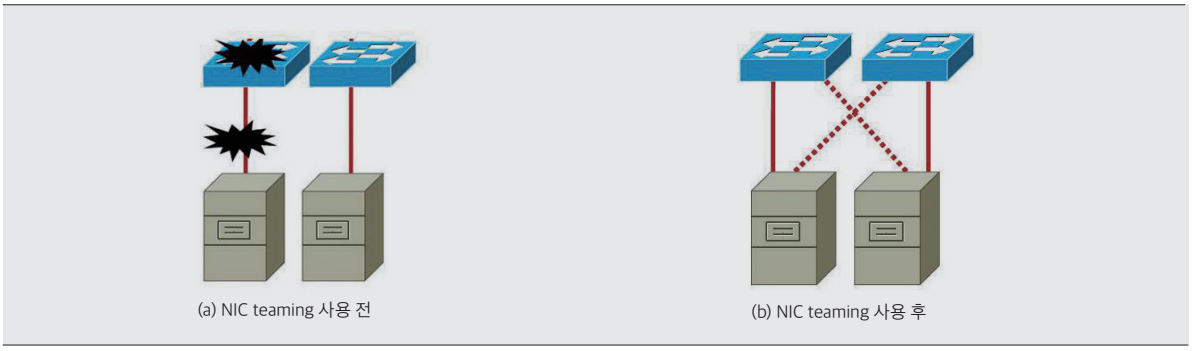
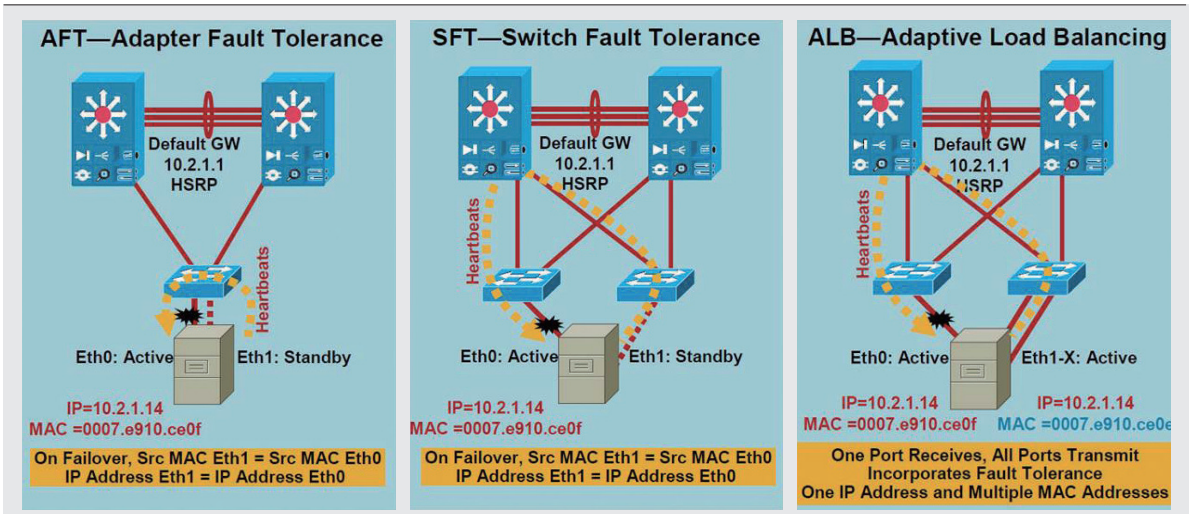


그림 17. NIC teaming

NIC teaming은 [그림 18]과 같이 3가지 방식이 존재합니다.

첫 번째 방식은 AFT(adapter fault tolerance)이며, [그림 18(a)]와 같이 2개의 NIC은 동일한 1대의 switch에 연결됩니다. 이때 1개 NIC은 active 역할을 하고, 나머지 NIC은 active NIC에 장애가 발생하면 active 역할을 대체하는 standby 역할입니다. 보통 때는 active NIC만 사용되는 문제점이 있지만, 장애 처리와 관리가 편하다는 장점이 있습니다.



(a) AFT (b) SFT (c) ALB

그림 18. NIC teaming 기본 3가지 방식

두 번째 방식은 [그림 18(b)]에 표현된 SFT(switch fault tolerance)이며, AFT와의 차이점은 각 NIC이 다른 switch에 연결된다는 것입니다. 이때 loop를 방지하기 위해 STP를 동작시켜야 합니다.

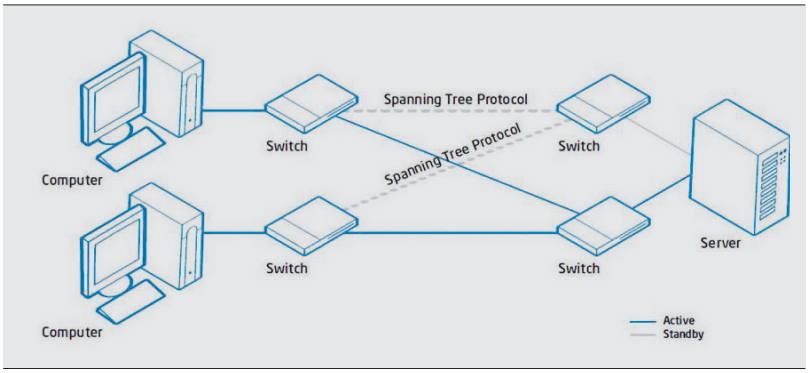


그림 19. SFT를 사용할 때는, loop 방지를 위해 STP를 사용해야 합니다

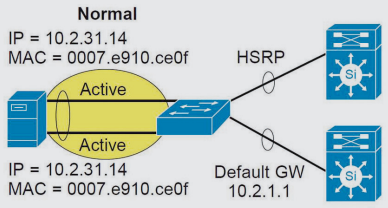
세 번째 방식은 ALB(adaptive load balancing)로, [그림 18(c)]처럼 2개의 NIC이 모두 송신이 가능한 active/active 방식입니다. 실제 모든 통신을 active/active로 처리하는 것이 아니라, 서로 다른 VLAN에 대한 unicast 패킷만을 active/active로 처리하고, 나머지(broadcast, multicast, 동일 VLAN에 대한 송신패킷)는 primary로 정의한 NIC에서만 실제 처리합니다.

ALB는 arp request를 받았을 때, primary로 선택한 NIC의 mac address로 arp reply를 보내, 모든 수신 패킷은 primary NIC으로만 수신되는 문제점이 존재합니다. 하지만 추가적으로 RLB(receive load balancing) 기능을 작동시키면, teaming 멤버 중 load가 적은 member의 mac 주소로 arp reply를 보내서, 수신 패킷도 load balancing이 가능합니다. 참고로 RLB의 동작방식은 cisco의 GLBP(gateway load balancing protocol)와 유사합니다.

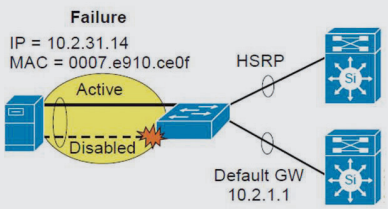
추가적으로 [그림 20]과 같이 스위치와 NIC간의 이더채널 teaming도 가능합니다. 이더 채널의 특징으로 인해, AFT, ALB 방식을 통합하는 특징을 가져서, 송신뿐만 아니라 수신까지도 active/active 처리가 가능해집니다. 하지만 단점으로는 물리적인 1대의 switch로만 연결이 가능하다는 한계가 존재합니다. 하지만 앞에서 설명한 stackwise와 virtual-chassis를 이용하면 여러 대의 switch가 1대의 switch로 인식되기에, 그 한계점마저도 극복이 됩니다.

이더채널 teaming을 구현 하려면, 당연히 switch에도 이더채널 세팅이 필요하며, static과 dynamic 방식 중의 1가지를 선택해야 합니다.

김해중
KBS보도기술국



(a) 평상시의 이더채널 teaming 동작



(b) 장애시의 이더채널 teaming 동작

그림 20. 이더채널 방식의 teaming

앞에서 설명한 4가지 방식의 teaming의 기능의 상관관계를 [그림 21]로 표현해 보았습니다. 이더채널 teaming이 가장 많은 feature를 제공함을 보여줍니다. 각 방식에서 사용되는 IP/MAC 주소의 개수를 [표 5]에 비교했습니다.

종 류	IP 개수	MAC-address 개수
AFT(adaptive fault tolerance)	1개	1개
SFT(switch fault tolerance)	1개	1개
ALB(adaptive load balaincing)	1개	2개
이더채널(link aggregation)	1개	1개

표 5. NIC teaming 에서 사용되는 IP/MAC 주소 개수 비교, 2NIC teaming 기준임

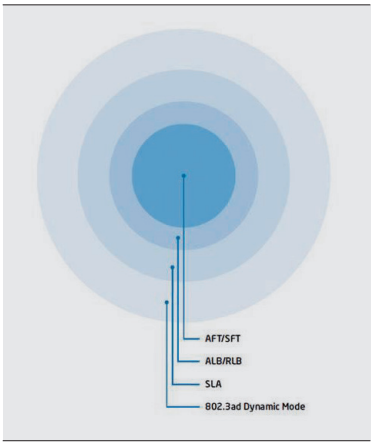
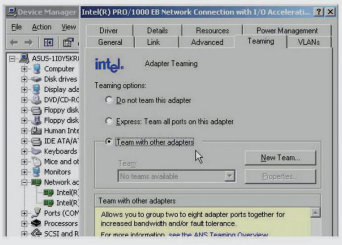
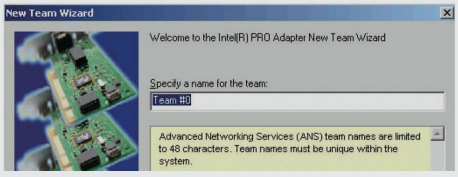


그림 21. 여러 가지 teaming 방식의 상관관계 (이더채널이 가장 많은 feature를 가진다)

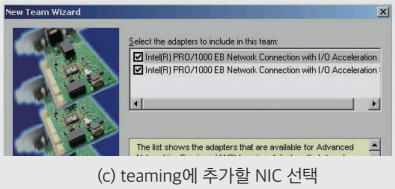
가장 널리 사용되는 teaming software인 INTEL PROset에서의 NIC teaming 세팅과정을 [그림 22]에 정리했습니다.



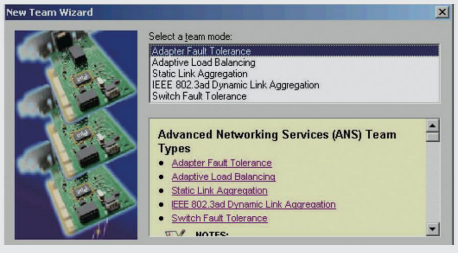
(a) teaming 활성화 하기



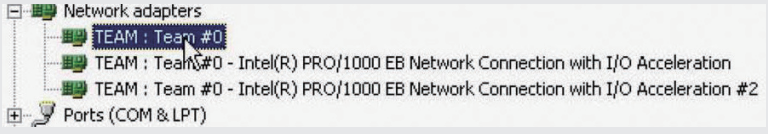
(b) teaming 이름 할당



(c) teaming에 추가할 NIC 선택



(d) teaming 방식 선택



(e) 최종 세팅 완료 후

그림 22. teaming software(Intel PROset)에서의 NIC teaming 세팅

만능으로 보이는 teaming도 1가지 문제점을 가집니다. [그림 23(a)]와 같이 switch의 uplink에 장애가 발생되면, 통신두절 현상이 발생합니다. 서버에 직접적으로 관련된 link의 down이 아니기에, teaming의 failover는 발생하지 않기 때문입니다. 이것에 대한 해결책으로 [그림 23(b)]와 같이 uplink에서 장애가 발생하면 down link도 자동으로 down 시키는 "trunk failover" 기술이 등장하였습니다.

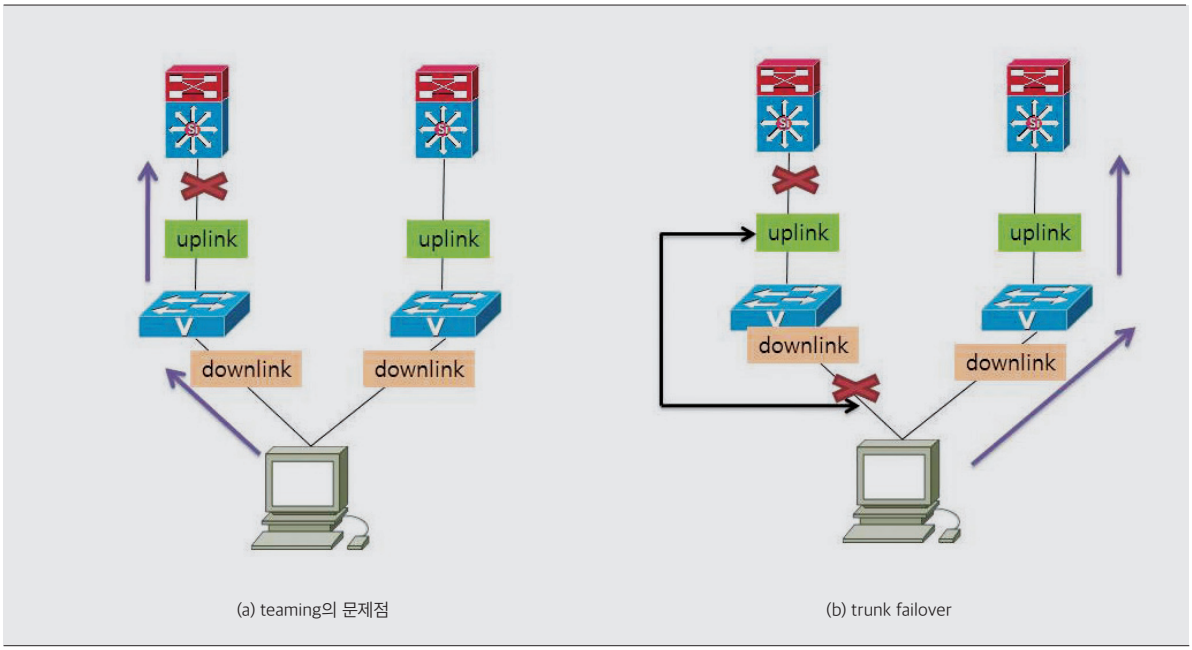


그림 23. trunk failover를 통해 teaming의 한계를 극복할 수 있습니다

서버에서 사용하는 NIC은 마더 보드에 포함된 NIC(on-board NIC)과 확장 슬롯에 증설할 "확장 NIC"으로 나눌 수 있습니다. 이 2가지를 믹스하여 Teaming을 하면, 가용성이 보장됩니다. 예를 들어 쿼드 포트(Quad Port)의 확장 NIC만으로 Teaming을 했을 때, 확장 슬롯에 고장이 나면 통신 두절이 발생하지만, on-board NIC과 확장 NIC을 함께 teaming을 하면, 장애에 대한 영향을 최소화 할 수 있습니다.

이번 시간에는 access, distribution layer에서 자주 사용되는 기술에 대해서 설명했다면, 다음 시간에는 회사의 core/distribution layer에서 주로 사용되는 스위치 가상화 기술인 VPC, VSS를 자세히 설명하도록 하겠습니다.