I/O Acceleration for VM

포티넷 코리아 이상훈

(우진) 이상훈

Fortinet SE

woojinlsh@gmail.com

KRDAG
(Korea Datacenter Architect)

Amway Korea
(Technical Architect, 인프라총괄)

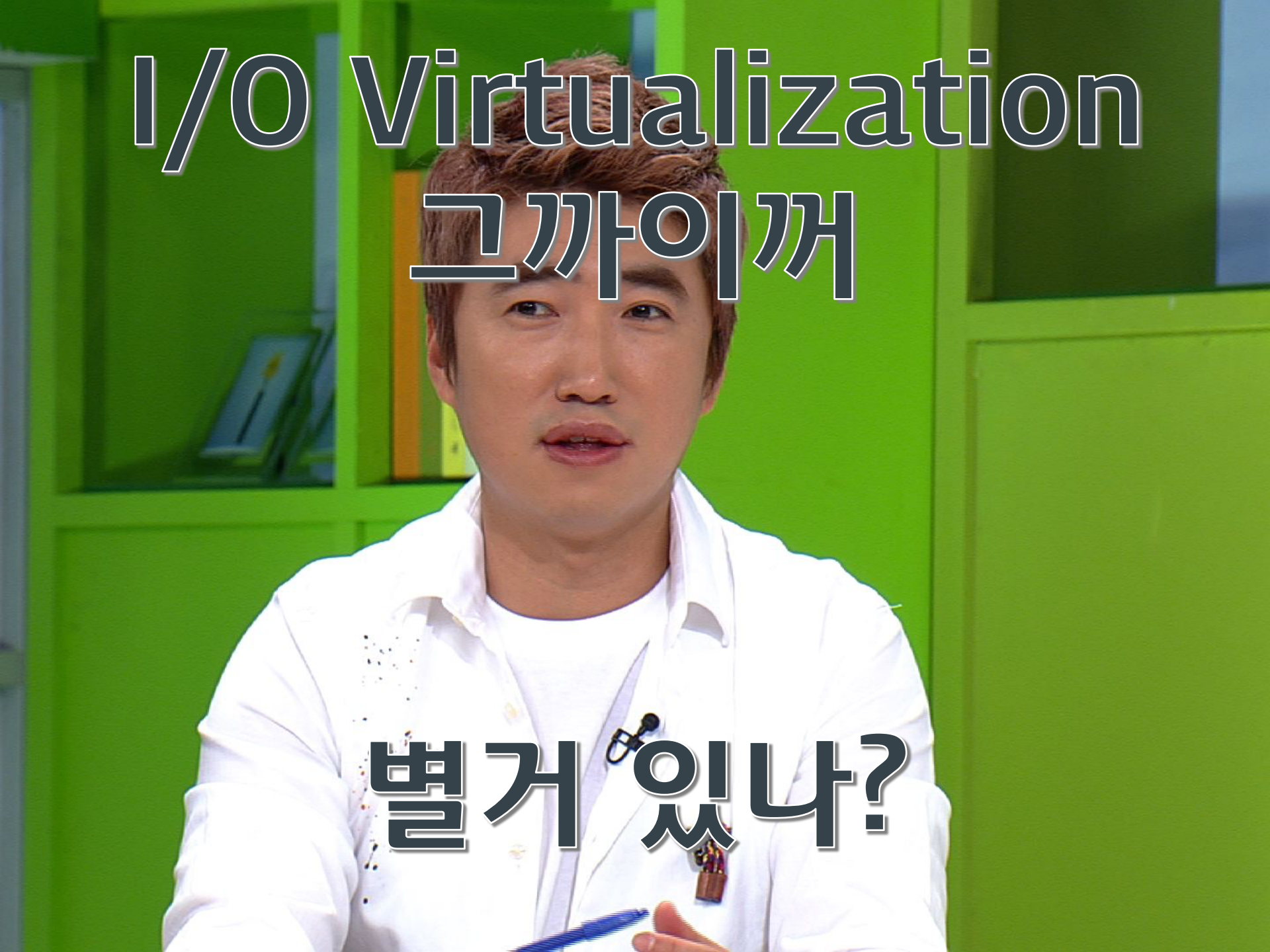
Infnis
(Solution Launch Support Engineer)

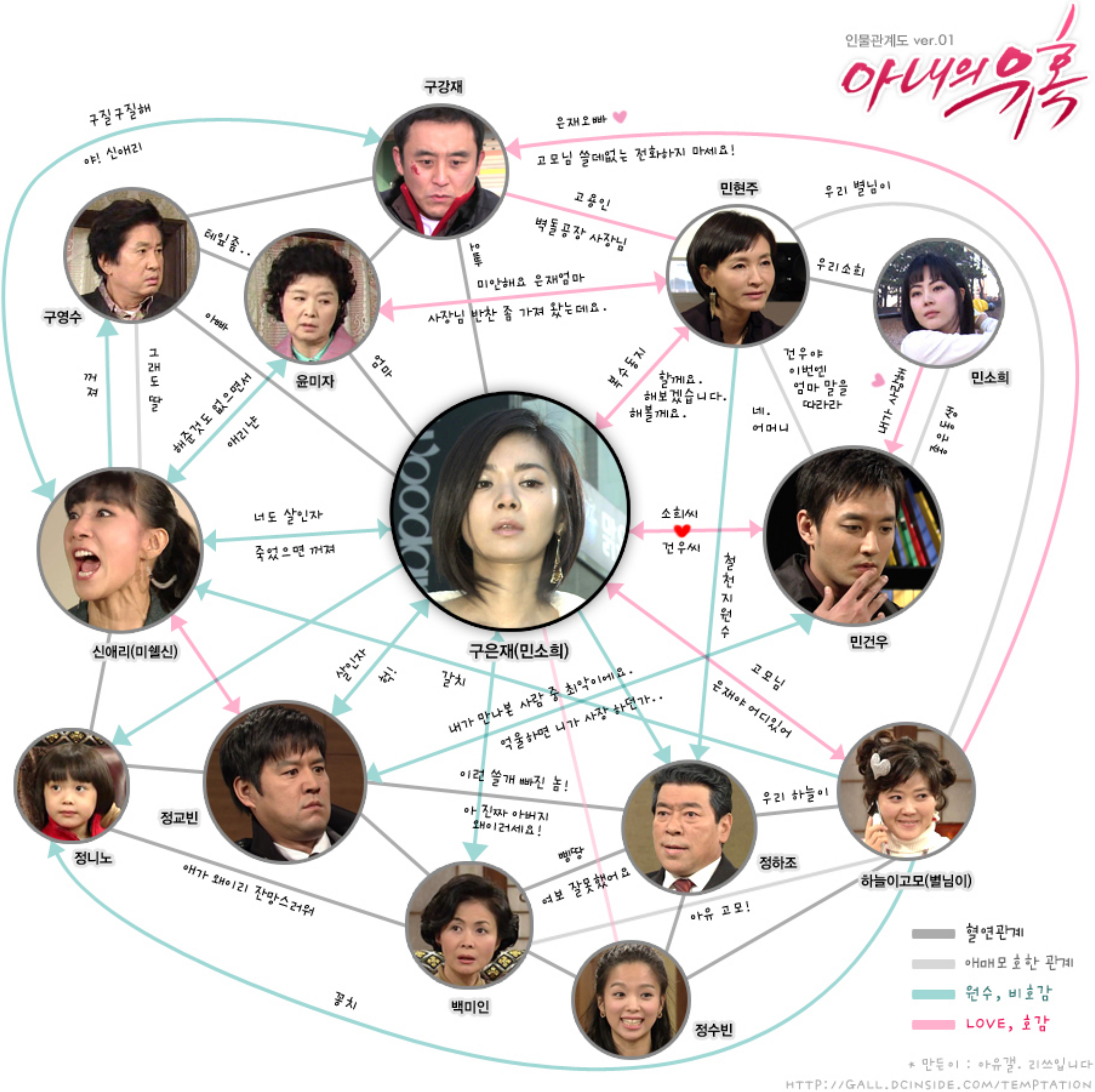
Agenda

1. Emulation & Hypervising
2. I/O Acceleration
    1. Onloading
    2. Offloading
3. I/O Virtualization Technologies
4. 마무으리


오늘은 그냥
달리는 걸로~!

본지의 편집방향과 일치하지 않을 수도 있습니다.


I/O Virtualization
그까이꺼
별거 있나?



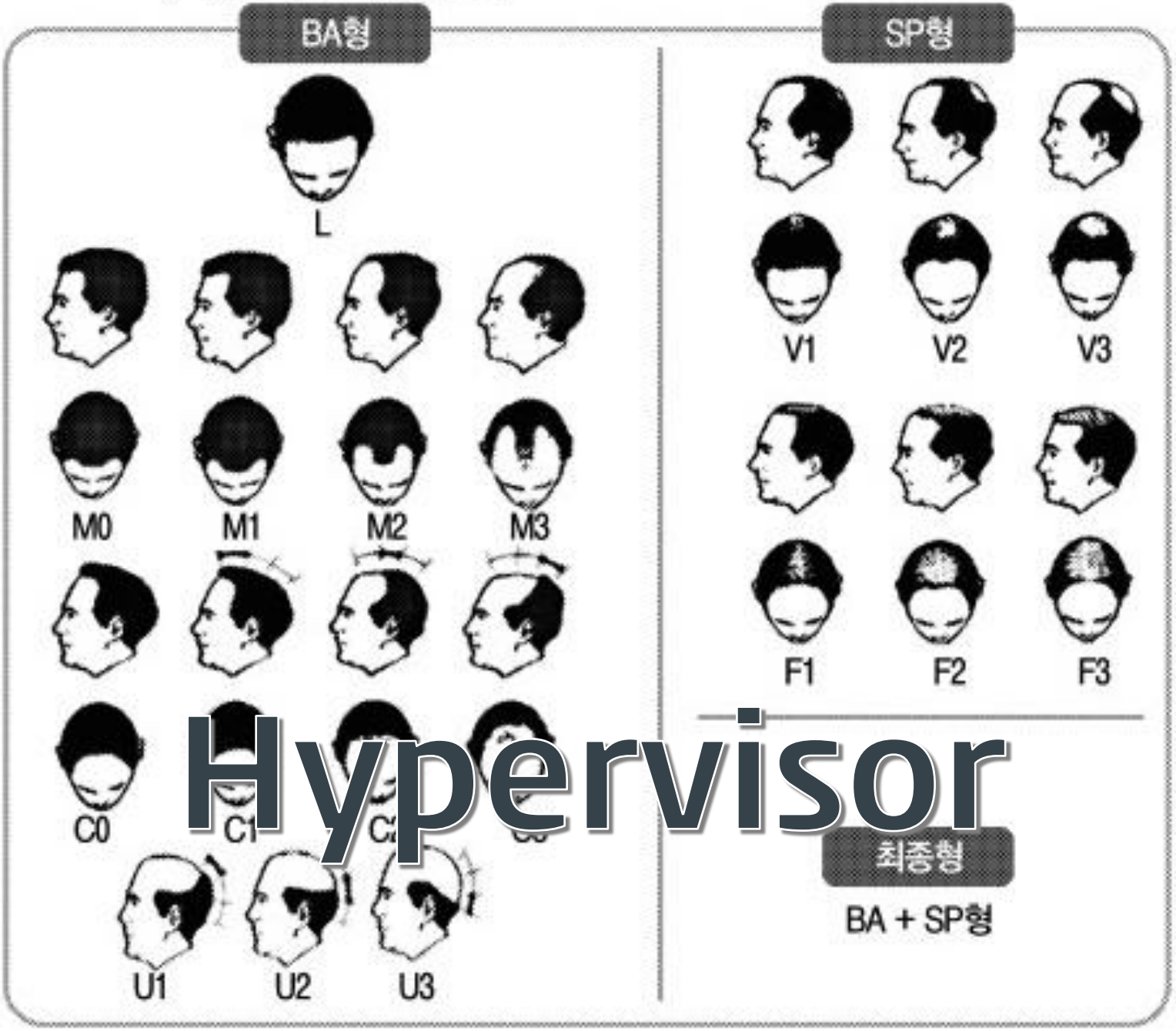
인물관계도 ver.01
아내의유혹
구강재
구영수
윤미자
민현주
민소희
구은재(민소희)
신애리(미쉘신)
민건우
정니노
정교빈
정하조
하늘이고모(별님이)
백미인
정수빈
구질구질해
야! 신애리
은재오빠
고모님 쓸데없는 전화하지 마세요!
고용인
벽돌공장 사장님
미안해요 은재엄마
사장님 반찬 좀 가져 왔는데요.
우리 별님이
우리소희
너도 살인자
죽었으면 꺼져
소희씨
건우씨
고모님
은재야 어디있어
내가 만나본 사람 중 최악이에요.
억울하면 내가 사장 하던가..
이런 쓸개 빠진 놈!
아 진짜 아버지 왜이러세요!
여보 잘못했어요
아유 고모!
우리 하늘이
애가 왜이리 칠칠맞스러워
꽁치
살인자
칫!
갈치
철천지 원수
혈연관계
애매모호한 관계
원수, 비호감
LOVE, 호감
* 만든이 : 아유갤. 리쓰입니다
HTTP://GALL.DCINSIDE.COM/TEMPTATION

그… 그게 좀 복잡…

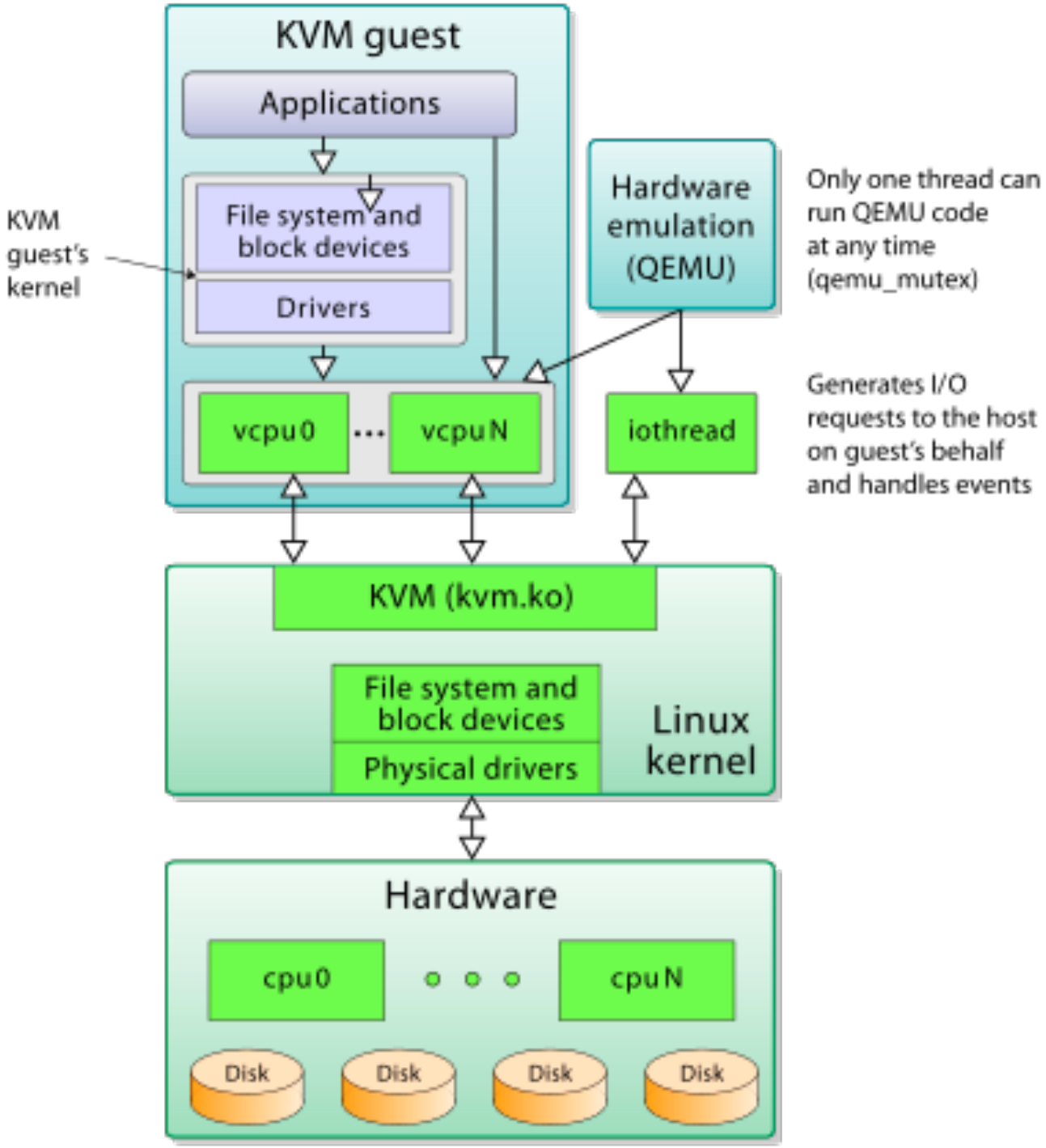

My Computer
ePSXe - Enhanced PSX emulator
File Run Config Options
ePSXe 1.6.0
Snes9X v1.51 for Windows
File Options Input Sound Window Cheat Netplay Help
SECRET of MANA
LEGENDS THROUGHOU
BUT ALL OF THESE S
JUST ONE WEAP
MAME32 0.103
File View Options Help
All Games
Available
CHD
Clones
CPU
Custom
Dumping Status
Favorites
Genre
Imperfect
Lightgun
Manufacturer
Multi-Monitor
Not Working
Orientation
Originals
Raster
Samples
Save State
Sound
Game
Bubble Memories - The
Crystal of Kings, The
Dark Seal (World revisio
Dragon Blaze
Dungeon Magic (World
Dungeon Magic (US)
Gaiapolis (ver JAF)
Gauntlet (2 Players, rev 3)
Gauntlet (2 Players, rev 6)
Gauntlet (rev 1)
Simpsons, The (4 Players)
Working
49 games
Gens - Genesis : SONIC THE
File Graphic CPU Sound Option Help
SONIC
THE HEDGEHOG
2
1 PLAYER
2 PLAYER VS © 1992 SEGA
Emulation
start
7:48 PM

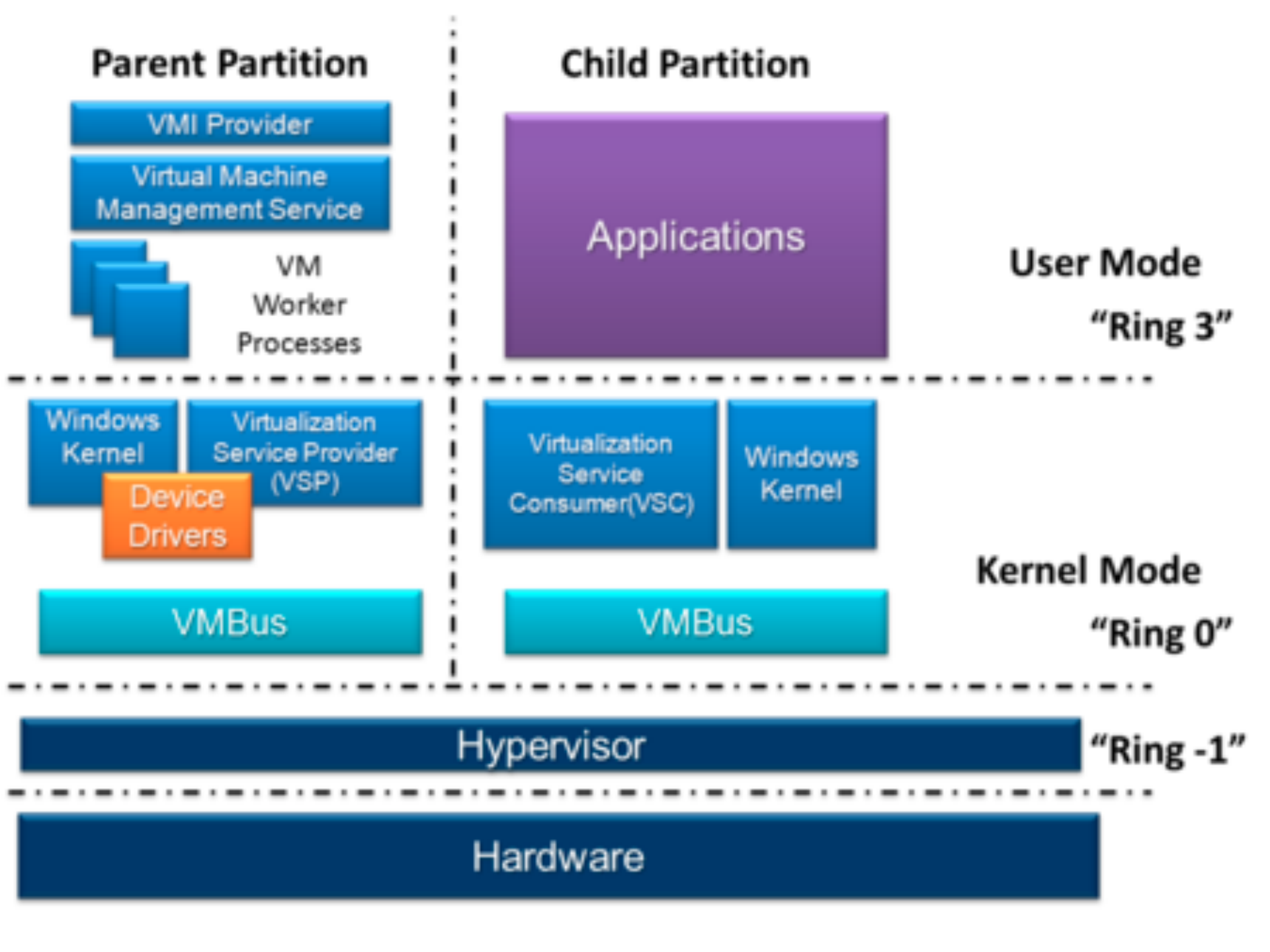

무에서
유를창조
MILK

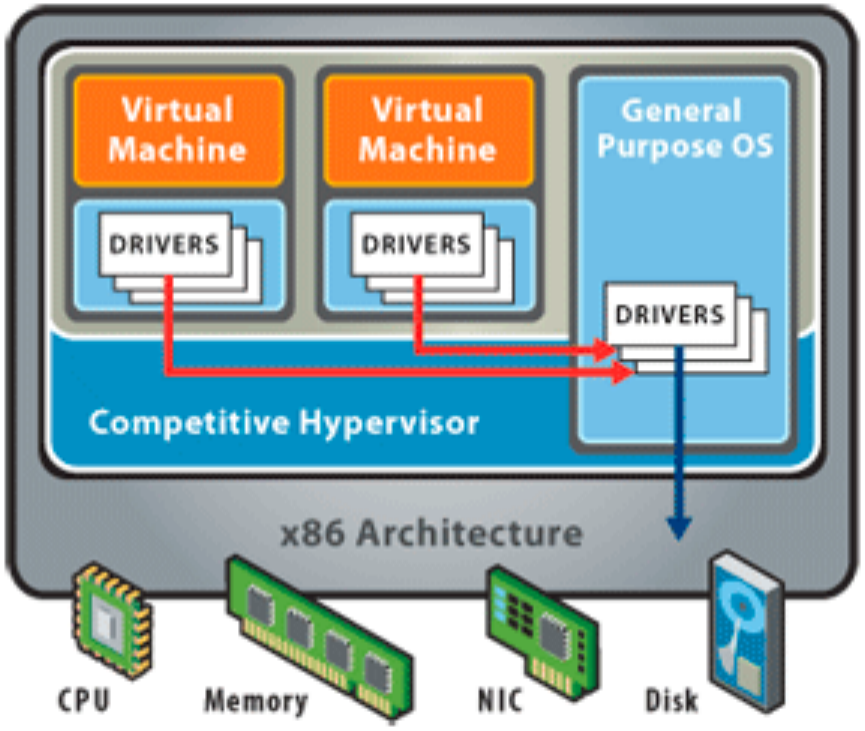
■ BASP에 따른 탈모 유형 분류

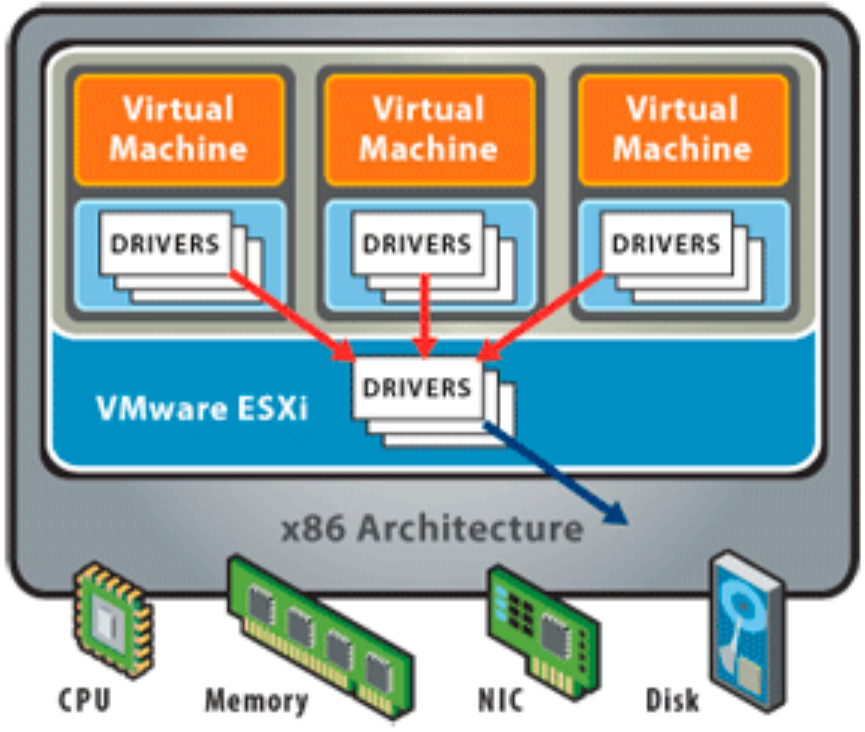


Hypervisor


KVM guest
Applications
File system and block devices
Drivers
KVM guest's kernel
Hardware emulation (QEMU)
Only one thread can run QEMU code at any time (qemu_mutex)
vcpu0
vcpuN
iothread
Generates I/O requests to the host on guest's behalf and handles events
KVM (kvm.ko)
File system and block devices
Physical drivers
Linux kernel
Hardware
cpu0
cpuN
Disk
Disk
Disk
Disk

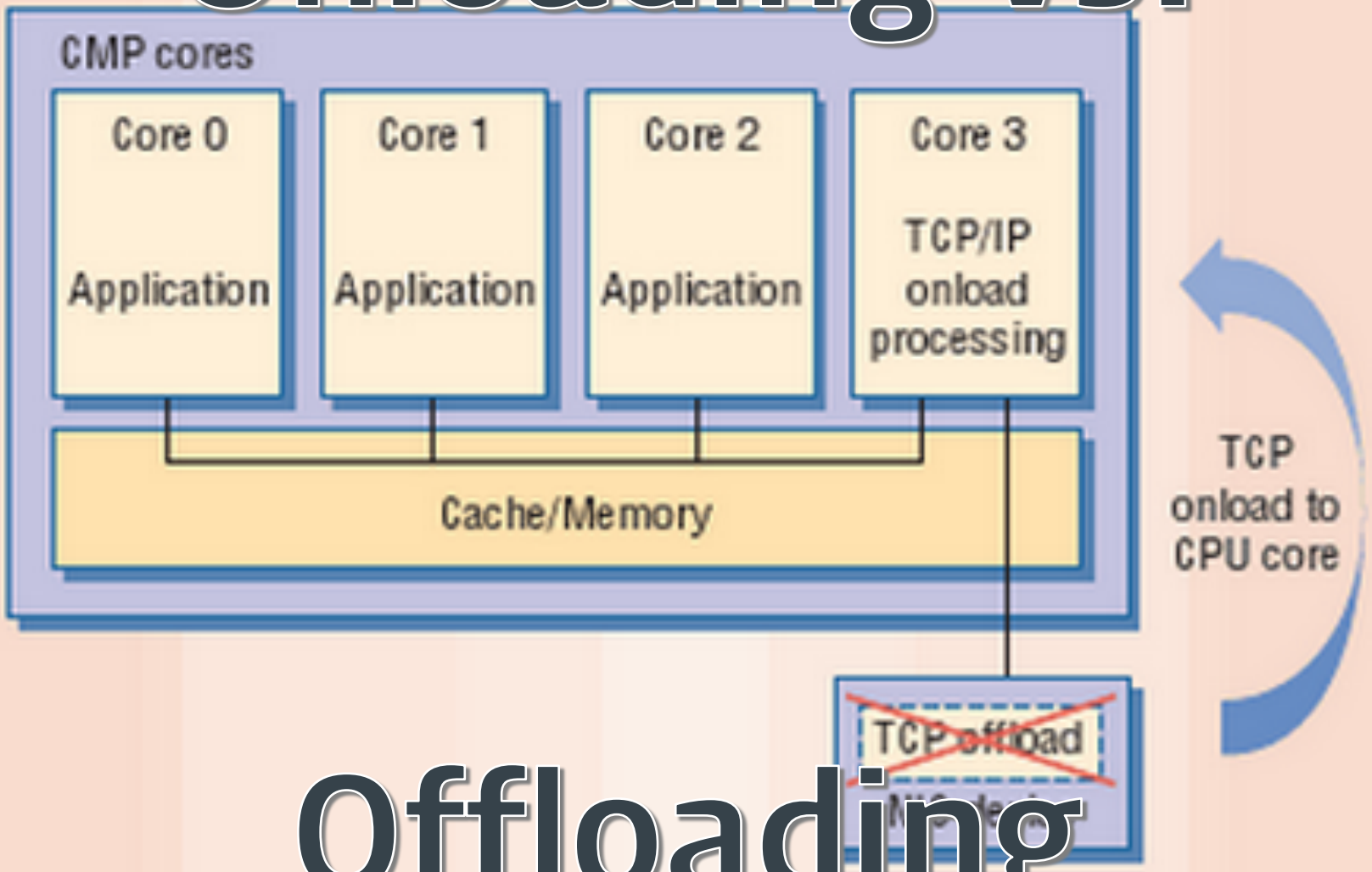

Parent Partition
Child Partition
VMI Provider
Virtual Machine Management Service
VM Worker Processes
Applications
User Mode
"Ring 3"
Windows Kernel
Virtualization Service Provider (VSP)
Device Drivers
Virtualization Service Consumer(VSC)
Windows Kernel
Kernel Mode
"Ring 0"
VMBus
VMBus
Hypervisor
"Ring -1"
Hardware


Virtual Machine
Virtual Machine
General Purpose OS
DRIVERS
DRIVERS
DRIVERS
Competitive Hypervisor
x86 Architecture
CPU
Memory
NIC
Disk

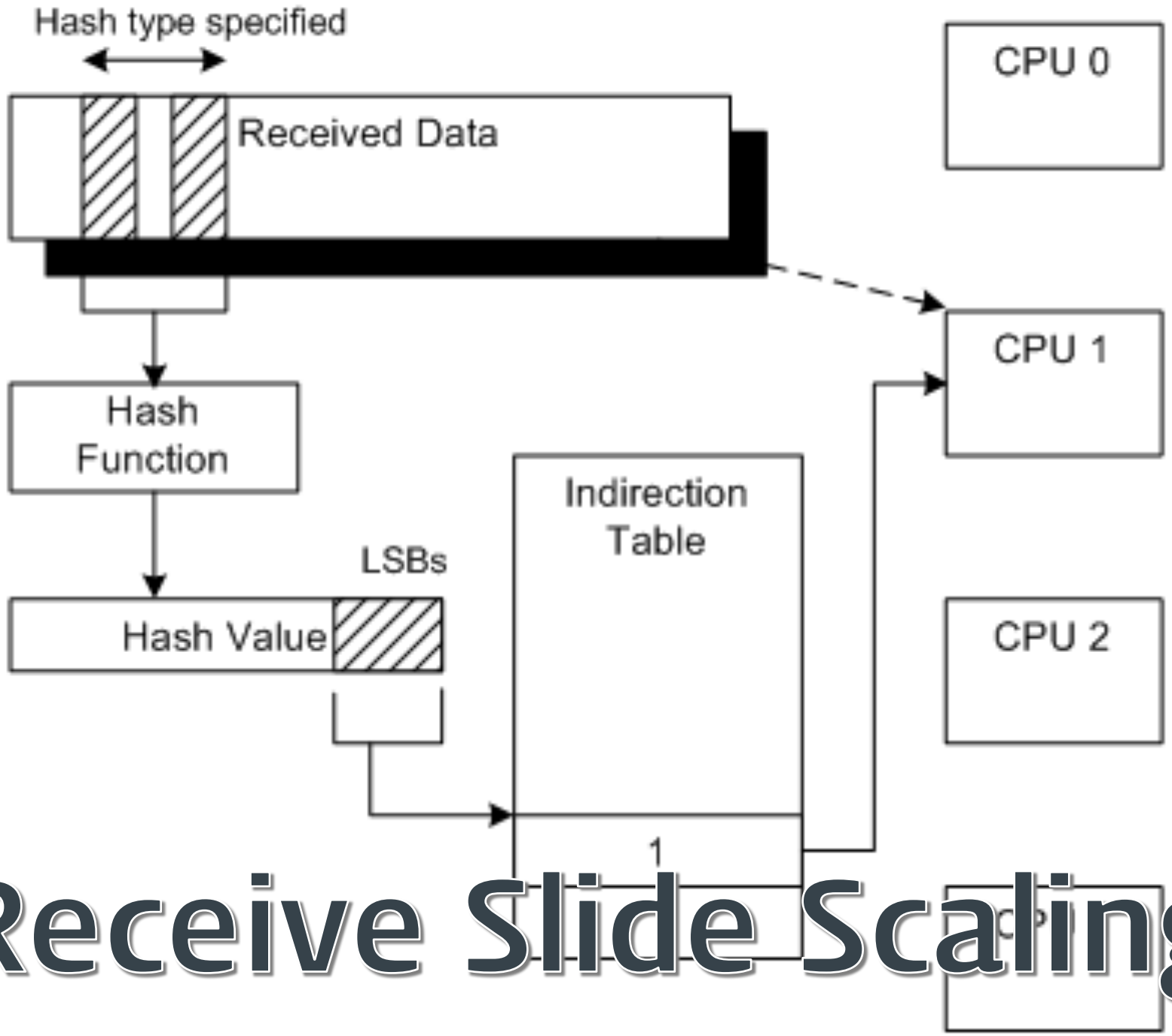

Virtual Machine
Virtual Machine
Virtual Machine
DRIVERS
DRIVERS
DRIVERS
DRIVERS
VMware ESXi
x86 Architecture
CPU
Memory
NIC
Disk

I/O Acceleration

Onloading vs. Offloading



Onloading

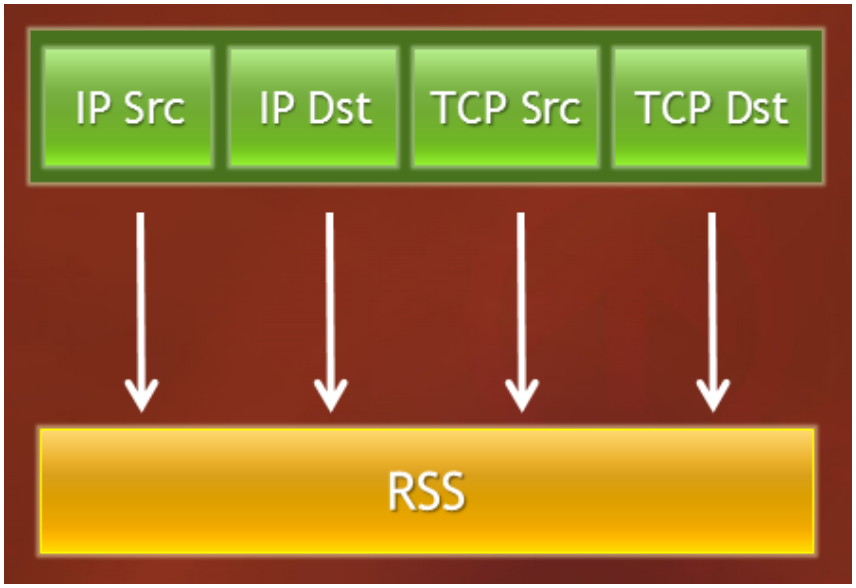


Receive Slide Scaling

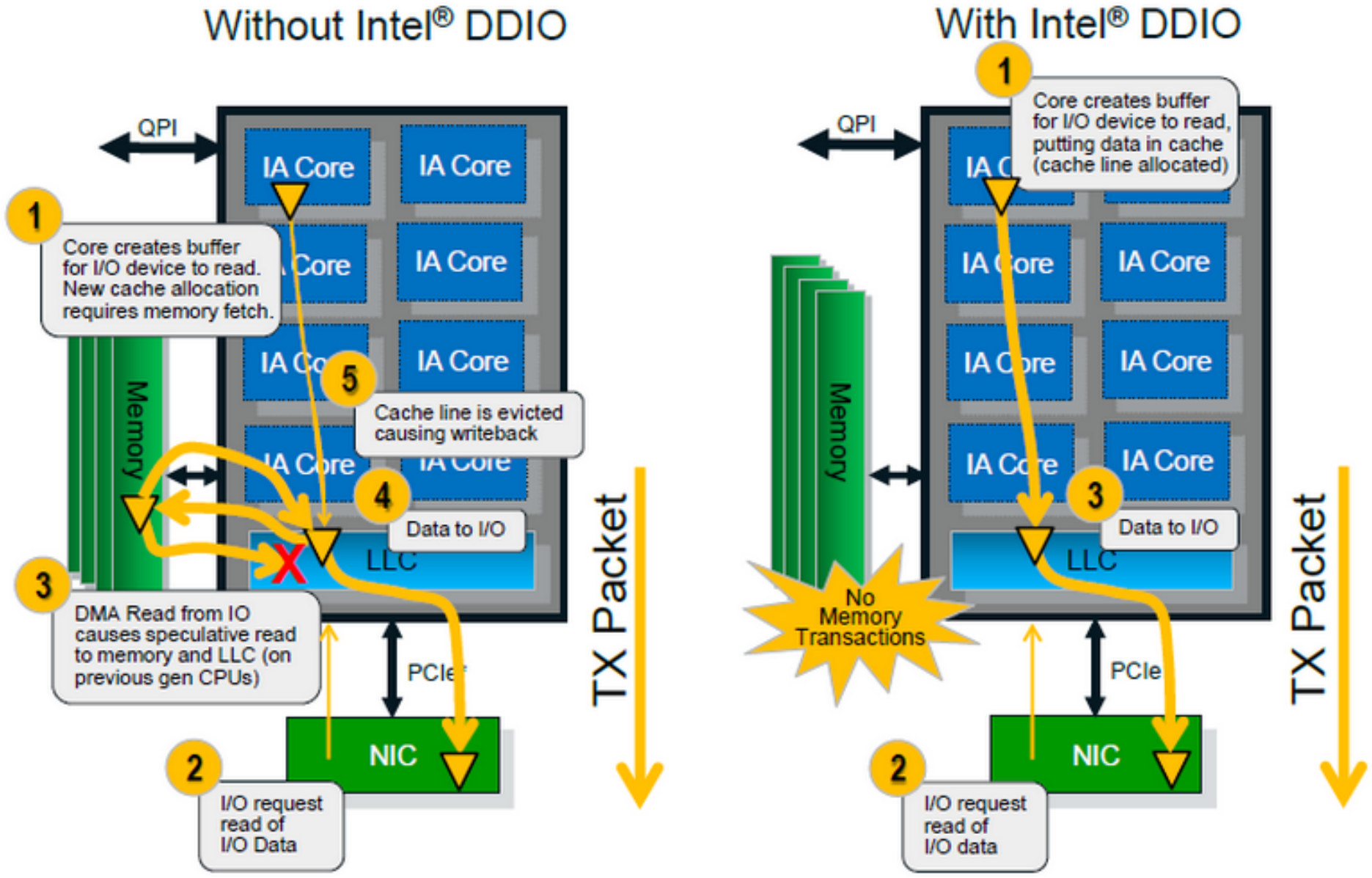

한놈만 패


17대 1

Receive Slide Scaling (RSS)

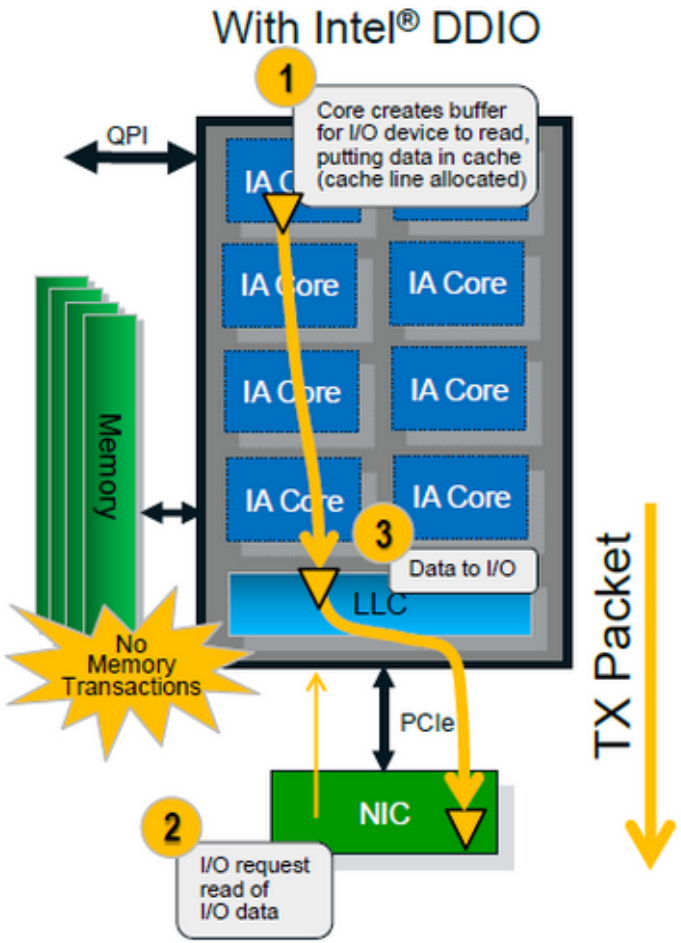


- Multicore 환경에서도 CPU 간 처리시간 차이에 따른 Packet 순서 변경으로 인한 성능 저하 문제 발생
- Flow 단위 (src, dstIP, src, dstPort) 로 Hashing 하여 그 값에 따라 CPU 처리
- (HASH TYPE = IPv4, IPv4+TCP, IPv6, IPv6_TCP, IPv6_EX, IPv6_EX_TCP)
- Windows 2012의 경우 64개 이상의 Core, NUMA Architecture 지원


Without Intel® DDIO
QPI
IA Core
IA Core
IA Core
IA Core
IA Core
IA Core
1
Core creates buffer for I/O device to read. New cache allocation requires memory fetch.
Memory
5
Cache line is evicted causing writeback
4
Data to I/O
LLC
3
DMA Read from IO causes speculative read to memory and LLC (on previous gen CPUs)
PCIe
NIC
2
I/O request read of I/O Data
TX Packet
With Intel® DDIO
1
Core creates buffer for I/O device to read, putting data in cache (cache line allocated)
QPI
IA Core
IA Core
IA Core
IA Core
IA Core
IA Core
Memory
3
Data to I/O
LLC
No Memory Transactions
PCIe
NIC
2
I/O request read of I/O data
TX Packet


김종관 감독 윤계상 | 정유미 | 윤희석 | 요조 장서원 오창석 염보리 김효서
조금만 더 가까이
Come, Closer
고장난 사랑에 관한 다섯가지 증상
너 때문에 나...
연애불구야...
10월 대개봉!

Intel Data Direct I/O (DDIO)



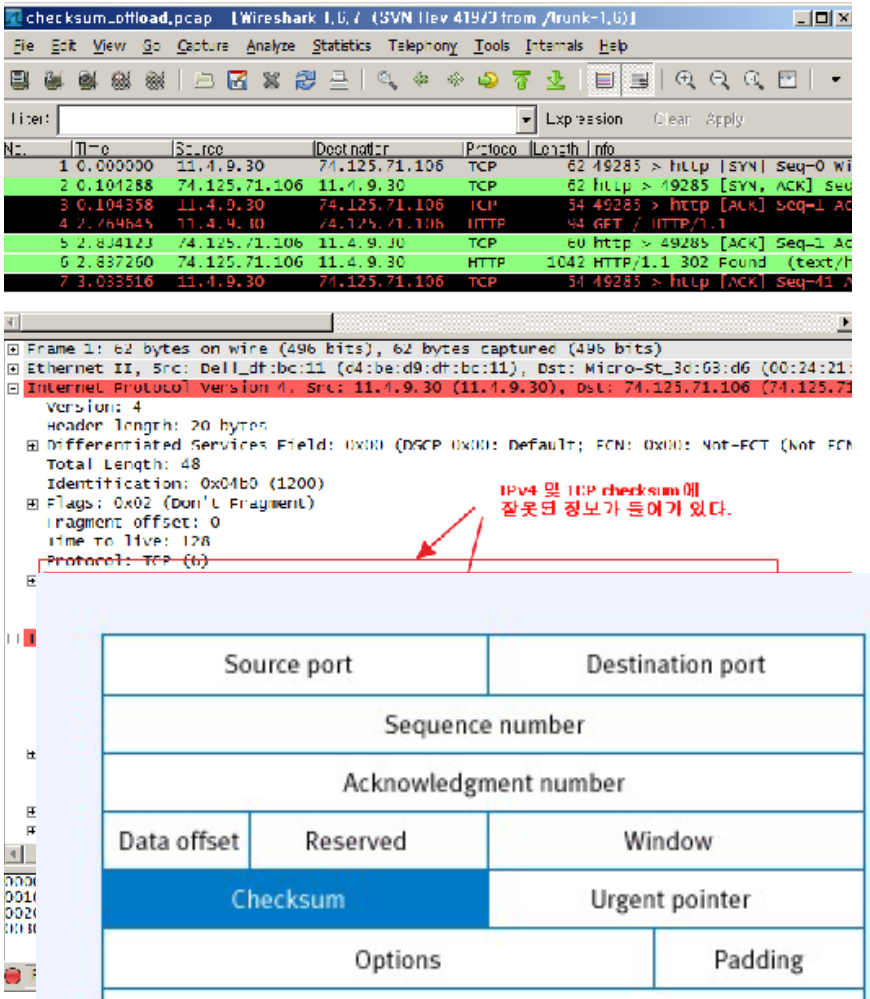
- I/O 시 Main Memory 를 사용하지 않고 CPU 의 Cache Memory를 사용함 (SRAM)
- Intel Sand Bridge Chipset 부터 지원
- LAN Card 상관 없음
- Ethernet 뿐 아니라 Infiniband, Fibre Channel, RAID 기술 까지 모두 사용 가능
- 현존하는 I/O Acceleration 중 가장 빠른 기술 (다른 가속기술과 중복 사용가능 하기 때문)

Offloading


928546
UPS Freight

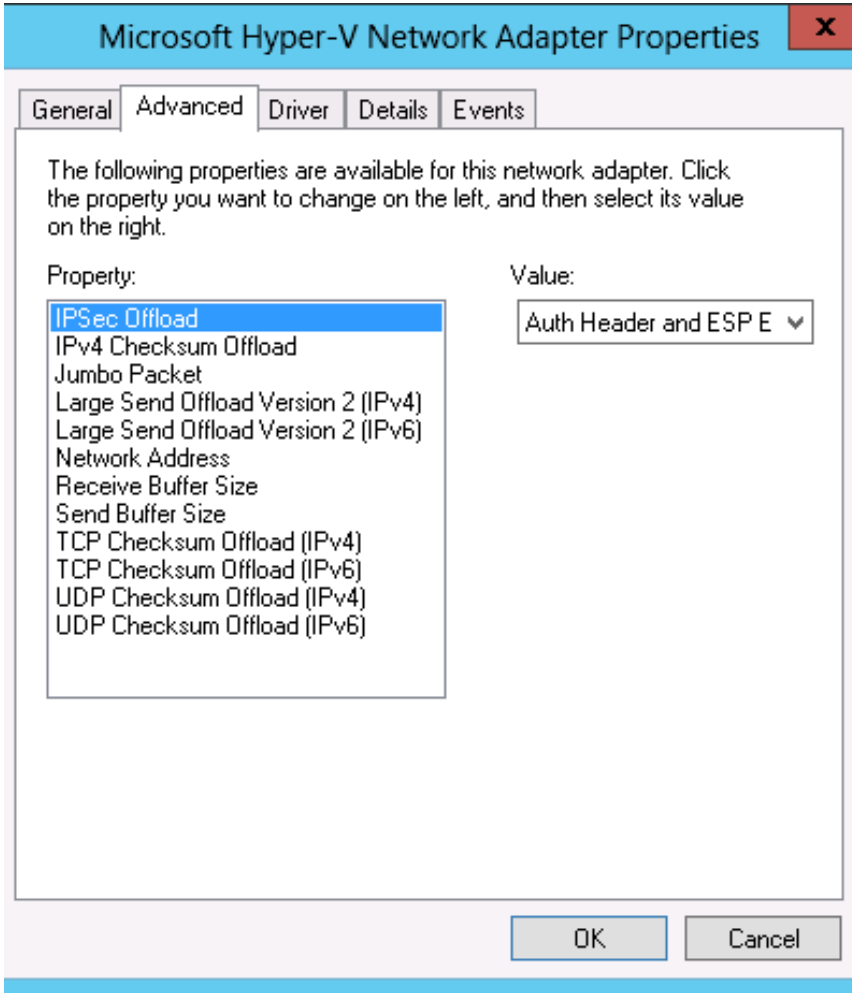
Checksum Offload
IPSEC Offload
Segmentation Offload
Chimney

Checksum Offload



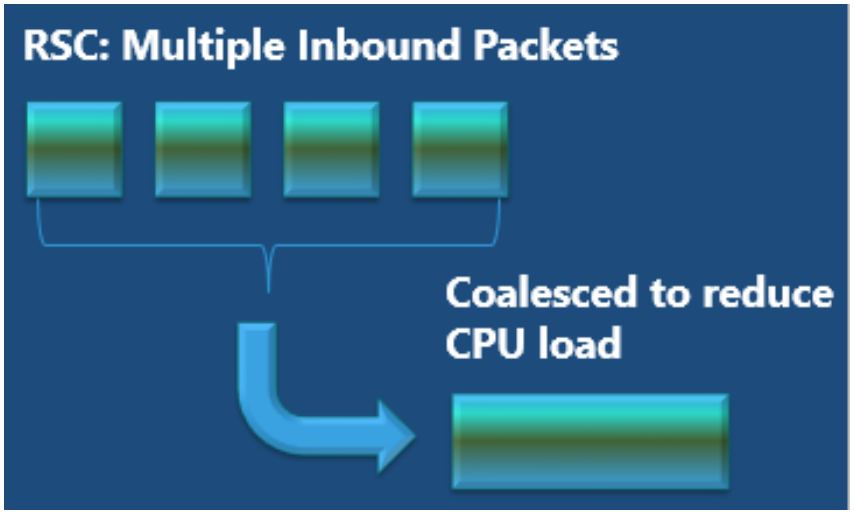
- 가장 기본적인 Offload 기능
- TCP Header 의 Checksum 을 NIC 에서 생성
- Checksum 로드를 NIC 에 할당하기 때문에 CPU 부하가 줄어듬
- Wireshark 에서 TCP Checksum 오류로 나타날 경우 Checksum 기능 제거 필요

IPSEC Offload



- Server NIC 에 Encryption ASIC 이 탑재된 경우 (Wireless NIC 등)
- CAVIUM Encryption Card 처럼 별도의 가속기를 PCI 에 장착한 경우
- CPU 내부에 가속 ASIC 이 장착된 경우 (IBM PowerPC, Oracle CPU, AMD Load 에 존재)
- 대부분 IPSEC Offload 의 경우 ESP, AH 패킷에 대한 가속만 지원하고 IKE에 대한 가속은 지원하지 못함

Segmentation Offload – Receive Segment Coalescing (RSC)



- Receive Segmentation Coalscing
  - » 일반적으로 Packet 당 Interrupt 를 처리 하지만 RSC 가 Enable 되어 있을 경우 Flow 단위로 한꺼번에 Packet 을 처리
  - » 대용량 Buffer 를 지원할 수 있어야 함
  - » 10G NIC 은 기본적으로 RSC를 지원
  - » 최대 64KB Payload 를 한꺼번에 처리할 수 있음 (v2는 256KB)
  - » Windows 2012에 소개
  - » Linux Kernel 2.6 부터 정식 지원 (2.4.x 소개)

Generic Segmentation Offload (GSO)

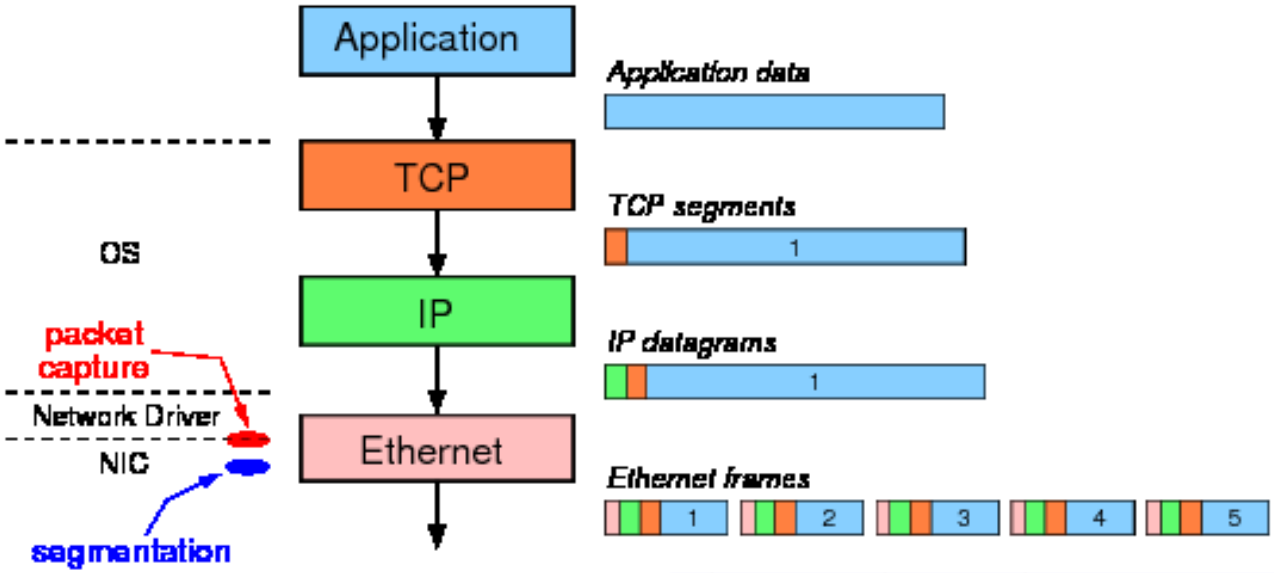
Large Send Offload (LSO)

Generic Receive Offload (GRO)

Large Receive Offload (LRO)

TCP Segmentation Offload (TSO)

Segmentation Offload – Large Send Offload v2 (LSOv2)



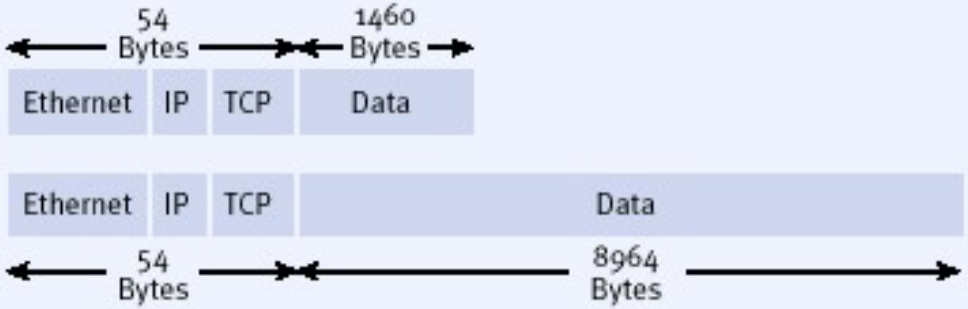
TCP	66	ndl-aas > 37295 [ACK]
TCP	2762	[TCP segment of a reas
TCP	66	37295 > ndl-aas [ACK]
TCP	4017	[TCP segment of a reas
TCP	66	37295 [ACK]
TCP	3538	[TCP
TCP	66	3729
TCP	3538	[TCP
TCP	66	37295 > nd aas [ACK]
TCP	98	[TCP segment of a reas
TCP	66	37295 > ndl-aas [ACK]
TCP	16242	[TCP segment of a reas

YES!

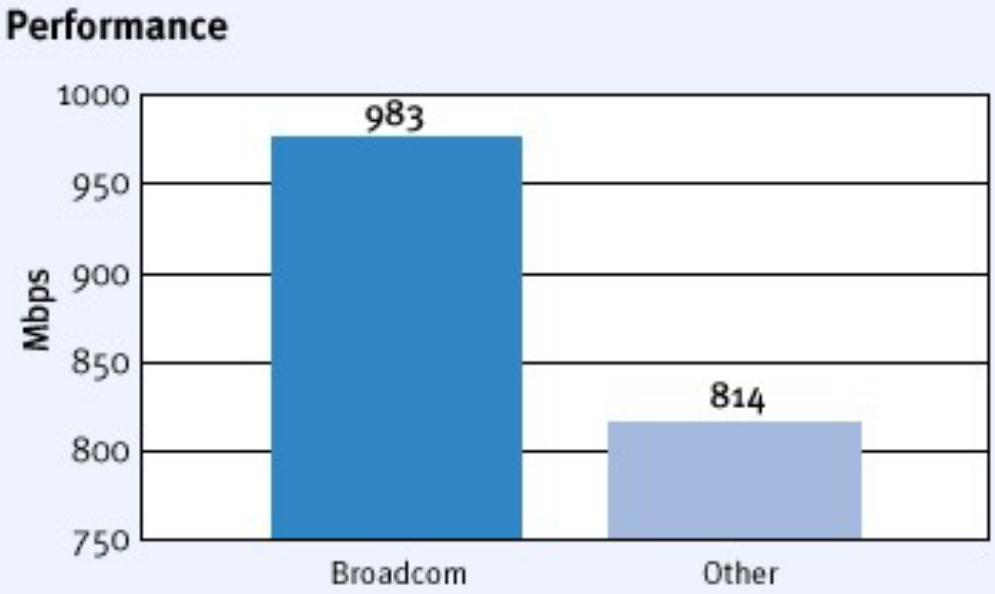
TCP	66	ndl-aas > 37293 [ACK]
TCP	1414	[TCP segment of a reas
TCP	66	37293 > ndl-aas [ACK]
TCP	1414	[TCP segment of a reas
TCP	66	37293
TCP	1414	[TCP
TCP	66	3729
TCP	1414	[TCP
TCP	66	37293 > ndl-aas [ACK]
TCP	1414	[TCP segment of a reas
TCP	66	37293 > ndl-aas [ACK]
TCP	1414	[TCP segment of a reas

No!!!

Jumbo Frame

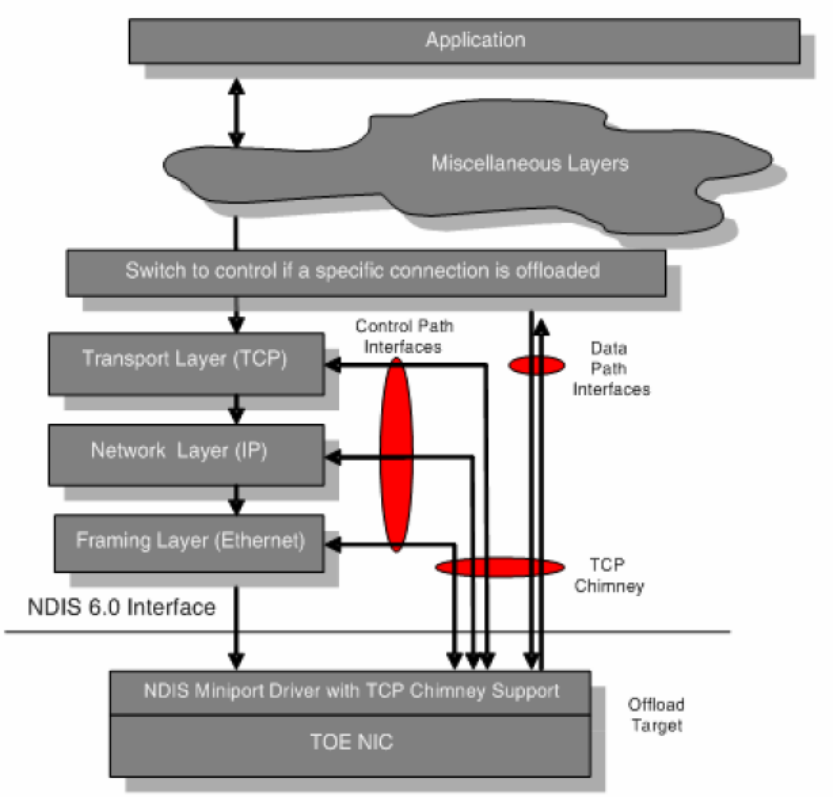

54 Bytes
1460 Bytes
Ethernet
IP
TCP
Data
Ethernet
IP
TCP
Data
54 Bytes
8964 Bytes


Throughput
Mbps
1000
950
900
850
800
750
874
983
Standard
Jumbo


Performance
Mbps
1000
950
900
850
800
750
983
814
Broadcom
Other


CPU utilization
Percent
100
80
60
40
20
0
100
50
Standard
Jumbo

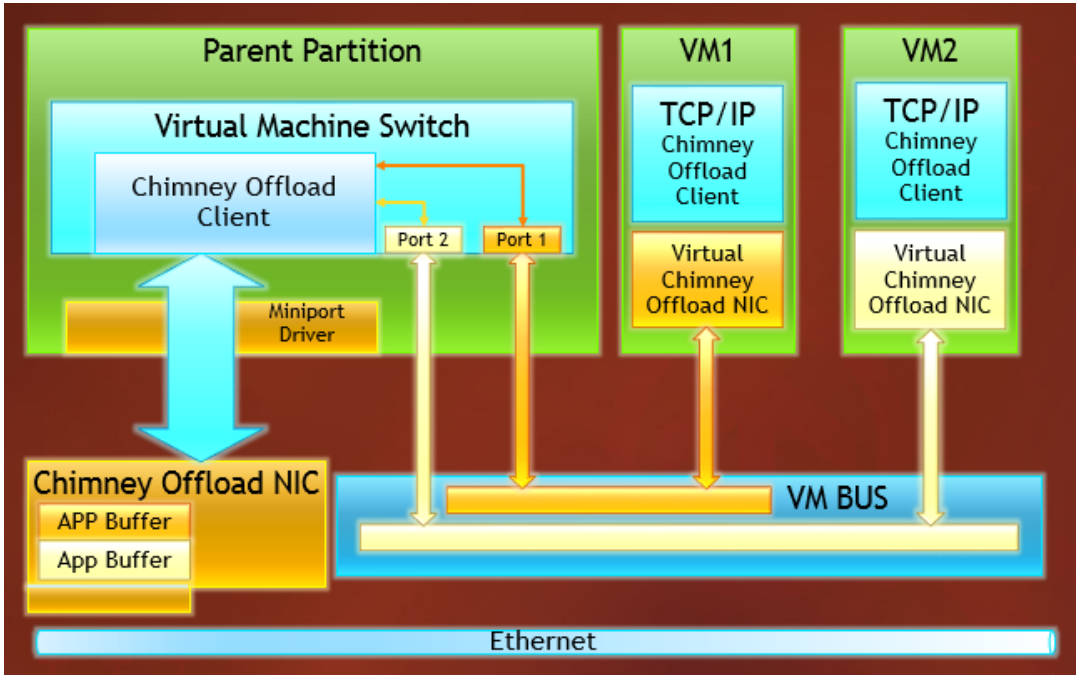
TCP Chimney Offload



- Application 에서 OS 의 TCP/IP Stack 을 거치지 않고 Direct 로 TOC NIC 에 직접 Connection
- 초기 NIC Driver bug 로 인하여 강제로 기능을 사용하지 않는 것을 추천 하였으나 최근 안정화 됨 (Driver 영향이 큼)
- Alacritech 사 특허


Merry!

TCP Chimney Offload on Hypervisor



- Hypervisor 에서도 Chimney Offload 기능을 사용할 수 있음
- Virtual Chimney Offload NIC 을 통해서 복제됨 (VM BUS 를 이용하므로 copy 부하는 적용)

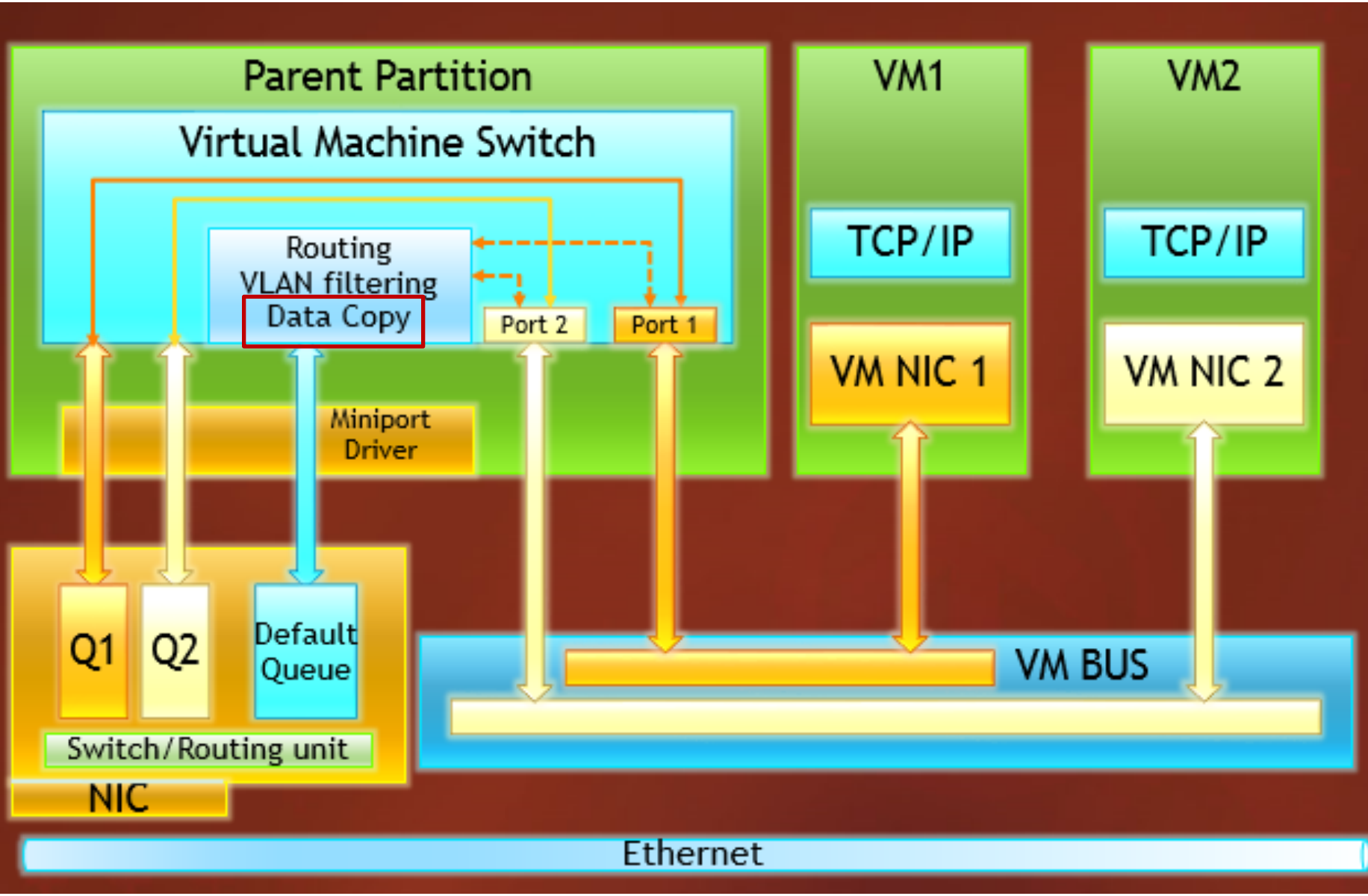
Considerstions

DATACENTER ARCHITECT GROUP KR

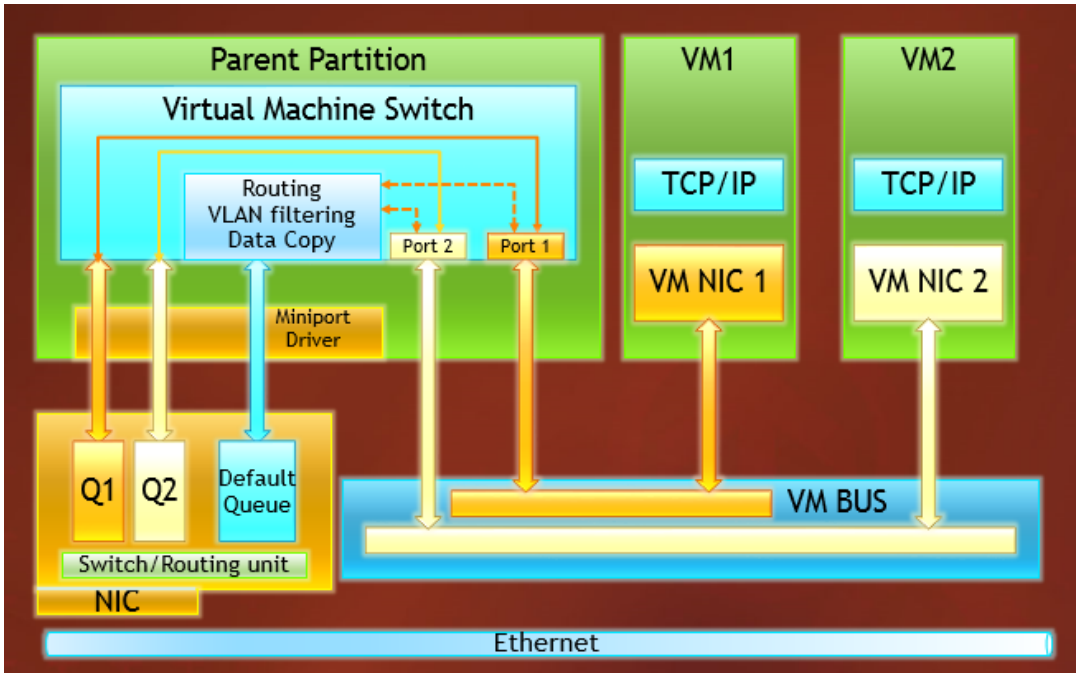
Performance Metric	Loopback Fast Path	Registered I/O (RIO)	Large Send Offload (LSO)	Receive Segmentation Offload (RSC)	Receive Side Scaling (RSS)	Virtual Machine Queues (VMQ)	Remote DMA (RDMA)	Single Root I/O Virtual (SR-IOV)
Lower End-to-End Latency	X	X					X	X
Higher Scalability		X			X	X		
Higher Throughput	X	X	X	X	X	X	X	X
Lower Path Length	X	X	X	X			X	X
Lower Variability		X						

I/O Virtualization Technologies

Virtual Machine Queue (VMQ)
SR-IOV
MR-IOV
NIC Partitioning

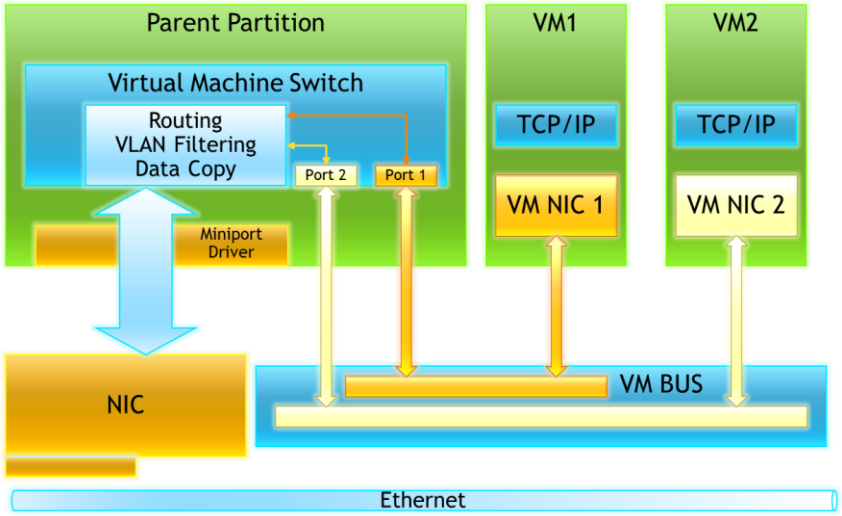

Parent Partition
Virtual Machine Switch
Routing
VLAN filtering
Data Copy
Port 2
Port 1
Miniport
Driver
VM1
TCP/IP
VM NIC 1
VM2
TCP/IP
VM NIC 2
Q1
Q2
Default
Queue
Switch/Routing unit
NIC
VM BUS
Ethernet

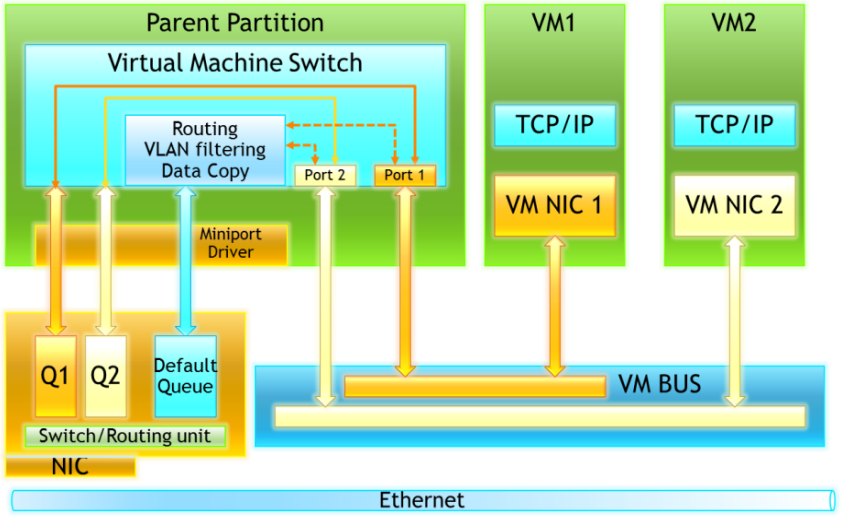
Virtual Machine Queue (VMQ)



- NIC 에 Virtual Machine Queue를 할당하여 GuestVM 과 구분되어 통신 하도록 지원하는 기능
- 각종 Offload 기능을 그대로 사용 가능
- Parent Partition 을 거쳐야 하므로 (Emulation) Direct I/O 기능을 지원하는 Virtual NIC 선택 필요
- Parent Partition 통과 여부가 SR-IOV, MR-IOV와 차별화 되는 점

Virtual Machine Queue (VMQ)

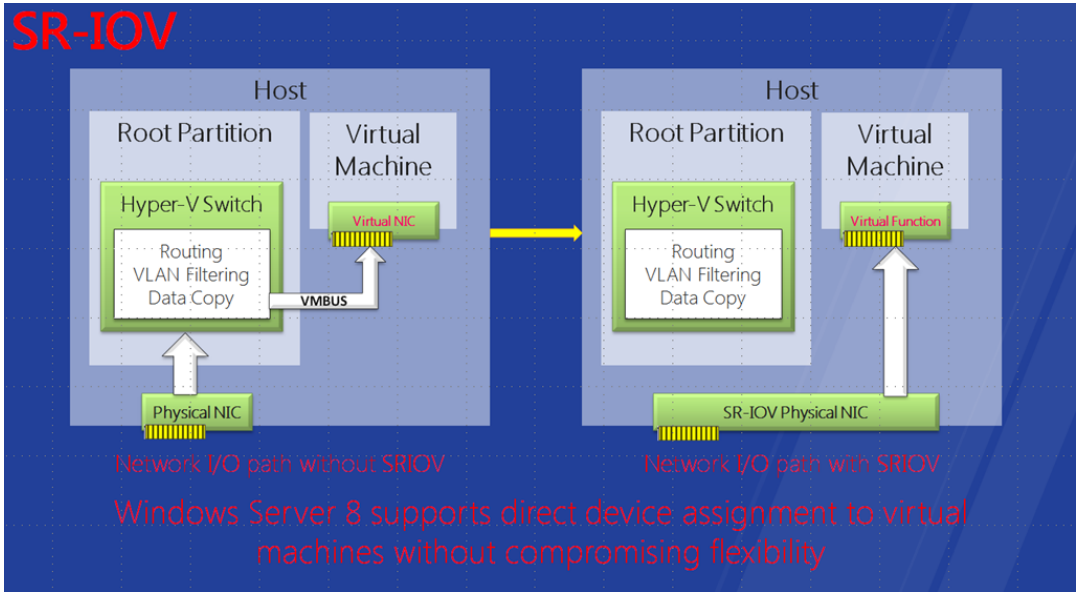




without VMQ

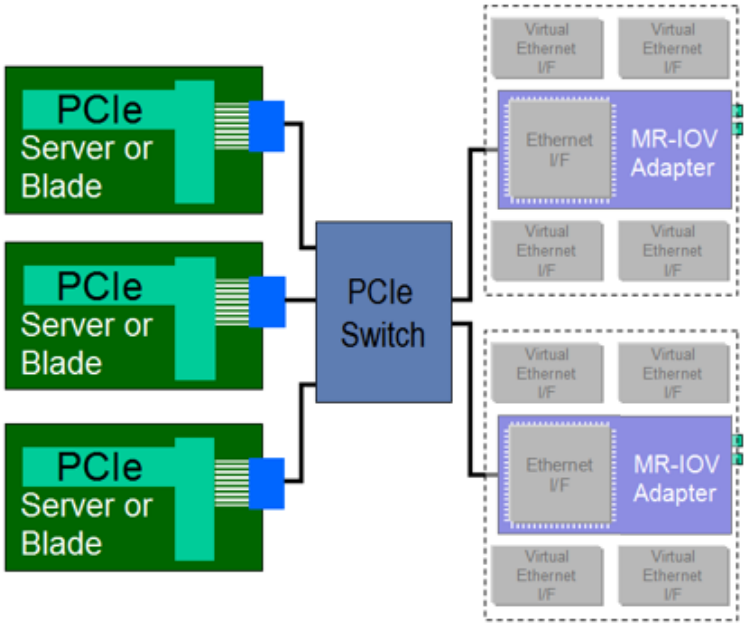
with VMQ

SR-IOV (Single Root IO Virtualization)



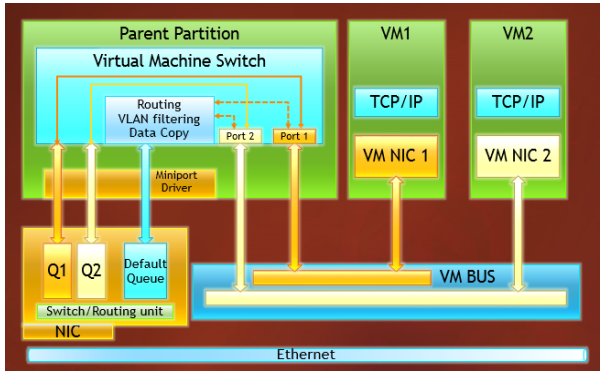
- Hypervisor 를 통하지 않고 I/O 장치에 Guest OS 가 Direct Access
- 기존 Hypervisor 에서 I/O 는 Emulation 으로 인한 속도 저하 문제가 있음
- Guest VM 에서는 SR-IOV Virtual NIC 으로 인식 됨

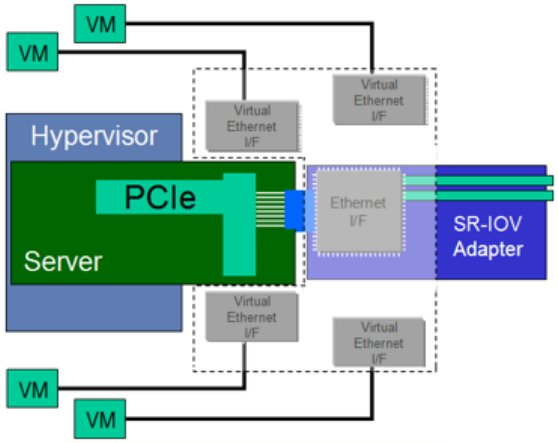
MR-IOV (Multi Root IO Virtualization)

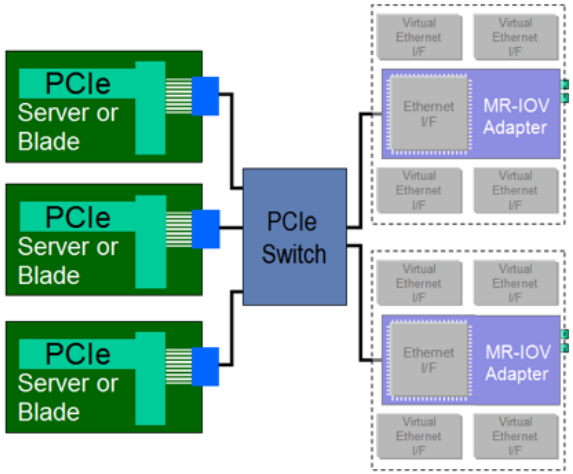


- Hypervisor 를 통하지 않고 I/O 장치에 Guest OS 가 Direct Access (SR-IOV 와 동일)
- 여러개의 Server 혹은 Blade 에서 PCIe Switch 를 이용하여 NIC에 접근
- PCIe 자체를 가상화 하므로 Guest VM 에서는 별도의 NIC 으로 보임

VMQ vs. SR-IOV vs. MR-IOV

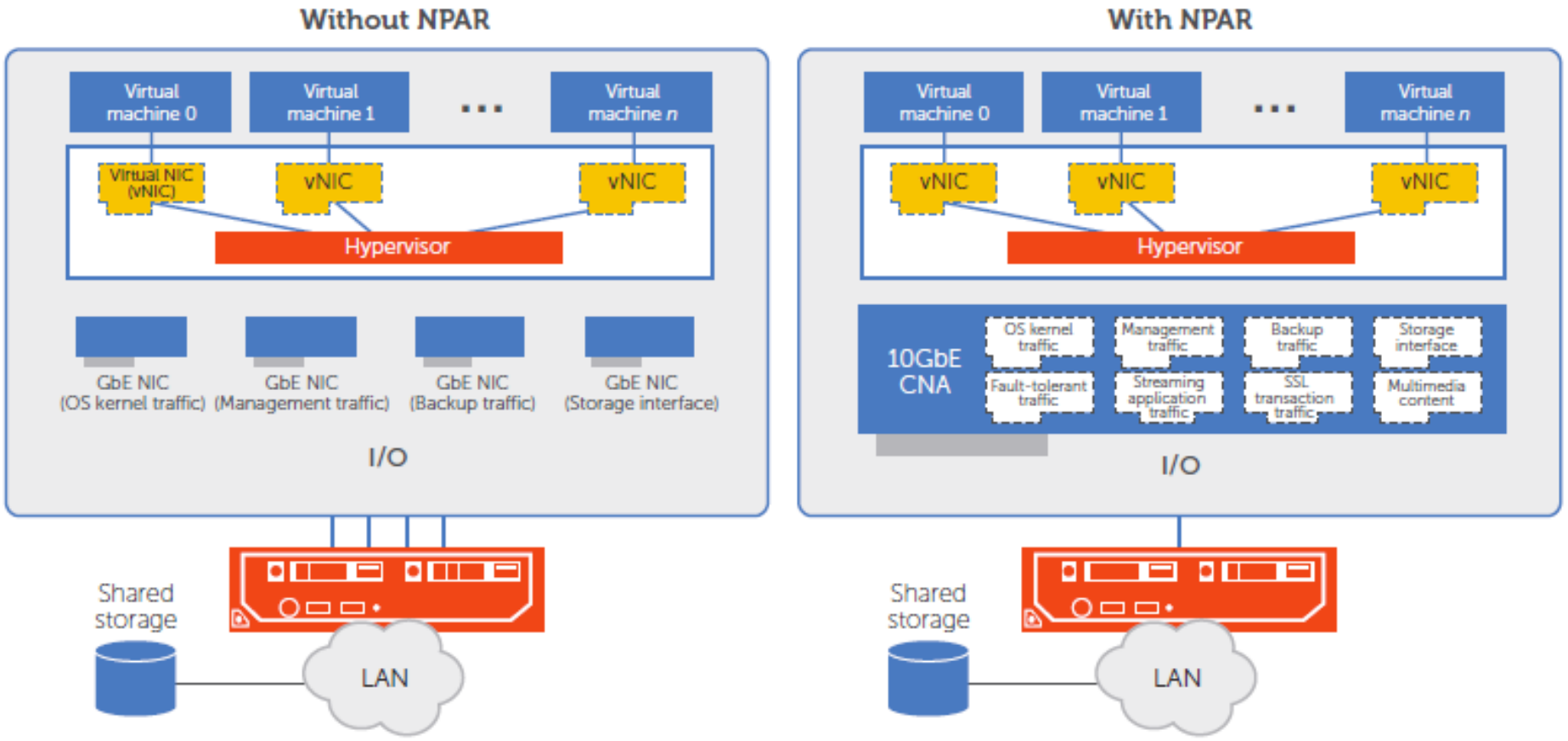






- VMQ & Non-I/O Virtualization
  - 장치메모리 ↔ PCI 버스(R) ↔ 하이퍼바이저 ↔ Guest VM 메모리
- SR-IOV
  - 장치메모리 ↔ 하나의 PCI 버스(R) ↔ Guest VM 메모리
- MR-IOV
  - 장치메모리 ↔ 여러 개의 PCI 버스(R) ↔ Guest VM 메모리

NIC Partitioning and etc. (Qlogic VMFlex, Broadcom)




마음으로

CPU, Memory Hypervisor

기타 장치
Emulation

달라요
다르다
Different
다르다니까
다를껄

Onloading
Offloading
가속 기술 들들들

VMQ
SR-IOV
MR-IOV