پیاده سازی سرویس SQL در محیط مجازی VSphere

مدرس: رضا اردانه

nikamooz;
آموزش برنامه نویسی و اجرای پروژه

معرفی رضا اردانه

- متخصص محیط مجازی سازی بر مبنای معماری VMware
- متخصص امنیت و شبکه
- مدیر ارشد سیستم بر پایه مایکروسافت
- رییس گروه زیر ساخت شرکت پرداخت الکترونیک سداد
- مدرس دوره‌های مجازی سازی، امنیت و شبکه

چه خواهیم آموخت؟

- نگاهی بر مجازی سازی VMware
- آشنایی با Workload های مختلف MS SQL Server
- آشنایی با ویژگی های بقای کسب و کار در سرویس SQL
- آشنایی با ویژگی های بقای کسب و کار در محیط vSphere
- امکانات محیط vSphere در زمینه مدیریت سرویس MS SQL
- بهروش های پیاده سازی سرویس SQL در محیط vSphere
- گذری بر Persistent Memory در محیط vSphere
- دمویی از محیط vROPS

مجازی سازی با VMware vSphere

- مدیریت بهینه منابع سخت افزاری
- مدیریت بهتر سیستم عامل ها
- سادگی در پیاده سازی سرویس ها

vmware®

آشنایی با Workload های مختلف MSSQL

- پایگاه های داده OLTP
- پایگاه های داده DSS
- پایگاه های داده ETL و Reporting

nikamooz;
آموزش برنامه نویسی و اجرای پروژه

آشنایی با ویژگی های بقای کسب و کار در MSSQL

- استفاده از Always On Availability Groups
- استفاده از Always On Failover Cluster Instance
- استفاده از Log Shipping و Database Mirroring

vmware®

آشنایی با ویژگی های بقای کسب و کار در vSphere

- استفاده از High Availability
- استفاده از DRS و SDRS
- استفاده از FT، NIOC و SIOC

vmware®

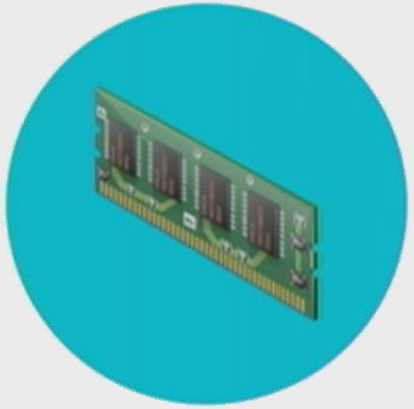
مدیریت سرویس MSSQL در محیط vSphere

- امکان Scale-Up و Scale-Out کردن
- امکان Right Sizing به منظور تثبیت کارآیی سرویس
- بهره مندی از ویژگی NUMA
- افزایش ضریب امنیت
- افزایش ضریب بقای سرویس
- افزایش ضریب بقای اطلاعات
- انجام فرآیندهای Maintenance با کارآیی بهتر

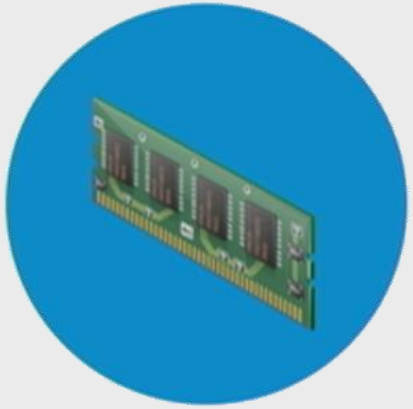
vmware®

Compute Resources

CPU



Memory



Disk

Network

بهروش های پیاده سازی سرویس SQL در محیط vSphere

- بهروش های مرتبط با تخصیص پردازنده و تاثیر NUMA بر عملکرد سرویس
- بهروش های مرتبط با حافظه اصلی و تاثیر الگوریتم های کنترلی ESXi
- بهروش های مرتبط با دیسک و تاثیر کنترلرها بر کیفیت سرویس
- بهروش های مرتبط با شبکه مجازی و تاثیر پیکربندی شبکه بر عملکرد سرویس
- بهروش های مرتبط با پیکربندی های مورد نیاز در سطح سیستم عامل
- بهروش های مرتبط با پیکربندی های مورد نیاز در سطح سرویس MS SQL

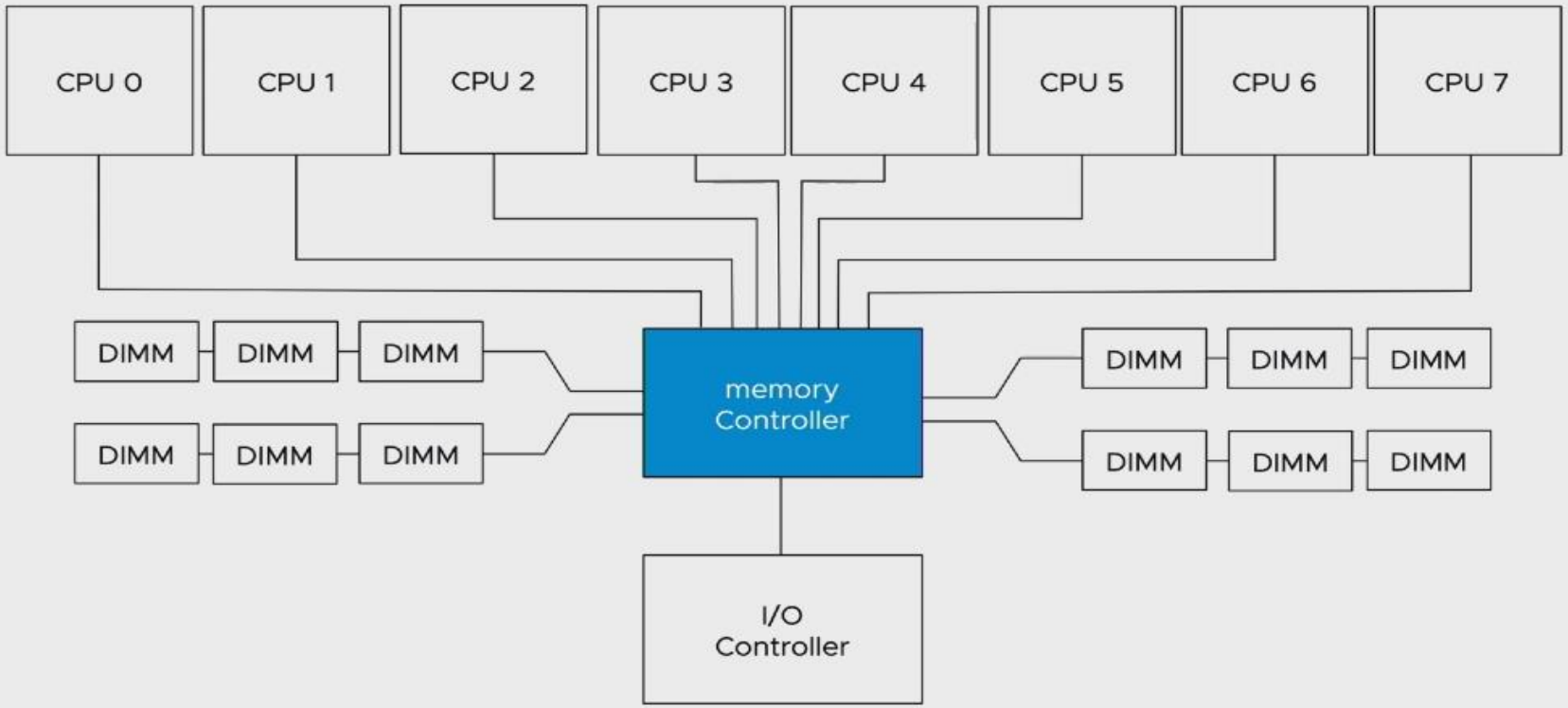
مفهوم Non-Uniform Memory Access

- لایه فیزیکی
- لایه VMKernel
- لایه Workload

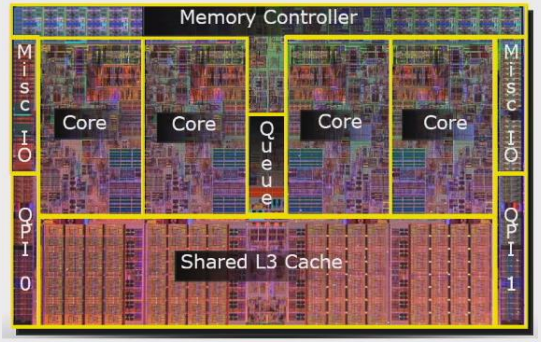
Physical Layer

Uniform Memory Access

Pre-AMD Opteron (2003) and Intel Nehalem (2007)

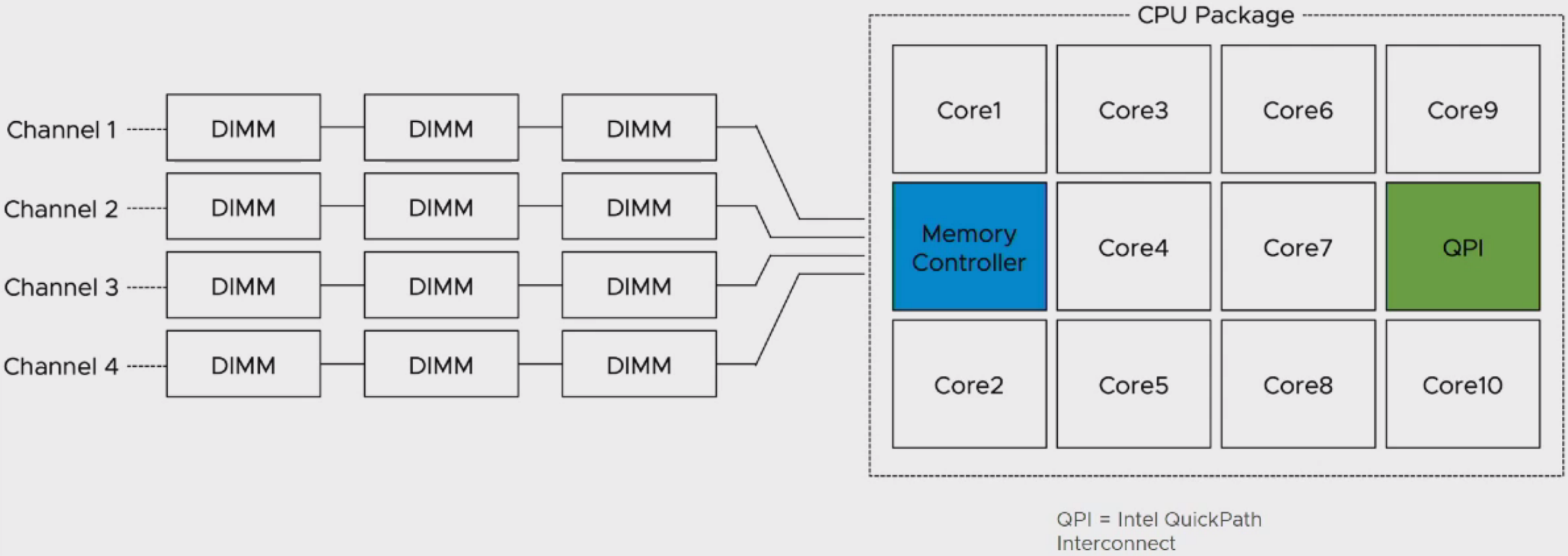


The First Nehalem Processor



On-Die Memory Channel

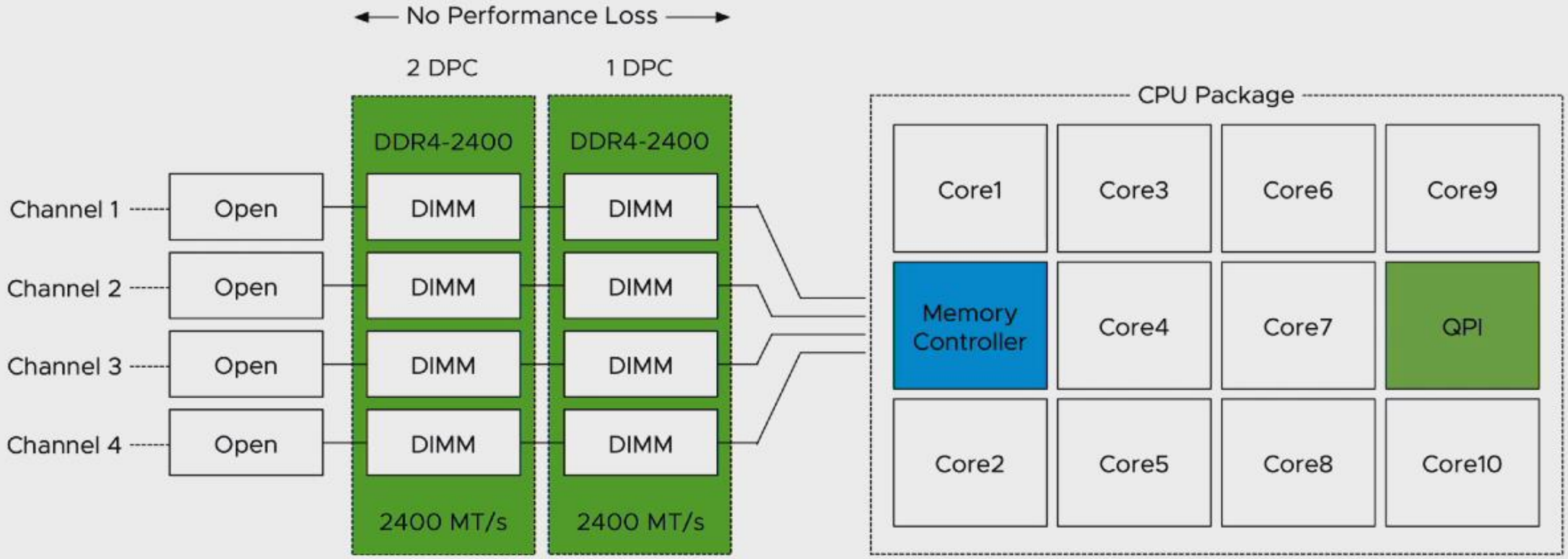
Low latency & high bandwidth connection to memory



QPI = Intel QuickPath Interconnect

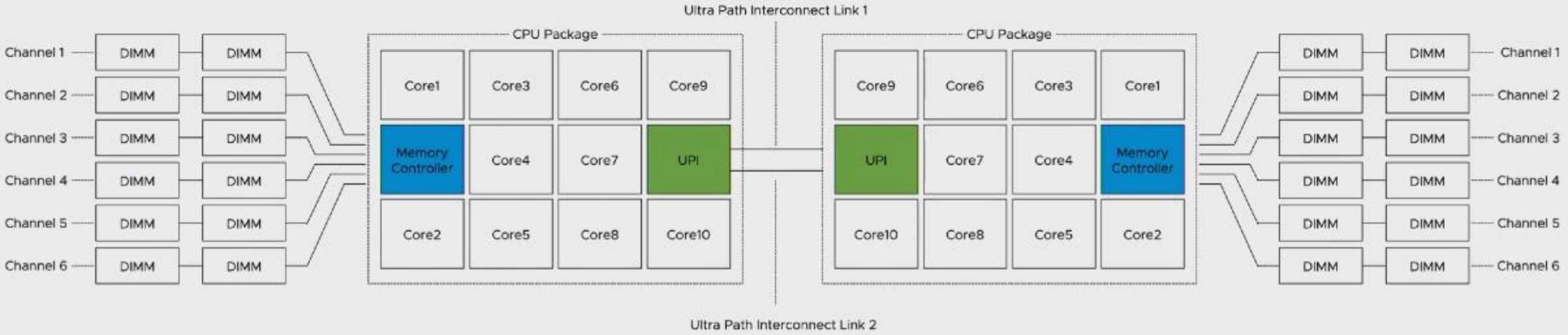
DIMMs per Memory Channel

2 DIMMS per Memory Channel = Optimal Memory Performance



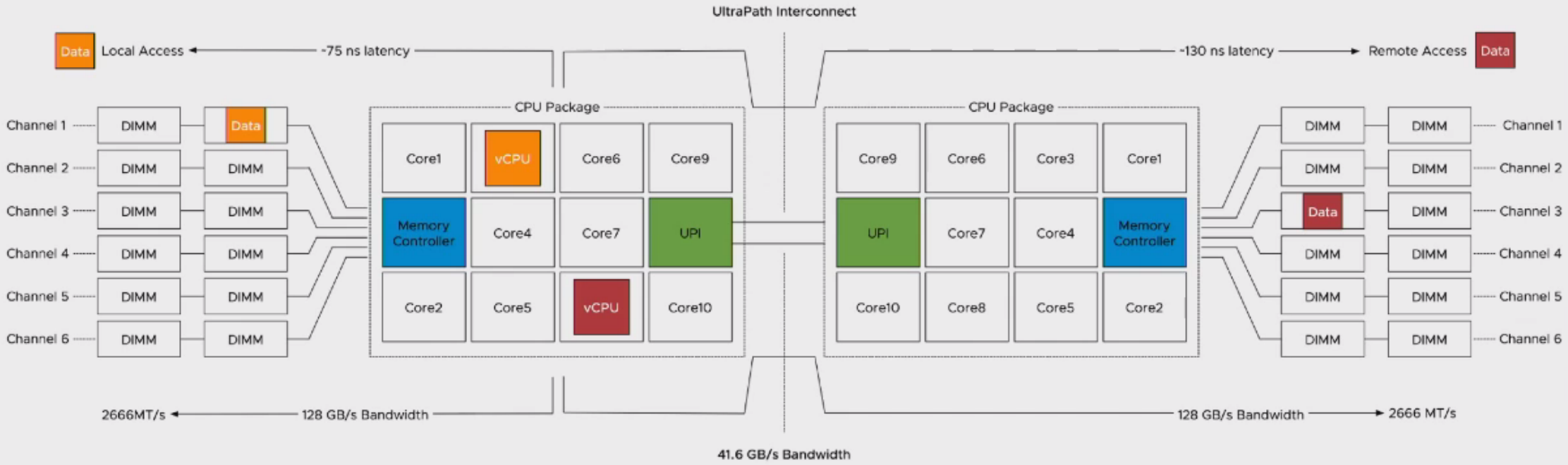
>90% of Data Center Servers are Dual-Socket Servers

Non-Memory Uniform Architecture



Local and Remote Memory Access in the Same Server

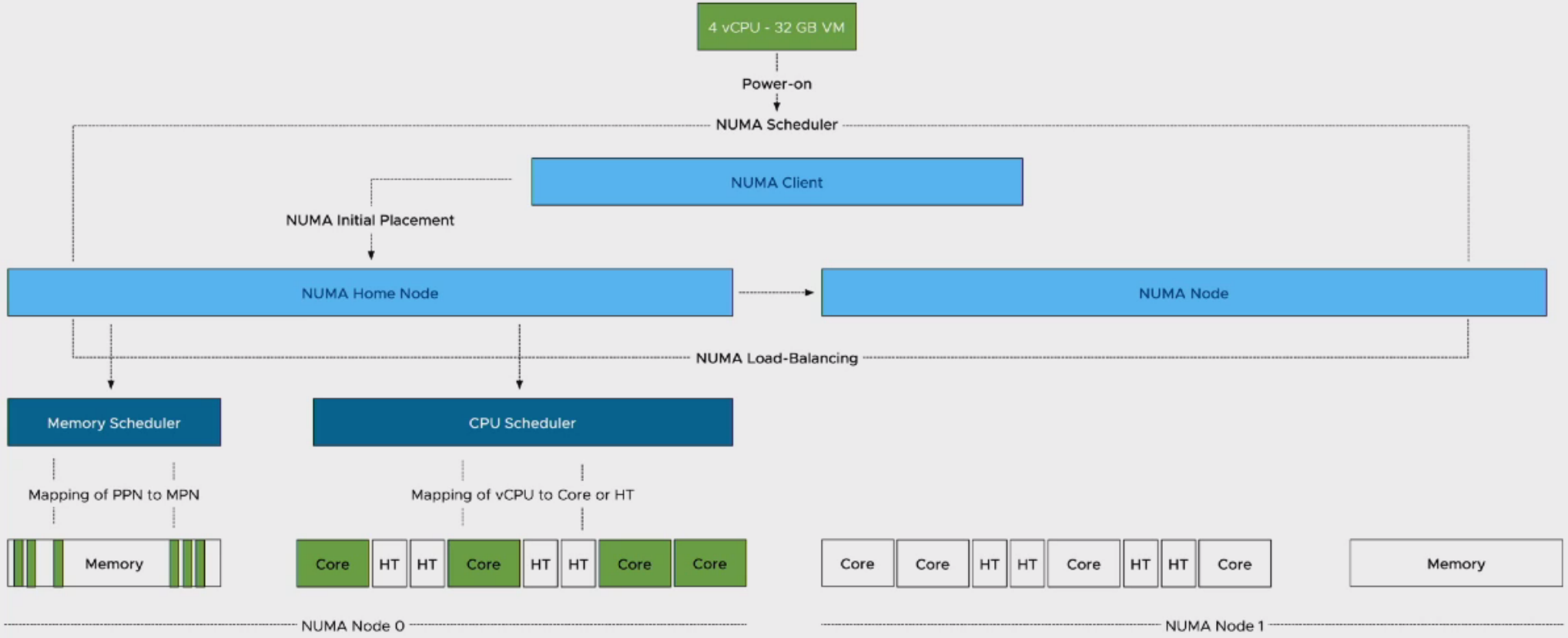
Inconsistent latency and bandwidth



VMKernel Layer

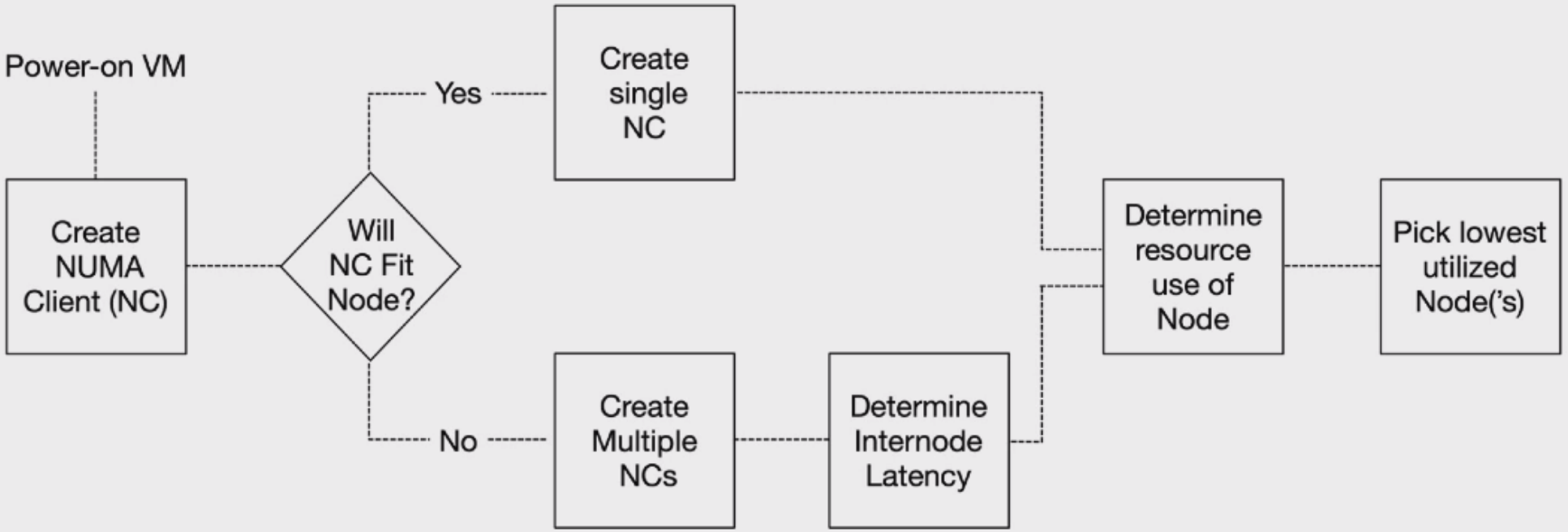
CPU, NUMA, and Memory Schedulers

How do these work together?



NUMA Initial Placement

Selecting the right NUMA node



Access Latencies

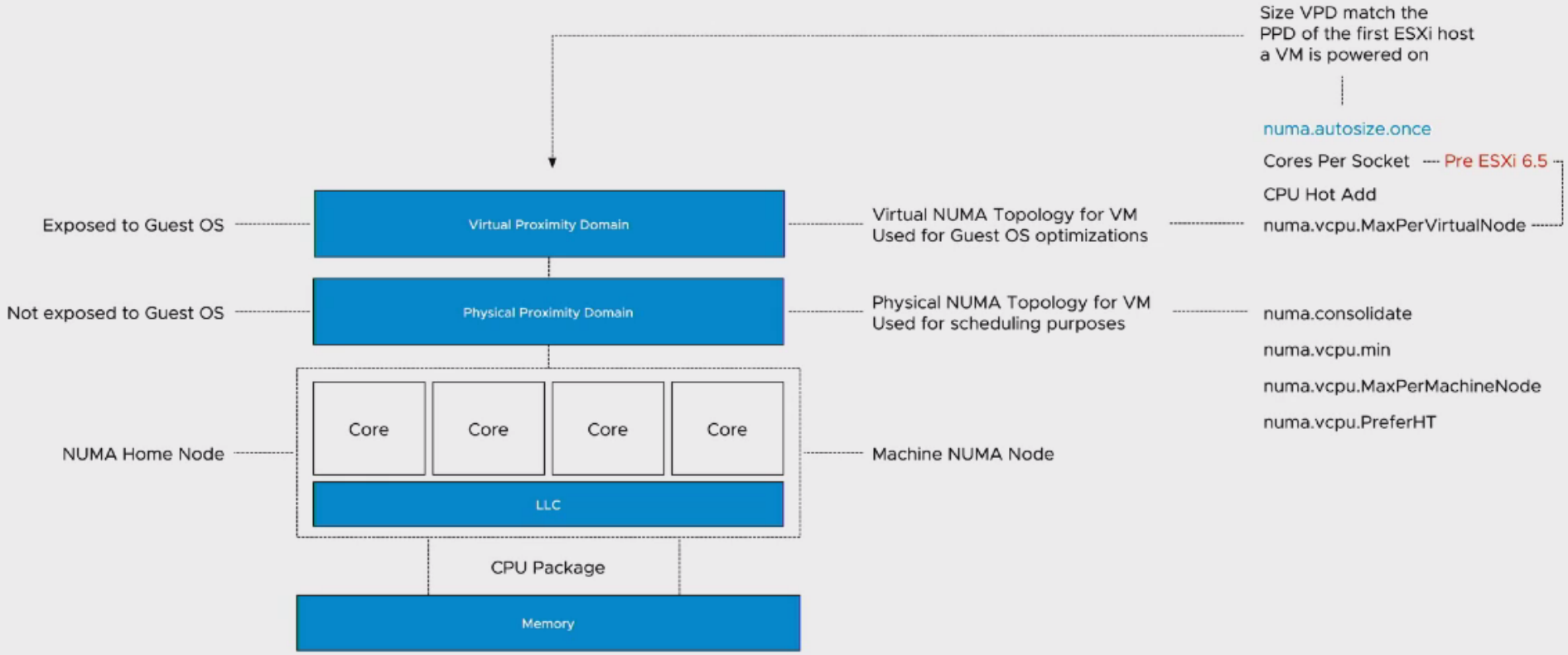
Why focus on LLC access?

System event	Actual latency	Human scaled latency
One CPU cycle (2.3 GHz)	0.4 ns	1 second
Level 1 cache access	1.6 ns	4 seconds
Level 2 cache access	4.8 ns	12 seconds
Level 3 cache access	15.2 ns	38 seconds
Remote L3 cache access	63 ns	157 seconds (2.5 minutes)
Local Memory Access	75 ns	188 seconds (3 minutes)
Remote Memory Access	130 ns	325 seconds (5 minutes)
Optane PMEM Access	~350 ns	875 seconds (15 minutes)
Optane SSD I/O	10 ms	7 Hours
NVMe SSD I/O	25 ms	17 Hours

Workload Layer

vNUMA Components

Advanced configurations impact the vNUMA composition



vNUMA Topology

```
[root@esxi00:~] vmdumper -l | cut -d \/ -f 2-5 | while
read path; do egrep -oi "DICT.*(displayname.*|numa.*|co
res.*|vcpu.*|memsize.*|affinity.*)= .*|numa:.*|numaHost
:.*" "/$path/vmware.log"; echo -e; done
DICT                 numvcpus = "10"
DICT                  memSize = "32768"
DICT        sched.cpu.affinity = "all"
DICT               displayName = "Large-VM"
numaHost: NUMA config: consolidation= 1 preferHT= 0
numa: coresPerSocket= 1 maxVcpusPerVPD= 10
numaHost: 10 VCPUs 1 VPDs 1 PPDs
numaHost: VCPU 0 VPD 0 PPD 0
numaHost: VCPU 1 VPD 0 PPD 0
numaHost: VCPU 2 VPD 0 PPD 0
numaHost: VCPU 3 VPD 0 PPD 0
numaHost: VCPU 4 VPD 0 PPD 0
numaHost: VCPU 5 VPD 0 PPD 0
numaHost: VCPU 6 VPD 0 PPD 0
numaHost: VCPU 7 VPD 0 PPD 0
numaHost: VCPU 8 VPD 0 PPD 0
numaHost: VCPU 9 VPD 0 PPD 0

[root@esxi00:~]
```

10 vCPU VM on 10 Core CPU Package

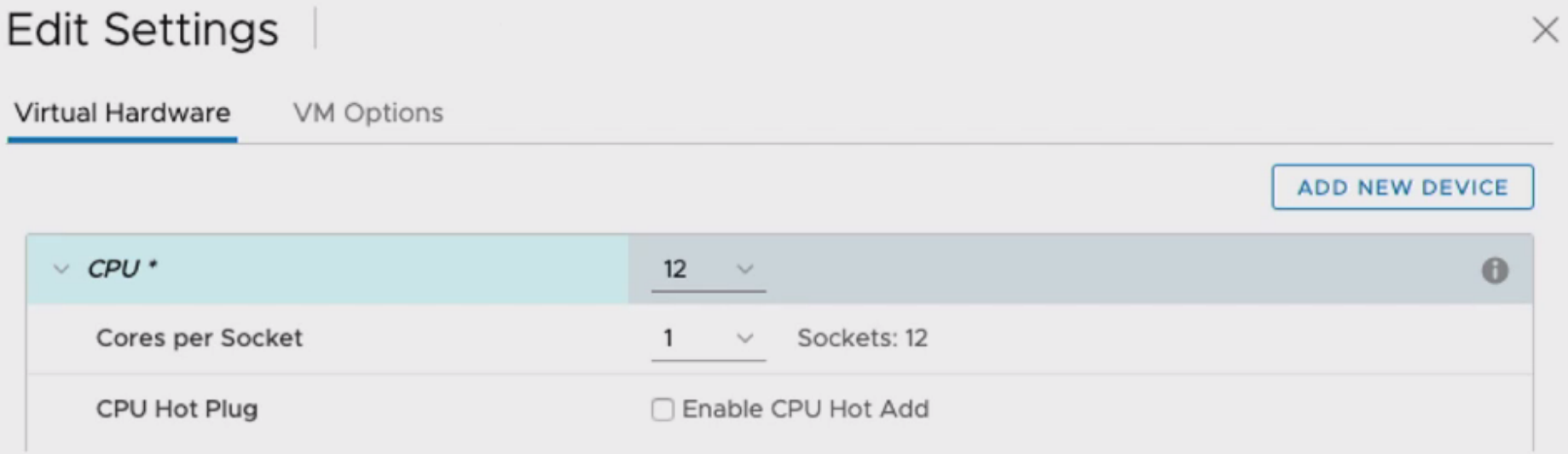
```
[root@esxi00:~] vmdumper -l | cut -d \/ -f 2-5 | while
read path; do egrep -oi "DICT.*(displayname.*|numa.*|co
res.*|vcpu.*|memsize.*|affinity.*)= .*|numa:.*|numaHost
:.*" "/$path/vmware.log"; echo -e; done
DICT                 numvcpus = "12"
DICT                  memSize = "32768"
DICT        sched.cpu.affinity = "all"
DICT               displayName = "Large-VM"
DICT      numa.autosize.cookie = "100001"
DICT numa.autosize.vcpu.maxPerVirtualNode = "10"
numaHost: NUMA config: consolidation= 1 preferHT= 0
numa: coresPerSocket= 1 maxVcpusPerVPD= 6
numaHost: 12 VCPUs 2 VPDs 2 PPDs
numaHost: VCPU 0 VPD 0 PPD 0
numaHost: VCPU 1 VPD 0 PPD 0
numaHost: VCPU 2 VPD 0 PPD 0
numaHost: VCPU 3 VPD 0 PPD 0
numaHost: VCPU 4 VPD 0 PPD 0
numaHost: VCPU 5 VPD 0 PPD 0
numaHost: VCPU 6 VPD 1 PPD 1
numaHost: VCPU 7 VPD 1 PPD 1
numaHost: VCPU 8 VPD 1 PPD 1
numaHost: VCPU 9 VPD 1 PPD 1
numaHost: VCPU 10 VPD 1 PPD 1
numaHost: VCPU 11 VPD 1 PPD 1

[root@esxi00:~]
```

12 vCPU VM on 10 Core CPU Package

Cores per Socket

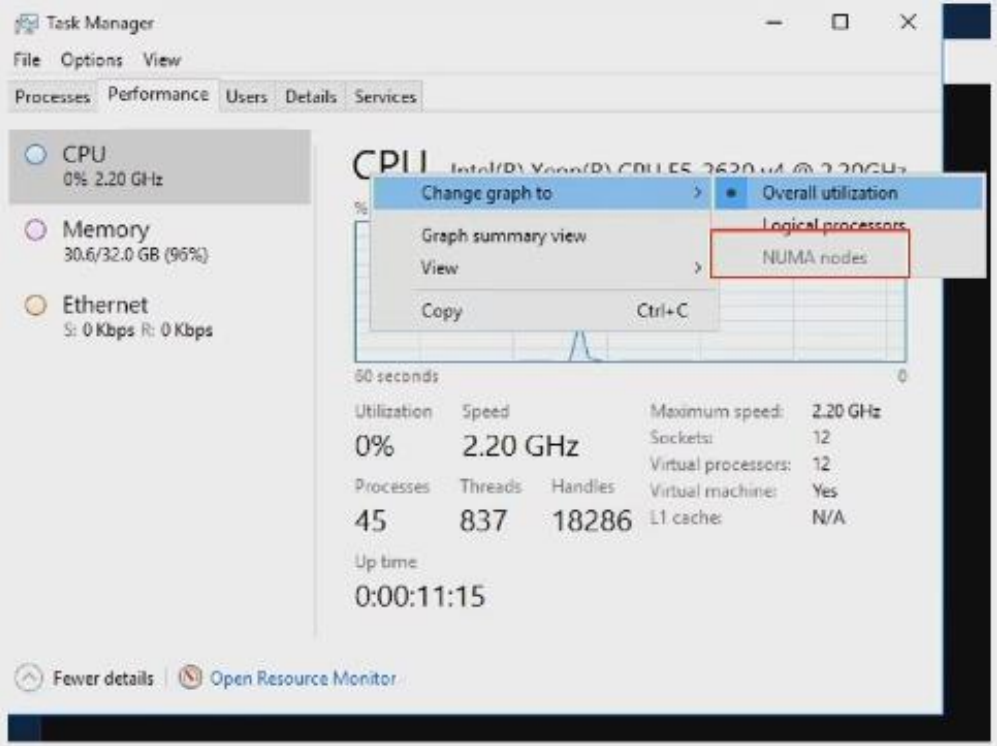
By default set to 1



CPU Hot Add

12 vCPU configuration. No vNUMA topology created

```
DICT                    numvcpus = "12"
DICT                     memSize = "32768"
DICT          sched.cpu.affinity = "all"
DICT                 displayName = "WX08"
DICT                 vcpu.hotadd = "TRUE"
DICT        numa.autosize.cookie = "40001"
DICT             numa.consolidate = "true"
numaHost: NUMA config: consolidation= 1 preferHT= 0
numa: Hot add is enabled and vNUMA hot add is disabled, forcing UMA.
numaHost: 12 VCPUs 1 VPDs 2 PPDs
numaHost: VCPU 0 VPD 0 PPD 0
numaHost: VCPU 1 VPD 0 PPD 0
numaHost: VCPU 2 VPD 0 PPD 0
numaHost: VCPU 3 VPD 0 PPD 0
numaHost: VCPU 4 VPD 0 PPD 0
numaHost: VCPU 5 VPD 0 PPD 0
numaHost: VCPU 6 VPD 0 PPD 1
numaHost: VCPU 7 VPD 0 PPD 1
numaHost: VCPU 8 VPD 0 PPD 1
numaHost: VCPU 9 VPD 0 PPD 1
numaHost: VCPU 10 VPD 0 PPD 1
numaHost: VCPU 11 VPD 0 PPD 1
```



Databases & CPU

How Many vCPUs / Rightsizing?

How many will be used?

Ignore the vendor req's

Knowledge of workload

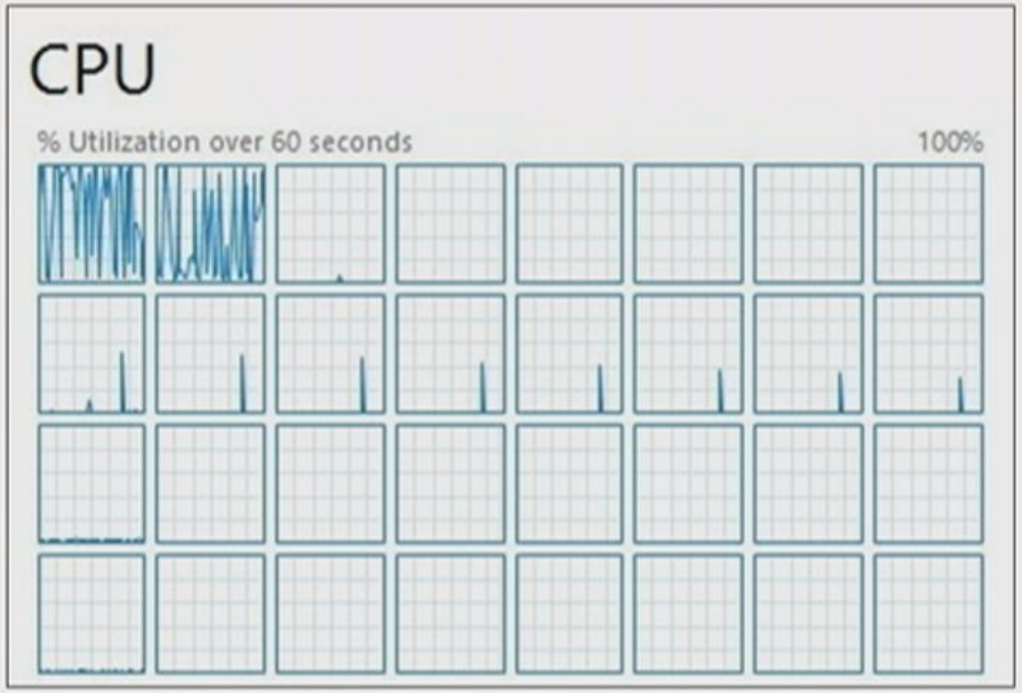
Size for now (plus headroom).

Can resize later

Target 40-60% utilization during biz hours

Monitoring & trending

Optimize NUMA placement (SQL Ent)

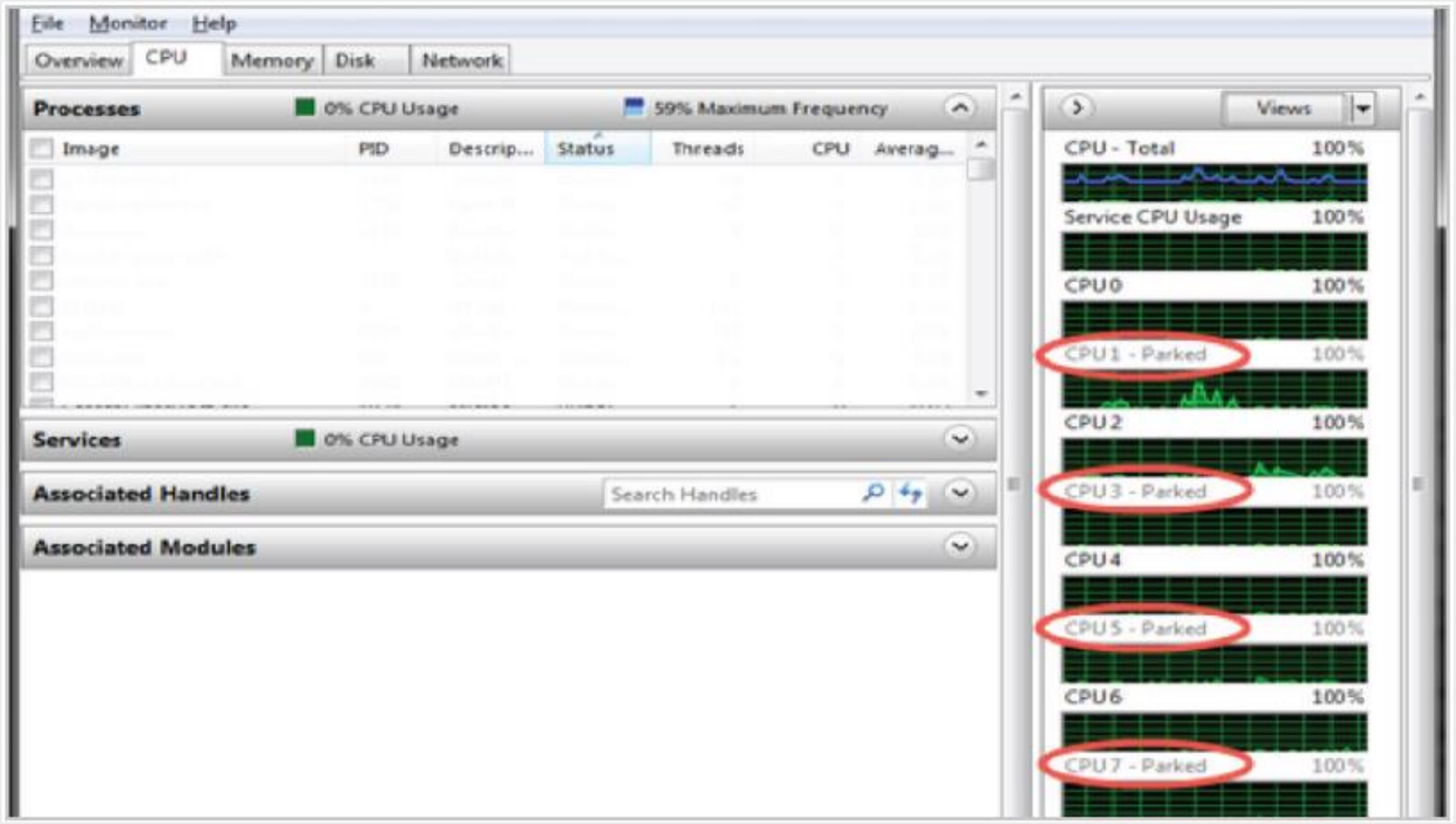


SQL Server Soft-NUMA

SQL Server 2016 and above

ProcessInfo	Text
Server	Machine supports memory error recovery. SQL memory protection is enabled to recover from memory corruption.
Server	Default collation: SQL_Latin1_General_CP1_CI_AS (us_english 1033)
Server	Automatic soft-NUMA was enabled because SQL Server has detected hardware NUMA nodes with greater than 8 physical cores.
Server	Buffer pool extension is already disabled. No action is necessary.
Server	InitializeExternalUserGroupSid failed. Implied authentication will be disabled.
Server	Implied authentication manager initialization failed. Implied authentication will be disabled.
Server	The maximum number of dedicated administrator connections for this instance is '1'
Server	This instance of SQL Server last reported using a process ID of 20172 at 8/15/2017 9:55:01 AM (local) 8/15/2017 1:55:01 PM (UTC)....

Server Hardware Power Considerations



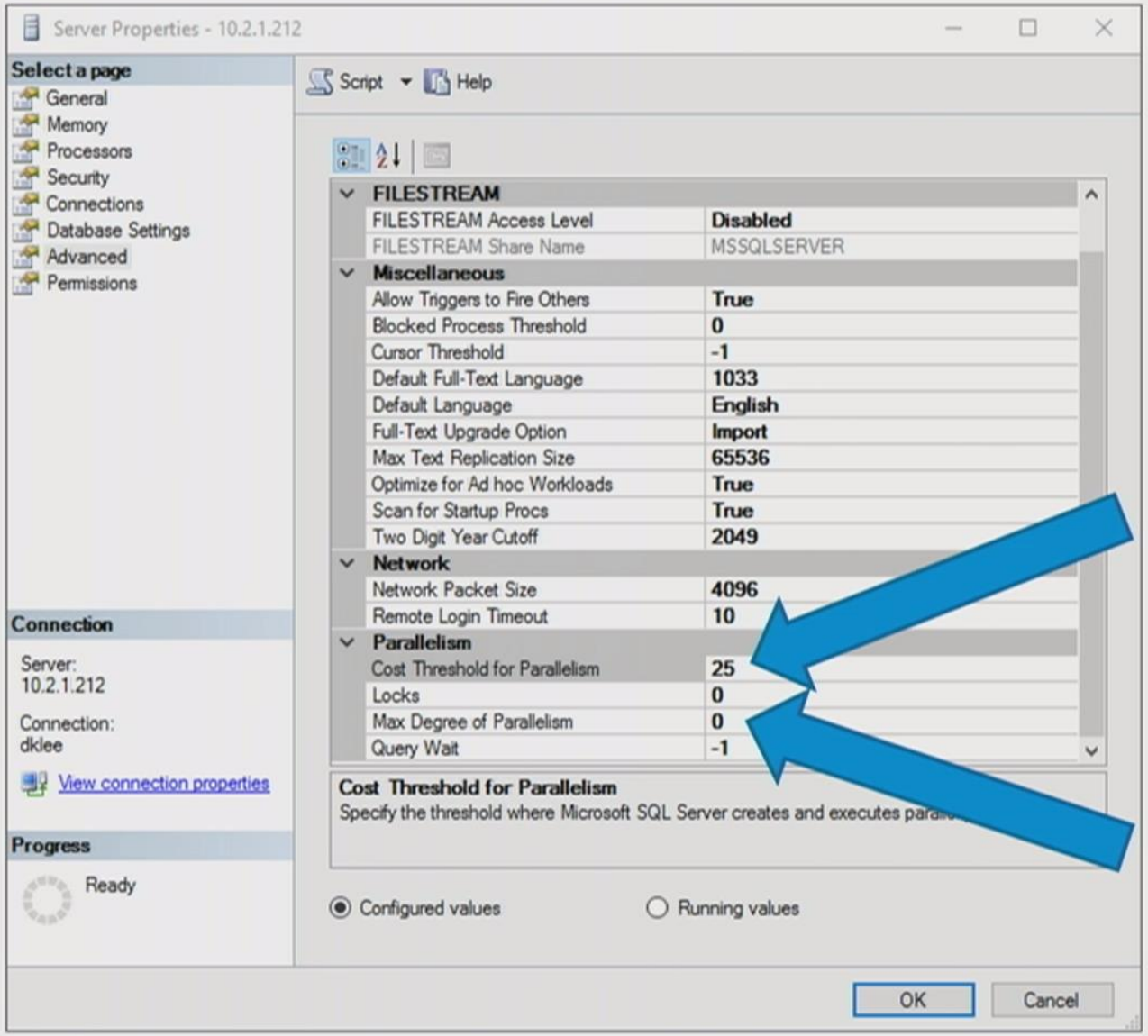
Performance Analysis

Metric	%Ready	%CoStop	%I/O Wait	%Demand
Investigation threshold	>10	>3	>5	>100

%Ready	If the threshold is exceeded, over-provisioning of vCPU, excessive usage of vSMP or a limit (check %MLMTD) has been set. This %RDY value is the sum of all vCPUs %RDY for a virtual machine. For example, if the max value of %RDY of 1vCPU is 100% and 4vCPU is 400%. If %RDY is 20 for 1 vCPU then this is problematic, as it means 1 vCPU is waiting 20% of the time for VMkernel to schedule it.
%CoStop	If the threshold is exceeded this indicates excessive usage of vSMP. Decrease amount of vCPUs for this particular virtual machine.
%I/O wait	If the threshold is >5, the virtual machine is waiting for swapped pages to be read from disk. You may have overcommitted memory.
% Demand	If >100%, the allocated ressources are not sufficient and additional capacity may be required

Parallelism

Microsoft® SQL Server®



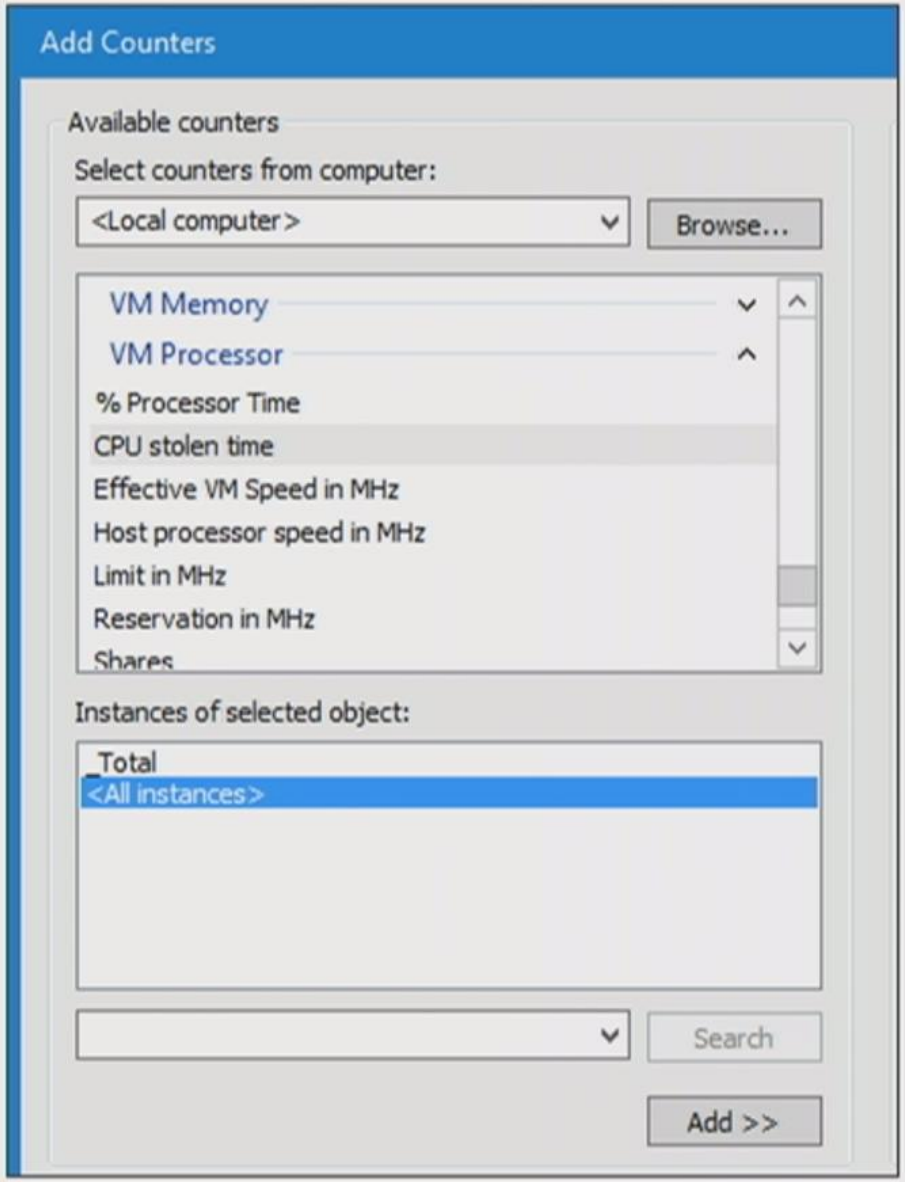
In-Guest Measurement

Windows Perfmon counter

Included with VMware Tools

VM Processor : CPU Stolen Time

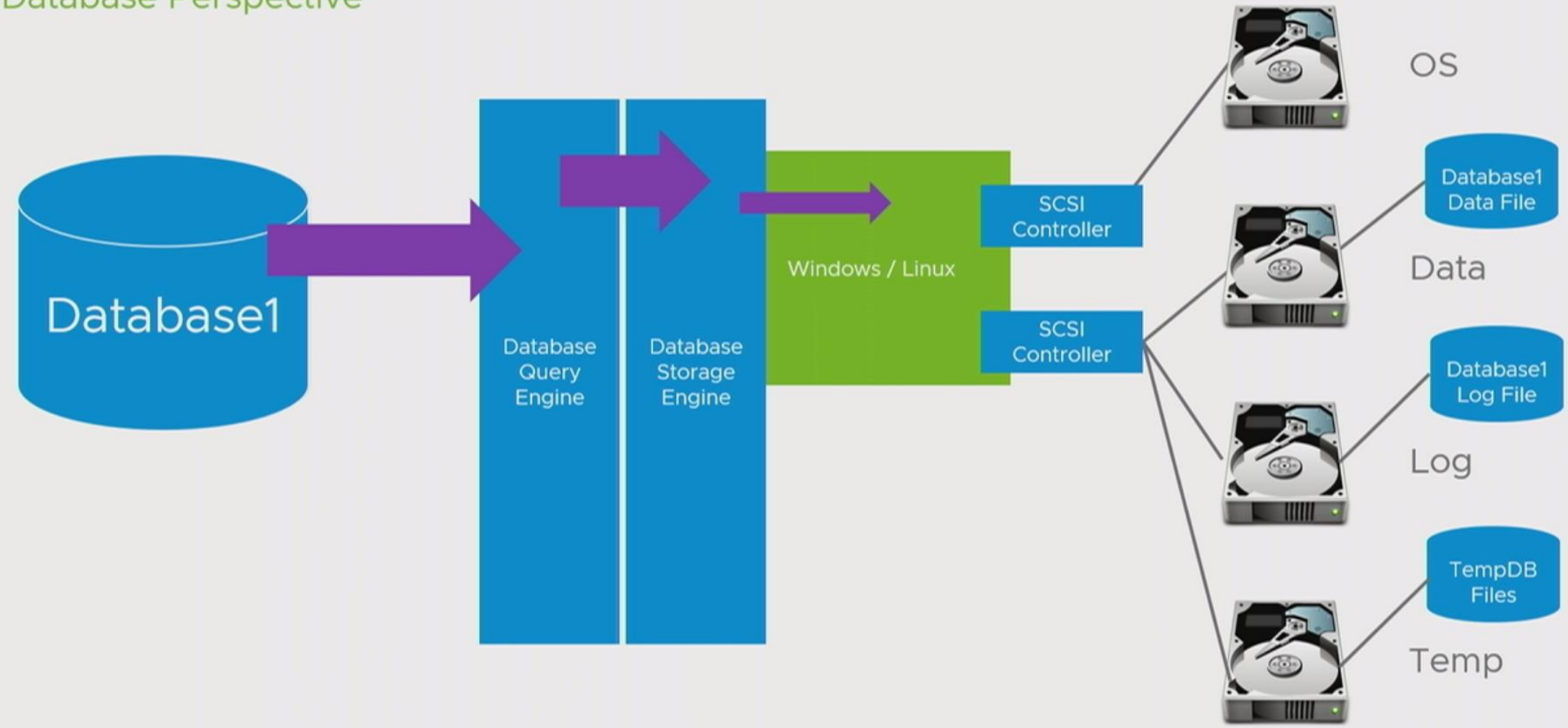
Convert to percent



Storage

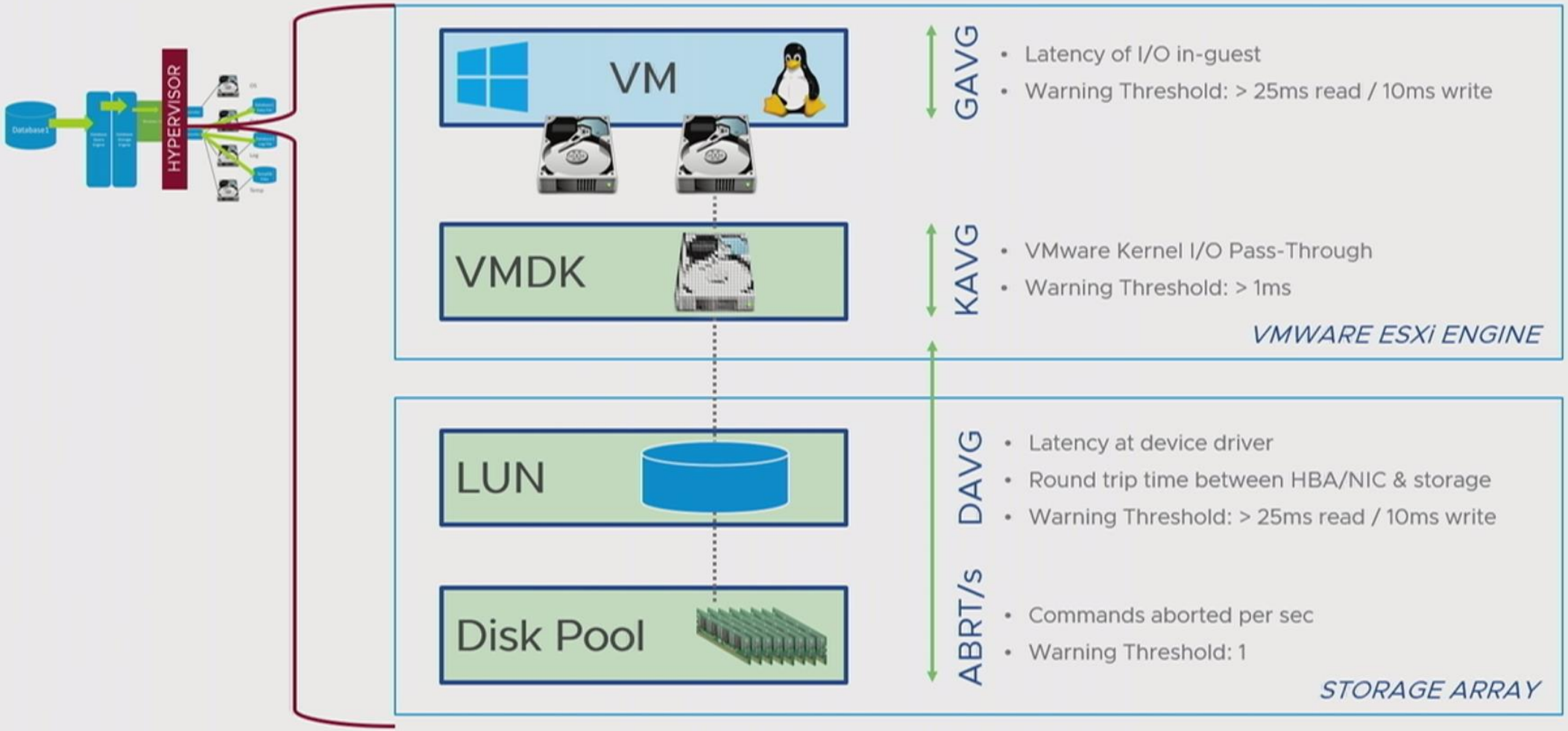
Virtual Disks & Queueing

Database Perspective



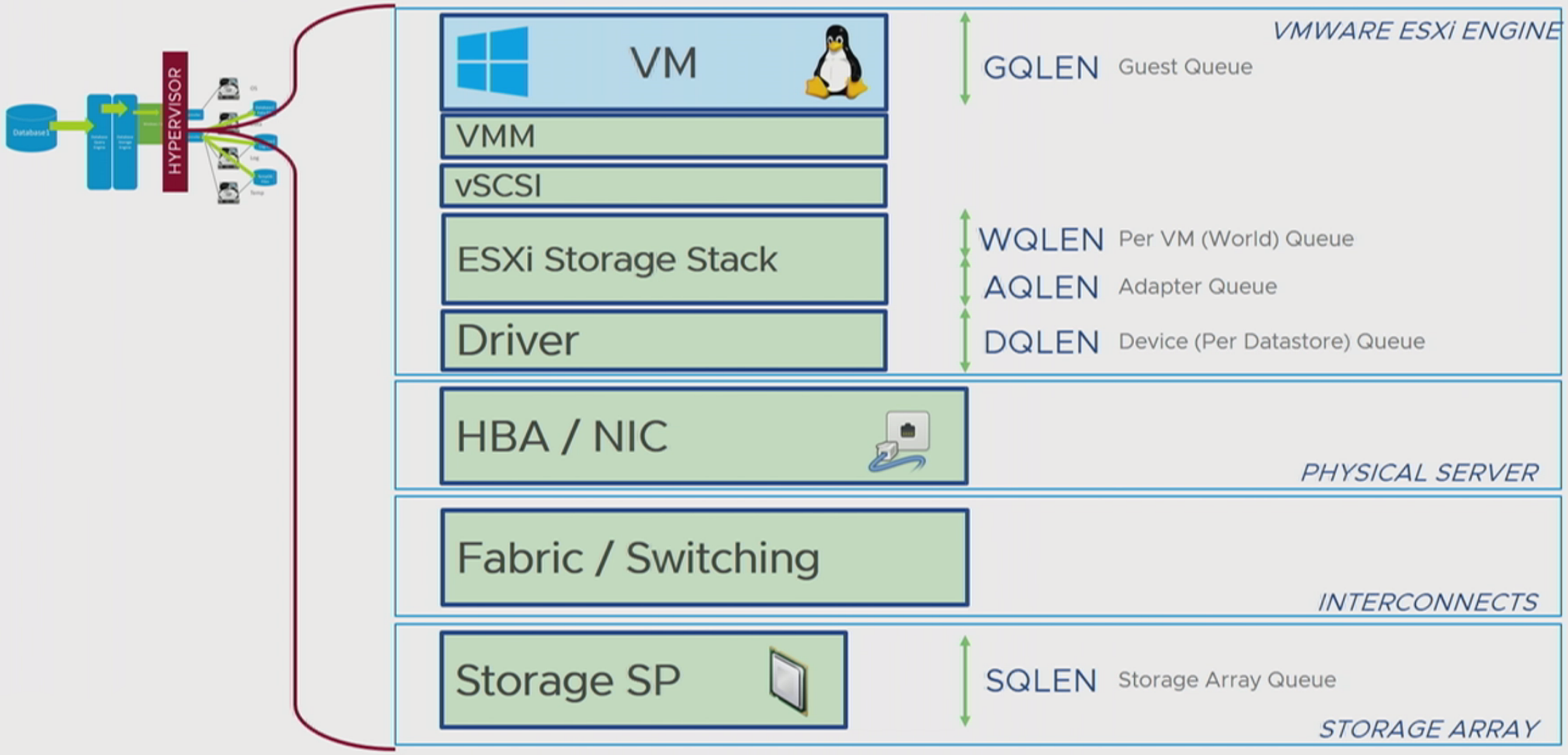
Virtual Disks Latencies

VMware's Logical Perspective



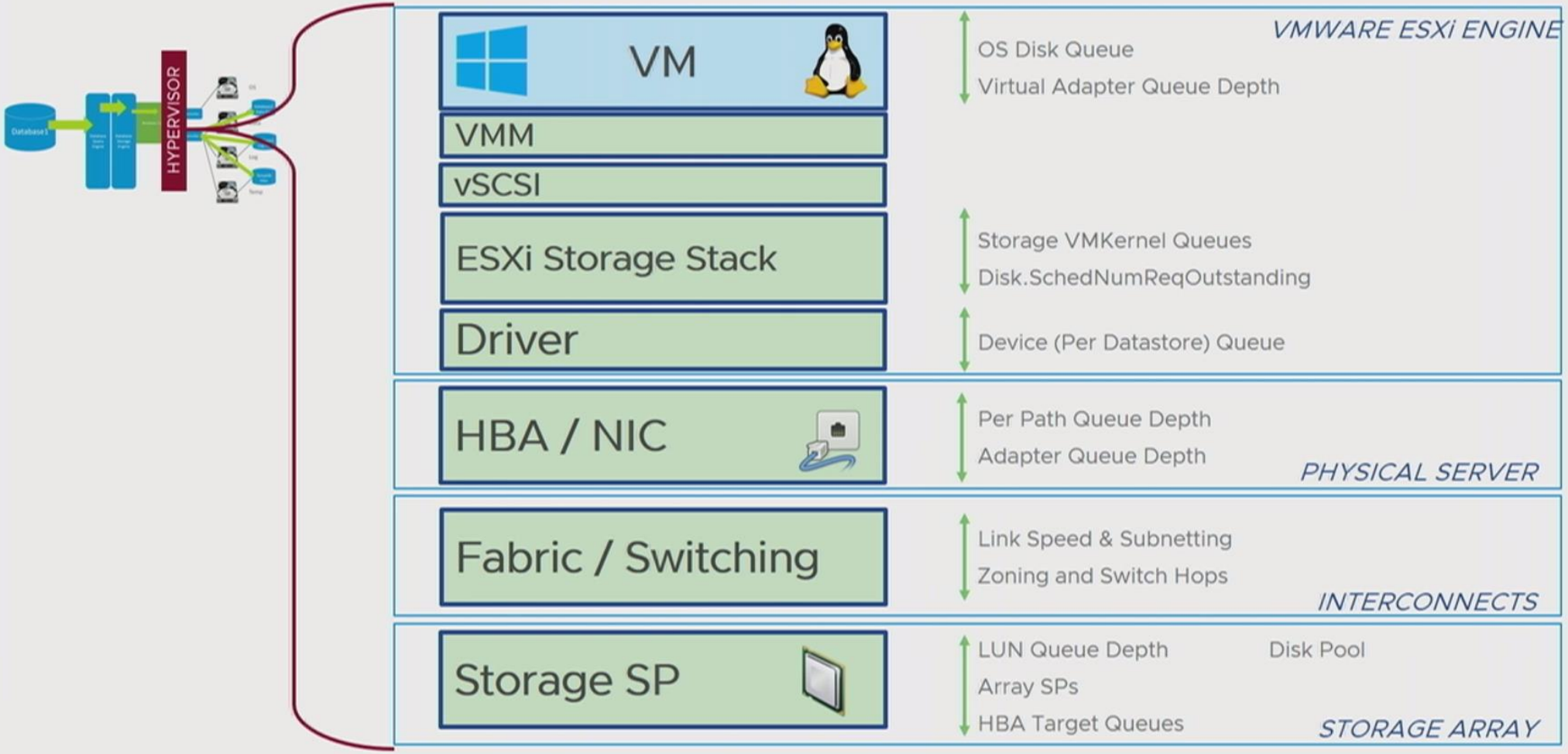
Virtual Disks Queueing

VMware's Logical Perspective



Virtual Disks Queue Depths

VMware's Logical Perspective



Change PVSCSI Queue Depths

Default device queue depth 64

Can override device to 254

Don't blindly override – *test* and validate gains

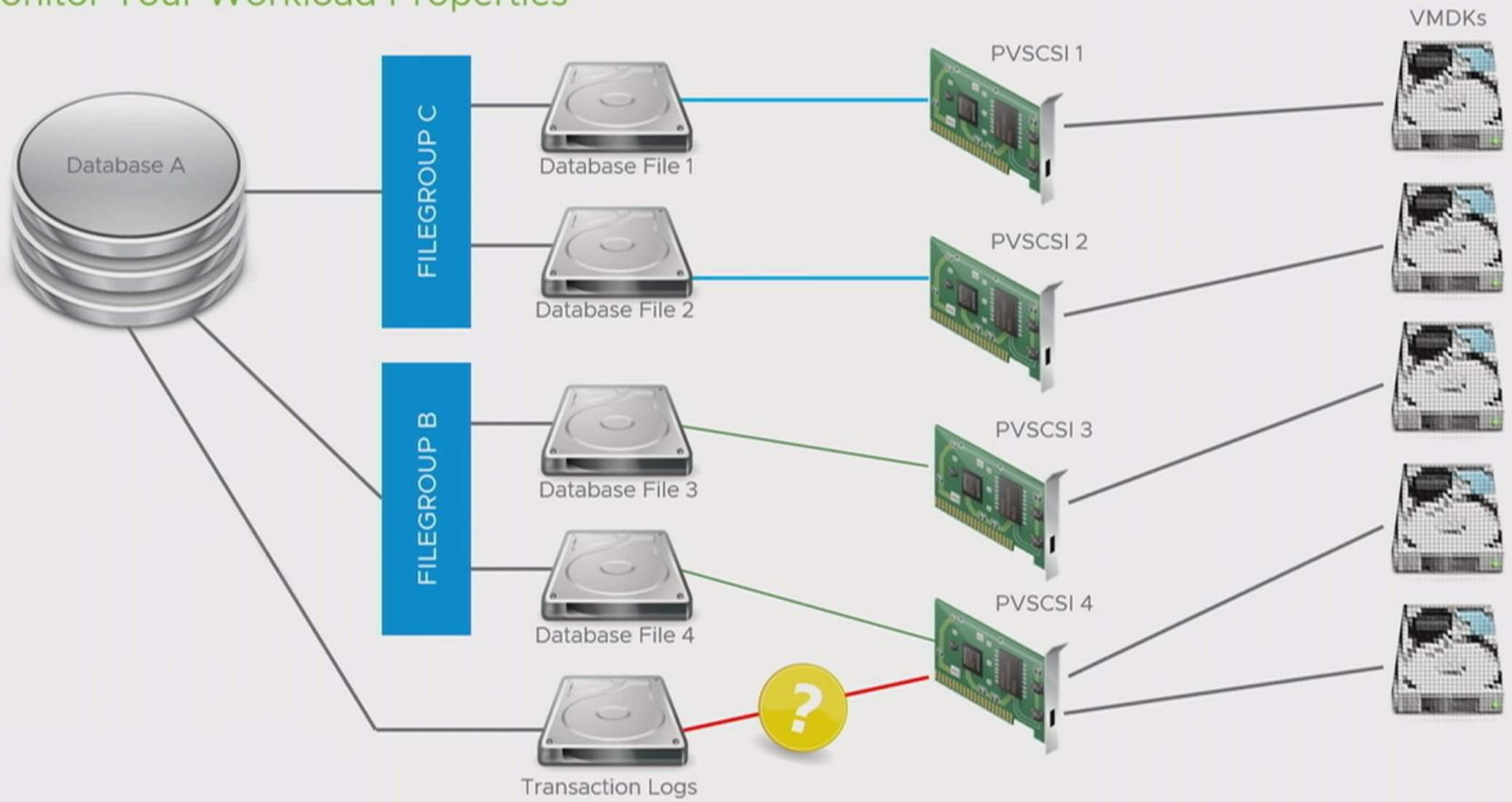
https://kb.vmware.com/s/article/1017423

Flash & hybrid benefit more than traditional spindle storage

```
REG ADD HKLM\SYSTEM\CurrentControlSet\services\pvscsi\Parameters\Device /v DriverParameter /t REG_SZ /d
"RequestRingPages=32,MaxQueueDepth=254"
```

SQL Server Object Placement

Monitor Your Workload Properties

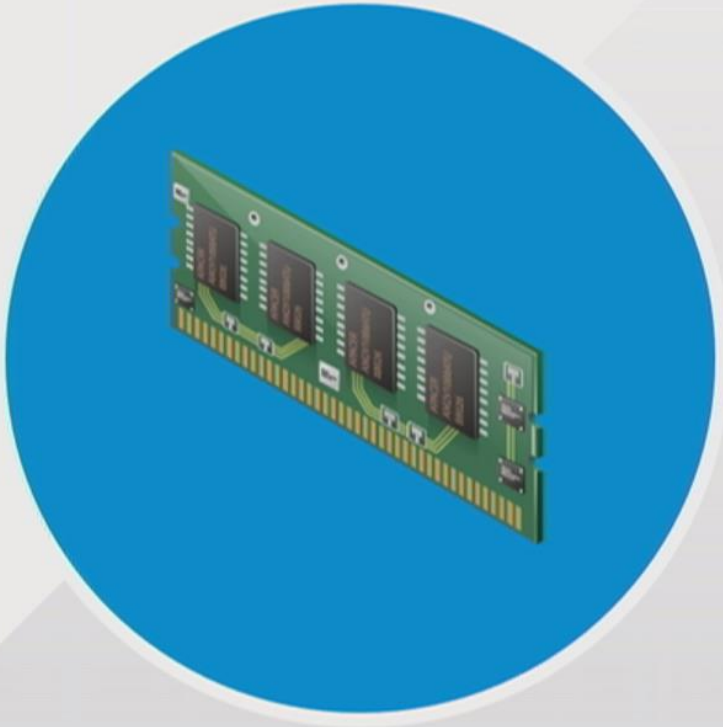


Performance Analysis

Metric	Read Latency (ms)	Write Latency (ms)	Used Space %	Snapshot age (days)
Investigation threshold	>10ms	>10ms	>80%	>3

Metric	Description
Read Latency (ms)	If higher than 10 ms on magnetic disks, >3ms on flash array : user may experience slowness
Write Latency (ms)	If higher than 10 ms on magnetic disks, >3ms on flash array : user may experience slowness
Used Space %	If higher than 80% you may face filesystem or datastore full. Snapshot space is necessary for backup activities or temporary maintenance needs
Snapshot age (days)	As a best practice avoid running production on older than 3 days snapshots. This my slow down the virtual machine and generate co-stop if the snapshot size is large.

Memory



Memory Management

Zero Overcommitment, No Problem

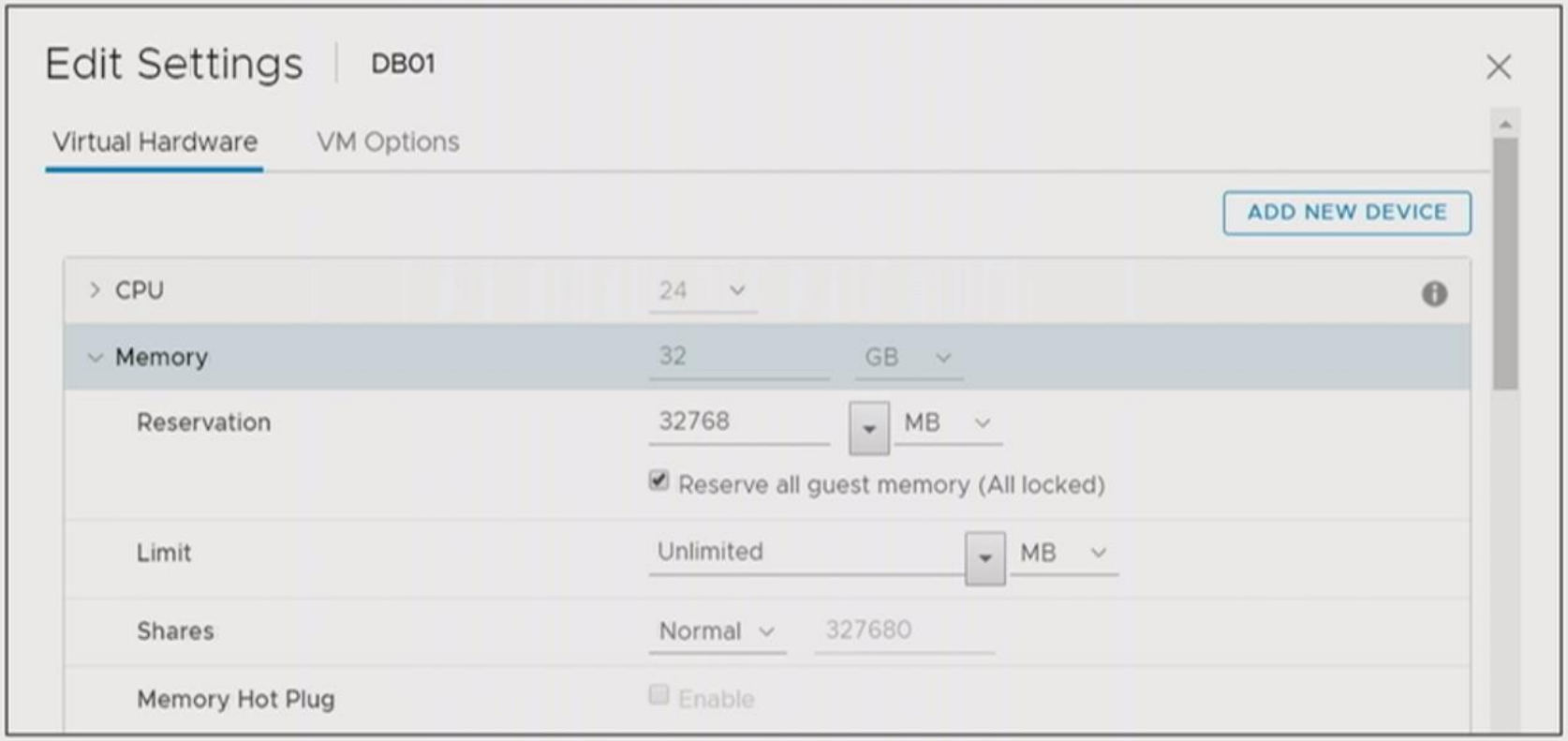
Easier (theoretically)

NO overcommitment of host RAM

Reserve all guest memory

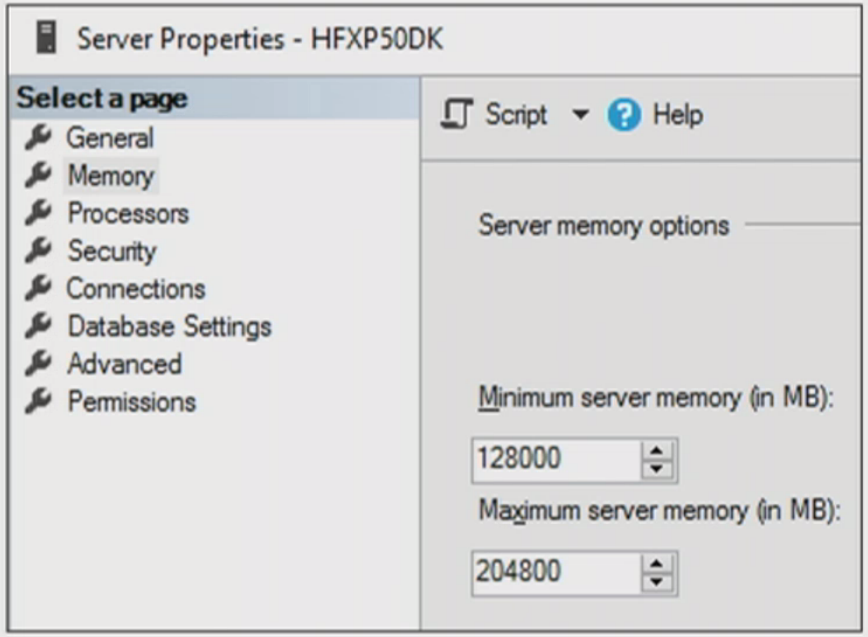
Active memory != memory ‘used’ by database engine

Do not use Active memory counter for any capacity estimation



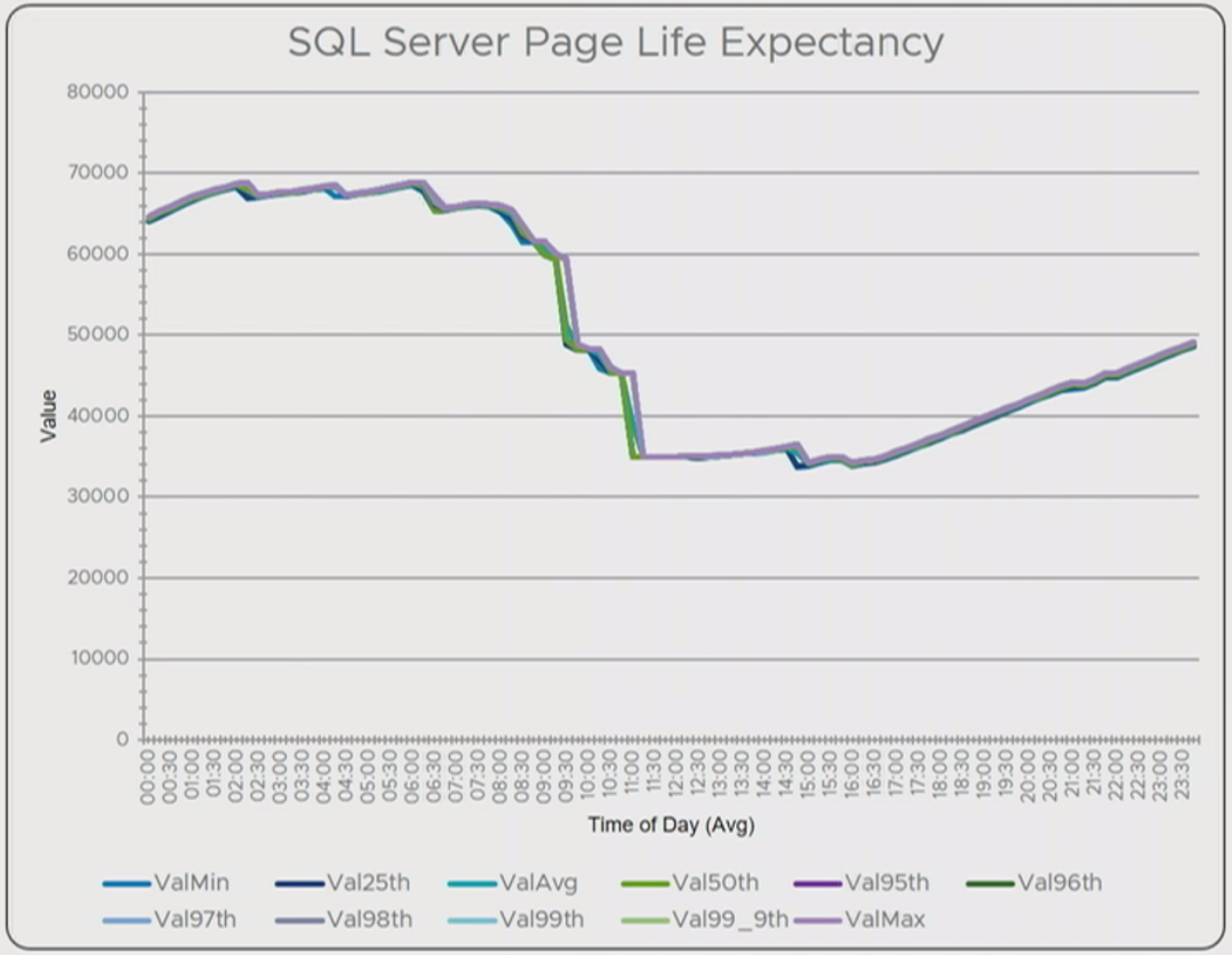
SQL Server Memory

Ask Your DBA About DB Memory Usage



SQL Server: Large memory pages

Oracle: Huge pages

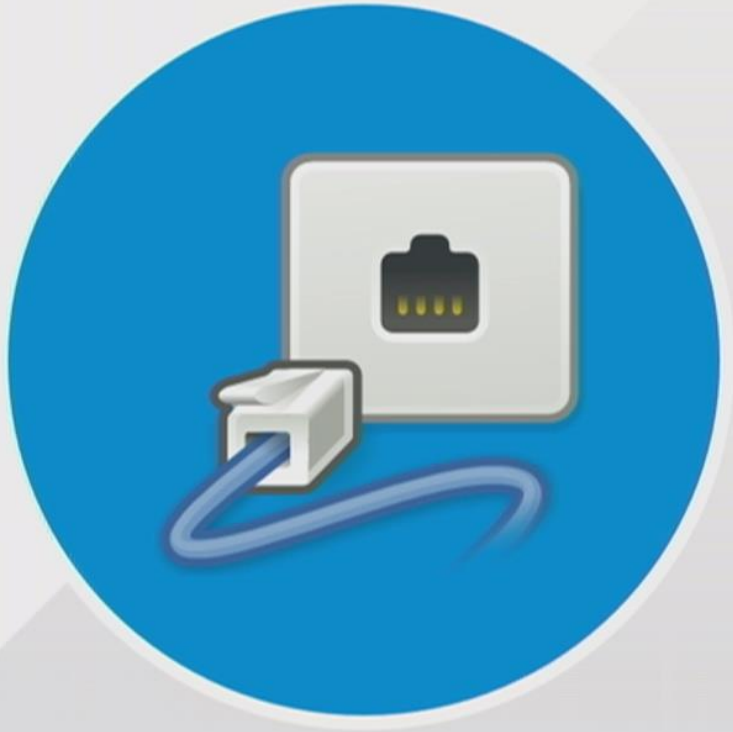


Troubleshooting Memory

Metric	Balloon %	Contention %	Swapped KB
Investigation threshold	>0	>0	>0

Metric	Description
Balloon(%)	If larger than 0, host is forcing virtual machine to inflate balloon driver to reclaim memory as the host is overcommitted
Contention(%)	If larger than 0 host has swapped memory pages and this is the % of time VMs are waiting to access swapped memory.
Swapped	Amount of guest memory beeing swapped. The memory will stay swapped even after congestion period. Need to be actively unswapped.

Networking



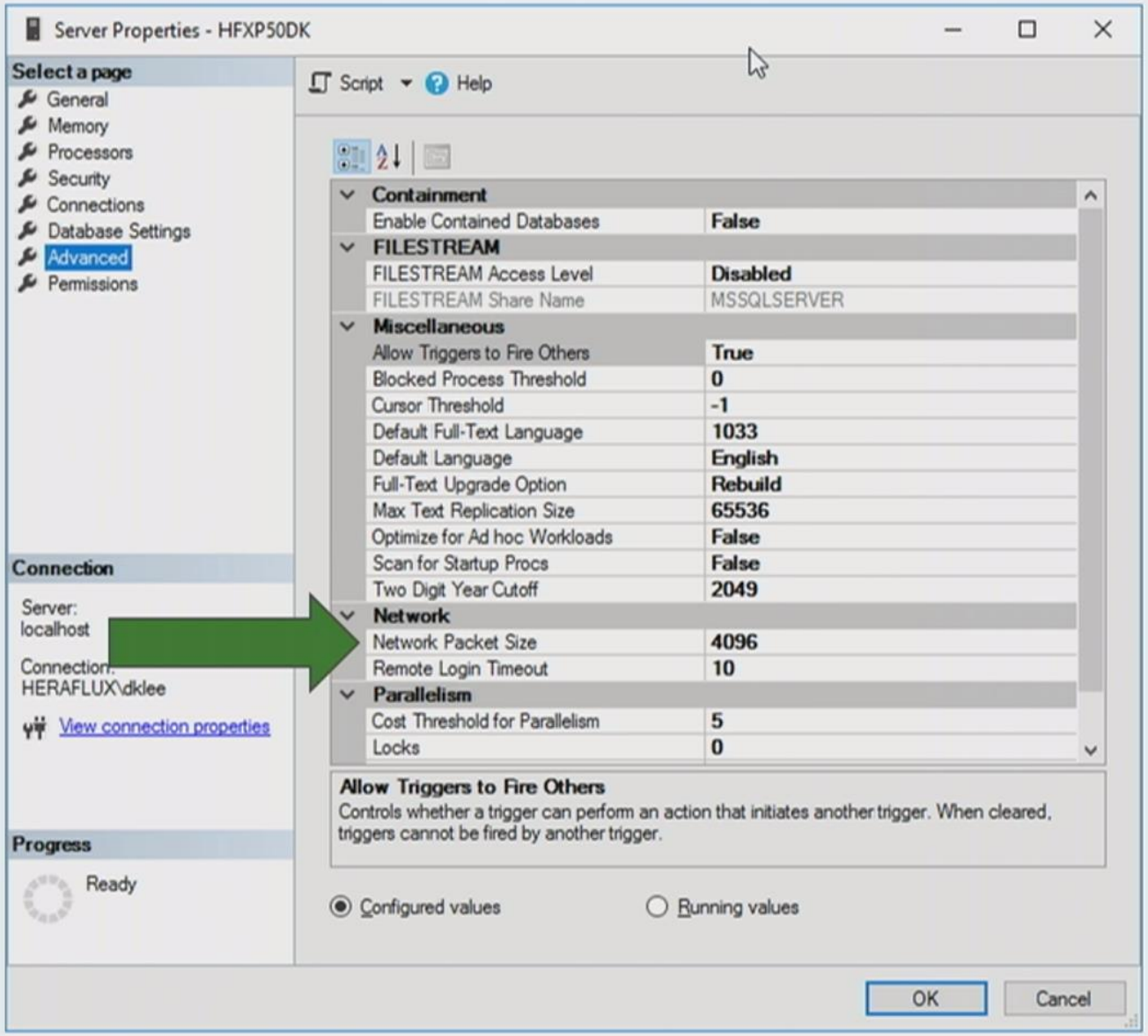
SQL Server Networking

Default packet size 4KB

Networking jumbo frames?

Does the workload demand it?

Network packet drop %



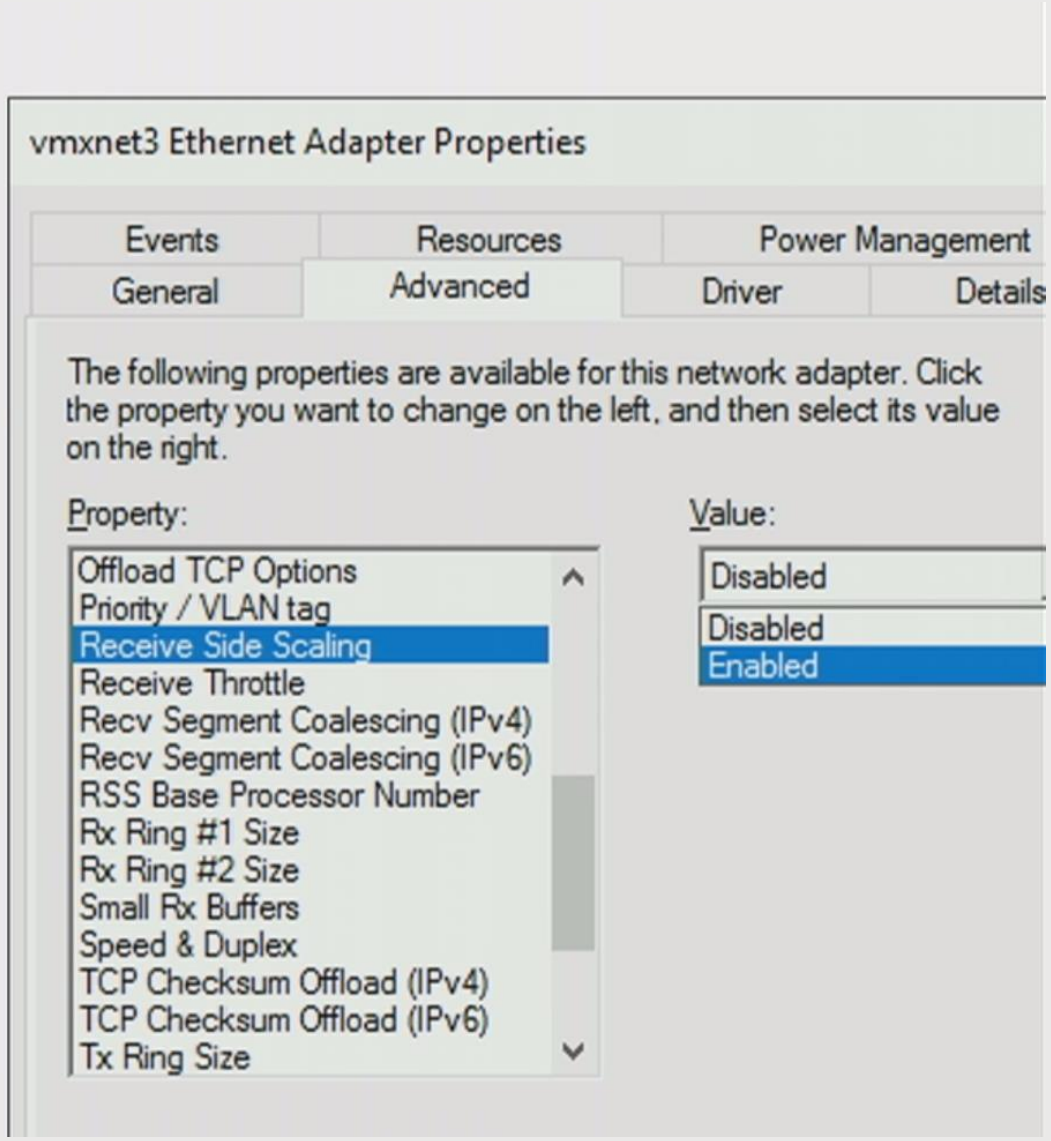
In-Guest Network Tuning

Use VMXNET3 driver
- 10GbE throughput

Jumbo frames?
- Usually not necessary in-guest

Enable Windows Receive Side Scaling
- https://kb.vmware.com/s/article/2008925

```
netsh interface tcp set global rss=enabled
```



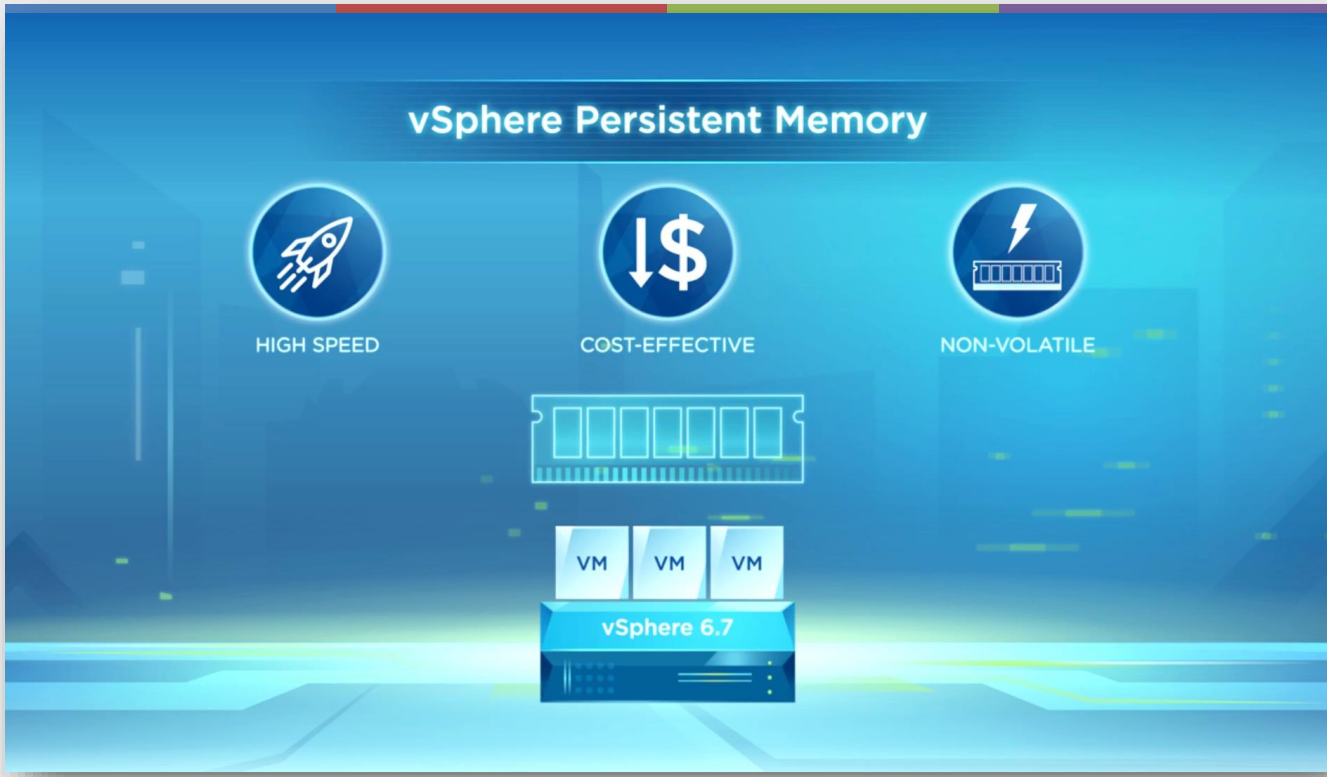
Performance Analysis

Metric	%DRPTX	%DRPRX
Investigation threshold	>0	>0

Metric	Description
%DRPTX	If larger than 0 transmit packets are being dropped, hardware is overworked due to high network utilization. From a network team standpoint investigations will be needed when >2%
%DRPRX	If larger than 0 receive packets are being dropped, hardware is overworked due to high network utilization. From a network team standpoint investigations will be needed when >2%

vSphere Persistent Memory

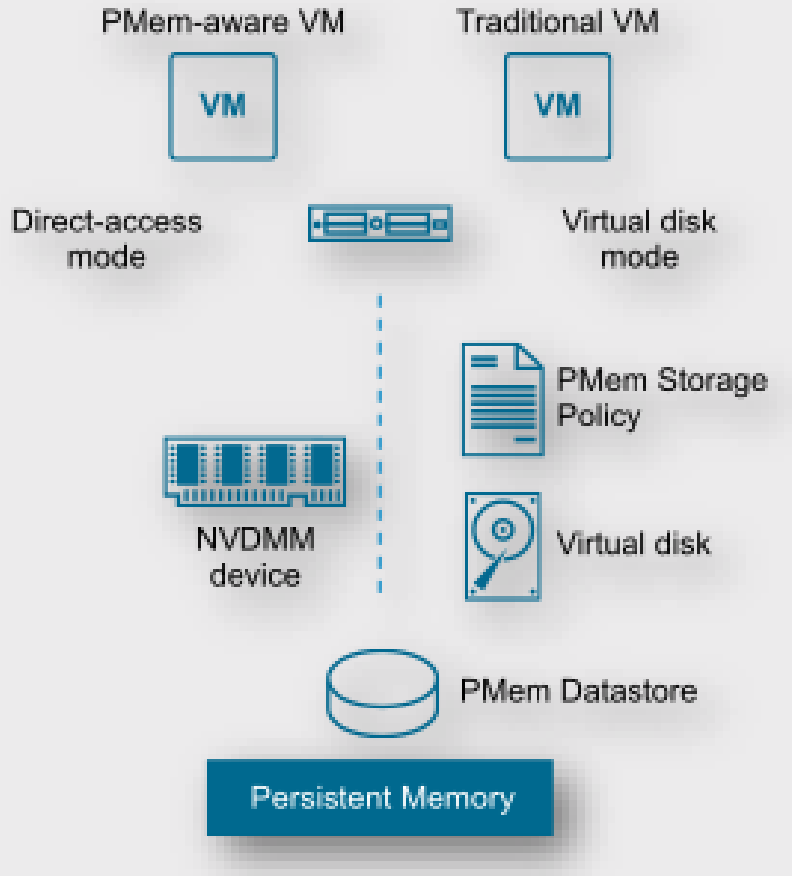
- پشتیبانی در سطح سخت افزار با دریافت تاییدیه از VCG
- نیاز به HW نسخه ۱۴ به بالا
- عدم پشتیبانی از Hot Plug
- نسخه vSphere 6.7 به بالا
- لایسنس Enterprise Plus
- ایجاد اتوماتیک تنها یک دیتااستور بعد از تنظیمات سطح BIOS
- پشتیبانی از هر دو حالت App-Direct Mode و Memory Mode
- اتصال مستقیم به هاست ESXi جهت مشاهده دیتااستور
- عدم امکان مدیریت دیتااستور PMEM
- عدم امکان قرار دادن فایل های مجازی مانند فایل های VMX و Log
- حداکثر فضای ۶ تا ۱۲ ترابایت براساس تعداد CPU مورد استفاده
- عدم امکان استفاده از ویژگی HA و Snapshot
- امکان استفاده از ویژگی DRS و DPM

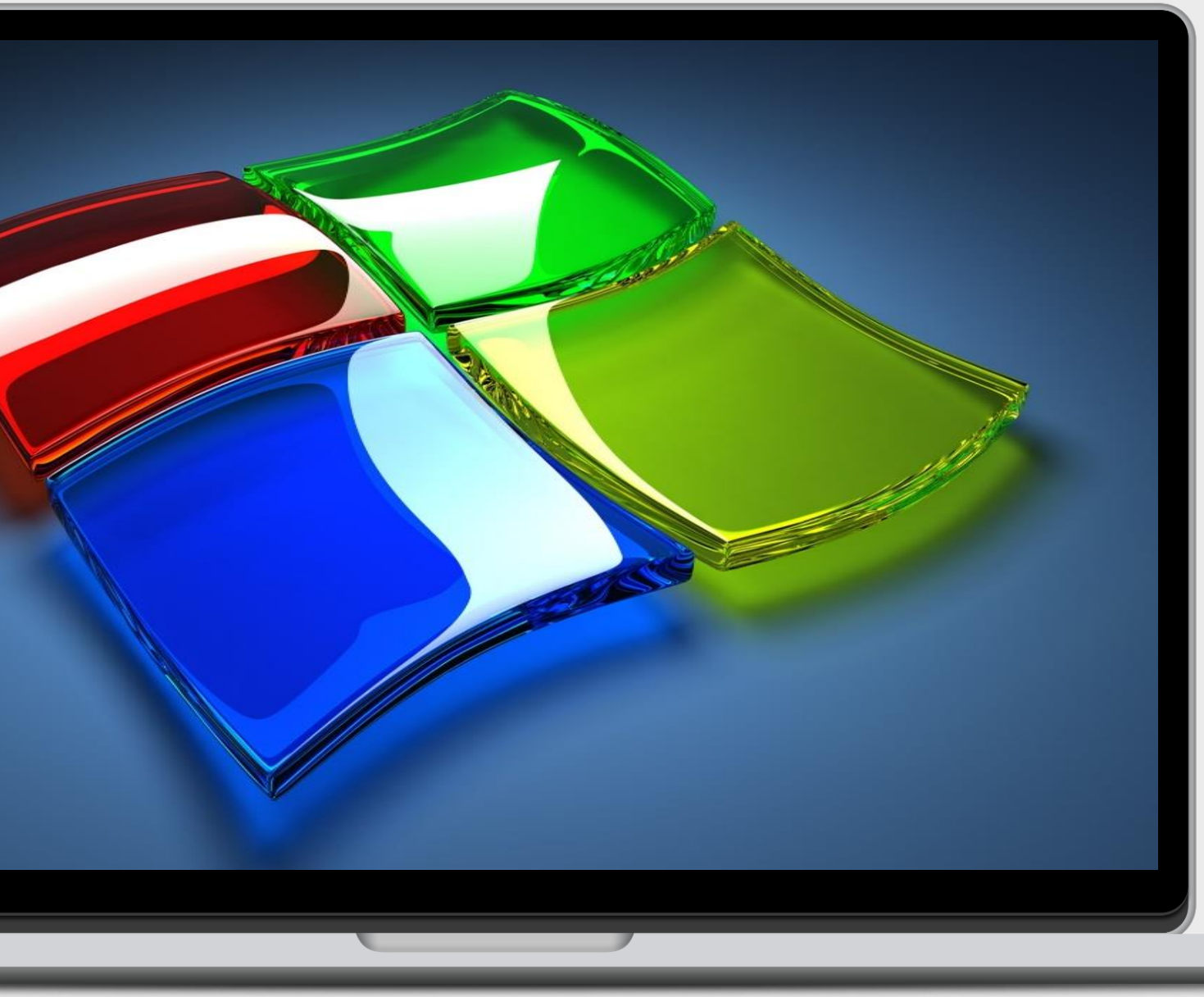


نحوه تخصیص PMEM به ماشین مجازی

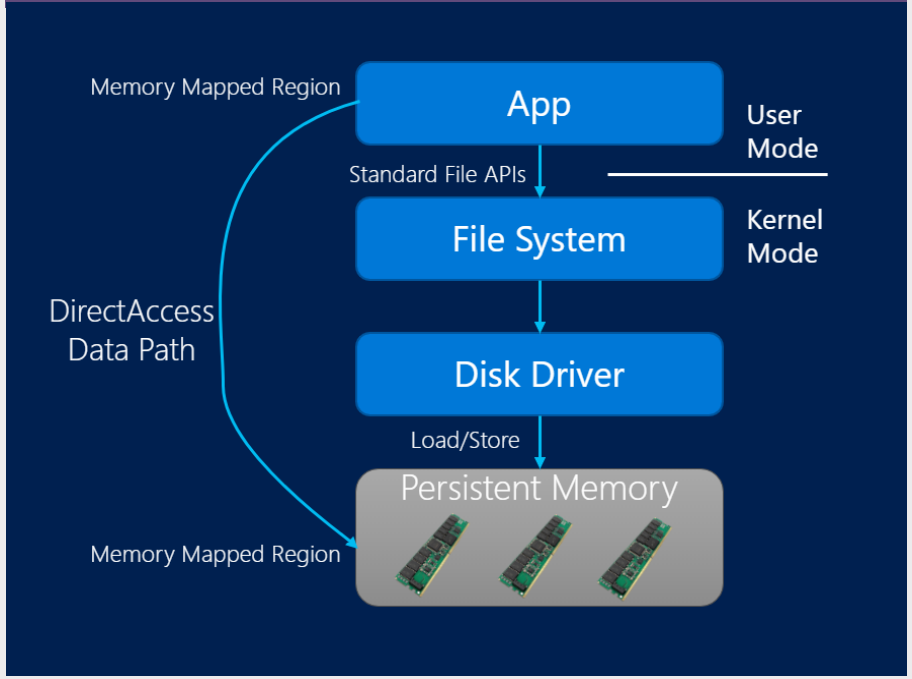
vPMemDisk – Block Mode
For PMEM Non-Aware Technologies

vNVDIMM – Byte Mode
For PMEM Aware Technologies





DAX – Direct Access Volume



- **No NTFS Software Encryption Support (EFS)**
- **No NTFS Software Compression Support**
- **No NTFS TxF Support (Transactional NTFS)**
- **No NTFS USN Range Tracking of Memory Mapped Files**
- **No NTFS Resident File Support**

دوره حرفه‌ای مجازی سازی

SQL Server با استفاده از VMware vSphere

مدرس: رضا اردانه