

vHighPerf

Planung und Optimierung
virtueller Infrastrukturen

Christian Stankowic

www.stankowic-development.net

VMUG West Germany

28.04.2016

Christian Stankowic

Messer Information Services GmbH

Linux-/vSphere-Administrator

Fachbuchautor

AGENDA

Agenda

Hardware-Auswahl

VM-Checkliste

Low Latency-Setups

Performance-Monitoring

HARDWARE



Scale-up?



Scale-out?



Scale-up vs. Scale-out

	Scale-up	Scale-out
	<i>vertikal</i>	<i>horizontal</i>
Hosts	weniger	mehr
Leistung/Host	stärker	schwächer
Lizenzen	weniger	mehr
Verwaltung	weniger	mehr
Use-cases	IMDB, CPU-intensiv	VDI, Server-Virtualisierung
Impact bei Ausfall	größer	kleiner

Blade?



Rack?



HCIA?



Blade-Server

Optimales Leistungs-/Effizienzverhältnis, da viel Leistung auf kleinem Platz

Nur bei min. 35% Chassis-Belegung profitabel!

Beschränkte Erweiterungsmöglichkeiten

Typische Kennzahlen:

- 2-4 CPU-Sockel

- 1-2 TB Arbeitsspeicher

- 8-16 Server pro Chassis

- 2-4 Disks pro Server

Rack-Server

Sehr hohe Leistungsdichte, jedoch geringere Effizienz

Oft dort eingesetzt, wo Blades zu schwach sind oder kein Blade-Konzept gerechtfertigt

Typische Kennzahlen:

- 2-8 CPU-Sockel

- 1.5-12 TB Arbeitsspeicher

- 2-6 HE

- 4-24 Disks pro Server

HCIA - Next Generation-IT?

AIO-Lösung, vereint als Appliance Hardware und essentielle Zusatzprodukte

Lokaler Speicher (*HDD/SSD*)

schnelle Installation und Erweiterung

Beispielprodukte:

EVO:RAIL™ / EVO:RACK™

VCE VxRail™ / VxRack™

Nutanix

VM CHECKLISTE

VM-Checkliste

vCPUs korrekt gesetzt?

NUMA/vNUMA beachtet?

Welcher SCSI-Controller wurde konfiguriert?

Welche NIC wurde ausgewählt?

VMware Tools aktuell?

vCPU-Sizing

Unterteilung analog zu pCPUs in Sockets und Cores

$\text{vCPUs} \leq \text{pCPUs} (+ HT)$

Empfehlung: 1 Core pro virtuellem Socket
(*bessere Skalierung*)

Pro vCPU wird ein VMM-Prozess gestartet

Lizenzregeln der Gäste/Software beachten
(*ggf. Sockets anpassen*)

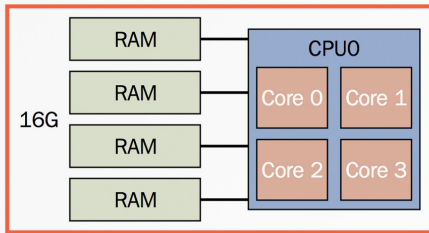
Hakuna-NUMA-ta

Server mit mehreren Sockets werden in **NUMA-Nodes** unterteilt

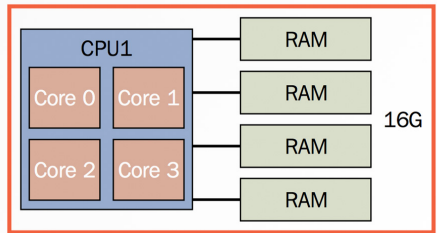
Pro Node erfolgt schnellerer Zugriff auf den eigenen Speicher

Auch möglich, den Speicher anderer Nodes zu verwenden, Zugriff jedoch langsamer

'*NUMA-aware*' Applikationen können von Performance-Boost profitieren



NUMA node 0



NUMA node 1

Pro Node 4 vCPUs, 64 GB RAM

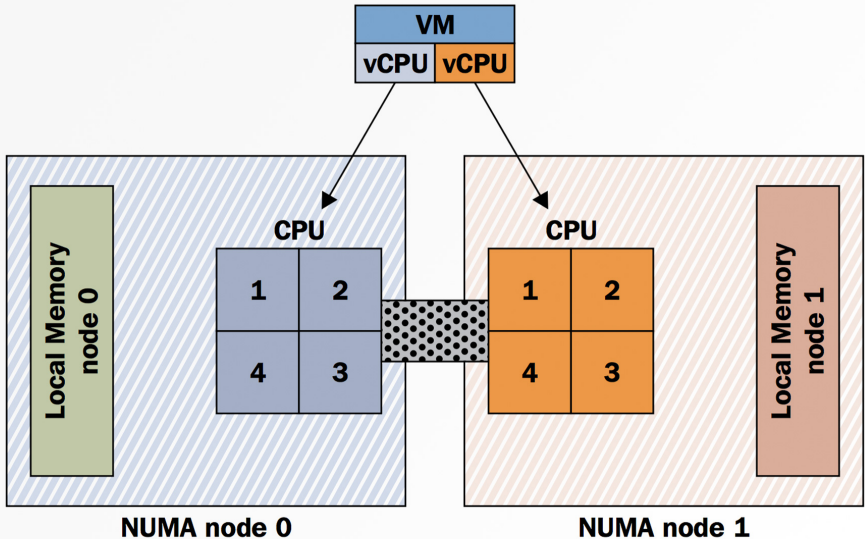
(v)NUMA-Crashkurs

vNUMA informiert VM über
NUMA-Architektur des Hosts (*vHW8+*)

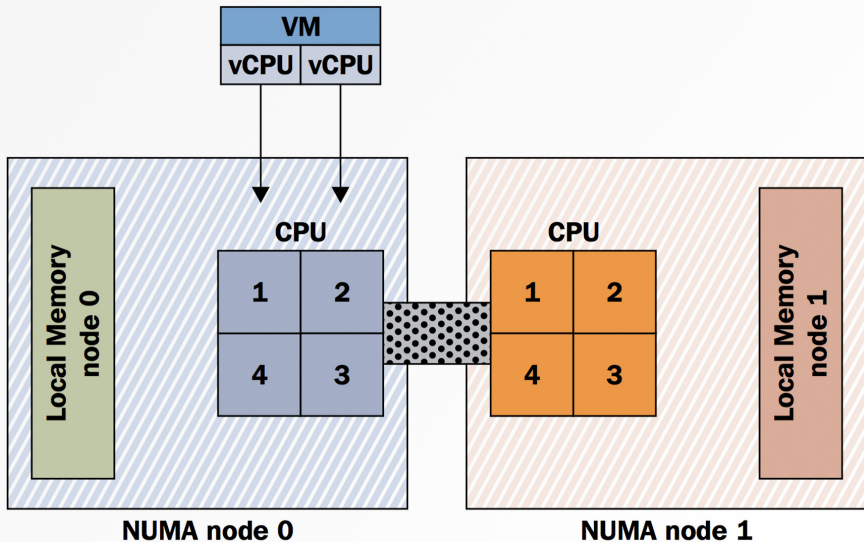
vNUMA wird automatisch bei 9+ vCPUs
aktiviert (`numa.vcpu.min`)

HT-Cores werden standardmäßig nicht
genutzt (`numa.vcpu.preferHT`)

Idealerweise 'passt' eine VM in eine
NUMA-Node



**VM erstreckt sich über zwei
NUMA-Nodes (*wide*)**



**VM befindet sich in NUMA-Node
(flat)**

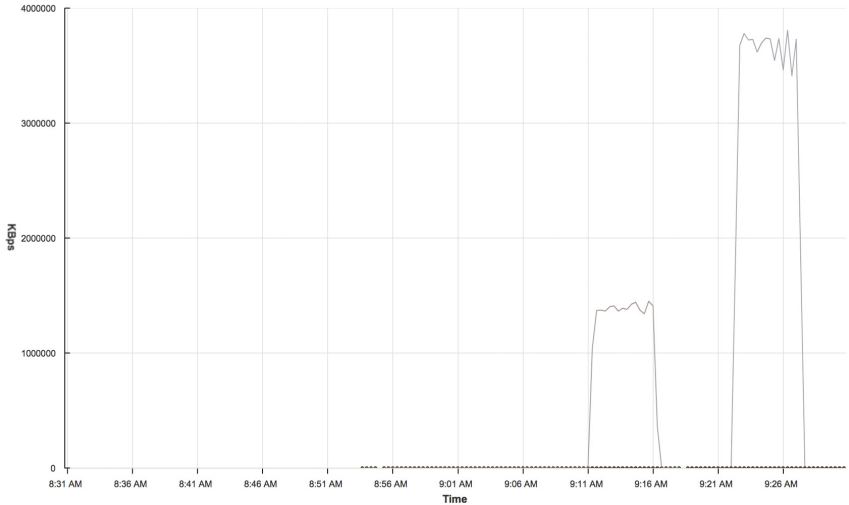
SCSI-Controller

	I/O Qs	OS-Support
PVSCSI	64	W2K3+, Linux 2.6.26+
LSI Logic SAS	32	W2K8+, Linux 2.6+
LSI Logic Parallel	32	-W2K3 R2, Linux 2.6+
Bus Logic	1	-W2K3, -Linux 2.4

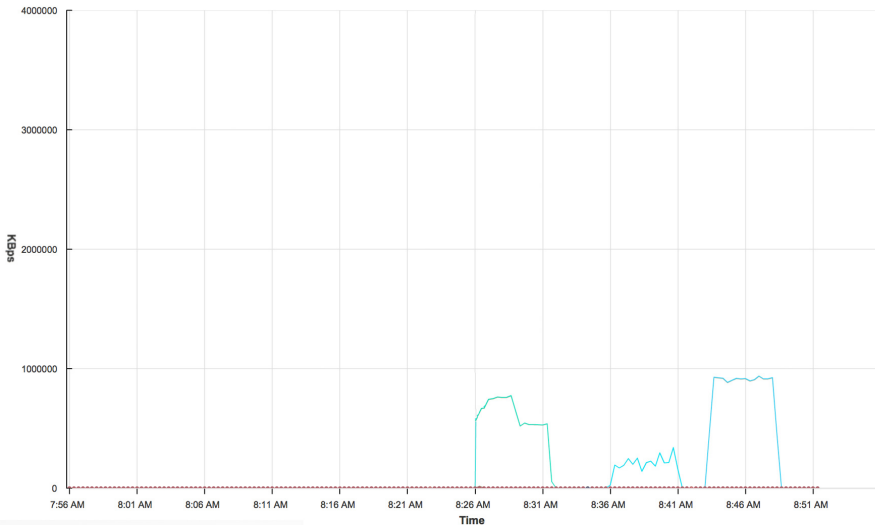
PVSCSI skaliert bei großen IOPS (500+)
deutlich besser
Bei herkömmlichen IOPS-Zahlen sind LSI
ebenbürtig

NIC-Typen

	vHW	Anmerkungen
VMXNET3	7+	Paravirtualisiert, beste Performance
E1000	4+	Standard GbE, für Legacy-OS empfohlen
E1000E	8+	Neuerer GbE, jedoch nicht schneller, primär für Desktop-OS konzipiert



E1000 und VMXNET3 unter EL7
(iperf, TCP)



E1000, E1000E und VMXNET3 unter W2K12 R2 (*iperf*, TCP)

VMware Tools

VMware Tools sind essentiell für virtuelle Workloads!

Enthalten u.a. Treiber (*VMXNET3*, *VMMEMCTL*,...)

Bieten saubere Snapshots

Auch als **open-vm-tools** für unixoide Systeme erhältlich

Immer aktuell halten!

VMware Tools vs. open-vm-tools

	VMT	OVT
Lizenz	closed-source	open-source
OS	Windows, Linux/UNIX,...	Linux/UNIX
Installation	ESXi / vUM	Distributor
Aufwand	mehr	weniger
Support	voll	voll

VMware Support-Matrix bzgl.
Gast-Betriebssysteme beachten!

LOW LATENCY

Low Latency pro VM

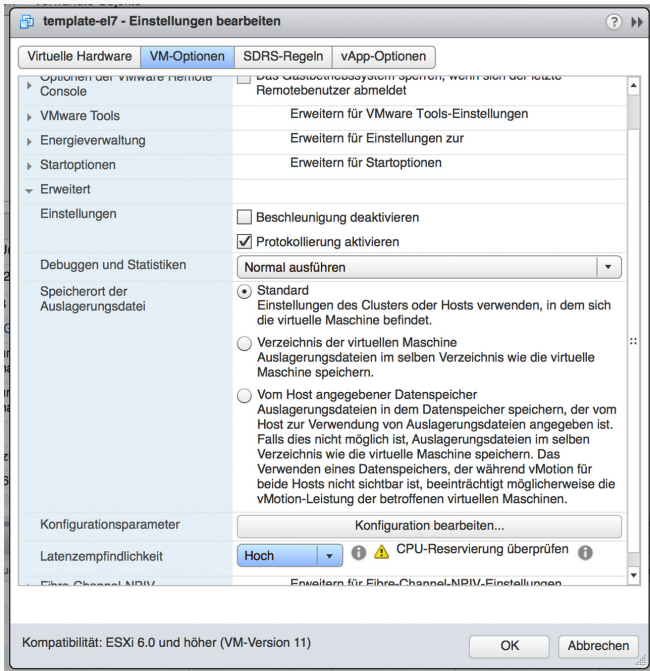
Ab vHW8: Latenz-Einstellung pro VM

Level: Normal, Low/Medium (*undokumentiert*), High

High: pro vCPU-Core wird ein pCPU-Core reserviert

Mit Bedacht für **wenige** VMs setzen, da **exklusiv** gesperrt!

Tipp: Reservierung des kompletten RAMs vornehmen



Low Latency-Netzwerk

Hardware-Passthrough (*VMkernel-Bypass*)
mit **DPIO** (1:1) bzw. **SR-IOV** (n:m)

DPIO unabhängig vom Typ der PCIe-Karte,
SR-IOV fokussiert auf NICs

SR-IOV muss von NIC unterstützt werden,
überschaubare Produktauswahl

SR-IOV NICs verfügen über **VFs**, die an **PFs**
gebunden sind - i.d.R. maximal 64

DPIO / SR-IOV

Einstellung wird pro VM und Karte aktiviert
(*Reboot erforderlich*)

Deaktiviert automatisch einige Features:

- Snapshots

- (*Storage*) vMotion und DRS

- High Availability, Fault Tolerance

I.d.R. geringer Mehrwert gegenüber
normalen Setups (*unter 10% Benefit*)

Network Offloading

Verlagern von Netzwerk I/O in Hardware
(z.B. *iSCSI-HBAs*, *NFS-Offloading*)

Aktivieren von **TCP Segmentation Offload**
und **Large Receive Offload**:

TSO verlagert Paket-Verarbeitung von
Host-CPU in NIC

LRO verwendet im Gast-Betriebssystem
größere Buffer, um weniger Pakete
verarbeiten zu müssen

PERFORMANCE MONITORING

Monitoring-Tools

vCenter Server-Graphen

esxtop bzw. resxtop

VisualEsxtop

vCenter-Graphen

vCenter Server bietet zahlreiche Metriken pro Cluster, Host, VM, etc. - u.a.:

- CPU

- Arbeitsspeicher

- Datenspeicher / Disk / Virtual disk

- Netzwerk

Werte können in Echtzeit oder archiviert ausgelesen werden

Tipp: erweiterte Graphen verwenden!

Bei Performance-Problemen...

CPU: Hohe Latenz bzw. Readyness?

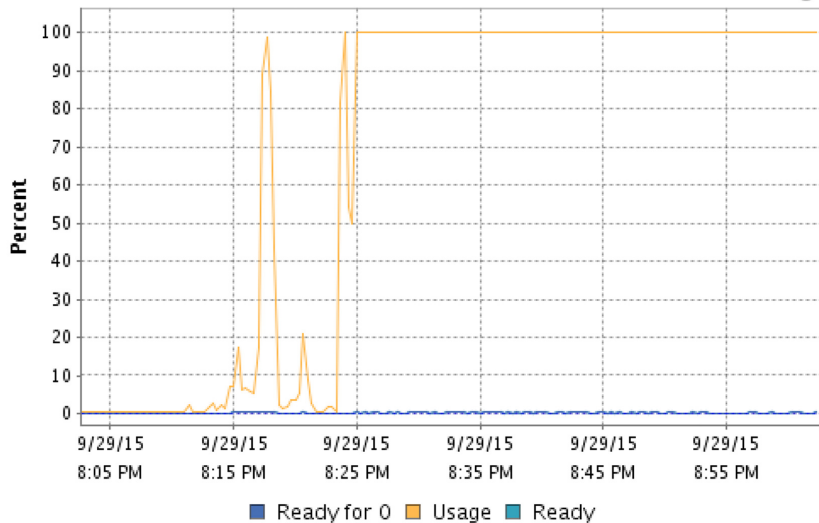
Memory: Swap-/Compression-Raten

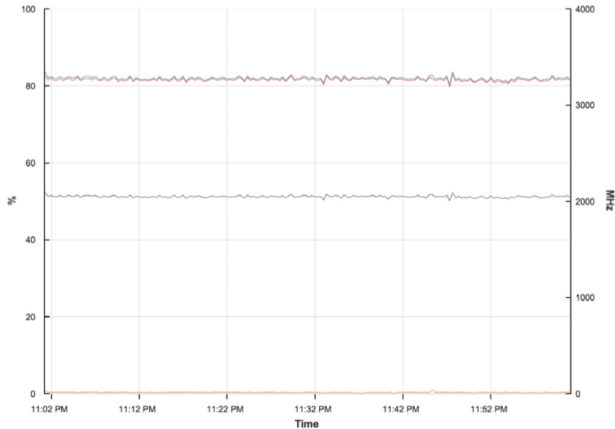
vDisk: Ausstehende Anfragen, ggf.
SCSI-Controller/Dateisystem korrigieren

Datastore: Latenz und Übertragungsraten

Disk: Bus-Resets, abgebrochene
Kommandos - ggf. Storage
ausgelastet/falsch konfiguriert

CPU (Top 10)





Performance Chart Legend

K...	Object	1 ▲ Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
0		Usage in MHz	Average	MHz	3262	3331	3192	3262.944
1		Usage in MHz	Average	MHz	12	39	5	11.678
tvm-centos01		Usage	Average	%	51.16	52.24	50.15	51.219
tvm-centos01		Usage in MHz	Average	MHz	3274	3343	3209	3277.861

```

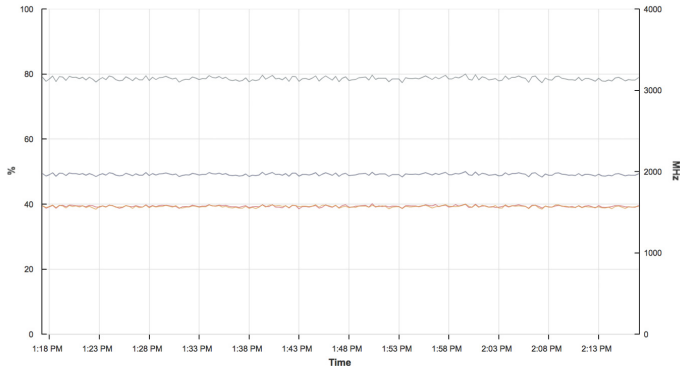
1 [|||||||||||||||||100.0%] Tasks: 48, 17 thr; 3 running
2 [|||||0.0%] Load average: 0.99 0.64 0.28
Mem[|||||211/7824MB] Uptime: 00:04:35
Swp[|||||0/2047MB]

```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
-----	------	-----	----	------	-----	-----	---	------	------	-------	---------

CPU/Real-time, 9/30/15, 1:17:20 PM GMT+2 - 9/30/15, 2:17:00 PM GMT+2 Chart Options

View:  



Performance Chart Legend

Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum
	web002	Usage	Average	%	49.31	50.01	48.31
	1	Usage in MHz	Average	MHz	1583	1604	1542
	0	Usage in MHz	Average	MHz	1571	1600	1536
	web002	Usage in MHz	Average	MHz	3156	3201	3092

leistungsfähigstes Monitoring-Tool

Aufruf über DCUI, SSH oder remote über
resxtop (*vCLI*, *vMA*)

Bietet zahlreiche Metriken und Ansichten,
u.a. für:

- CPU

- Arbeitsspeicher

- Datenspeicher / Disk / Virtual disk

- Netzwerk

```
1. ssh
1:40:52pm up 78 days 18:37, 626 worlds, 10 VMs, 24 vCPUs; CPU load average: 0.0
9, 0.12, 0.12
PCPU USED(%): 4.6 2.7 2.7 1.8 2.7 2.7 1.9 2.9 AVG: 2.7
PCPU UTIL(%): 5.4 4.3 4.5 3.0 5.0 4.3 3.4 4.5 AVG: 4.3
CORE UTIL(%): 9.2      7.1      8.9      7.6      AVG: 8.2
```

ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%VMWAI
35553651	35553651	st-vcsa	10	5.68	8.48	0.05	1000.00	2.0
346247	346247	st-win7	9	3.62	5.85	0.03	900.00	0.0
27740283	27740283	st-backup02	12	3.28	4.50	0.04	1200.00	0.0
38080567	38080567	esxtop.5279209	1	2.98	1.94	0.01	100.00	
19138	19138	st-spacewalk02	8	1.59	2.49	0.02	800.00	0.1
20614	20614	st-web04	7	0.86	1.50	0.03	700.00	0.1
21145934	21145934	tvm-spacewalk02	8	0.76	1.41	0.02	800.00	0.0
18049	18049	st-ipa	7	0.59	1.05	0.01	700.00	0.2
18063	18063	st-devel02	7	0.59	0.99	0.01	700.00	0.9
17774166	17774166	st-mon02	7	0.57	1.03	0.01	700.00	0.4
29117	29117	sfcB-ProviderMa	6	0.51	0.62	0.00	600.00	
16021	16021	st-storage02	7	0.46	0.83	0.01	700.00	1.4
2	2	system	138	0.45	0.79	0.00	13800.00	
29400	29400	sfcB-ProviderMa	9	0.41	0.47	0.00	900.00	
1076163	1076163	vpXa.182680	32	0.30	0.50	0.01	3200.00	
29154	29154	sfcB-ProviderMa	10	0.29	0.48	0.00	1000.00	
7257	7257	hostd.34074	24	0.07	0.10	0.00	2400.00	
15588	15588	dcui.35159	4	0.06	0.11	0.00	400.00	

esxtop-Basics

Taste	Bedeutung
?	Hilfe
V	Nur VM-Informationen, keine System-Prozesse
SPACE	Aktualisieren
s	Aktualisierungsintervall
f	Felder auswählen
o / O	Sortierungsreihenfolge ändern
k	Prozess beenden

esxtop-Metriken

%RDY	Ready time, warten auf Core
%CSTP	Co-Stop, VM für SMP-Scheduling pausiert - zu viele vCPUs?
%VMWAIT	VM Wait Time, warten auf VMkernel - Host ausgelastet?
SWR/s	Swap-Leserate, i.d.R. 0
UNZIP/s	Dekompressionsrate für Memory Pages
N%L	NUMA locality, Ressourcenanteil in Home Node
GAVG/cmd	Antwortzeit Gast-Kernel-Storage
TEAM-PNIC	Aktiv verwendete NIC des Teams

FRAGEN?



Professional Expertise Distilled

vSphere High Performance Essentials

Improve the performance of VMware vSphere using this fast-paced guide

Christian Stankowic

[PACKT] enterprise
PUBLISHING professional expertise distilled

Christian Stankowic

[PACKT] enterprise
PUBLISHING professional expertise distilled



Google play

Vielen Dank für die Aufmerksamkeit!

<http://www.stankowic-development.net>



@stankowic_devel



Christian Stankowic



stankowicdevel