Virtuelle Welten -
Netzwerk Virtualisierung in Solaris 11

Thomas Nau
Universität Ulm – kiz
Ulm

Schlüsselworte:
Netzwerk, Solaris 11, Virtualisierung

Einleitung:
Eine der gravierendsten Änderung bei der Weiterentwicklung von Solaris betrifft den unter dem Projektnamen Crossbow komplett überarbeiteten Netzwerkstack des Kernels. Dieser schafft nicht nur die Voraussetzungen für die effizientere und damit performantere Nutzung moderner Netzwerkkarten, sondern stellt darüber hinaus das Fundament bereit, das für heutige und zukünftige Virtualisierungstechniken unabdingbar ist. Diese Arbeiten umfassen neben der selbstverständlichen Virtualisierung der zu Grunde liegenden Hardware auch die Überarbeitung hinsichtlich der Skalierung auf potentiell hunderte mögliche Gast-Systeme. Der letztgenannte Punkt, häufig auch unter dem Stichwort "multi-tenant" geführt, umfasst zwingend das notwendige Ressourcen Management ohne das die Co-Existenz einer solch großen Anzahl von Gast-Systemen nicht realisierbar ist.
Besondere Bedeutung erlangt der Netzwerkstack auch durch die Konvergenz, die im Bereich der Anbindung von Speichersystemen immer deutlicher wird. Sie nutzen verstärkt im Netzwerkumfeld seit Jahren etablierte Technologien als Transportmechanismus. Beispiele hierfür sind FCoE, iSCSI, jedoch auch iSER und SRP. Dieser Konvergenz trägt auch die IEEE mit den *Data Center Bridging* (DCB) Standards Rechnung. Diese sollen die Zuverlässigkeit von Ethernet-basierten Netzwerken hinsichtlich einer gemeinsamen Nutzung durch Anwendungen mit unterschiedlichsten Anforderungen durch Priorisierung, Bandbreiten-Reservierung, Flusskontrolle usw. verbessern.
Einen wesentlicher Vorteil beim Einsatz von Solaris ergibt sich aus der Tatsache, dass Zonen – die in Solaris integrierte Virtualisierungstechnik – den Kernel des Wirts-System nutzen und daher mit diesem noch enger verflochten sind als para-virtualisierte Treiber dies vermögen.
Im folgenden werden die grundlegenden Netzwerk Virtualisierungstechniken vorgestellt.

Hintergrund des Autors:
Das Kommunikations- und Informationszentrum (kiz) der Universität Ulm trägt unter anderem die Gesamtverantwortung für deren IT-Infrastruktur, inklusive Telefonie, sowie die Versorgung der Wissenschaftler und Studenten sowohl mit elektronischen als auch mit Print-Medien. Die Kernaufgaben der Abteilung Infrastruktur, deren Leiter der Autor ist, umfassen hierbei insbesondere Planung, Weiterentwicklung und den Betrieb der Netzwerke, sowie aller zentralen Server. Zu diesen zählen neben Backup- und HPC-Systemen insbesondere auch die "virtuellen Welten" und die auf HA-Clustern basierenden Mail-, LDAP-, Portal-, Datenbank- und File-Server der Universität Ulm.

Historie:
Einfachste Formen der Netzwerkvirtualisierung, etwa der parallele Einsatz logischer IP-Adressen für individuellen Dienste eines Servers, sind seit vielen Jahren fester Bestandteil in Rechenzentren. Im letzten Jahrzehnt haben sich jedoch die Anforderungen im Betrieb einer zentralen Infrastruktur durch den Einsatz von Virtualisierungstechniken grundlegend geändert. Diese sind, derzeit in aller Regel noch auf den Server Bereich fokussiert, nunmehr fester Bestandteil des Portfolios. Den Ausschlag für den Einsatz geben hierbei oft eine TCO (Total Cost of Ownership) Betrachtung sowie die in

entsprechende Lösungen integrierten Redundanz-Mechanismen. Mit ihnen lassen sich relativ einfach high-availability (HA) Lösungen realisieren, die für eine Vielzahl der angebotenen Services im Allgemeinen ausreichend sind. Allerdings gehen nahezu alle am Markt befindlichen Lösungen davon aus, dass sowohl die Storage-Kapazitäten als auch die Netzwerkanbindung redundant und ausfallsicher bereit gestellt werden.
Durch die Weiterentwicklung der Hardware in den letzten Jahren, hierbei ist in erster Linie der maximale Speicherausbau entscheidend, lassen sich heute bereits mit wenigen mid-range Geräten viele Dutzend Server konsolidieren. Diesen haben jedoch meist eines gemein: sie stellen nur geringe Ansprüche an dauerhafte Netzwerk- und Storage-Performance. Ist dies für die Netzanbindung nicht gegeben werden entsprechende Schutzmechanismen für die anderen Gast-Systeme, etwa Ressourcen Capping bzw. Reservierung, zwingend.

Netzwerktechnik in Solaris 11.1:
Die folgende Tabelle fasst den aktuellen Stand der Netzwerk Virtualisierung und des zugehörigen Ressource Managements in Solaris 11.1 zusammen. Ein data-link bezeichnet hierbei entweder eine physikalische Netzwerkkarte oder einen sogenannten etherstub, ein entsprechendes Pseudo-Device das nicht mit virtuellen NICs (VNICs) verwechselt werden sollte.
Auszug aus dem Handbuch:

> An Ethernet stub can be used instead of a physical NIC to create VNICs. VNICs created on an etherstub will appear to be connected through a virtual switch, allowing complete virtual networks to be built without physical hardware.

Aus Sicht der Anwendung ist eine VNIC nicht von einem physikalischen Interface oder etherstub zu unterscheiden.

Netzwerk Virtualisierung	• virtuelle Interfaces (VNIC) • virtuelle Switches, Router, Load-Balancer und Firewalls • "network in a box" • Schutz vor spoofing
Bandbreiten Verteilung	• Quality of Service (QoS) ist fester Bestandteil und erlaubt die Zuordnung auf Basis von links oder flows
Ressourcen Kontrolle	• die zur Verarbeitung notwendigen CPU Kapazitäten lassen sich auf CPU, CPU-Pool oder Zonen Basis einschränken
erweiterte Funktionalität	• Routing, Firewall, VRRP, Load Balancing • "freie" Wahl von Interface Namen • "sniffing" auch auf lokalen Interfaces
Monitoring	• DTrace provider für IP, TCP, UDP, SRP, iSCSI, NFSv3, NFSv4, … • real-time Daten für VNICs und flows • *kstat* Interface
Skalierung	• automatischer Wechsel zwischen polling und Interrupt • parallele Nutzung von CPU und NIC Ressourcen • Nutzung des Hardware Supports von NICs

Abbildung 1 verdeutlicht den generellen Aufbau der neuen Solaris Netzwerk-Virtualisierungs-Architektur. Diese stellt jeder Zone individuelle Ressourcen bereit und nutzt dabei die Solaris Mechanismen für Hardware- und Software-Skalierung.
Die angesprochenen Änderungen schlagen sich auch in den notwendigen Kommandos zur Administration nieder. *ifconfig(1m)* war in der Vergangenheit zusammen mit Konfigurationsdateien das Mittel der Wahl ist jedoch den heutigen viel komplexeren Anforderungen nicht mehr gewachsen. Seit Solaris 11 übernehmen die nachfolgend genannten Kommandozeilen-Tools nicht nur die aktuelle Konfiguration sondern auch die persistente hinsichtlich Neustarts des Systems. Das Bearbeiten der zugehörigen Dateien mit *vi & friends* sollte tunlichst unterlassen werden.

dladm(1m)	verwaltet data-links also NICs, virtuelle NICs (VNICs), VLANs, link-aggregations, virtuellen bridges und etherstubs
dlstat(1m)	liefert Statistiken über data-links, wie etwa übertragene Bytes, ...
ipadm(1m)	konfiguriert IP Interfaces sowie TCP/IP Parameter einschließlich IP multipathing
flowadm(1m)	Bandbreiten Verwaltung auf Basis von flows – definiert durch Attribute der Layer 3 und 4 sowie von Zonen
flowstat(1m)	liefert Statistiken über flows, wie etwa übertragene Bytes, …

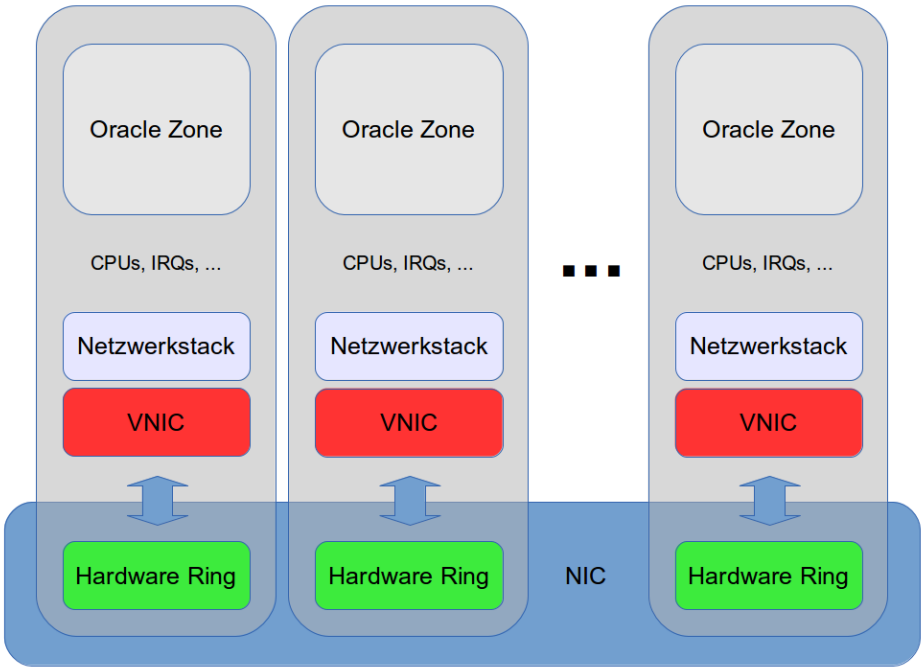

Abbildung 1: Netzwerk-Virtualisierungs-Architektur

Durch die flexible Implementierung lassen sich über einfache building-blocks (Abbildung 2) und durch die individuell zugewiesenen Ressourcen auch komplexe Strukturen abbilden. Diese umfassen, wie in Abbildung 3 teilweise ersichtlich, auch virtuelle Bridges, Router oder Firewalls etwa auf Basis

von *ipfilter(5)*. Die notwendigen administrativen Kommandos zur Begrenzung der maximal verfügbaren Bandbreite gestalten sich einfach. Diese Konfiguration ist persistent bezüglich reboots.

```
obi-wan# dladm create-vnic -l net0 -p maxbw=300M myvnic0
obi-wan# dladm create-vnic -l net0 -p maxbw=100M myvnic1
obi-wan# dladm create-vnic -l net0 -p maxbw=300M myvnic2
```

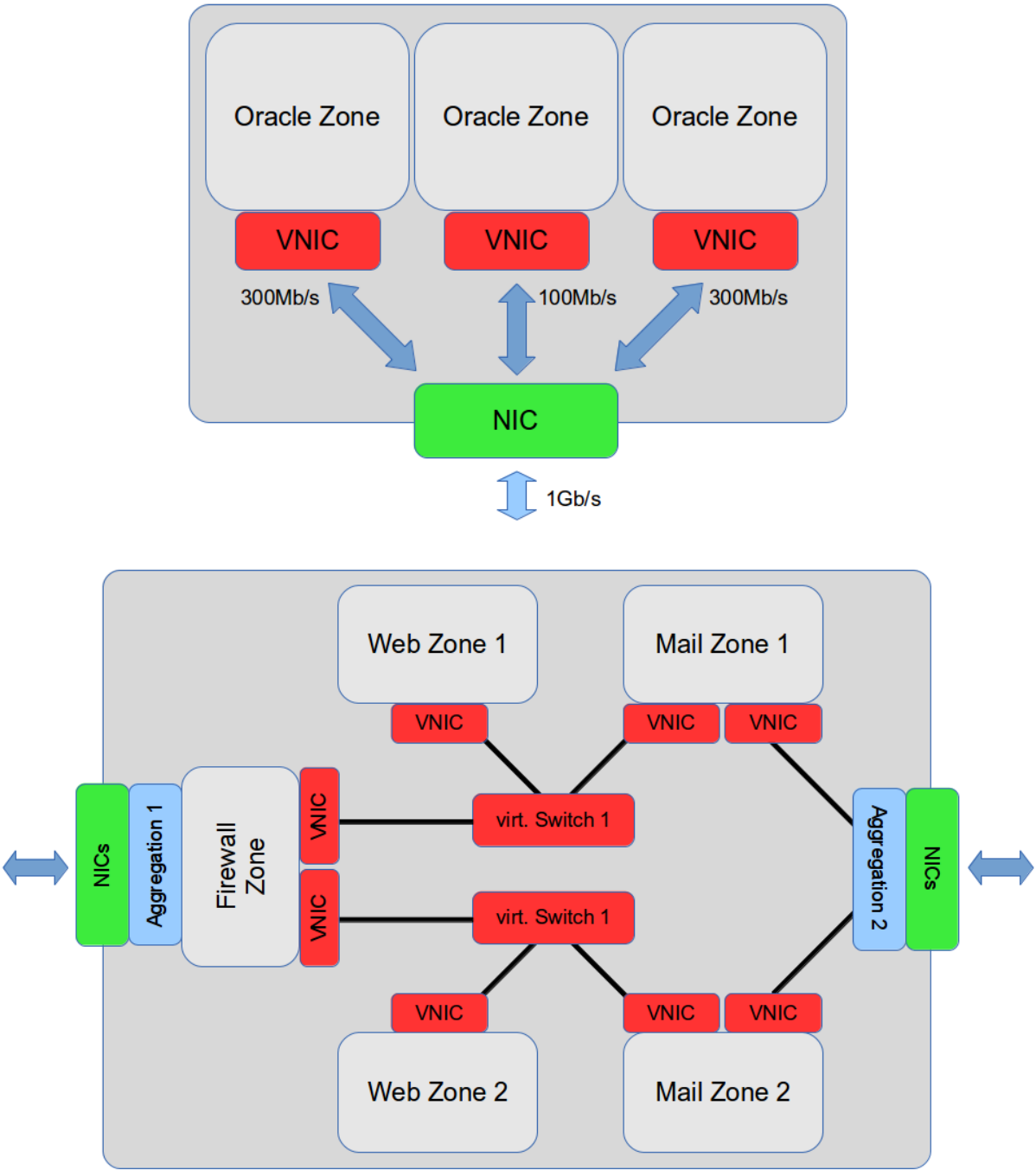

Abbildung 3: Virtuelles Rechenzentrum

Waren einige der bisher beschriebenen Funktionalitäten bereits Bestandteil von Solaris 11 so brachte das erste Update, 11.1, einige wesentliche Neuerungen im Bereich der Ausfallsicherheit. Es ist nun möglich, einzelne oder alle einem data-link zugeordneten VNICs an einen anderen data-link zu binden. Dies ist auf den jeweiligen Server beschränkt und transparent für die Anwendung.

```
obi-wan# dladm modify-vnic -l net1 myvnic0    # myvnic0 → net1
obi-wan# dladm modify-vnic -l net1 -L net0    # alle net0 VNICs → net1
```

Link Aggregation, Trunking und IP-Multipathing (IPMP), wie in Abbildung 4 dargestellt, bilden seit längerem die Basis für Ausfallsicherheit bzw. load-balancing im Solaris Netzwerk Stack. Beide Ansätze unterschieden sich jedoch gravierend in den jeweiligen Vor- und Nachteilen.

Technik	**Pro**	**Contra**
Link Aggregation Trunking	transparent für Zonen und virtuelle Maschinen; einfache Administration	erfordert spezielle Konfiguration der Switches
	erhöht die verfügbare Bandbreite	Beschränkt die Konnektivität auf einen einzelnen Switch oder verlangt proprietäre Protokolle
	automatisches failover/fallback	alle verwendeten Interfaces müssen identische Duplex Modi und Geschwindigkeiten haben
IP Multipathing (IPMP)	failover über mehrere Switches hinweg ohne Nutzung proprietärer Protokolle	erfordert individuelle Konfiguration für jede Zone oder virtuelle Maschine
	Switch Konfiguration nicht notwendig	

Mit data-link multipathing (DLMP, Abbildung 5) ist seit Solaris 11.1 eine Technik verfügbar die für die meisten Anwendungsszenarien das Beste aus beiden Welten vereinigt. Sie liefert Hochverfügbarkeit über mehrere Switches hinweg jedoch ohne auf herstellerspezifische und proprietäre Protokolle zurückgreifen zu müssen. Die Transparenz hinsichtlich der VNICs aus Anwendungssicht bleibt hierbei ebenfalls erhalten. Derzeitiger Nachteil der Lösung: load-balancing wird nur begrenzt unterstützt in dem die zugeordneten VNICs auf die physikalischen Interfaces verteilt werden.

Um failover Entscheidungen treffen zu können ist data-link multipathing derzeit auf die Link Status Informationen der Interface Karten angewiesen d.h. es findet keine aktive Überprüfung wie etwa bei IPMP statt.

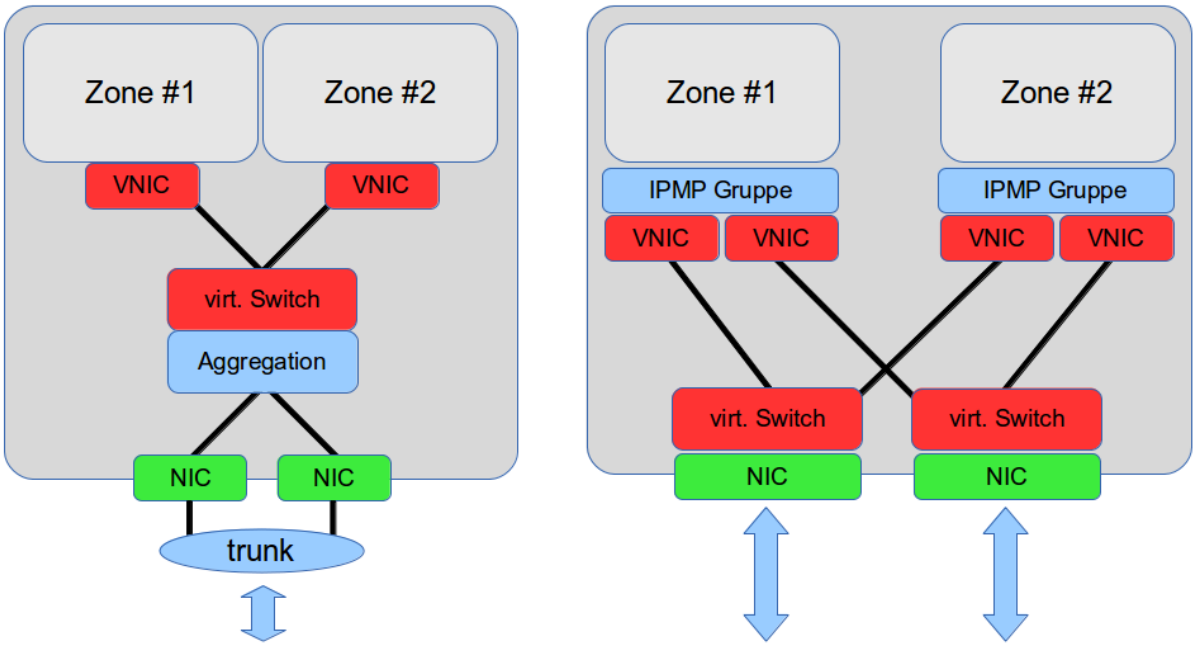

Abbildung 4: Link Aggregation Trunk versus IP-Multipathing (IPMP)

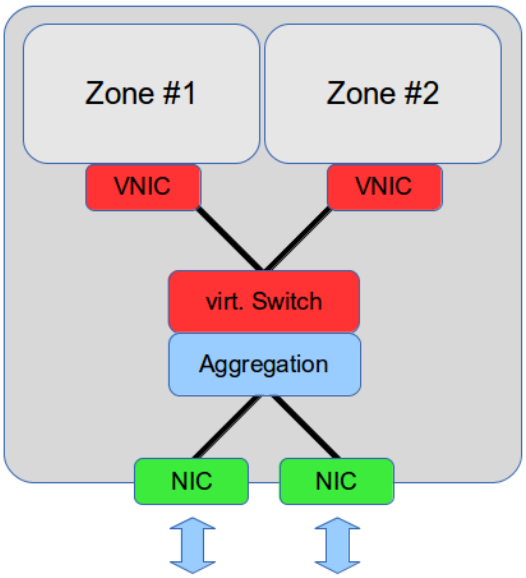

Abbildung 5: Data-link Multipathing

Beispiel:
Zur Konfiguration des in Abbildung 3 gezeigten virtuellen Rechenzentrums sind nur wenige Schritte aus Sicht des virtuellen Netzwerks notwendig.

Anlegen der DLMP aggregations basierend auf den physikalischen Interfaces net2/3 sowie net4/5

```
dladm create-aggr -m dlmp -l net2 -l net3 aggr1
dladm create-aggr -m dlmp -l net4 -l net5 aggr2
```

Erzeugung der virtuellen Switches (etherstubs)

```
dladm create-etherstub vsw1
dladm create-etherstub vsw2
```

VNICs (#0) für die Mail- und Web-Zonen erzeugen und mit dem Switch verbinden

```
dladm create-vnic -l vsw1 vnic_wz1_0
dladm create-vnic -l vsw1 vnic_mz1_0
dladm create-vnic -l vsw2 vnic_wz2_0
dladm create-vnic -l vsw2 vnic_mz2_0
```

Weitere VNICs (#1) der Mail-Zonen werden nun mit der zweiten DLMP aggregation verbunden; diese kann z.B. zur Anbindung von iSCSI Storage dienen

```
dladm create-vnic -l aggr2 vnic_mz1_1
dladm create-vnic -l aggr2 vnic_mz2_1
```

Die Firewall-Zone mit beiden virtuellen Switches verbinden. Hier wird oft die globale Zone verwendet und die erzeugte aggregation direkt mit IP Adressen, … konfiguriert

```
dladm create-vnic -l vsw1 vnic_fw_0
dladm create-vnic -l vsw2 vnic_fw_1
```

Das Ergebnis:

```
obi-wan# dladm
LINK                CLASS     MTU    STATE    OVER
net0                phys      1500   up       --
net1                phys      1500   down     --
net2                phys      1500   down     --
net3                phys      1500   down     --
net4                phys      1500   down     --
net5                phys      1500   down     --
vsw1                etherstub 9000   unknown  --
vsw2                etherstub 9000   unknown  --
aggr1               aggr      1500   unknown  net2 net3
aggr2               aggr      1500   unknown  net5 net4
vnic_wz1_0          vnic      9000   up       vsw1
vnic_mz1_0          vnic      9000   up       vsw1
vnic_wz2_0          vnic      9000   up       vsw2
vnic_mz2_0          vnic      9000   up       vsw2
vnic_fw_0           vnic      9000   up       vsw1
vnic_fw_1           vnic      9000   up       vsw2
```

Bitte beachten: alle etherstubs und zugehörigen VNICs werden als Voreinstellung mit einer MTU (Maximum Transfer Unit) von 9000 angelegt.
Das gezeigte Beispiel lässt sich nun noch einfach um Ressource Management, also etwa CPU Zuweisungen, Bandbreitenbegrenzung oder auch flows) erweitern. Diese Übung sei jedoch dem geneigten Leser überlassen

Fortschreitende Konsolidierung:
Wie bereits einleitend erwähnt gibt es einen klaren Trend, im Datacenter etablierte Netzwerk Infrastrukturen wie Ethernet und Infiniband (IB) auch für die Anbindung von Storage Systemen zu nutzen. Im Vordergrund steht hier neben Kostenersparnis und einheitlichem Management vor allem auch die rasante Entwicklung hinsichtlich verfügbarer Bandbreite. Gilt QDR (netto 32Gb/s) bei Infiniband seit langem als etablierter Standard, hinkt Fiber-Channel (FC) mit aktuell 16Gb/s deutlich hinterher. Auch dessen weiteres Entwicklungspotential wird oft kritisch betrachtet.
Einigkeit herrscht jedoch über die Tatsache, dass nicht alleine technische Parameter entscheidend sind, sondern die transparente Verflechtung von Netzwerk- und Virtualisierung-Technologien mit Management-Tools und Protokollen zu einem *Software Defined Network* (SDN).
Basierend auf standardisierten Protokollen wie *Edge Virtual Bridging (EVB)* werden in Zukunft Hypervisor und auch Anwendungen den jeweiligen Netzwerkkomponenten Informationen hinsichtlich Priorisierung oder gewünschter Bandbreite übermitteln. *Data Center Bridging (DCB)* übernimmt hierbei die Aufgabe, basierend auf Ethernet Technologie isolierte Pfade für sensible Anwendungen bereitzustellen. Diese Pfade sind beispielsweise durch garantierte Bandbreiten und höhere Priorität gekennzeichnet. Gleichzeitig sorgt die individuelle Flusskontrolle pro Pfad dafür, dass keine Pakete auf Grund von Überlast verworfen werden. Mit Hilfe von *EVB* lassen sich dann auch die die Server versorgenden Switches transparent mit einbinden und entsprechend den SLA (service level agreement) Anforderungen der jeweiligen virtuellen Maschinen konfigurieren.

Zusammenfassung:
Solaris 11.1 bietet viele Technologien und aktuelle Protokolle um Netzwerke innerhalb eines Servers zu virtualisieren und dies auch im Sinne eines *Software Defined Network* in das Datacenter zu tragen. Neben dem Produktionsbetrieb, der physikalische Netzwerk Komponenten voraussetzt, unterstützt Solaris mit seiner Netzwerk Implementierung auch große Anwendungsumgebungen, die vollständig virtualisiert sind. Diese kommen zu Tests oder im Rahmen der Qualitätssicherung zum Einsatz, entsprechend leistungsfähige Server vorausgesetzt.

Danksagung:
Mein Dank gilt insbesondere meinem Kollegen Dr. Harald Däubler für all seine Anregungen, Tipps und Korrekturen. Darüber hinaus danke ich Nicolas Droux – Senior Principal Software Engineer bei Oracle – und seinen Kollegen für die andauernde Unterstützung und Bereitstellung von Informationen.

Kontaktadresse:

Thomas Nau
Universität Ulm – kiz
Albert Einstein Allee 11
D-89081 Ulm

Telefon: +49 (0) 731 50-22464
Fax: +49 (0) 731 50-12-22464
E-Mail Thomas.Nau@uni-ulm.de
Internet: http://www.uni-ulm.de/einrichtungen/kiz