ulm university universität ulm

Solaris 11.2
Erfahrungen aus der Praxis

Thomas Nau, kiz (Thomas.Nau@uni-ulm.de)

Kommunikations- und Informationszentrum (kiz)

- Die Aufgaben der „Abteilung Infrastruktur“ umfassen u.a.
  - Azubi Ausbildung
  - Cluster basierende Universitäts weite Mail-, LDAP-, Portal-, Datenbank- und File-Services, ...
  - Betreuung von ca. 600 Desktop und Laptop Arbeitsplätzen
    - 25% Linux, 75% Windows
  - Backup Service (Bacula Enterprise) für mehrere Universitäten in Baden-Württemberg
  - Landes HPC Cluster mit Schwerpunkt „Theoretische Chemie“
  - 4 lokale Netzwerke plus flächendeckendes Campus WLAN und MAN im Ulmer Stadtbereich
  - Telefonanlage mit ca. 14.000 Anschlüssen unter Einsatz von VoIP und 2-Draht Technik

„BW“ Projekte

- bwIDM
- Belwue
- bwHPC
- bwBackup
- bw100G
- bwCloud

Unsere Mission

- ein Team, ein Ziel
  - IT-Security, Netz- und TK-Gruppen sind Bestandteil des Teams
    - kein Elfenbeinturm
    - hilft realistische und umsetzbare Ziele zu definieren
- tolles Arbeitsumfeld mit vielen Freiheiten für Leute die bereit sind Verantwortung zu übernehmen
  - „erlaubt ist was funktioniert und wartbar ist“
  - natürlich gibt es auch Ausnahmen wie Datenschutz, ...
- Einsatz technischer Lösungen die einen besseren Service erlauben
  - 11.2 early adopters

Solaris als bewusste strategische Entscheidung

- Stabilität und Datensicherheit sind Schlüsselfaktoren; nahezu alle zentralen Server basieren auf Solaris
  - SAP/Oracle, MySQL, PostgreSQL, Apache/PHP, Typo3, Cyrus IMAP, NFS, CIFS, iSCSI, Bacula Enterprise Backup, LDAP, …
  - Java, Tomcat und Solaris Zonen sowie auch Apache mit PHP in Zonen sind Standardszenarien
  - Ausnahmen: Windows Active Directory und Citrix XenServer
- 90% aller Server verwenden Solaris 11.2
  - meist x86 aber auch SPARC basierte T4 Systeme
  - sehr wenige (3 SAP Server, …) physikalisch unter Solaris 10
  - S10 Zonen für TSM Server da DB2 unter S11 nicht funktionsfähig
    - Ablösung durch Bacula Enterprise in Kombination mit PostgreSQL

Apropos Stabilität...

```
# ssh very-dedicated-isolated-v240-s10
…

# uptime
10:35am  up 2716 days  0:14,  0 users,  load ...
```

Apropos Stabilität...

7 Jahre 5 Monate

Solaris 11.2 Erfahrungen

- ca. 20 Monate erstklassige Erfahrungen insbesondere auch im Rahmen des 11.2 Platin Beta Programms
  - Einsatz im Produktionsumfeld von Anfang an; spricht für sich
  - viele Verbesserungen und neue Features die ein Upgrade empfehlenswert machen
- Beta Programm hat auch etliche Fehler aufgedeckt
  - Umgang damit zeigt daß Oracle Solaris ernst nimmt

Bekannte und unbekannte Highlights

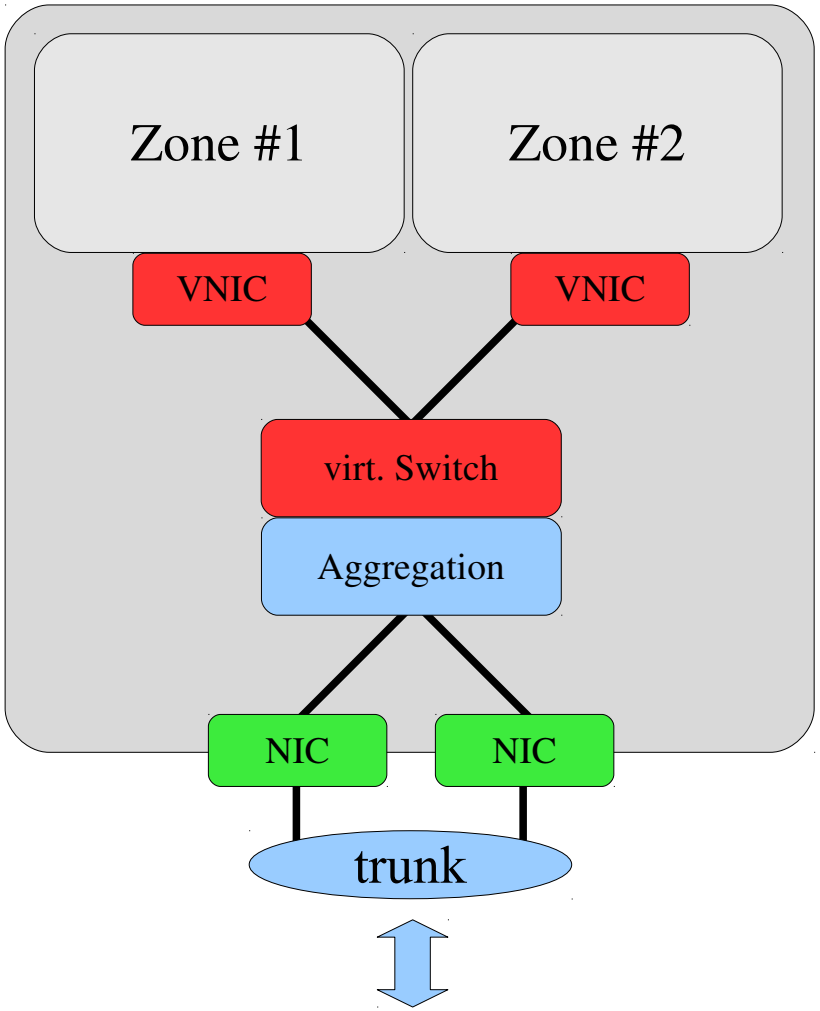
- Puppet
- OpenStack
- Kernel Zones und Virtualisierung
- Netzwerk Stack, Software Defined Networking (SDN)
- IPS, Unified Archives
- Compliance Checking und Reporting
- SMF (System Management Facility) Verbesserungen
- ZFS Verbesserungen
- Admin-Tools

Veränderungen in Ulm: Netzwerk (DLMP)

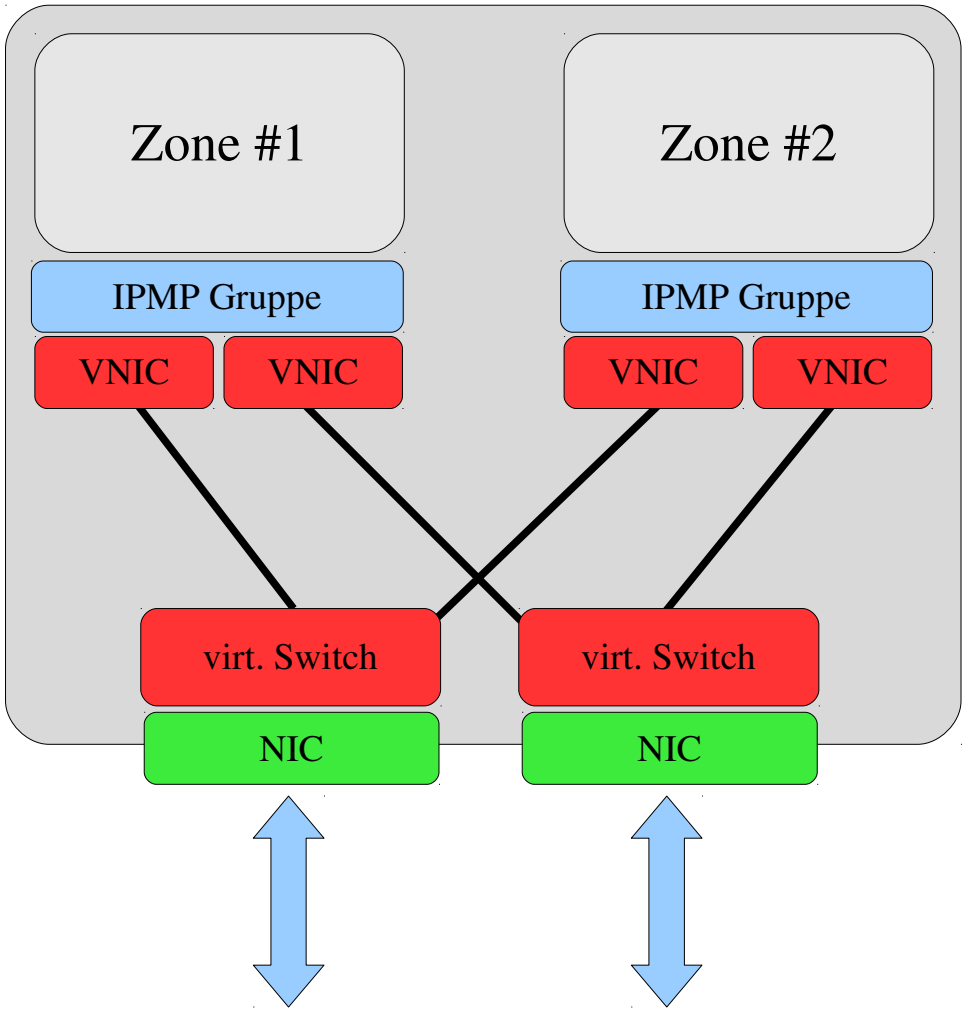
Technik	Pro	Contra
Link Aggregation Trunking	transparent für Zonen und virtuelle Maschinen; einfache Administration	erfordert spezielle Konfiguration der Switches
	erhöht die verfügbare Bandbreite	Beschränkt die Konnektivität auf einen einzelnen Switch oder verlangt proprietäre Protokolle
	automatisches 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	

Veränderungen in Ulm: Netzwerk (DLMP)

Link-Aggregation / Trunking



IP-Multipathing (IPMP)

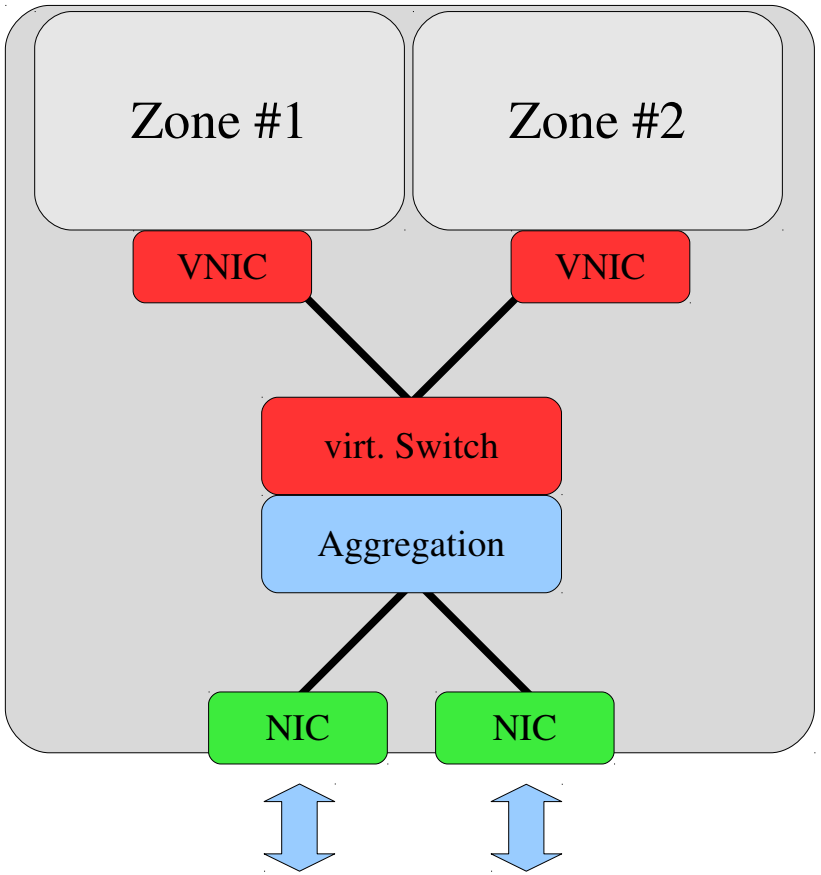


Veränderungen in Ulm: Netzwerk (DLMP)

- Data-link Multipathing (DLMP)
  - verbindet das beste aus den Welten IP-Multipathing (IPMP) und Link-Aggregation / Trunking (LACP)
  - funktioniert Switch übergreifend ohne spezielle Software auf den Netzwerk Devices
  - nur „grobes“ load-balancing über VNIC Zuweisung
    - VNICs werden dennoch bei Ausfall eines physikalischen Interfaces migriert
  - seit 11.2 „probe-based“ Fehler Erkennung durch *in.dlmpd(1m)*
    - ICMP basiert im lokalen Subnetz
- alle wichtigen NFS, CIFS, iSCSI und Zonen Server verwenden den neuen Mechanismus
  - 10GE Infrastruktur benötigt kein Trunking mehr

Veränderungen in Ulm: Netzwerk (DLMP)

Data-Link Multipathing




```
dladm create-aggr -m dlmp -l net0 -l net1 aggr0
```

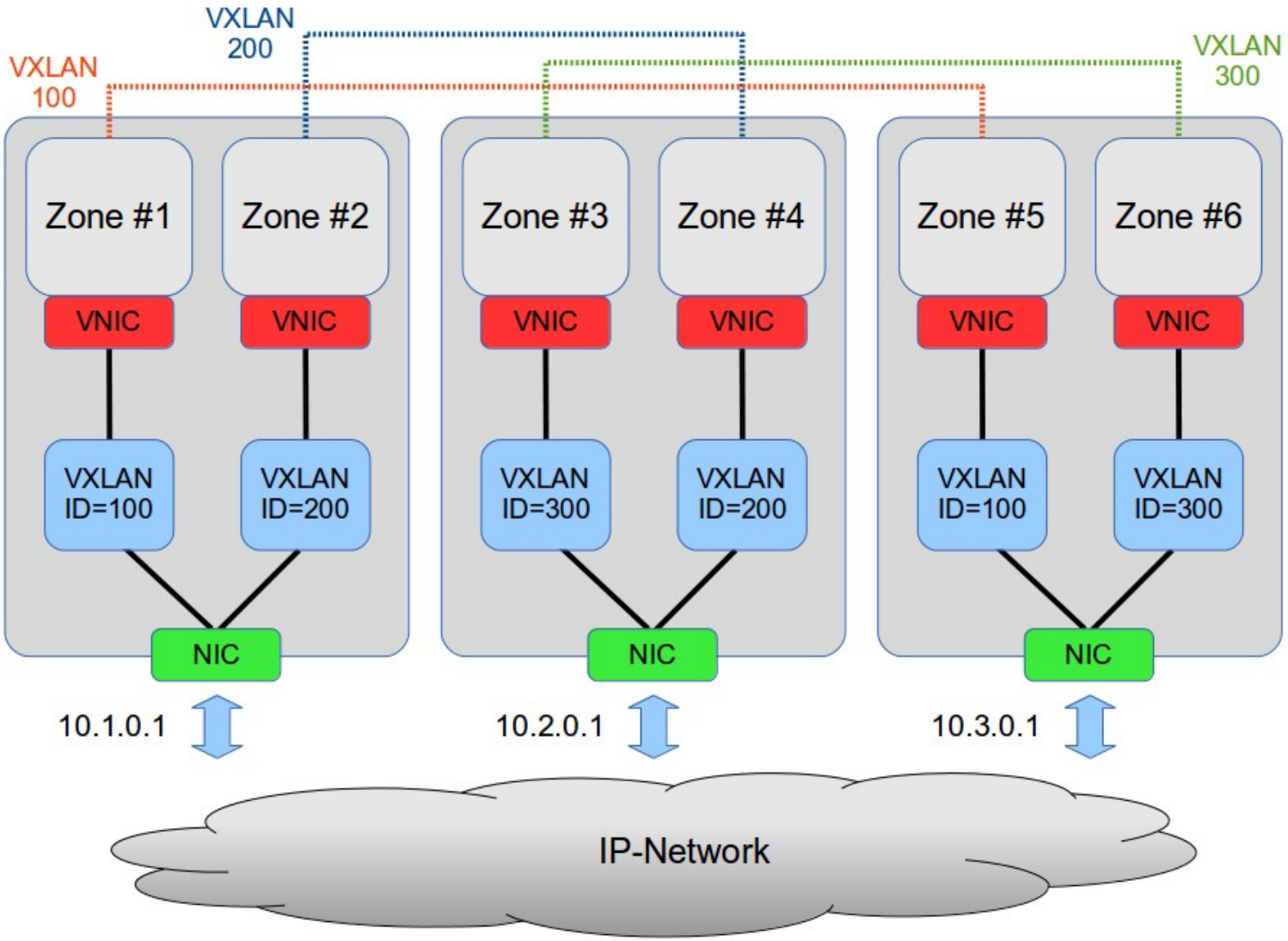
Veränderungen in Ulm: Netzwerk (DLMP)

```
obi-wan# dladm set-linkprop -p probe-ip=+192.168.2.1 aggr0

obi-wan# dladm show-aggr -S
LINK        PORT            FLAGS STATE   TARGETS          XTARGETS
aggr0       net2            u-2-  active  --               net3
--          net3            u--3  active  hydra-storage    net2

obi-wan#  dlstat show-aggr -n -P all
 TIME   AGGR     PORT           LOCAL          TARGET   PROBE  NETRTT     RTT
0.06s  aggr0     net2             net2           net3    t24      --      --
0.06s  aggr0     net2             net2           net3    t24  1.96ms  2.08ms
0.32s  aggr0     net3    192.168.2.21    192.168.2.1    i24      --      --
0.32s  aggr0     net3    192.168.2.21    192.168.2.1    i24  0.12ms  0.86ms
0.84s  aggr0     net3             net3           net2    t25      --      --
0.84s  aggr0     net3             net3           net2    t25  1.86ms  1.95ms
1.20s  aggr0     net3    192.168.2.21    192.168.2.1    i25      --      --
1.20s  aggr0     net3    192.168.2.21    192.168.2.1    i25  0.13ms  1.10ms
...
```

Weitere Netzwerk Verbesserungen

- „reflective relays“ in EVB (Edge Virtual Bridging)
  - erlaubt es Pakete die im Normalfall intern über einen virtuellen Switch geführt werden und daher das physikalische System nicht verlassen über einen externen Switch „umzuleiten“
    - externer Switch wertet z.B. ACLs aus
- Virtual Extensible Local Area Networks (VXLANs)
  - „Cloud-Technologie“
  - Transport von Layer-2 Paketen über Layer-3 Infrastruktur und kann mit IPsec kombiniert werden
  - tools wie *wireshark(1)*, *tshark(1*) und *snoop(1m)* sind „VXLAN aware“
  - wird Bestandteil unseres bwCloud Testbeds

Weitere Netzwerk Verbesserungen: VXLAN



Verbesserungen in der virtuellen Welt

- „*immutable zones*“ (IMZ)
  - haben den Nachteil daß Nutzer mit root Rechten in der globalen Zone, unabhängig von den Einstellungen der NGZ, dort Daten verändern können

file-mac-profile	**Beschränkungen**
none	keine; normale Zone
strict	read-only Zone (immutable)
fixed-configuration	Verzeichnisse unterhalb von */var* sind schreibbar sofern diese keine Konfigurationsdateien enthalten
flexible-configuration	wie *fixed-configuration* aber auf */etc* ausgedehnt

Verbesserungen in der virtuellen Welt

- in 11.2: „trusted-path“
  - definiert Zugangsweg der als sicher eingestuft ist, etwa Konsole oder für NGZ mit „`zlogin -T`“
  - login über diesen Mechanismus erlaubt Änderungen einer IMZ ohne in den update-mode booten zu müssen
- Erweiterungen auf globale Zone
  - „*immutable global zone*“, trusted login nur via Konsole
  - kann mit „*verified boot*“ kombiniert werden
  - Beispiele im Blog von Casper Dik unter

    https://blogs.oracle.com/casper/entry/solaris_11_2_immutable_global
  - gute Ergänzung für key-server, management hosts oder sensible gateways (DMZ)

Verbesserungen in der virtuellen Welt

- „*kernel zones*“
  - verwenden einen eigenen Solaris Kernel und sind damit aus administrativer Sicht vollkommen von der globalen Zone unabhängig
    - Uni Ulm Vorteil: mehrere CIFS Server mit unterschiedlicher AD Anbindung auf einer physikalischen Hardware
  - es existieren Versionsabhängigkeiten
  - Installation wie üblich aber es existieren zusätzliche properties für Speicher, CPUs, Host-ID, ...

```
zonecfg:NerdHerd> select capped-memory
zonecfg:NerdHerd:capped-memory> set physical=8G
zonecfg:NerdHerd:capped-memory> end
zonecfg:NerdHerd> add virtual-cpu
zonecfg:NerdHerd:virtual-cpu> set ncpus=2
zonecfg:NerdHerd:virtual-cpu> end
```

Verbesserungen in der virtuellen Welt

- „*kernel zones*“
  - setzen Hardware (CPU) Support voraus
    http://docs.oracle.com/cd/E36784_01/html/E37629/gnwoi.html

Sun Fire x4150	Sun Fire x4170-M3

`# virtinfo`
`NAME CLASS`
`non-global-zone supported`
	`# virtinfo`
	`NAME CLASS`
	`non-global-zone supported`
	`kernel-zone supported`

Verbesserungen in der virtuellen Welt

- native Solaris 11 Systeme oder Zonen können in Kernel Zonen migriert/konsolidiert werden
- Migration von Zonen mit gemeinsamen Storage Bereichen zwischen kompatiblen physikalischen Systemen ist möglich
  - auch suspend/resume bei Kernel Zonen
- mit `zonecfg -r` können Parameter von laufenden Zonen on-the-fly bis zum nächsten reboot der Zone geändert werden

Verbesserungen in der virtuellen Welt

- zusätzliches Interface in Zone bereitstellen

```
jedi# zlogin yoda
[Connected to zone 'yoda' pts/6]

yoda:.../~# dladm
LINK                CLASS     MTU    STATE    OVER
net0                vnic      1500   up       ?

jedi# zonecfg -z yoda -r \
        "add anet;set linkname=net1;set lower-link=net3;end;commit"
zone 'yoda': Checking: Adding anet linkname=net1
zone 'yoda': Applying the changes

...
yoda:.../~# dladm
LINK                CLASS     MTU    STATE    OVER
net0                vnic      1500   up       ?
net1                vnic      1500   down     ?
```

Unified Archives (UA)

- schließen die durch den Wegfall der *flash-archives* entstandene Lücke aber bieten deutlich mehr
- basieren auf ZFS streams
- können mehrere Instanzen enthalten
- werden vom Automated Installer unterstützt (AI)
- erlauben das Erzeugen boot-fähiger Disaster Recovery Medien (USB, DVD)

```
jedi# archiveadm create --recovery no_more_disasters.uar
```

Unified Archives (UA)

```
morpheus# zoneadm list -cv
  ID NAME        STATUS      PATH                    BRAND      IP
   0 global      running     /                       solaris    shared
   1 wiki        running     /zoss/wiki              solaris    excl
   2 flare       running     /zoss/flare             solaris    excl
   6 ftp         running     /zoss/ftp               solaris    excl
  15 proxy       running     /zoss/proxy             solaris    excl
   - idp-test    configured  /zoss/idp-test          solaris    excl
  ...

morpheus# archiveadm create \
              --exclude-zone=ftp \
              --exclude-zone=proxy \
              --exclude-dataset=flare_rpool/rpool/data/ai \
              --exclude-dataset=rpool/uar \
              --skip-capacity-check \
              /rpool/uar/testing.uar
Initializing Unified Archive creation resources...
Unified Archive initialized: /rpool/uar/testing.uar
```

Unified Archives (UA)

```
morpheus:# archiveadm info -v /rpool/uar/testing.uar
Archive Information
          Creation Time:  2014-09-28T11:31:55Z
            Source Host:  morpheus
           Architecture:  i386
       Operating System:  Oracle Solaris 11.2 X86
       Recovery Archive:  No
              Unique ID:  e0dbf57f-b579-c6c6-c64e-da8850304310
         Archive Version:  1.0

Deployable Systems
           'global'
              OS Version:  0.5.11
               OS Branch:  0.175.2.1.0.5.2
               Active BE:  s11u2-1_5_0
                   Brand:  solaris
            Size Needed:  6.5GB
              Unique ID:  f33ece58-2562-e9b4-ecd2-fafa328adb25
               AI Media:  0.175.2_ai_i386.iso
              Root-only:  No
```

Unified Archives (UA)

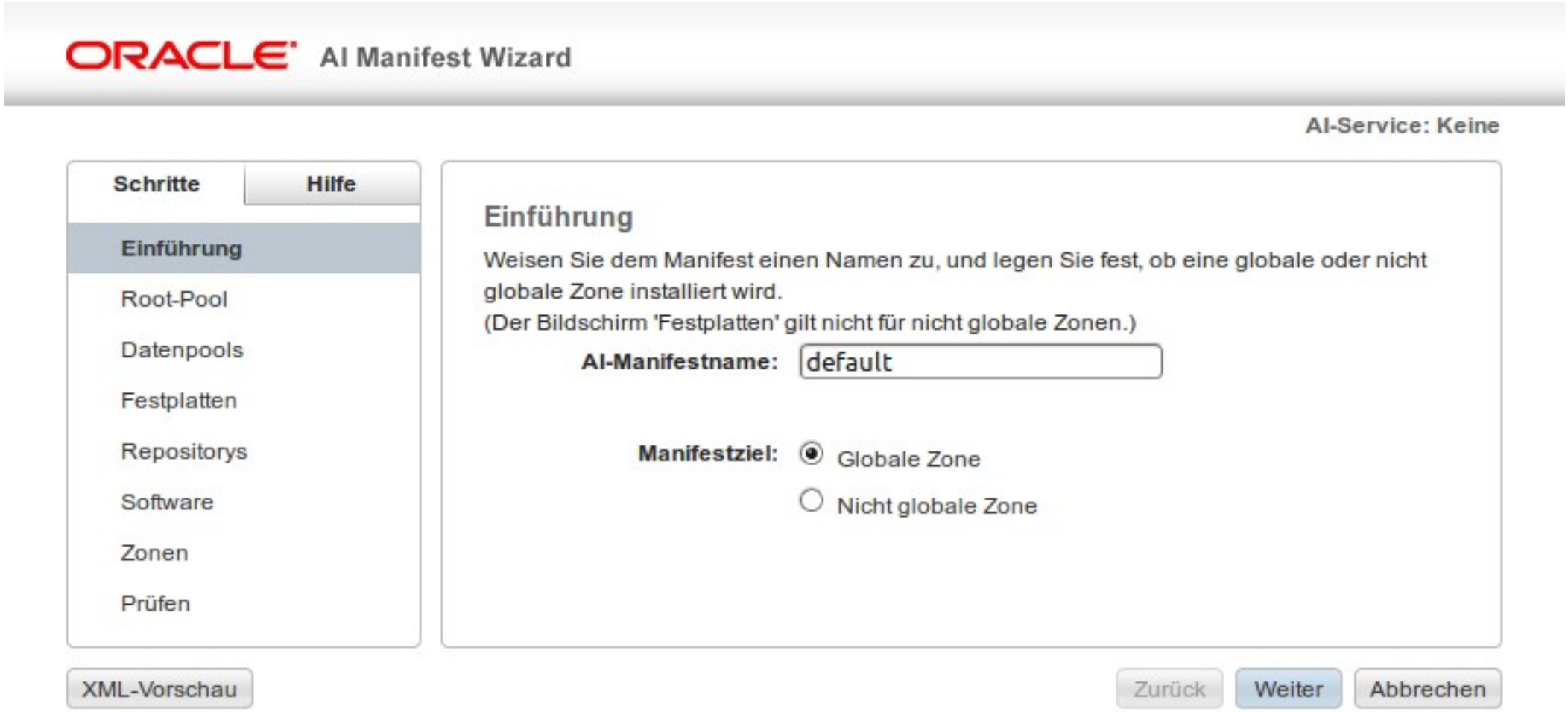
```
'flare'
  OS Version:  0.5.11
   OS Branch:  0.175.2.1.0.5.2
   Active BE:  solaris-8
       Brand:  solaris
 Size Needed:  31.1GB
   Unique ID:  a96d3fd6-9578-605f-99ab-b3adb7b497ae
    AI Media:  0.175.2_ai_i386.iso
   Root-only:  Yes

 'wiki'
  OS Version:  0.5.11
   OS Branch:  0.175.2.1.0.5.2
   Active BE:  zone-4
       Brand:  solaris
 Size Needed:  4.7GB
   Unique ID:  e94bb6c6-c0c4-4692-8b4e-dfd030c4a6ee
    AI Media:  0.175.2_ai_i386.iso
   Root-only:  Yes
```

Unified Archives (UA)

```
jedi# archiveadm info testing.uar
Archive Information
          Creation Time:  2014-09-28T11:31:55Z
            Source Host:  morpheus
           Architecture:  i386
       Operating System:  Oracle Solaris 11.2 X86
      Deployable Systems: global,flare,wiki

jedi# zonecfg -z wisdom create  -a /root/testing.uar -z wiki

jedi# zoneadm -z wisdom install -a /root/testing.uar -z wiki
The following ZFS file system(s) have been created:
    rpool/zoss
    rpool/zoss/wiki
Progress being logged to
/var/log/zones/zoneadm.20140928T131636Z.wisdom.install
    Installing: This may take several minutes...
...
       Done: Installation completed in 282.104 seconds.
 Next Steps: Boot the zone, then log into the zone console
             (zlogin -C) to complete the configuration process.
```

Automated Install Manifest Wizard

- Aktivierung mit *installadm(1m)*



Die kleinen Freuden der Administratoren

- SMF (Service Management Facility)
  - das in 11.1 bereit gestellte Tool *svcbundle (1m)* erlaubt es einfach und schnell Dienste „SMF ready“ zu erstellen
  - viele unserer Services wurden umgestellt
  - Problem der Konfigurationsdateien wird in 11.2 durch „stencils“ adressiert
  - Apache Beispiel findet sich im Blog von Jörg Möllenkamp unter

    http://www.c0t0d0s0.org/archives/7715-New-Solaris-11.2-features-SMF-stencils.html
- SMF Instanzen sind wenig beachtet aber sehr hilfreich
  - Bsp.: mehrere DB-Instanzen, …
  - siehe Tagungsunterlagen bzgl. „HowTo“

Die kleinen Freuden der Administratoren

```
alderaan# svcbundle -o mydaemon.xml \
            -s service-name=mysite/mydaemon \
            -s model=daemon \
            -s start-method="/usr/local/bin/mydaemon"

#  Nach ggf. notwendigen Anpassungen etwa hinsichtlich der
#  Service Abhängigkeiten genügen zwei weitere Kommandos um
#  den Dienst zu aktivieren.

alderaan# cp mydaemon.xml /lib/svc/manifest/site
alderaan# svcadm restart manifest-import
```

```
jedi# svcs */bacula*
STATE          STIME    FMRI
disabled       Sep_01   svc:/application/backup/bacula-fd:readonly
disabled       Sep_01   svc:/application/backup/bacula-fd:default
online         Sep_05   svc:/application/backup/bacula-fd:home
online         Sep_05   svc:/application/backup/bacula-fd:cifs
```

Die kleinen Freuden der Administratoren

- SMF Kommandozeilen Option „-L“

```
obi-wan# svcs -L bacula-fd:www
/var/svc/log/application-backup-bacula-fd:www.log

obi-wan# svcs -xL bacula-fd:www
svc:/application/backup/bacula-fd:www (Bacula FileDaemon)
 State: online since September 29, 2014 08:23:29 AM MEST
   See: /var/svc/log/application-backup-bacula-fd:www.log
Impact: None.
   Log:
[ Sep 18 12:59:58 Enabled. ]
[ Sep 18 12:59:58 Executing start method ("...shortend to fit..."). ]
[ Sep 18 12:59:59 Method "start" exited with status 0. ]
[ Sep 29 08:23:27 Executing start method ("...shortend to fit..."). ]
[ Sep 29 08:23:29 Method "start" exited with status 0. ]

Use: 'svcs -Lv svc:/application/backup/bacula-fd:www' to view the
complete log.
```

Die kleinen Freuden der Administratoren

- Netzwerktools
  - *dladm(1m), tcpstat(1m), ipstat(1m)* und *dlstat(1m)*

```
alderaan# tcpstat -c -l 5 10 1
Please wait...
ZONE     PID PROTO  SADDR     SPORT DADDR              DPORT   BYTES
global   965 TCP    mosel.rz.   666 home.rz.uni-ulm.   2049    3.5M
global   965 TCP    buche.rz.   874 home.rz.uni-ulm.   2049  190.0K
global   965 TCP    hannover.   865 home.rz.uni-ulm.   2049   19.6K
global   965 TCP    home.rz.u  2049 mosel.rz.uni-ulm    666   16.2K
global   965 TCP    hof.rz.un   799 home.rz.uni-ulm.   2049   14.4K
Total: bytes in:  3.7M bytes out: 41.8K
```

Die kleinen Freuden der Administratoren

- parallelisierte Tools
  - *pbzip(1)* und *pixy(1)*
    - Ausgangsbasis 1GB großes tar-Archiv */usr/lib*

```
obi-wan# time bzip2 < usr_lib.tar > /dev/null
real    2m19.602s
user    2m19.184s
sys     0m0.397s

obi-wan# time pbzip2 -p8 < usr_lib.tar > /dev/null
real    0m19.102s
user    2m30.868s
sys     0m0.720s
```

Die kleinen Freuden der Administratoren

- wer nutzt eigentlich welchen socket?
  - Ausgabe wurde für die Folie neu formatiert

```
obi-wan# netstat -f inet -u

TCP: IPv4
   Local Address        Remote Address    User      Pid    Command
---------------       -------------    ----      ---    -------
   obi-wan.ssh          remote.39849     root      961    sshd
   obi-wan.61654        spica.41917      root      784    nscd
   obi-wan.62880        home.nfsd        root      896    automountd
   obi-wan.50203        home.nfsd        root      896    automountd
   obi-wan.798          home.nfsd        root      0      <kernel>
   obi-wan.60030        cifs.46203       daemon    12886  nfs4cbd
...
```

ZFS Freuden

- erhebliche Verbesserung der resilver-Performance für RAIDZ* Pools
- Pool import ohne mount oder export (NFS/CIFS) durch Verwendung der Kommandozeilen Option „-N“
- `zfs send` liefert Schätzung über die Größe des Streams
  - korreliert meist gut mit der Übertragungsdauer

ZFS Freuden

```
yoda# zfs snapshot -r ai@sync-02
yoda# zfs send -v -R -I ai@sync-01 ai@sync-02 |
        rsh windu zfs receive -x compression -v -d -F mirror/yoda
sending @sync-01 to ai@sync-02
sending @sync-01 to ai/iso@sync-02
sending @sync-01 to ai/packages@sync-024
sending @sync-01 to ai/repo@sync-02
sending @sync-01 to ai/repo/s11u2-ga@sync-02
sending full stream to ai/repo/s11u2-2_8_0@sync-02
sending @sync-01 to ai/repo/s11u2-1_5_0@sync-02
sending @sync-01 to ai/repo/s11u2b42@sync-02
sending @sync-01 to ai/repo/kiz@sync-02
sending @sync-01 to ai/target@sync-02
estimated stream size: 5.72G
```

Zusammenfassung

Danke für's Zuhören
und
Danke an Harald Däubler