

# **Virtualisierungscluster mit Ganeti, KVM und DRBD**

### **Aufbau / Ziele:**

- Einführung
- Überblick Ganeti
- Instanzen
- Installation
- Administration
- Web-Administration
- Ausblick
- Links / Hilfen

- Stefan Neufeind
- Mit-Geschäftsführer der SpeedPartner GmbH aus Neuss ein Internet-Service-Provider (ISP)
  - Individuelle TYPO3-Entwicklungen
  - Hosting, Housing, Managed Services
  - Domains / Domain-Services
  - IPv6, DNSSEC, ...
- Aktive Mitarbeit im Community-Umfeld (PHP/PEAR, TYPO3, Linux)
- Freier Autor für z.B. t3n, iX, Internet World, ...

### Warum virtualisieren?

- Gemeinsame, effizientere Nutzung von Ressourcen eines Hostsystems
- Einheitliche (virtuelle) Systemumgebungen
- Vereinfachung bei Ausrollen und Skalierung von (virtuellen) Maschinen
- Einsparungen (Zeit und Kosten)

### Herausforderungen:

- Sicherheit
- Performance
- Stabilität
- Verfügbarkeit  
(Single point of failure?)

### Lösungsansätze:

- Erprobte Lösungen, Abwägung Risiken/Möglichkeiten
- Zentrale Zuteilung umfangreicher Host-Ressourcen
- Erprobte Lösungen, „schlanke“ Standard-Lösungen
- Replikation, Failover, N+1 Redundanz,  
Livemigration für geplante Wartungsarbeiten

### Verfügbare Systeme (Auswahl Voll- und Para-Virtualisierung):

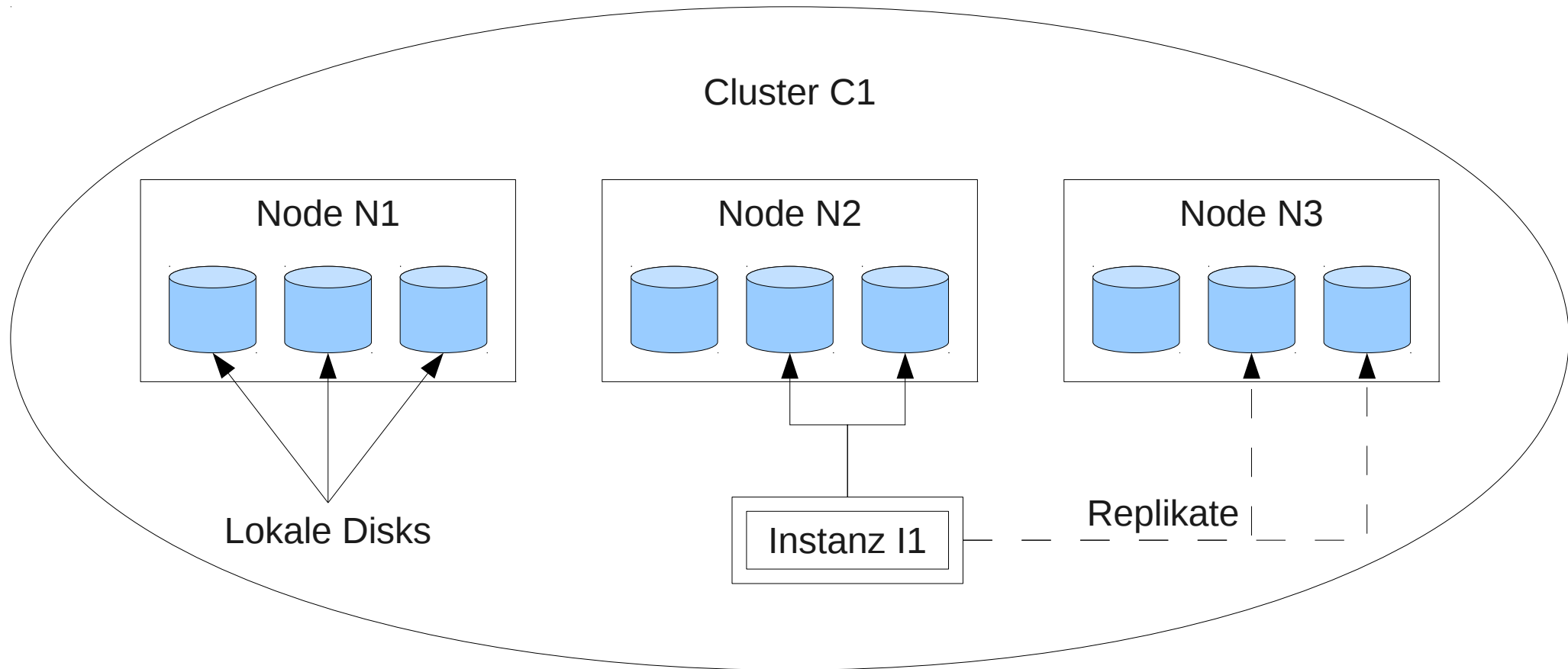
- VMware, VirtualBox, Xen, KVM, ...

### Ganeti:

- Virtual-Server-Management für Cluster
- Nutzung von Xen und KVM als Hypervisor
- Verwendung von lokalem Storage (+DRBD)
- Nutzung von „Standard-Hardware“
- Open-Source-Projekt, aktive Community
  - Durch Google initiiert / geleitet
  - Open-Source seit 2007
  
- Basiert auf Standard Linux-/OSS-Komponenten
- Geringe Abhängigkeiten
- Modular und erweiterbar (Hooks, REST-API, ...)
- „Simpel“ Storage-Lösung
- Gutes Preis-Leistungs-Verhältnis

### Abgrenzung zu anderen Lösungen:

- Lösungen wie „libvirt“ für einzelnen Knoten
- Eigene Lösungen Hypervisor-Lösungen
- Nutzung von NAS-/SAN-Lösungen
- Teilweise spezialisierte Hardwarelösungen
- Teilweise kommerziell

**Bestandteile:**

### Bestandteile:

- Cluster: 1+n Knoten („nodes“)
- Knoten:
  - Können dynamisch zu Cluster hinzugefügt/entfernt werden
  - Enthalten Dienste zur Verwaltung sowie je nach Rolle zentrale Dienste (Master)
  - Führen Instanzen (VMs) aus
  - Lokales Storage, optional repliziert
    - Volume-Verwaltung per LVM
    - Replikation per DRBD
- Instanz: Virtuelle Maschine (VM), optional fehlertolerant innerhalb des Clusters
- Ganeti-Cluster mit Fehlertoleranz „by design“
  - Jeder Knoten autark
  - Cluster-Konfiguration auf allen Nodes bekannt
  - Cluster-Kontrolle- und API-Dienste auf „master“-Knoten, jedoch jederzeit failover möglich

### Dienste:

- |  |                                 |
|--|---------------------------------|
| • ganeti-noded (auf allen Knoten):                     | Verwaltet Ressourcen            |
| • ganeti-confd (auf allen Knoten, aktiv auf „master“): | Verwaltet Cluster-Konfiguration |
| • ganeti-rapi (auf „master“):                          | HTTP-basierte REST-API          |
| • ganeti-masterd (auf „master“):                       | Zentrale Cluster-Verwaltung     |

### Periodische aufgerufene Skripte:

- ganeti-watcher
  - Auf Master:
    - Als aktiv markierte Instanzen automatisch starten (begrenzte Anzahl)
    - DRBD-Links reaktivieren, falls Secondary rebootet wurde
    - Alte Jobs archivieren
  - Auf jedem Node:
    - Node-Daemons neustarten falls notwendig
    - Führt Hook-Skripte aus
    - Stoppt Instanzen/DRBD-Links für „offline“ markierte Nodes (falls „maintain\_node\_health“)
- ganeti-cleaner: löscht alte, archivierte Jobs und abgelaufene Zertifikate/Keys



### Tools zur Verwaltung:

- gnt-cluster: Cluster-bezogene Kommandos
  - Erstellen/löschen, clusterweit Dateien kopieren oder Kommandos ausführen, Cluster-Parameter (Default-Werte für Instanzen) setzen, Cluster-Status prüfen, ...
- gnt-node: Node-bezogene Kommandos
  - Hinzufügen zum/entfernen aus Cluster, Übersicht Ressourcen (Storage/RAM) und Instanzen, Migration/Failover (primäre Instanzen), Evakuierung (sekundäre Instanzen)
- gnt-instance: Instanz-bezogene Kommandos
  - Erstellen/verändern/löschen, starten/stoppen, migrieren, Übersicht
- gnt-backup: Import/Export
  - Instanz (Daten und Konfiguration) importieren/exportieren
    - z.B. für Migration zwischen Clustern
- gnt-job: Job-Verwaltung
  - Status über laufende/geplante Jobs, Detaillierte Log-Informationen, Beobachten der Ausgaben laufender Jobs

### Parameter:

- Allgemeine Parameter (beparams; Backend-Parameter)
  - Speicher
  - CPUs
- Netzwerk-Parameter (nicparams)
  - Bridged / routed
  - MACs
  - Verbindungen (z.B. Bridge-Zugehörigkeit)
- Hypervisor-Parameter (hvparams)
  - ACPI
  - Boot-Reihenfolge
  - Festplatten-/CDROM-/Floppy-Typ etc.
  - Netzwerk-Typ
  - Parameter für VNC, Serialkonsole, ...
  - Kernel-Parameter (je nach Hypervisor)

Konfiguriert über:

Cluster-Standardwerte

oder

Instanz-spezifische Werte

### Parameter:

- Disk-Template
    - diskless: keine Festplatten, für Sonderfälle
    - file: reguläre Dateien
    - plain: LVM-Device
    - drbd: LVM-Device mit DRBD
- keine Redundanz
- Spiegelung auf 2 Nodes (aktiv/passiv)

### Basissystem:

- (Fast) beliebige, aktuelle Linux-Distribution
  - Debian / Ubuntu
  - Gentoo
  - RHEL / CentOS
    - Mit ELRepo für DRBD 8.3-Pakete
    - Ganeti-Pakete (leider nur) aus einer inoffiziellen Quelle und selbst rebuilden
- Xen oder KVM
- LVM (empfohlen; notwendig für Redundanz)
- DRBD 8.2 / 8.3 (notwendig für Redundanz)
  - DRBD 8.4 bisher nicht unterstützt; wesentliche Syntax-Änderungen (<http://code.google.com/p/ganeti/issues/detail?id=212>)
  - Bei Bedarf Anzahl DRBD-Instanzen erhöhen; ab DRBD 8.3. `usermode_helper` umsetzen (<http://code.google.com/p/ganeti/issues/detail?id=65>)

```
echo drbd minor_count=128 usermode_helper=/bin/true >> /etc/modules
```

**Vorbereitungen:**

- Basis-Installation der Server inkl. LVM sowie Paketen für DRBD, Python-Module, Ganeti
- Planung Netzwerk
  - Primäres Netz für Kommunikation benötigt
  - Empfehlung (aber optional):  
Sekundäres Netz für Replikation und Kommunikation zwischen den Nodes
- Einrichtung Netzwerkbridge(s)
- root-Login von Mastern-Servern(n) auf Nodes muss erlaubt sein (PermitRootLogin)
- DNS für Nodes und Instanzen (VMs) sauber konfigurieren
  - Prüfungen / DNS-Abfragen teilweise deaktivierbar,  
jedoch leichter saubere DNS-Einträge zu verwenden
  - DNS-Eintrag Cluster-Name = Master-IP

**Primäres Netz / Extern:**

```
192.168.1.100 cluster.example.com
192.168.1.101 node1.example.com
192.168.1.102 node2.example.com
192.168.1.103 node3.example.com
192.168.1.201 vm1.example.com
```

**Sekundäres Netz:**

```
10.0.1.1
10.0.1.2
10.0.1.3
```

**Initialisierung:**

- Auszuführen auf Master, hier node1.example.com

```
gnt-cluster init --enabled-hypervisors=kvm --master-netdev=br0 --vg-name=virvg \  
-s 10.0.1.1 democluster.example.com  
gnt-node add -s 10.0.1.2 node2.example.com  
gnt-node add -s 10.0.1.3 node3.example.com
```

**Instanz erzeugen:**

- Standardmäßig startet Installation automatisch (abhängig vom OS-Type, z.B. per debootstrap)
- Hier beispielhaft manueller Start mit einmalig anderen Boot-Parametern

```
gnt-instance add --disk-template=drbd -B memory=1G --os-size=10000 \  
--os-type=linux+default --no-start --no-install -n node1.example.com:node2.example.com \  
vm1.example.com  
gnt-instance start -H boot_order=cdrom,cdrom_image_path=/tmp/CentOS-6.0-x86_64-  
minimal.iso vm1.example.com
```

- Konsole per VNC-Port erreichbar (bei Bedarf evtl. durch SSH tunneln),  
automatisch vergebener Port siehe Instanz-Infos

```
gnt-instance info vm1.example.com
```

### Standardkommandos:

- Starten, stoppen (sendet ACPI-Event und wartet), sofortiges stoppen („destroy“)

```
gnt-instance startup <instance>  
gnt-instance shutdown <instance>  
gnt-instance shutdown --shutdown-timeout=0 <instance>
```

- Auf serielle Konsole zugreifen
  - Konsole muss je nach Distribution in VM erst aktiviert werden
  - Verlassen mit Ctrl+] ]

```
gnt-instance console <instance>
```

- Netzwerk anpassen
  - MAC anpassen, weiteres Netzwerk-Interface hinzufügen
  - Falls keine MAC angegeben, wird automatisch eine zufällige generiert

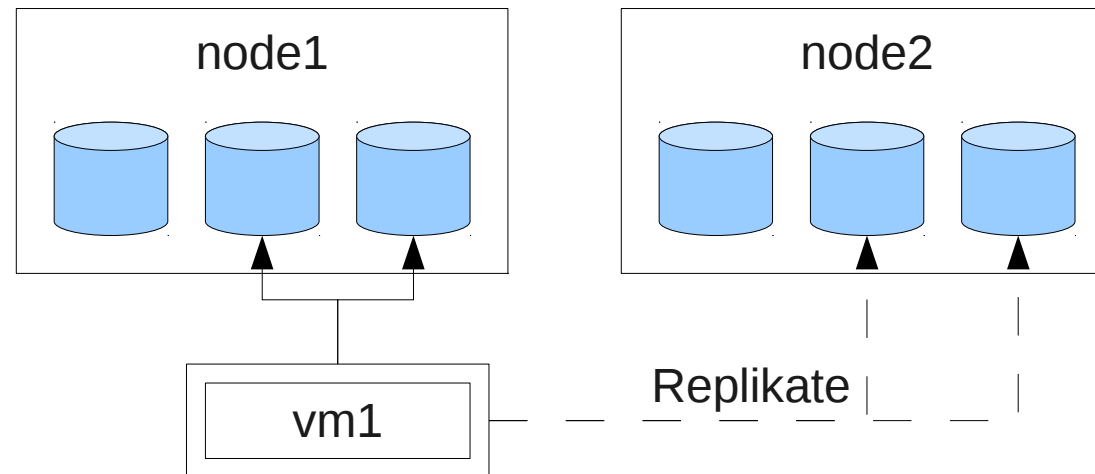
```
gnt-instance modify --net 0:mac=52:54:00:d0:0f:00 <instance>  
gnt-instance modify --net add:mac=52:54:00:d0:0f:01,mode=bridged,link=br1 <instance>
```

- Disk hinzufügen

```
gnt-instance modify --disk add:size=5G <instance>
```

### Kommandos für Fortgeschrittene:

- VM (Instanz) von 1 Node (LVM) auf 2 Node (DRBD, primary/secondary) erweitern und zurück
  - Instanz muss vorher Änderungen gestoppt werden
  - Es werden alle zugehörigen Disks angepasst

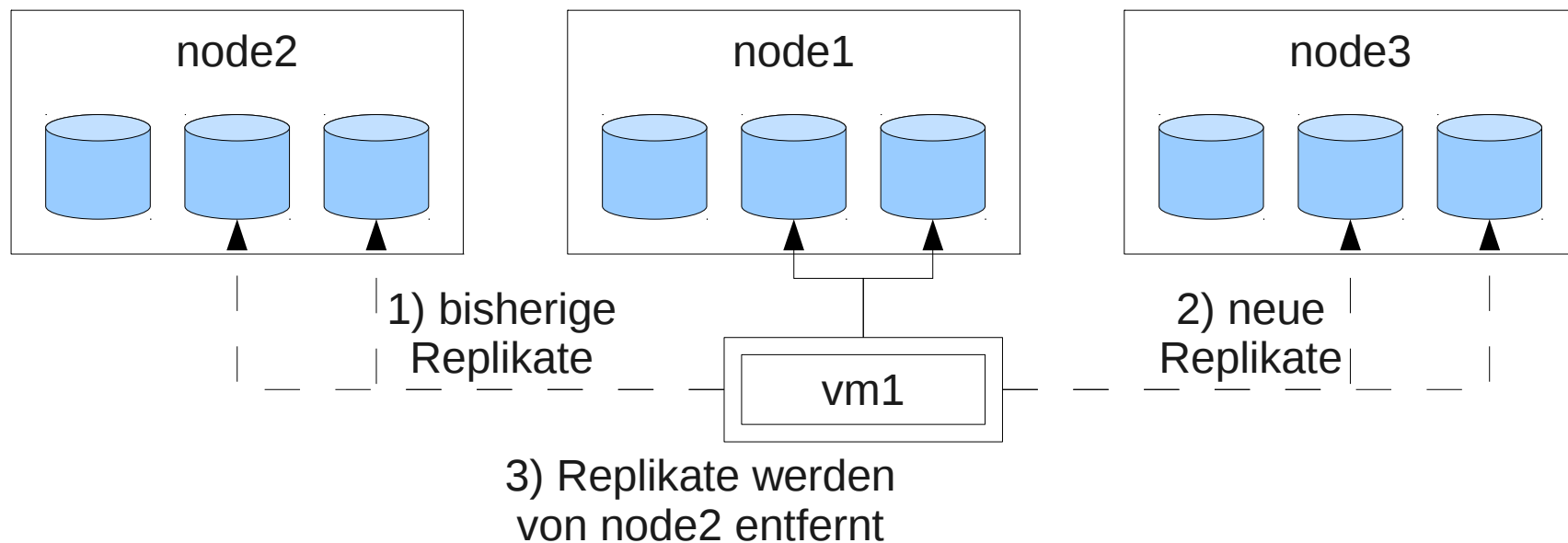


```
# Erweiterung auf primary/secondary  
gnt-instance modify -t drbd -n node2 vm1.example.com  
# zurück zu Single-Node  
gnt-instance modify -t plain vm1.example.com
```



**Kommandos für Fortgeschrittene:**

- Secondary einer Instanz auf einen anderen Node migrieren



```
gnt-instance replace-disks --new-secondary node3 vm1.example.com
```

### Kommandos für Fortgeschrittene:

- Status des Clusters prüfen
  - Instanzen und Ressourcen
  - failover-Szenarien

```
gnt-cluster verify
```

```
Tue Jan 01 00:01:03 2012 * Verifying global settings
Tue Jan 01 00:01:05 2012 * Gathering data (2 nodes)
Tue Jan 01 00:01:08 2012 * Gathering disk information (2 nodes)
Tue Jan 01 00:01:12 2012 * Verifying node status
Tue Jan 01 00:01:12 2012 - ERROR: node node2.example.com: file '/etc/hosts' has wrong
checksum
Tue Jan 01 00:01:12 2012 * Verifying instance status
Tue Jan 01 00:01:12 2012 * Verifying orphan volumes
Tue Jan 01 00:01:12 2012 * Verifying orphan instances
Tue Jan 01 00:01:12 2012 * Verifying N+1 Memory redundancy
Tue Jan 01 00:01:12 2012 - ERROR: node node2.example.com: not enough memory to
accomodate instance failovers should node node1.example.com fail (6144MiB needed,
1791MiB available)
Tue Jan 01 00:01:12 2012 - ERROR: node node1.example.com: not enough memory to
accomodate instance failovers should node node2.example.com fail (6656MiB needed,
2282MiB available)
Tue Jan 01 00:01:12 2012 * Other Notes
Tue Jan 01 00:01:13 2012 * Hooks Results
```

**Kommandos für Fortgeschrittene:**

- Migration einer Instanz

```
gnt-instance migrate <instance>
```

- Falls von Hypervisor unterstützt Livemigration
- Optional Non-Live-Migration möglich (VM vorübergehend eingefroren)

```
Tue Jan 01 00:01:20 2012 Migrating instance vm1.example.com
Tue Jan 01 00:01:20 2012 * checking disk consistency between source and target
Tue Jan 01 00:01:21 2012 * switching node node2.speedpartner.de to secondary mode
Tue Jan 01 00:01:22 2012 * changing into standalone mode
Tue Jan 01 00:01:23 2012 * changing disks into dual-master mode
Tue Jan 01 00:01:25 2012 * wait until resync is done
Tue Jan 01 00:01:26 2012 * preparing node2.speedpartner.de to accept the instance
Tue Jan 01 00:01:27 2012 * migrating instance to node2.speedpartner.de
Tue Jan 01 00:01:04 2012 * switching node node1.speedpartner.de to secondary mode
Tue Jan 01 00:01:10 2012 * wait until resync is done
Tue Jan 01 00:01:14 2012 * changing into standalone mode
Tue Jan 01 00:01:15 2012 * changing disks into single-master mode
Tue Jan 01 00:01:17 2012 * wait until resync is done
Tue Jan 01 00:01:18 2012 * done
```

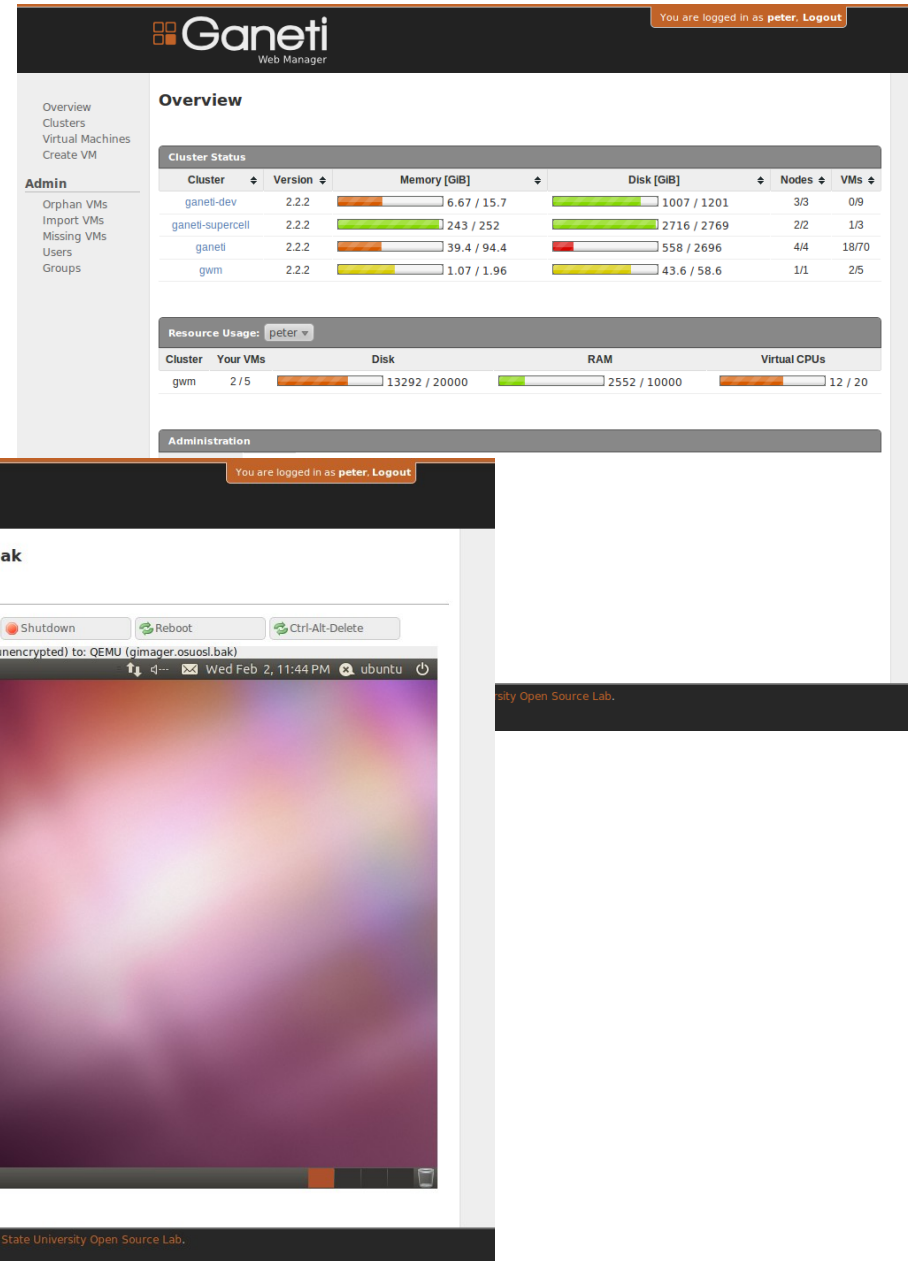
### Remotezugriff per http(s) möglich:

- REST-API für viele Funktionen
- Rückgaben per JSON

```
# wget --no-check-certificate -O - https://127.0.0.1:5080/2/instances/vm1.example.com
{
  "admin_state": true,
  "beparams": {
    "auto_balance": true,
    "memory": 1024,
    "vcpus": 1
  },
  "os": "linux+default",
  "pnode": "node1.example.com",
  "serial_no": 14,
  "snodes": [
    "node2.example.com"
  ],
  "status": "running",
  "tags": [],
  "uuid": "32a0fd4b-2056-8427-8339-894c7b234c70"
}
```

### Administration über Webinterface möglich:

- Unabhängig entwickeltes Add-On
- Greift per REST-API auf Ganeti zu
- Installation je nach Distribution evtl. schwierig
- Inkl. Rechte- und Quota-Verwaltung
- Grafische Konsole ohne Plugins, dank noVNC per HTML5 (Canvas, WebSockets)



The screenshot displays the Ganeti Web Manager interface. The top navigation bar includes 'Overview', 'Clusters', 'Virtual Machines', and 'Create VM'. The 'Admin' section lists 'Orphan VMs', 'Import VMs', 'Missing VMs', 'Users', and 'Groups'. The main content area is divided into two sections: 'Overview' and 'Administration'.

**Overview Section:**

**Cluster Status Table:**

Cluster	Version	Memory [GiB]	Disk [GiB]	Nodes	VMs
ganeti-dev	2.2.2	6.67 / 15.7	1007 / 1201	3/3	0/9
ganeti-supercell	2.2.2	243 / 252	2716 / 2769	2/2	1/3
ganeti	2.2.2	39.4 / 94.4	558 / 2696	4/4	18/70
gwm	2.2.2	1.07 / 1.96	43.6 / 58.6	1/1	2/5

**Resource Usage:** peter

Cluster	Your VMs	Disk	RAM	Virtual CPUs
gwm	2 / 5	13292 / 20000	2552 / 10000	12 / 20

**Administration Section:**

The 'Administration' section shows a specific virtual machine named 'ganeti-test : gimager.osuosl.bak'. It includes tabs for 'Overview', 'Users', and 'Console'. Below the tabs are control buttons: 'Encrypt', 'Disconnect', 'Shutdown', 'Reboot', and 'Ctrl-Alt-Delete'. The console window displays a QEMU window titled 'Connected (unencrypted) to: QEMU (gimager.osuosl.bak)' with a desktop environment showing 'Examples' and 'Install Ubuntu 10.10' icons. The system tray shows 'Wed Feb 2, 11:44 PM' and 'ubuntu'.

A Project by the Oregon State University Open Source Lab.

### Dinge aktuell in Arbeit / ungelöste Probleme (eine Auswahl):

- Bisher nur letzte Festplatte einer Instanz entfernbar, nicht beliebige (Lösung bereits realisiert; wird ab Ganeti 2.6 verfügbar sein)  
<http://code.google.com/p/ganeti/issues/detail?id=188>
- Verschieben von Festplatten zwischen Instanzen ermöglichen  
<http://code.google.com/p/ganeti/issues/detail?id=172>
- Erweiterung um Funktionalität für "External Storage" (z.B. SAN, Ceph, ...)  
<http://www.mail-archive.com/ganeti-devel@googlegroups.com/msg21522.html>  
<http://docs.ganeti.org/ganeti/master/html/design-shared-storage.html>
- Zugriff auf Serial-Konsole für Ganeti Web Manager (voraussichtlich per jsTerm)  
<https://code.osuosl.org/issues/2217>

- Ganeti-Website, inkl. HowTos (im Wiki)  
<http://code.google.com/p/ganeti/>
- Ganeti-Source-RPMs (für RHEL/CentOS)  
<http://jfut.integ.jp/linux/ganeti/>
- ELRepo (DRBD-Pakete für RHEL/CentOS)  
<http://elrepo.org/>
- Remote API (RAPI)  
<http://docs.ganeti.org/ganeti/current/html/rapi.html>
- Ganeti Web Manager  
<https://code.osuosl.org/projects/ganeti-webmgr>

Danke fürs Zuhören  
sowie  
viel Erfolg und Spaß  
mit Ganeti!

Link zu den Slides: <http://talks.speedpartner.de/>

Bei Fragen stehen wir selbstverständlich gerne zur Verfügung:

Stefan Neufeind, [neufeind@speedpartner.de](mailto:neufeind@speedpartner.de)  
SpeedPartner GmbH, <http://www.speedpartner.de/>