KVM fürs Rechenzentrum

fork eines google-Projekts

Sascha Lucas

GISA GmbH Halle (Saale)

10. März 2018 Chemnitzer Linux Tage

GISA® IT. Mehr als Standard.

Vorstellung

Sascha Lucas

- aktuelle Tätigkeit
  - Linux Admin Team
  - SLES & Ubuntu
  - ansible
- Virtualisierung mit KVM seit $\approx$ 2010

GISA® IT. Mehr als Standard.

Gliederung

GISA®
IT. Mehr als Standard.

Gliederung

GISA®
IT. Mehr als Standard.

KVM: für wen und womit?

Rechenzentrum: Räumlichkeiten, in denen die zentrale Rechentechnik ... untergebracht ist (wikipedia)

Einsteiger

Mittlere

Größere

GISA® IT. Mehr als Standard.

KVM: für wen und womit?

Rechenzentrum: Räumlichkeiten, in denen die zentrale Rechentechnik ... untergebracht ist (wikipedia)

Einsteiger

Mittlere

Größere

Ganeti von Google entwickelt: verwaltet VMs in einem Cluster
KVM/XEN; CLI/REST; Python2/Haskell; BSD-2-Clause

GISA® IT. Mehr als Standard.

Ganeti: Eigenschaften für ein breites Spektrum

- mini-RZ für Einsteiger: Cluster aus 3 PCs
  - std. PC, lokale Platten
    $\rightarrow$ ideal für kleine Infrastrukturen (Startups, Außenstandorte)
  - Ganeti ist „selfcontained“
  - keine management-Systeme notwendig
  - keine weiteren Software-Stacks (keine DB etc.)
  - im Debian/(Ubuntu) enthalten

GISA®
IT. Mehr als Standard.

Ganeti: Eigenschaften für ein breites Spektrum

- mini-RZ für Einsteiger: Cluster aus 3 PCs
  - std. PC, lokale Platten
    → ideal für kleine Infrastrukturen (Startups, Außenstandorte)
  - Ganeti ist „selfcontained“
  - keine management-Systeme notwendig
  - keine weiteren Software-Stacks (keine DB etc.)
  - im Debian/(Ubuntu) enthalten
- mit „Rechenzentrums“-Eigenschaften
  - hohe Verfügbarkeit der VMs (N+1 Redundanz)
  - Cluster hat kein SPOF
  - scale-out (compute/storage)
  - VMs migrieren live
  - automatische Ressourcen-Verwaltung
  - upgrades ohne downtime (für die VMs)
- 1 Cluster skaliert in mittlere Region: 50-100 HW-Server, 1000 VMs
  → KMUs (ecommerce, Dienstleister, UNIs/Institute, ...)
- Anwendung bei „Großen“ (Google): hunderte von Clustern

GISA® IT. Mehr als Standard.

Gliederung

GISA®
IT. Mehr als Standard.

Ganeti-Nutzer

Große und Mittlere

Google interne Dienste (virt. Desktops, Entwicklungssysteme, ...)
sehr viele Cluster über Standorte verteilt

grnet griechisches Wissenschafts- und Technologie-Netzwerk → Synnefo Cloud Okeanos (≈ 5k VMs)

VMs are not cattle, they are pets.

GISA® IT. Mehr als Standard.

Ganeti-Nutzer

Große und Mittlere

Google interne Dienste (virt. Desktops, Entwicklungssysteme, ...)
sehr viele Cluster über Standorte verteilt

grnet griechisches Wissenschafts- und Technologie-Netzwerk → Synnefo Cloud Okeanos (≈ 5k VMs)

VMs are not cattle, they are pets.

OSUOSL hostet bekannte OSS-Projekte: busybox, inkscape, jenkins, kde, musicbrainz, mythtv, openstreetmap, qemu

debian.org Debian Infrastruktur

JULIE Lab Uni Jena Computer-Linguistik (≈ 16 VMs Infra und Sci)

GISA® IT. Mehr als Standard.

Ganeti-Nutzer

Große und Mittlere

Google interne Dienste (virt. Desktops, Entwicklungssysteme, ...)
sehr viele Cluster über Standorte verteilt

grnet griechisches Wissenschafts- und Technologie-Netzwerk → Synnefo Cloud Okeanos (≈ 5k VMs)

VMs are not cattle, they are pets.

OSUOSL hostet bekannte OSS-Projekte: busybox, inkscape, jenkins, kde, musicbrainz, mythtv, openstreetmap, qemu

debian.org Debian Infrastruktur

JULIE Lab Uni Jena Computer-Linguistik (≈ 16 VMs Infra und Sci)

skroutz gr. Preisvergleichs-Site (= Debian Paket Maintainer)

sprd.net AG spreadshirt T-Shirt-Druck (≈ 380 VMs e-com Plattform, ≈ 190 VMs Infra und Dev Hauptsitz Leipzig)

GISA GmbH IT-Dienstleister Halle (Saale) (≈ 200 VMs)

GISA® IT. Mehr als Standard.

Ganeti-Nutzer

Einsteiger

Brasilien, West-Afrika, Bangladesch, Bhutan, ...
... via Network Startup Ressource Center (NSRC) Uni Oregon

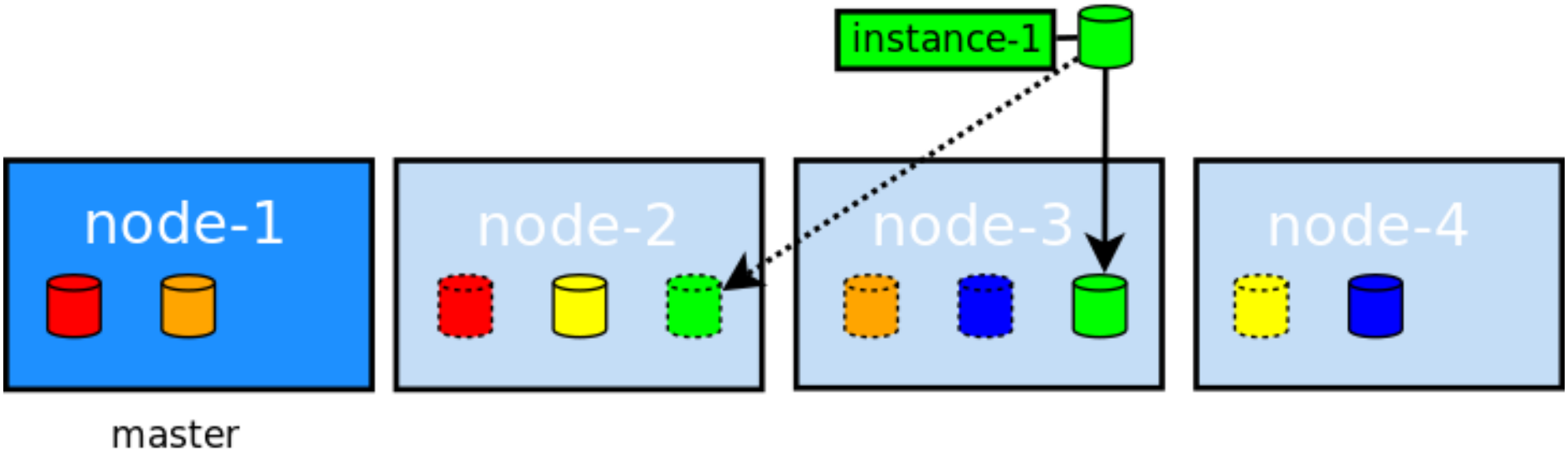
GISA® IT. Mehr als Standard.

Gliederung

GISA®
IT. Mehr als Standard.

Ganeti Cluster: Architektur

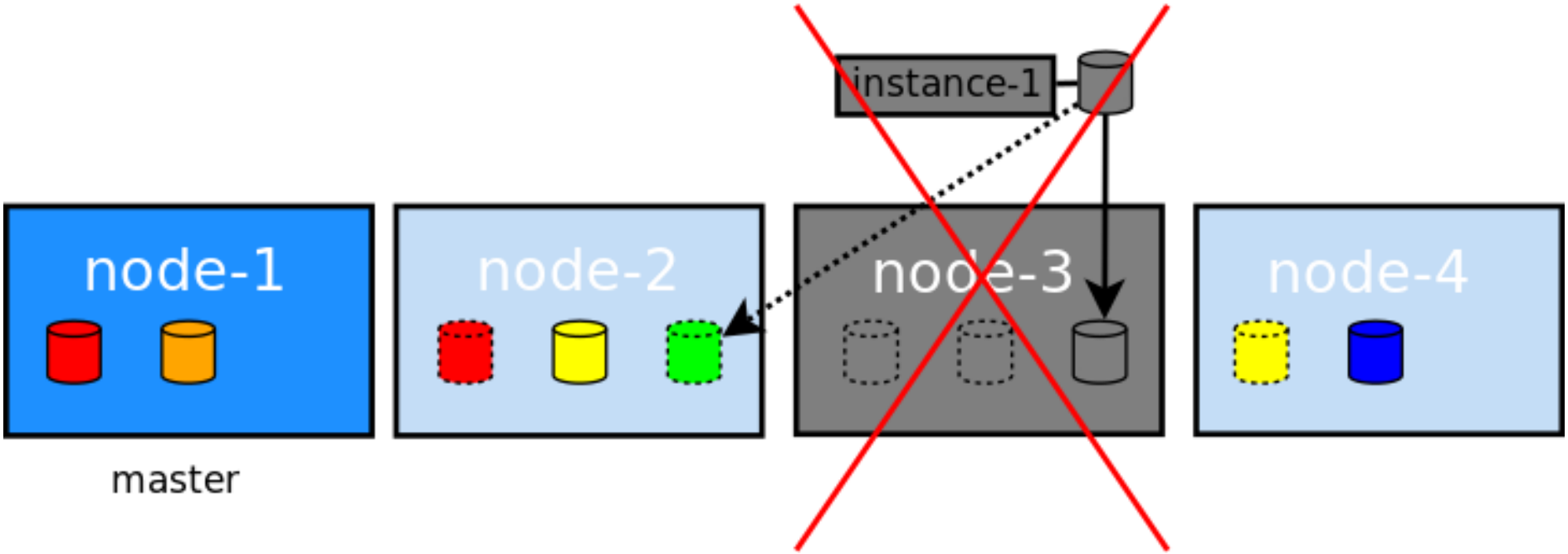
DRBD



- HW-Server = node
- master-node steuert Cluster
- Instanzen = VMs
- hier DRBD-Setup: Spiegelung der Daten je VM auf sekundärem node

Ganeti Cluster: Redundanzprinzipien

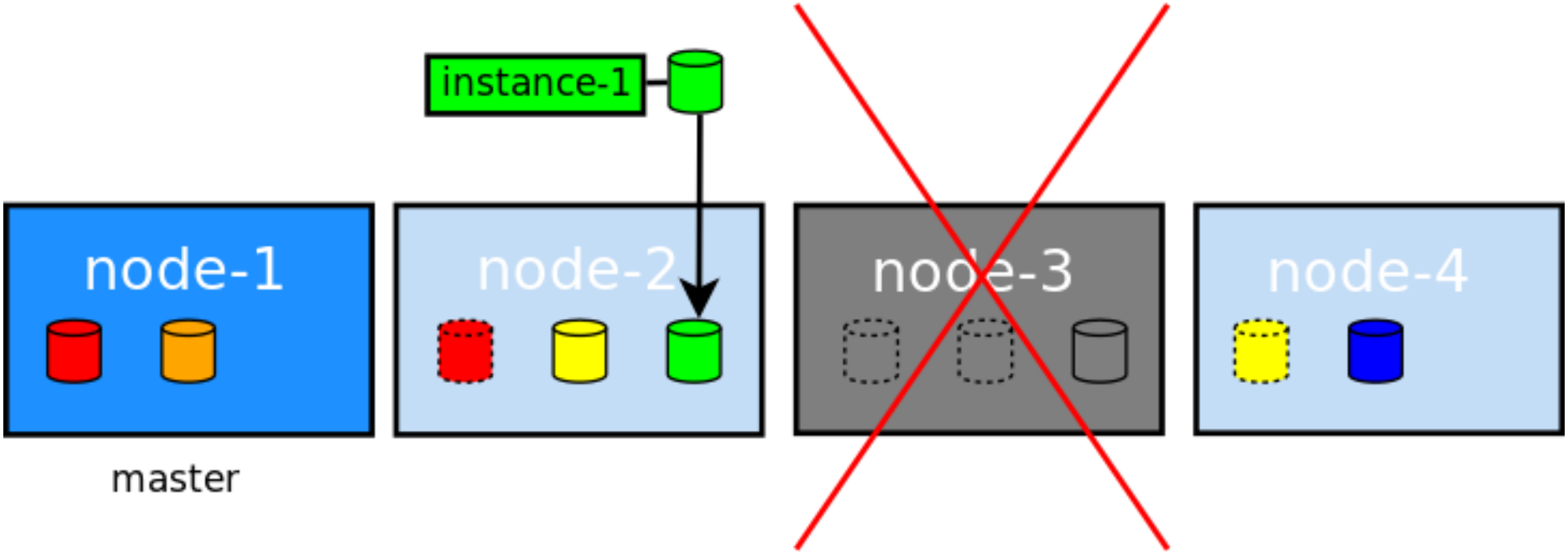
Ausfall node-3



- Ausfall node-3

Ganeti Cluster: Redundanzprinzipien

Ausfall node-3

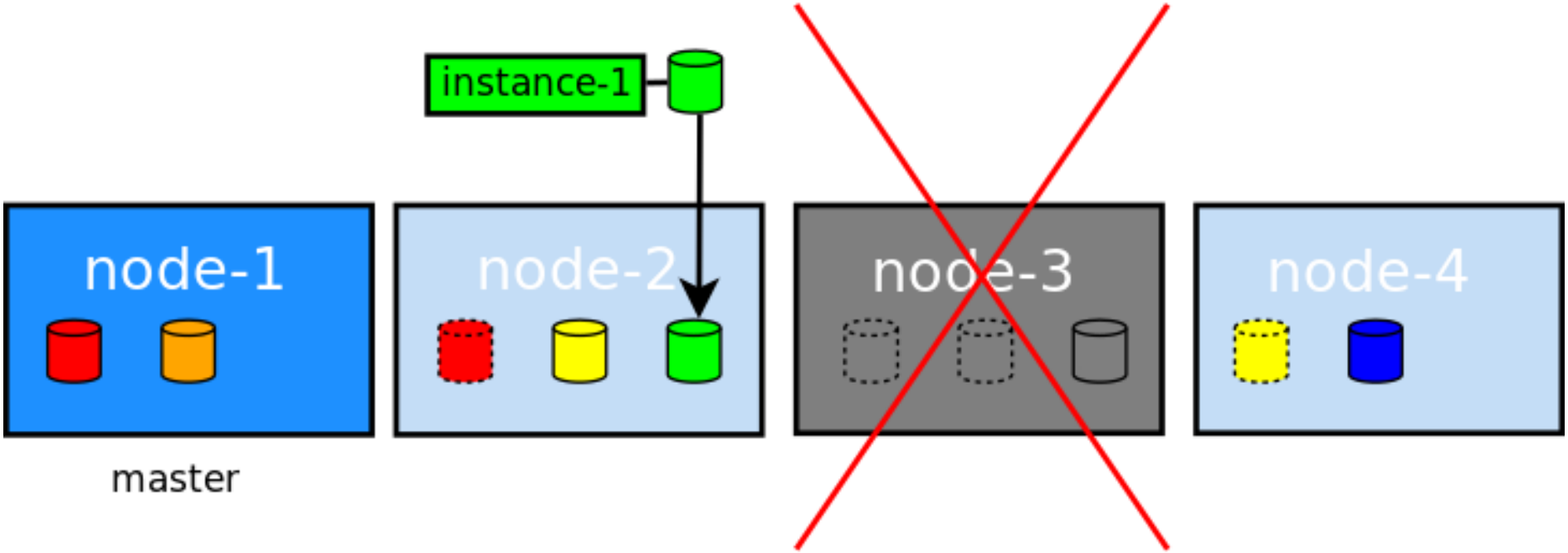


- Ausfall node-3
- failover aller Instanzen von node-3 (von Hand anzustoßen)
- Frage: was ist begrenzender Faktor für failover?

GISA® IT. Mehr als Standard.

Ganeti Cluster: Redundanzprinzipien

Ausfall node-3

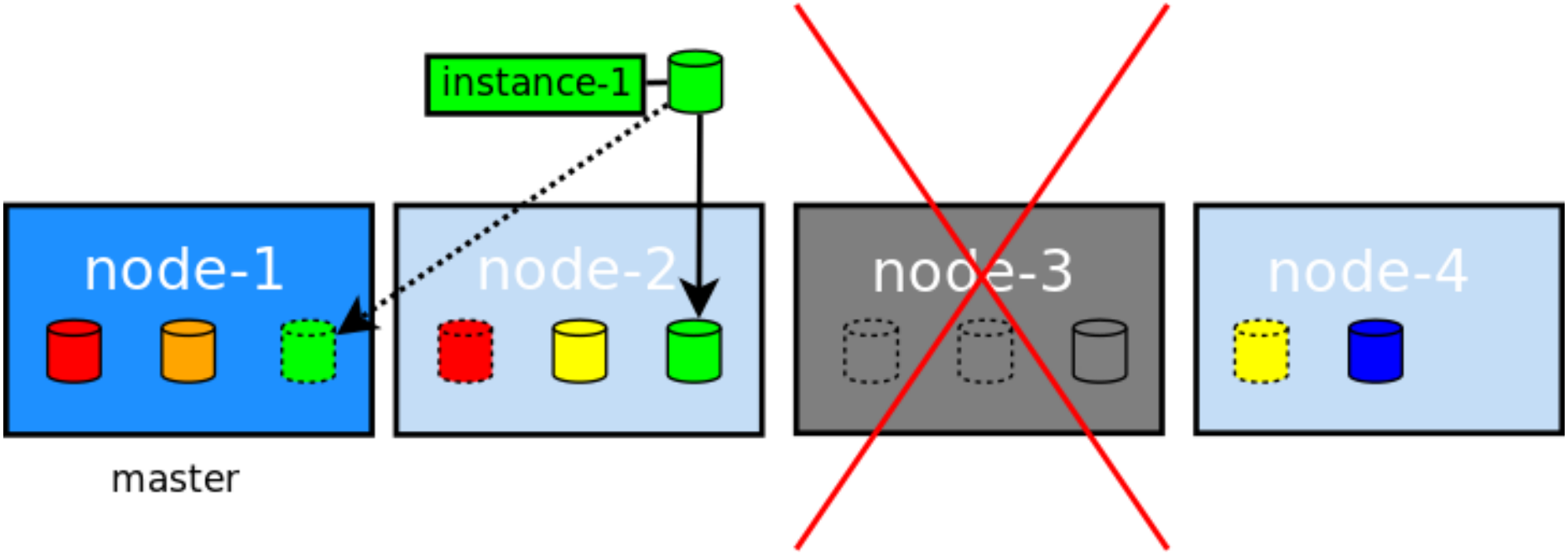


- Ausfall node-3
- failover aller Instanzen von node-3 (von Hand anzustoßen)
- Antwort: hier RAM node-2: N+1 Redundanz wird garantiert

GISA® IT. Mehr als Standard.

Ganeti Cluster: Redundanzprinzipien

Ausfall node-3

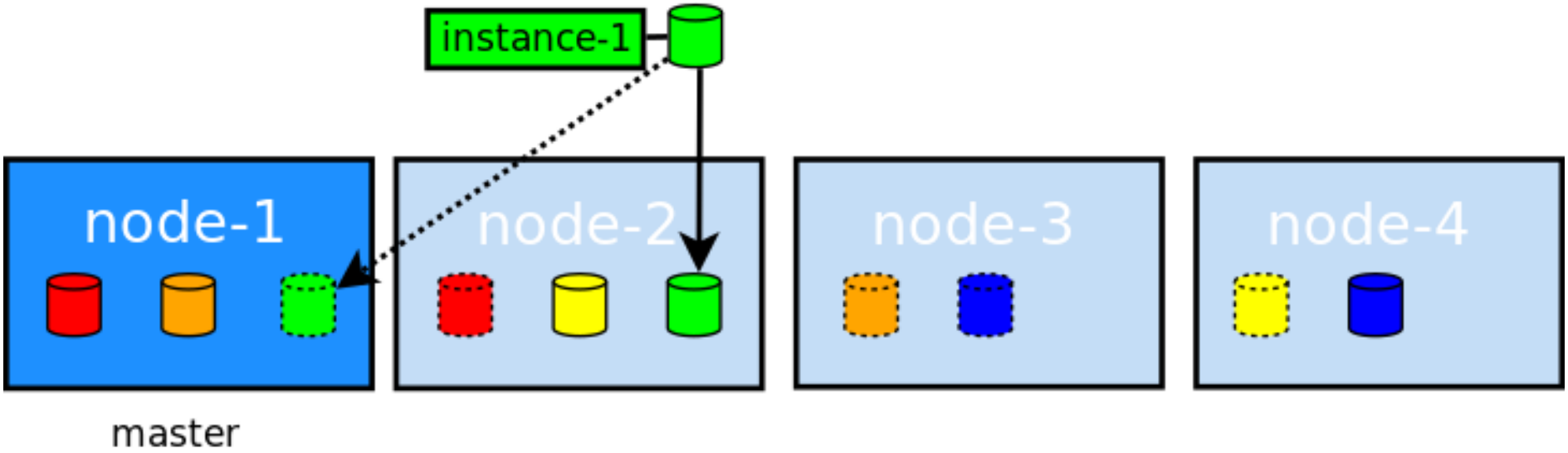


- Ausfall node-3
- failover aller Instanzen von node-3 (von Hand anzustoßen)
- Antwort: hier RAM node-2: N+1 Redundanz wird garantiert
- Wiederherstellen der Redundanz Instanz-1

GISA® IT. Mehr als Standard.

Ganeti Cluster: Redundanzprinzipien

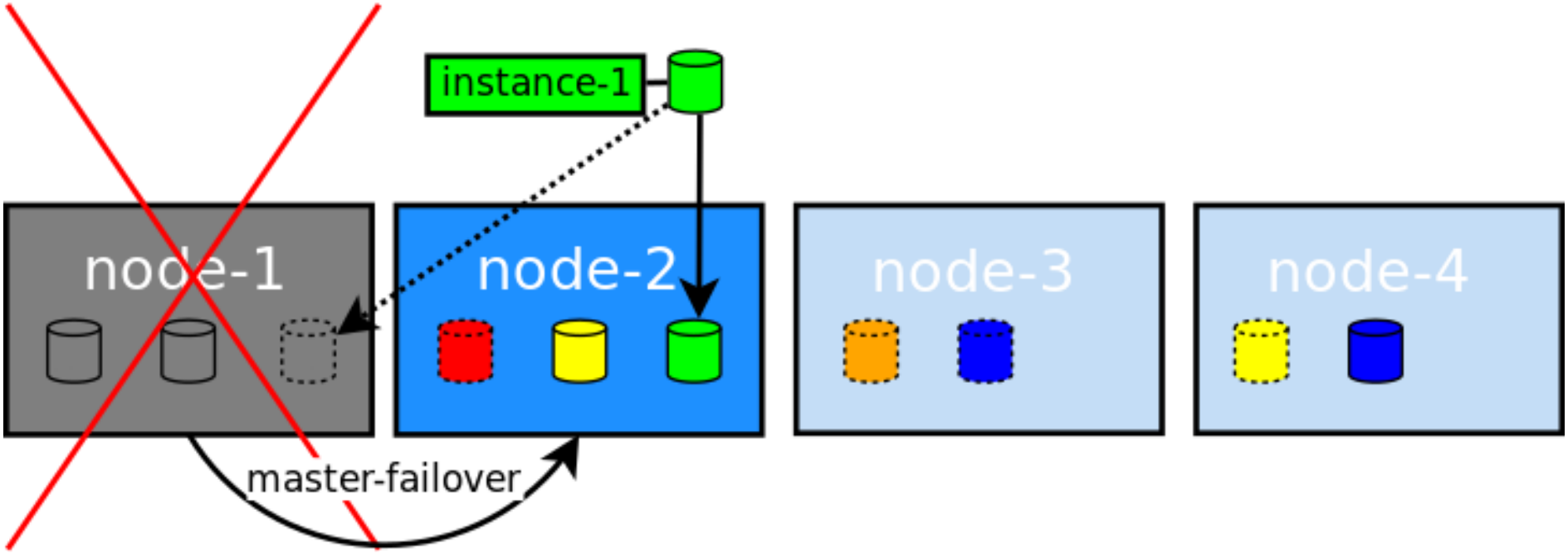
Reparatur node-3



- Wiederherstellen der Redundanz Instanz Orange und Blau
  - Delta-Synchronisation für Ausfallzeit node-3

Ganeti Cluster: Redundanzprinzipien

Ausfall master-node

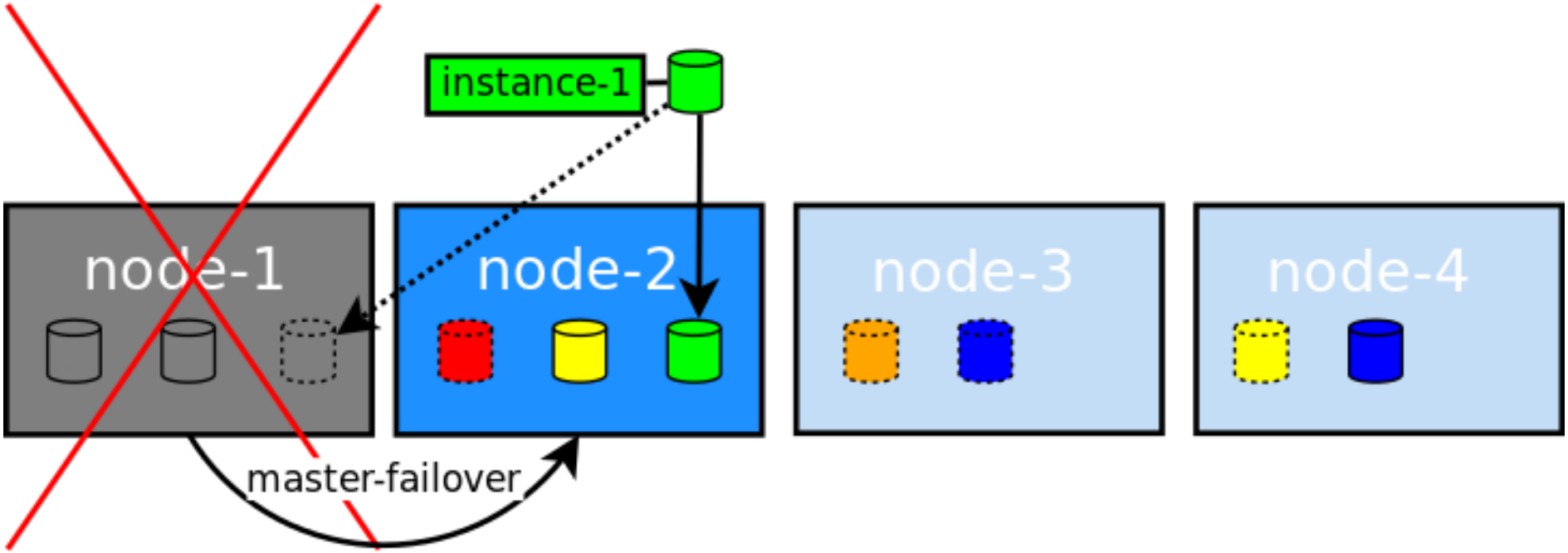


- zuerst master-Rolle umschalten (von Hand anzustoßen)
- default 10 master-Kandidaten

GISA® IT. Mehr als Standard.

Ganeti Cluster: Redundanzprinzipien

Ausfall master-node

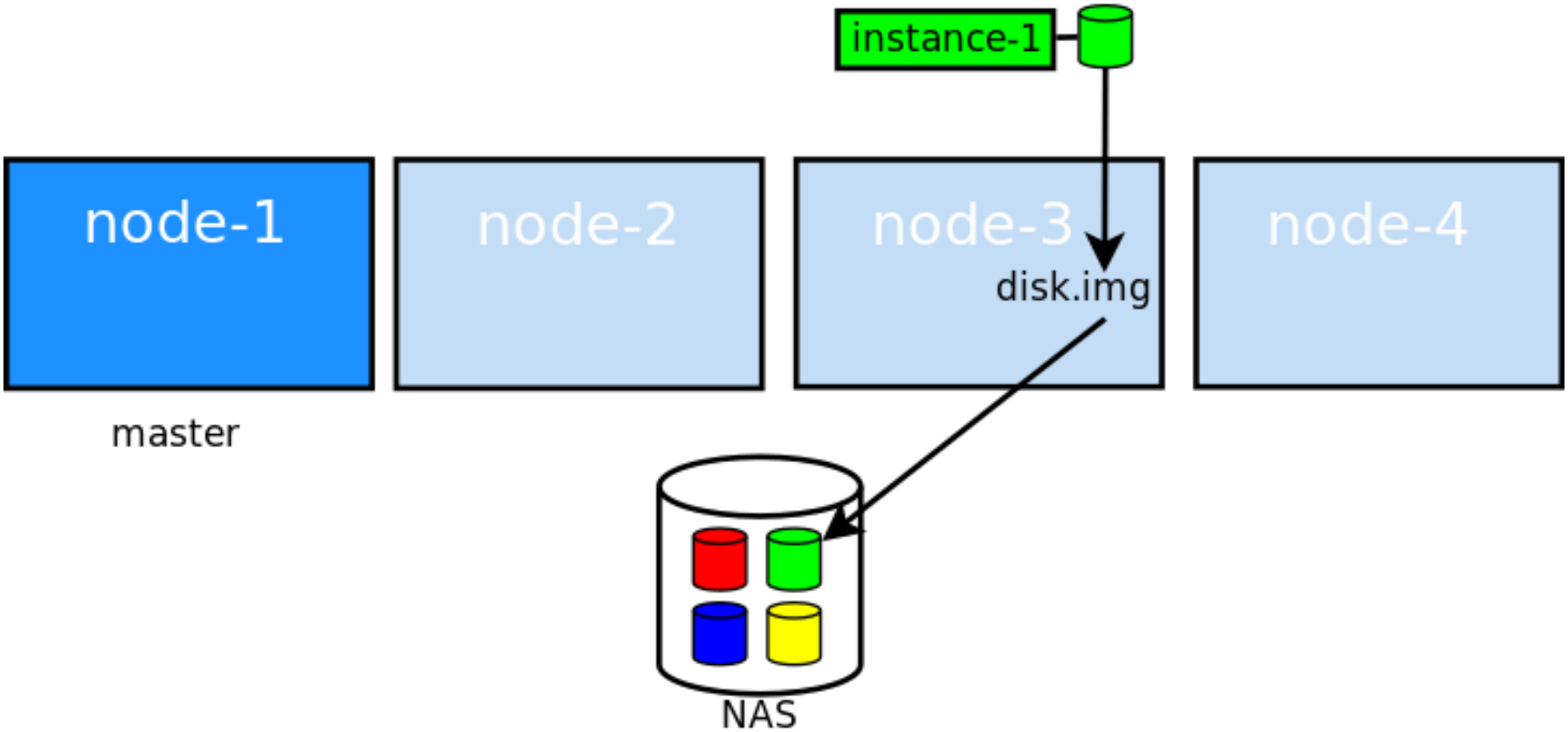


- zuerst master-Rolle umschalten (von Hand anzustoßen)
- default 10 master-Kandidaten
- danach failover node-1 (analog „Ausfall node-3“)
  - betrifft konkret Instanz Rot und Orange

GISA® IT. Mehr als Standard.

Ganeti Cluster: Architektur

NAS

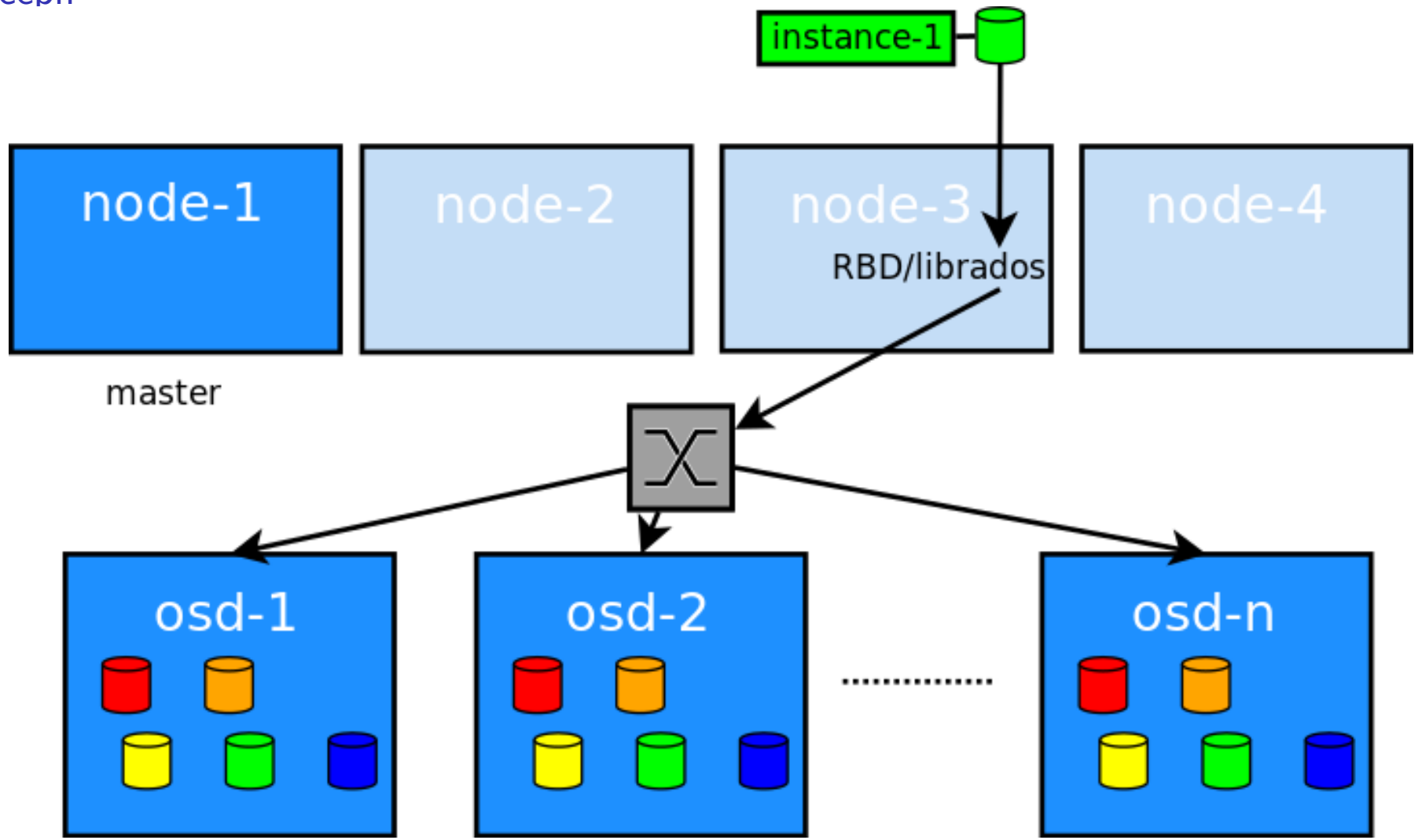


- Netzwerk-Dateisystem (z.B. NFS)
- raw disk images (sparse=thin), kein qcow2
- failover/migration jeder node möglich

GISA® IT. Mehr als Standard.

Ganeti Cluster: Architektur

ceph

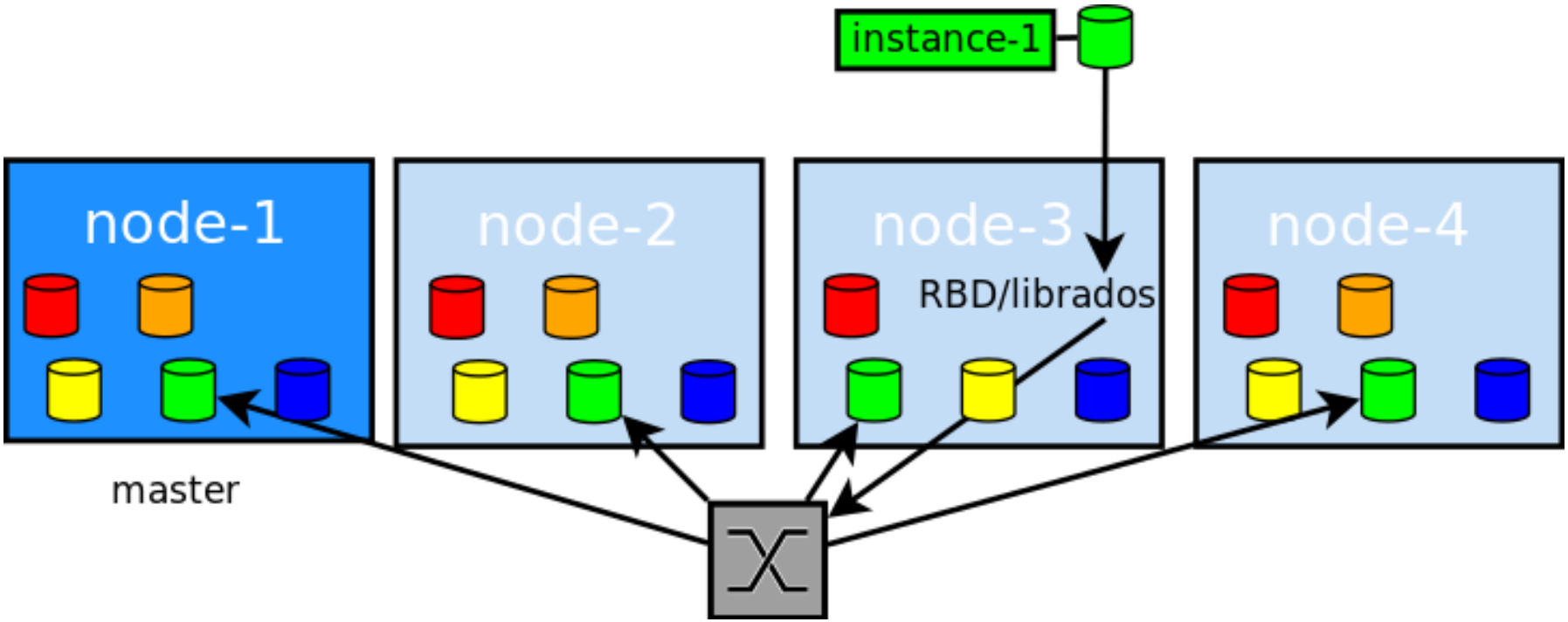


- ceph/rados cluster (dediziert, nicht via Ganeti verwaltet)
- RBD oder librados

GISA® IT. Mehr als Standard.

Ganeti Cluster: Architektur

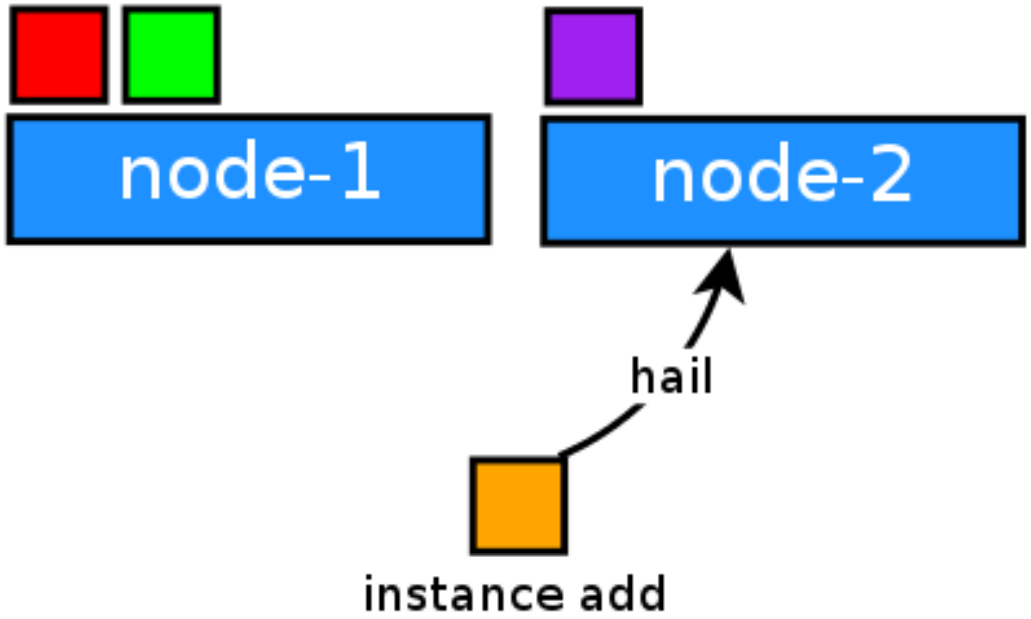
ceph



- ceph/rados cluster (convereged, nicht via Ganeti verwaltet)
- RBD oder librados

Ganeti Cluster: Ressourcenverwaltung

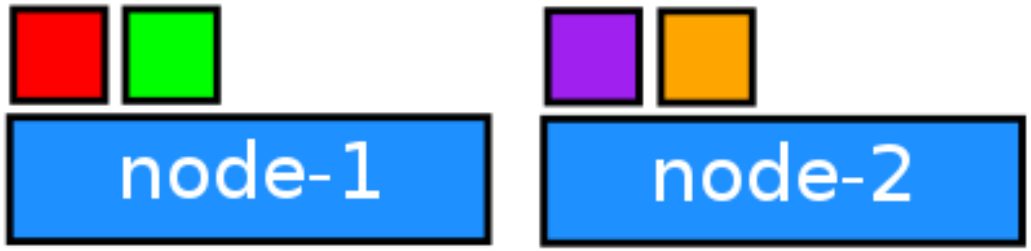
Instanz hinzufügen



- Anlegen einer neuen Instanz

Ganeti Cluster: Ressourcenverwaltung

Instanz hinzufügen

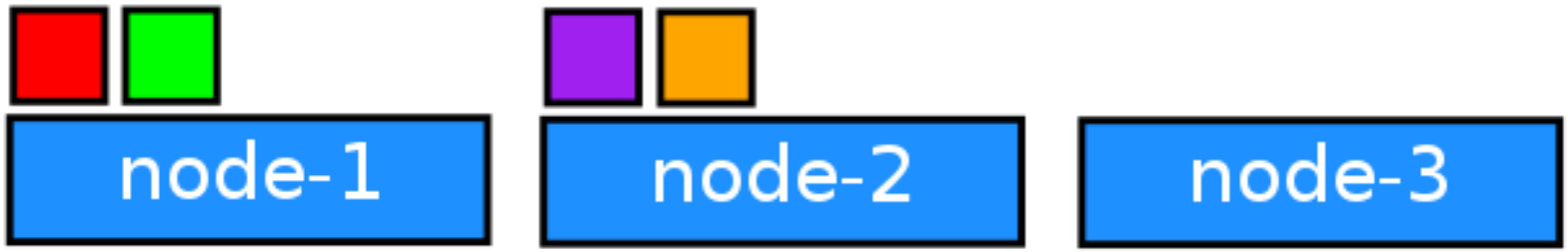


- Anlegen einer neuen Instanz
- wird automatisch auf einem passenden node platziert

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

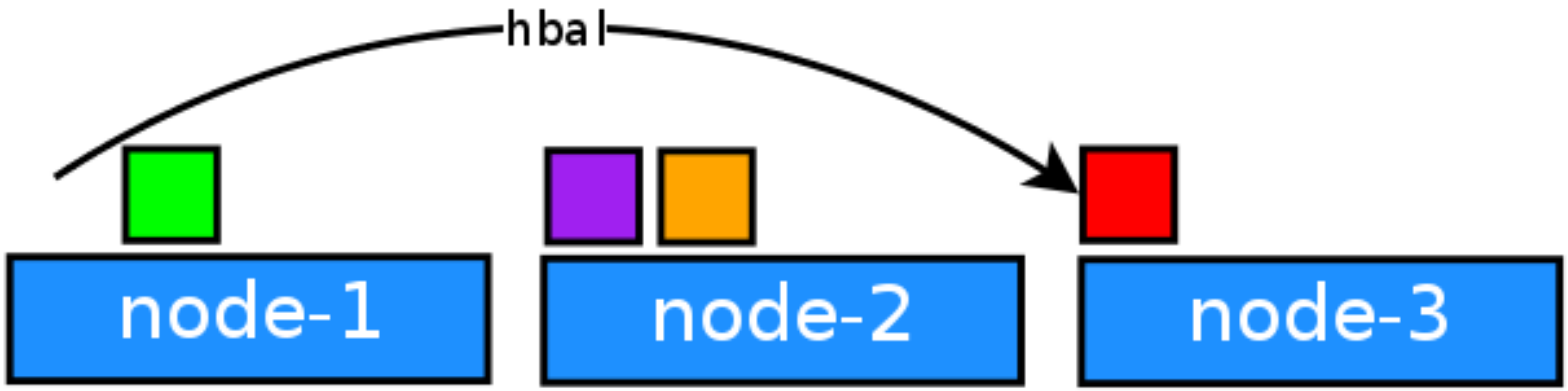
node hinzufügen



- Cluster um einen neuen node erweitern

Ganeti Cluster: Ressourcenverwaltung

node hinzufügen

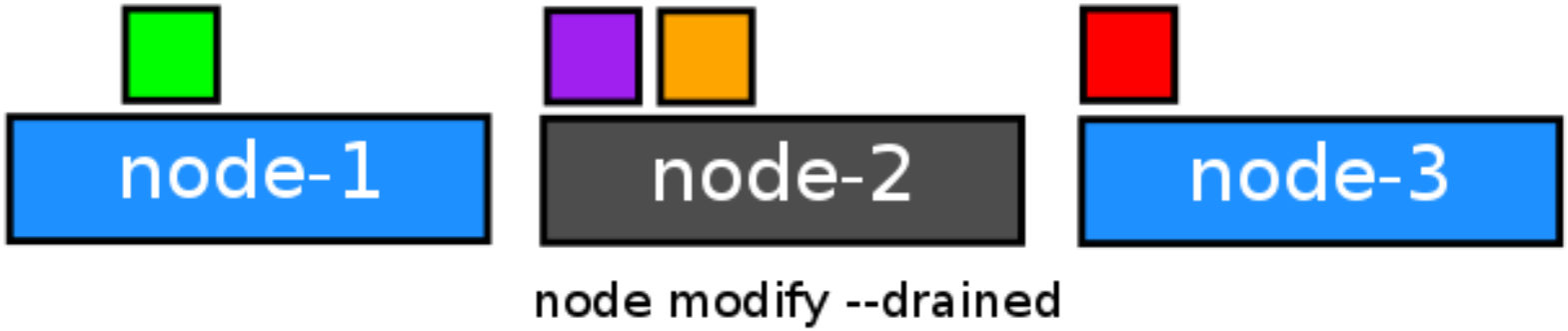


- Cluster um einen neuen node erweitern
- Ausbalancieren der Instanzen

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

node Wartung

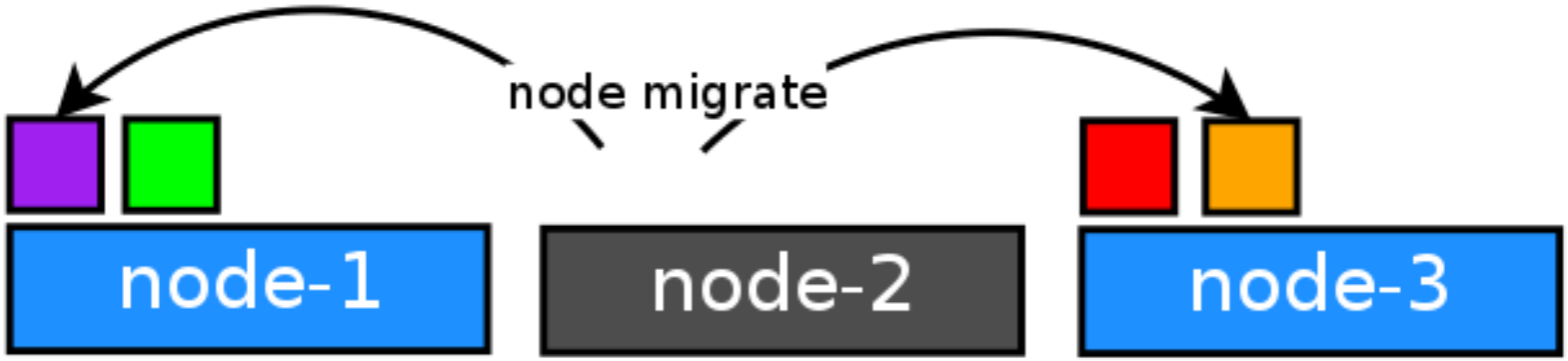


- node für neue Allokationen sperren (drain)

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

node Wartung

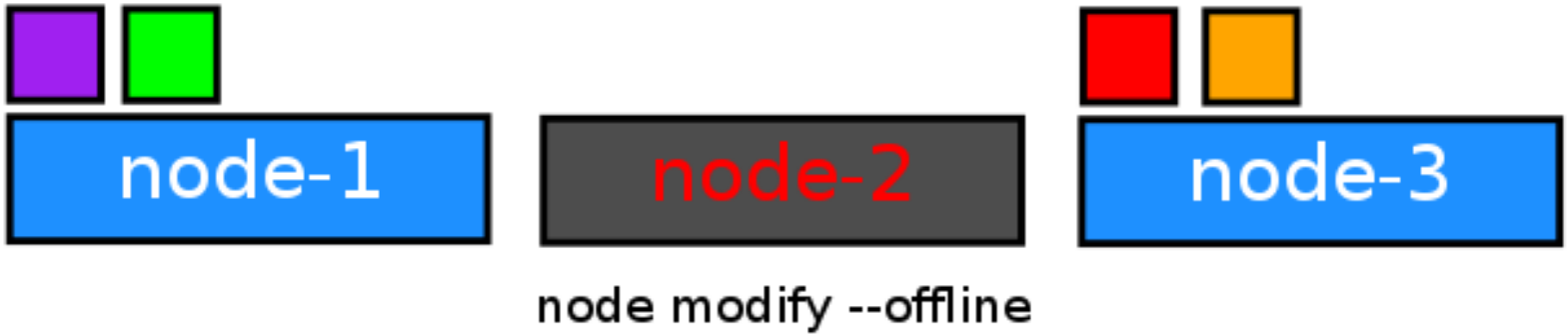


- node für neue Allokationen sperren (drain)
- alle Instanzen von node-2 migrieren

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

node Wartung



- node für neue Allokationen sperren (drain)
- alle Instanzen von node-2 migrieren
- node als offline markieren → Wartung durchführen

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

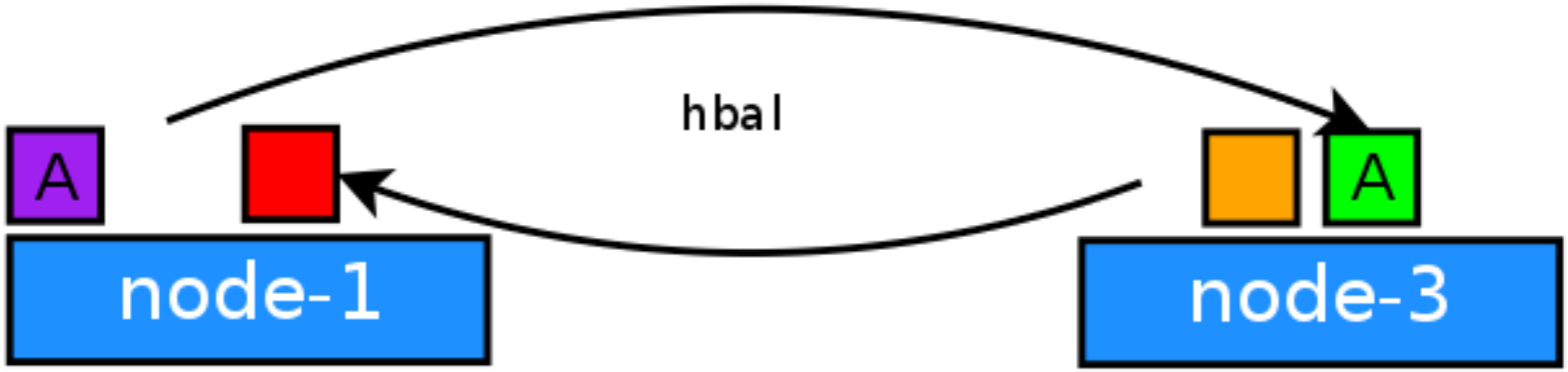
Instanz-Auschluss-Tags

- App A, z.B. 2 DNS-Server, sollen nicht auf gleichem node laufen
- Tags: Cluster-Ebene htools:iextags:app, Instanz app:A

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

Instanz-Auschluss-Tags



- App A, z.B. 2 DNS-Server, sollen nicht auf gleichem node laufen
- Tags: Cluster-Ebene htools:iextags:app, Instanz app:A
- Ausbalancieren

GISA® IT. Mehr als Standard.

Ganeti Cluster: Ressourcenverwaltung

Zusammenfassung

- N+1 Redundanz wird garantiert (1 node darf ausfallen)
- `hail`: neue Instanzen werden automatisch im Cluster verteilt
- `hbal`: Cluster kann ausbalanciert werden
  - node hinzufügen, failover, migrieren
  - Applikationen gleichen Typs auf Wunsch über mehrere nodes verteilt
- `hspace`: Kapazitätsplanung: wieviele Instanzen noch frei
- `hroller`: optimierte rolling node reboots
- `hsqueeze`: Cluster maximal verdichten (green IT)

GISA® IT. Mehr als Standard.

OS-Interface

Frage: Wie kommt das OS in die VM?

GISA®
IT. Mehr als Standard.

OS-Interface

Frage: Wie kommt das OS in die VM? Antwort: Mit dem Erstellen.

GISA®
IT. Mehr als Standard.

OS-Interface

Frage: Wie kommt das OS in die VM? Antwort: Mit dem Erstellen.

Ganeti OS interface

Funktionen ($\widehat{=}$ executables) auf node:

`create`, `export`, `import`, `rename`, `verify`

OS-Interface

Frage: Wie kommt das OS in die VM? Antwort: Mit dem Erstellen.

Ganeti OS interface

Funktionen ($\widehat{=}$ executables) auf node:

`create`, `export`, `import`, `rename`, `verify`

`create`

ENV	**Verwendungzweck**
`OS_VARIANT`	welches OS soll installiert werden
`DISK_%N_PATH`	auf welchen Platten
`INSTANCE_NAME`	Hostname der Instanz (FQDN)
`NIC_%N_X`	`X`=`IP`,`NETWORK_{SUBNET,GATEWAY}` $\rightarrow$ eth%N
`OSP_*`	benutzerdefinierte Parameter

Ermöglicht OS-Provisionierung mit Option zur Anpassung. Bsp.:

- VM ist mit Erstellung online (hook-Skript)

GISA® IT. Mehr als Standard.

OS-Interface

bekannte OS-Interfaces

noop	macht „nichts“ → Installation via ISO etc.
debootstrap	nutzt debootstrap (Debian, Ubuntu)

ganeti-os-noop und ganeti-instance-debootstrap sind in Debian enthalten

GISA® IT. Mehr als Standard.

OS-Interface

bekannte OS-Interfaces

noop macht „nichts“ → Installation via ISO etc.

debootstrap nutzt debootstrap (Debian, Ubuntu)

ganeti-os-noop und ganeti-instance-debootstrap sind in Debian enthalten

snf-image cloned Disk-Images, Anpassungen in Hilfs-VM

- Handhabung für `$most_linux_distros` und *BSD
- und sogar M$-windows (Vgl. sysprep)

GISA® IT. Mehr als Standard.

OS-Interface

bekannte OS-Interfaces

noop macht „nichts“ → Installation via ISO etc.

debootstrap nutzt debootstrap (Debian, Ubuntu)

ganeti-os-noop und ganeti-instance-debootstrap sind in Debian enthalten

snf-image cloned Disk-Images, Anpassungen in Hilfs-VM

- Handhabung für `$most_linux_distros` und *BSD
- und sogar M$-windows (Vgl. sysprep)

„your OSI“ selber machen

- z.B. mit < 100 Zeilen bash
- benutzt `dd` und `guestfish` für Anpassung SLES/Ubuntu (in VM isoliert)
- s. Vortrag Mini GanetiCon 2017

GISA® IT. Mehr als Standard.

Ganeti Cluster: Redundanz und Ressourcenverwaltung

Zusammenfassung

- geringe Einstiegshürde: 3(1) Server, lokaler Storage, keine Managemement-Systeme
- in Debian/(Ubuntu) enthalten $\rightarrow$ nur leichtgewichtige Abhängigkeiten
- skaliert problemlos: $\approx$ 1000 VMs pro Cluster
- VMs sind hoch verfügbar: DRBD-Spiegel, N+1 Redundanz
- Ressourcen werden automatisch verwaltet
- VM ist provisioniert und als Option online

GISA® IT. Mehr als Standard.

Gliederung

GISA® IT. Mehr als Standard.

Projekt-Status

Geschichte

- 16.07. 2007: Initial commit
- 14.02. 2009: Debian GNU/Linux 5.0 released (ganeti-1.2.6)
- um 2012 bis 2015 rege Aktivitäten seitens Google
  - Entwicklungs-Team in München
  - Stellenausschreibung Entwicklungs-Leiter
  - mehrere GSoC
  - externe Beiträge durch GRNET (synnefo cloud)

  häufige Releases, viele Neuerungen $\rightarrow$ Grundlagen für heutiges Ganeti

  Bis dahin: Google entwickelt, kleine Nutzer-Gemeinschaft

GISA® IT. Mehr als Standard.

Projekt-Status

Geschichte

- 16.07. 2007: Initial commit
- 14.02. 2009: Debian GNU/Linux 5.0 released (ganeti-1.2.6)
- um 2012 bis 2015 rege Aktivitäten seitens Google
  - Entwicklungs-Team in München
  - Stellenausschreibung Entwicklungs-Leiter
  - mehrere GSoC
  - externe Beiträge durch GRNET (synnefo cloud)

  häufige Releases, viele Neuerungen $\rightarrow$ Grundlagen für heutiges Ganeti
  Bis dahin: Google entwickelt, kleine Nutzer-Gemeinschaft
- 28.07. 2015: letztes feature release 2.15
- 16.12. 2015: letztes bugfix release 2.15.2

GISA® IT. Mehr als Standard.

Projekt-Status

Geschichte

- 16.07. 2007: Initial commit
- 14.02. 2009: Debian GNU/Linux 5.0 released (ganeti-1.2.6)
- um 2012 bis 2015 rege Aktivitäten seitens Google
  - Entwicklungs-Team in München
  - Stellenausschreibung Entwicklungs-Leiter
  - mehrere GSoC
  - externe Beiträge durch GRNET (synnefo cloud)

  häufige Releases, viele Neuerungen → Grundlagen für heutiges Ganeti
  Bis dahin: Google entwickelt, kleine Nutzer-Gemeinschaft
- 28.07. 2015: letztes feature release 2.15
- 16.12. 2015: letztes bugfix release 2.15.2
- was ist passiert?
  - in 2016 Auflösung Team München → zurück nach Dublin
  - weiterhin Bugfix- und Optimierungs-commits

GISA® IT. Mehr als Standard.

Projekt-Status

Geschichte

- 16.07. 2007: Initial commit
- 14.02. 2009: Debian GNU/Linux 5.0 released (ganeti-1.2.6)
- um 2012 bis 2015 rege Aktivitäten seitens Google
  - Entwicklungs-Team in München
  - Stellenausschreibung Entwicklungs-Leiter
  - mehrere GSoC
  - externe Beiträge durch GRNET (synnefo cloud)

  häufige Releases, viele Neuerungen → Grundlagen für heutiges Ganeti
  Bis dahin: Google entwickelt, kleine Nutzer-Gemeinschaft
- 28.07. 2015: letztes feature release 2.15
- 16.12. 2015: letztes bugfix release 2.15.2
- was ist passiert?
  - in 2016 Auflösung Team München → zurück nach Dublin
  - weiterhin Bugfix- und Optimierungs-commits
- 15.12. 2017: mini GanetiCon Leipzig → tuwat!

GISA® IT. Mehr als Standard.

Community

„Jeder fängt mal an“

auf mini GanetiCon 2017 wird klar:

- Beteiligung von Google beschränkt auf eigene Interessen
  → langfristig wird ein `fork` notwendig sein
- Aufbau einer eigenen Community

GISA® IT. Mehr als Standard.

Community

„Jeder fängt mal an“

auf mini GanetiCon 2017 wird klar:

- Beteiligung von Google beschränkt auf eigene Interessen
  $\rightarrow$ langfristig wird ein `fork` notwendig sein
- Aufbau einer eigenen Community
- ad-hoc Maßnahmen
  - Erstellung des release 2.16 (Google)
  - Sicherung des Debian-Paketes für 2.16 (Debian Maintainer)
  - Verbesserung Paket-Qualität Ubuntu/LTS (Community)

GISA® IT. Mehr als Standard.

zukünftige Ausrichtung

langfristig

Festhalten an positiven Eigenschaften

- selfcontained: keine Management-Systeme, keine weiteren SW-Stacks
- fester Bestandteil der Debian/(Ubuntu) Distribution

GISA®
IT. Mehr als Standard.

zukünftige Ausrichtung

langfristig

Festhalten an positiven Eigenschaften

- selfcontained: keine Management-Systeme, keine weiteren SW-Stacks
- fester Bestandteil der Debian/(Ubuntu) Distribution

Was muss/soll/kann sich ändern?

- PEP 373: python-2.7 EOL 2020 → Migration zu Python 3 notwendig
- etwas beizutragen soll „einfach“ sein
  - Python ist attraktiv, Haskell weniger → Migration zu Python?
  - allg. Code-Basis mit nur einer Sprache wünschenswert
- Bekanntheit und Attraktivität steigern
  - Vorträge wie CLT
  - bessere „defaults“ (kvm:security_model, drbd:dynamic-resync)
- Feature: DRBD-9?

GISA® IT. Mehr als Standard.

Kontaktdaten

home	http://www.ganeti.org
doc	http://docs.ganeti.org
wiki	https://ganeti.googlesource.com/wiki/+/master
code	https://github.com/ganeti/ganeti
mail	https://groups.google.com/forum/#!forum/ganeti
dev-mail	https://groups.google.com/forum/#!forum/ganeti-devel
IRC	#ganeti on Freenode

Einsteiger: nutzt Debian, oder fragt auf der ML nach dem Ubuntu Status

GISA® IT. Mehr als Standard.

Gliederung

GISA® IT. Mehr als Standard.

Ubuntu Paket Status

allgemein

- ganeti und ganeti-instance-debootstrap sind in universe-Sektion
- universe-Sektion wird durch community (MOTU) betreut
- Ubuntu merged/synced mit Debian unstable/sid bis FeatureFreeze
  - bionic/18.04: 1. März

GISA®
IT. Mehr als Standard.

Ubuntu Paket Status

allgemein

- ganeti und ganeti-instance-debootstrap sind in universe-Sektion
- universe-Sektion wird durch community (MOTU) betreut
- Ubuntu merged/synced mit Debian unstable/sid bis FeatureFreeze
  - bionic/18.04: 1. März

Ubuntu universe-Sektion

- „zufälliger“ Schnappschuss von Debian unstable/sid
- ohne paket-spezifische Maintainer

→ Debian Maintainer muss Ubuntu release-Zyklen beachten

GISA® IT. Mehr als Standard.

Ubuntu Paket Status

xenial/16.04

- ganeti-2.15.2-3: schlägt fehl wg. openssh DSA-Key
  - RSA-Key Unterstützung im debain Paket ganeti-2.15.2-7 enthalten
- ganeti-instance-debootstrap-0.15-2: schlägt fehl wg. `sfdisk` Parameteränderung (LP #1577346)
  - mit v0.16 gefixed → in Debian stretch

triviale Fehler, stören Einsteiger-Wahrnehmung → Fix sollte möglich sein

GISA® IT. Mehr als Standard.

Ubuntu Paket Status

autopkgtests

ganeti

	xenial	yakkety	zesty	artful	bionic
amd64	✖ fail	✖ fail	✖ fail	✖ fail	✔ pass
arm64	✖ fail		✖ fail	😖 error	✔ pass
armhf	✖ fail	✖ fail	✖ fail	✔ pass	✖ fail
i386	✔ pass	✔ pass	✖ fail	✔ pass	✖ fail
ppc64el	✖ fail	✖ fail	✖ fail	✖ fail	✔ pass
s390x	✖ fail	✖ fail	✖ fail	✖ fail	✖ fail

Ubuntu autopkgtests = Debian CI

- artful/17.10 build-Fehler arm64
  - Überschneidung FeatureFreeze + Debian build queue
- bionic/18.04 autopkgtest-VM zu klein: 4 Anläufe IRC #ubuntu-devel
  - 2.16rc2, ↯ i386, armhf, s390x → nur in „proposed“
    Fix ebenfalls möglich

GISA® IT. Mehr als Standard.

Ende

Fazit

- Ganeti sorgt für entspannten RZ-Betrieb → Weiterführung lohnt sich
- Unterstützung seitens Google, Debian und grnet vorhanden
- Ubuntu Probleme erkannt und Lösung denkbar (Ansgar, ich)
  - PPA: https://launchpad.net/~ansgarj/+archive/ubuntu/ganeti

neue Nutzer und Beiträge sind herzlich willkommen

GISA® IT. Mehr als Standard.

Ende

Fazit

- Ganeti sorgt für entspannten RZ-Betrieb → Weiterführung lohnt sich
- Unterstützung seitens Google, Debian und grnet vorhanden
- Ubuntu Probleme erkannt und Lösung denkbar (Ansgar, ich)
  - PPA: https://launchpad.net/~ansgarj/+archive/ubuntu/ganeti

neue Nutzer und Beiträge sind herzlich willkommen

DANKE

Fragen?

GISA® IT. Mehr als Standard.