Ceph Distributed Storage for the Cloud

An update of enterprise use-cases at BMW

Andreas Pöschl, BMW
Senior Solutions Architect
andreas.poeschl@bmw.de

Michael Vonderbecke, BMW
Solutions Architect
michael.vonderbecke@bmwmc.com

The environment at BMW

Where we come from

- Huge Linux experience and always driving innovation
  - Using Linux since 2002
  - Migrated from RISC servers to x86 (mostly Linux) until 2010
  - Virtualizing production workloads on SLES®/XEN since 2008
    - Using Fibre Channel SAN storage and host based mirroring (HBM)
    - Unique combination of HBM and live migration ability (block-dmmd)
    - Management via homegrown scripting framework
  - Approx. 4500 Linux instances in place
    - 1300 physical servers, 350 virtualization hosts, 2800 VMs
    - Also virtualized most business critical workloads (production control systems, SAP, databases (PostgreSQL/Oracle)

The environment at BMW

Where we come from (2)

- Huge variety of compute workloads
  - VMs from 1 vCPU / 1GB to 16 vCPU / 64 GB
  - Physical servers up to 60 cores (+HT) and 1 TB memory
- SAN storage on enterprise arrays, connected via 8/16GBit fibre channel SAN (approx. 4,5 PB)
  - Only mirrored setups in separate datacenters
  - Separate LUNs for application storage and VM instances
- NAS storage (NFS) for parallel access (especially in the web server environment (approx. 5,7 PB)

The environment at BMW

Storage issues …

- Currently used (enterprise) storage is safe, fast and worth it’s price, but lacks agility on migrations and changing environment
- Currently we are trying to gain more flexibility by using NPIV with FC based SAN
- Migrations and necessary reconfigurations are still an issue in an FC based SAN environment, even when using automation
- NFS is widely used and might resolve some of the issues, but is not a cloud storage system (from our perspective)

The environment at BMW

… how to resolve them

We are looking for:

- An IP based storage architecture,
- Which keeps us independent from storage vendors (and underlying technology at all),
- Is disaster safe,
- Has no single point of failure
- And is “native” to us (integration into Linux distribution)

Ceph/Rados Comes Into Play

Our expectations

How we plan to benefit from Ceph/Rados

- Increased flexibility by using Ethernet instead of Fibre Channel (easy routing in case of changes)
- No necessity for resource intensive migration tasks in the future as Ceph handles the (re-)placement
- Possibilities like thin provisioning, cloning and snapshotting without vendor specific storage extensions
- Future proof concept (developed for the cloud) and ready also for future use cases

Our goals – 2 years ago

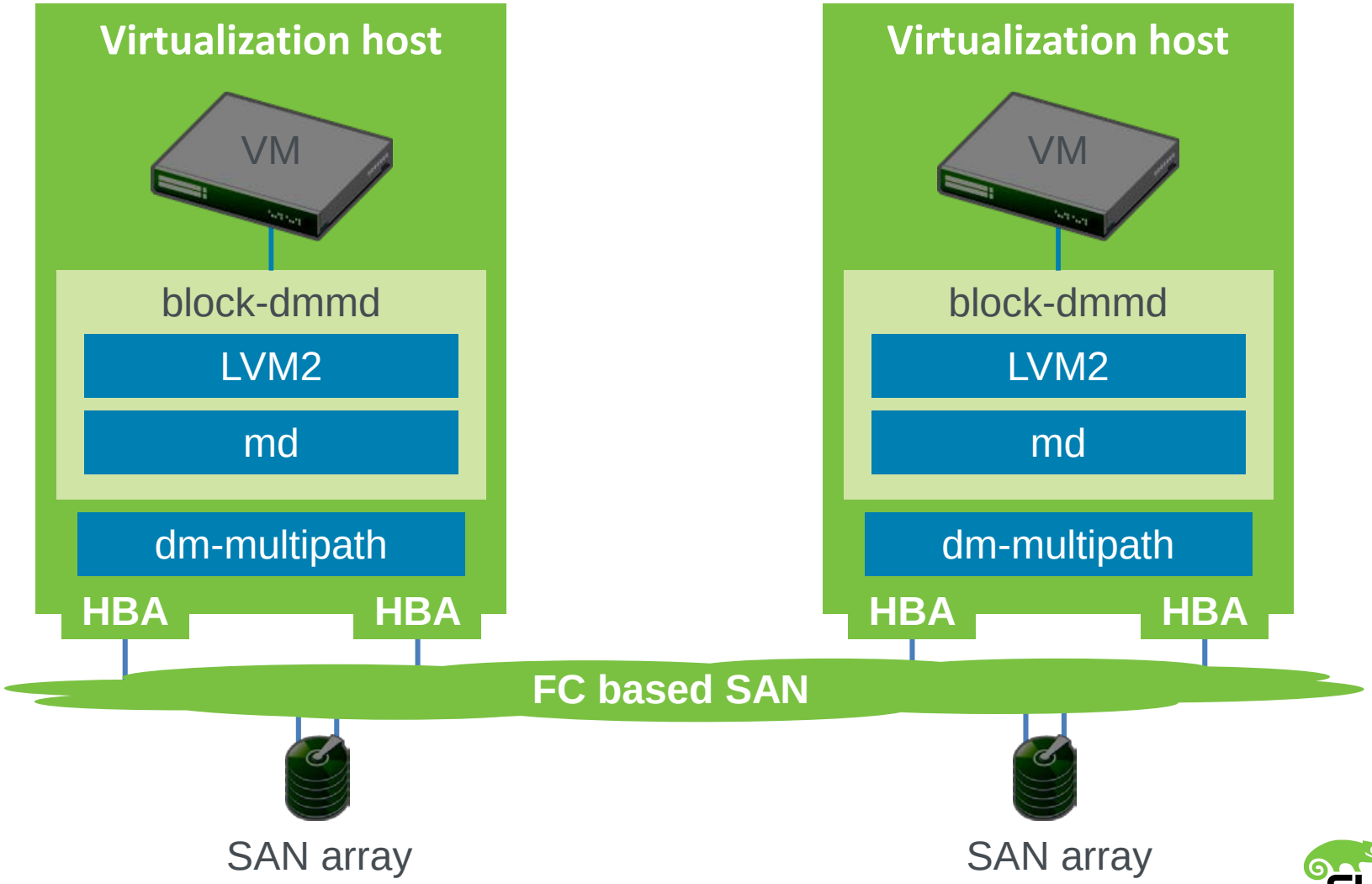
How we planned to use Ceph/Rados

- We wanted to limit the scope of Ceph/Rados for the first implementation (just use the Rados block layer (rbd) to store images for virtual servers)
- No CephFS or NFS replacement intended (yet)
- Leverage the possibilities only, where you can live with the drawbacks (in terms of maturity and admin experience)

➔ Try to replace at least a part of the SAN storage used for VMs with low criticality and load

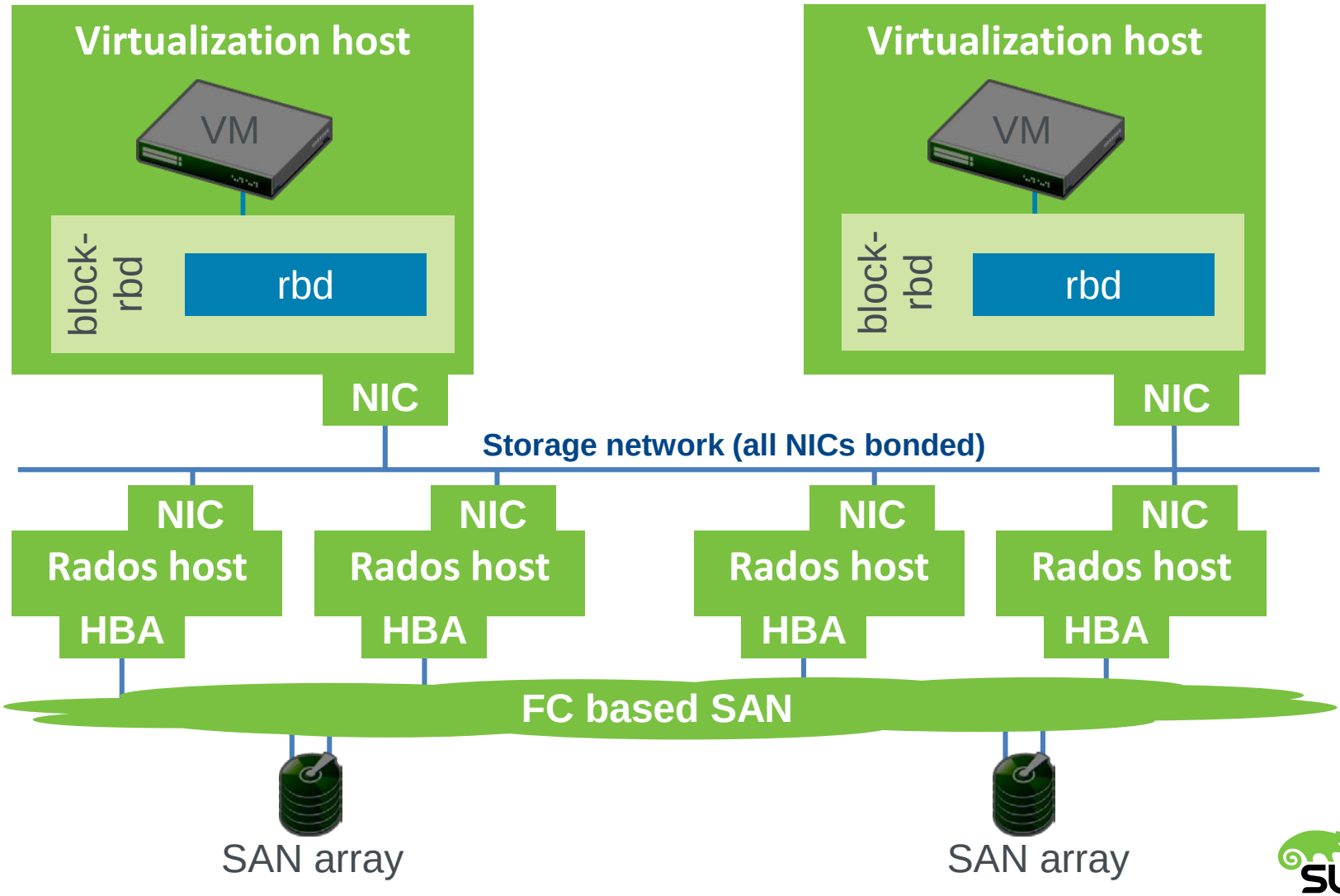
Setup

What we have at the moment



Setup

First idea for the future (legacy based)



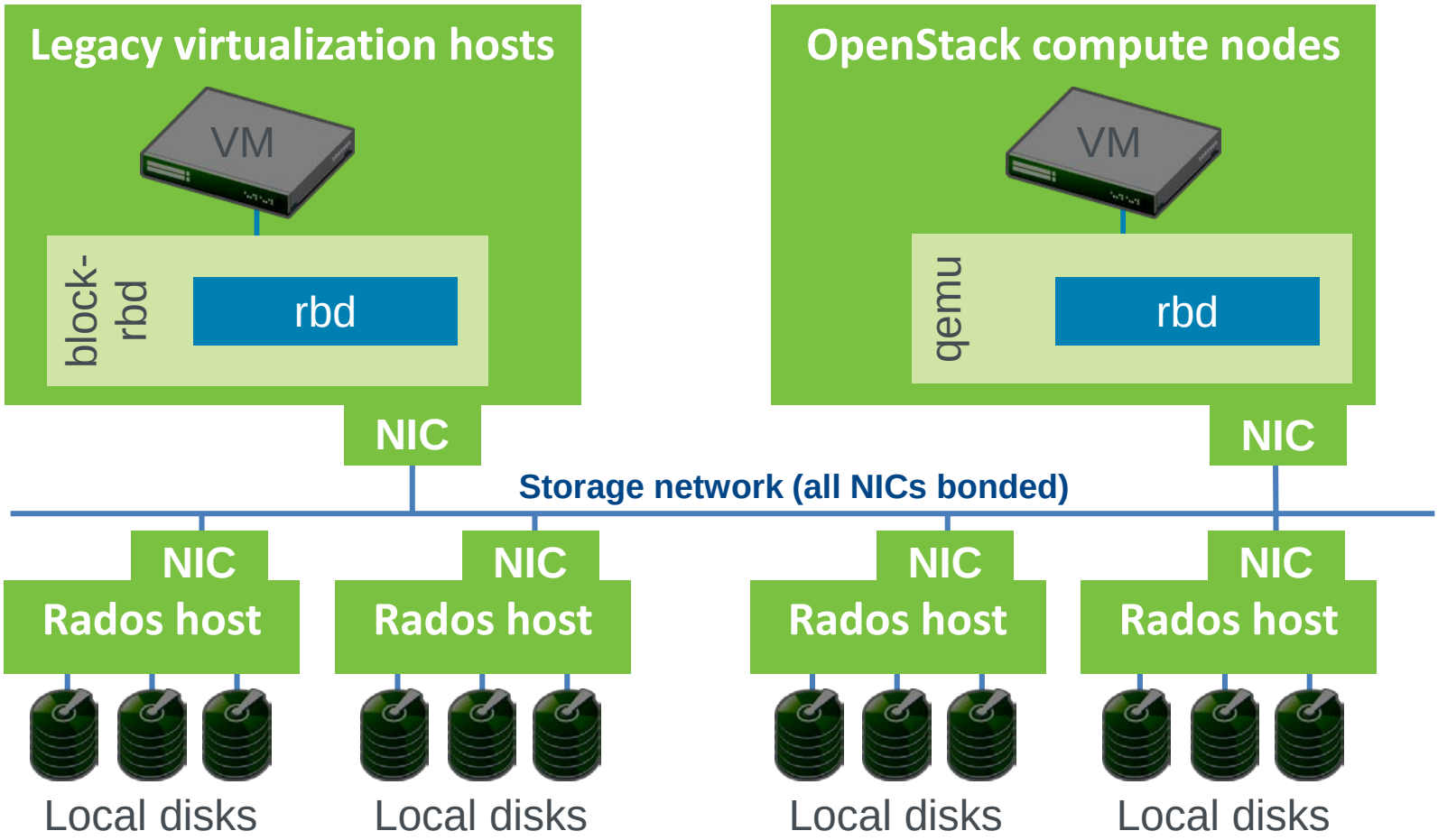
Our goals – now

How we plan to use Ceph/Rados

- Ceph is still a hot option – but is it sufficient to just to introduce a new storage?
- We have to rethink not only the storage type, but also the storage usage!
- If a workload is suitable for Ceph/Rados, it might as well be suitable for further evolution (-> Cloud operations)
- Use Ceph/Rados for non-critical workloads (performance and business impact) in the legacy environment (to learn) and completely bet on it within OpenStack

Setup

Plan for the future (cloud based)



Results

What we found out – the technical part

- Installation via SUSE® Cloud 4 was pretty easy, but up to now, you cannot configure it very well via crowbar
- Automatic distribution of data over failure domains (on creation of storage and on addition and removal of storage nodes) works well
- If any node fails, storage access continues after some seconds of safety period
- First performance tests showed about 45 MB/s (read) and 30 MB/s (write) on a 1G network

Results

What we found out – the non-technical part

Prepare yourself for a different mindset

- You lose control in terms of knowing the exact location and distribution of your data – Ceph/Rados takes care of that
- You need to trust the system, that it manages the data distribution correctly
- There are several (new) ways to access your data with Ceph/Rados – chose wisely
- Even if Ceph/Rados takes care of the technical distribution, put effort into a wise definition of the usage concept (pools, placement groups, crush map, …)

Conclusion

How to continue

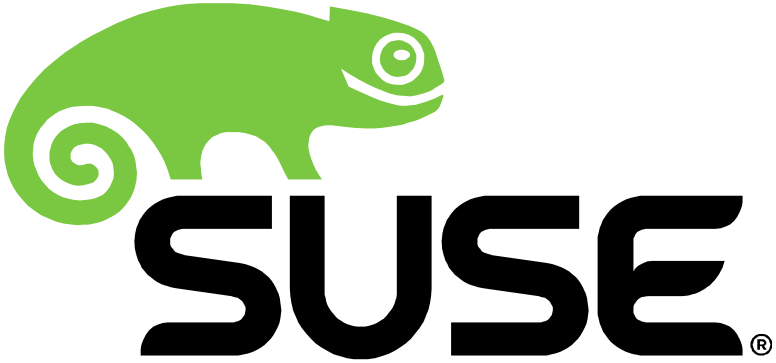
- We are currently building a new lab setup with 8 storage nodes, 15TB each and 10Gbit network
- We currently prepare a production setup in the datacenter to serve the OpenStack environment
- Provide the Ceph based object storage to applications and help them to change their way of storing data
- Have an eye on the CephFS, as it might be a successor for NFS in specific areas

➔ Try it, use it, challenge it.

Questions?

Annotations?

Thank you.


SUSE®
We adapt. You succeed.