Using Ganeti for running highly available virtual services

2016-04-21, HEPiX Spring 2016, Zeuthen

NordForsk

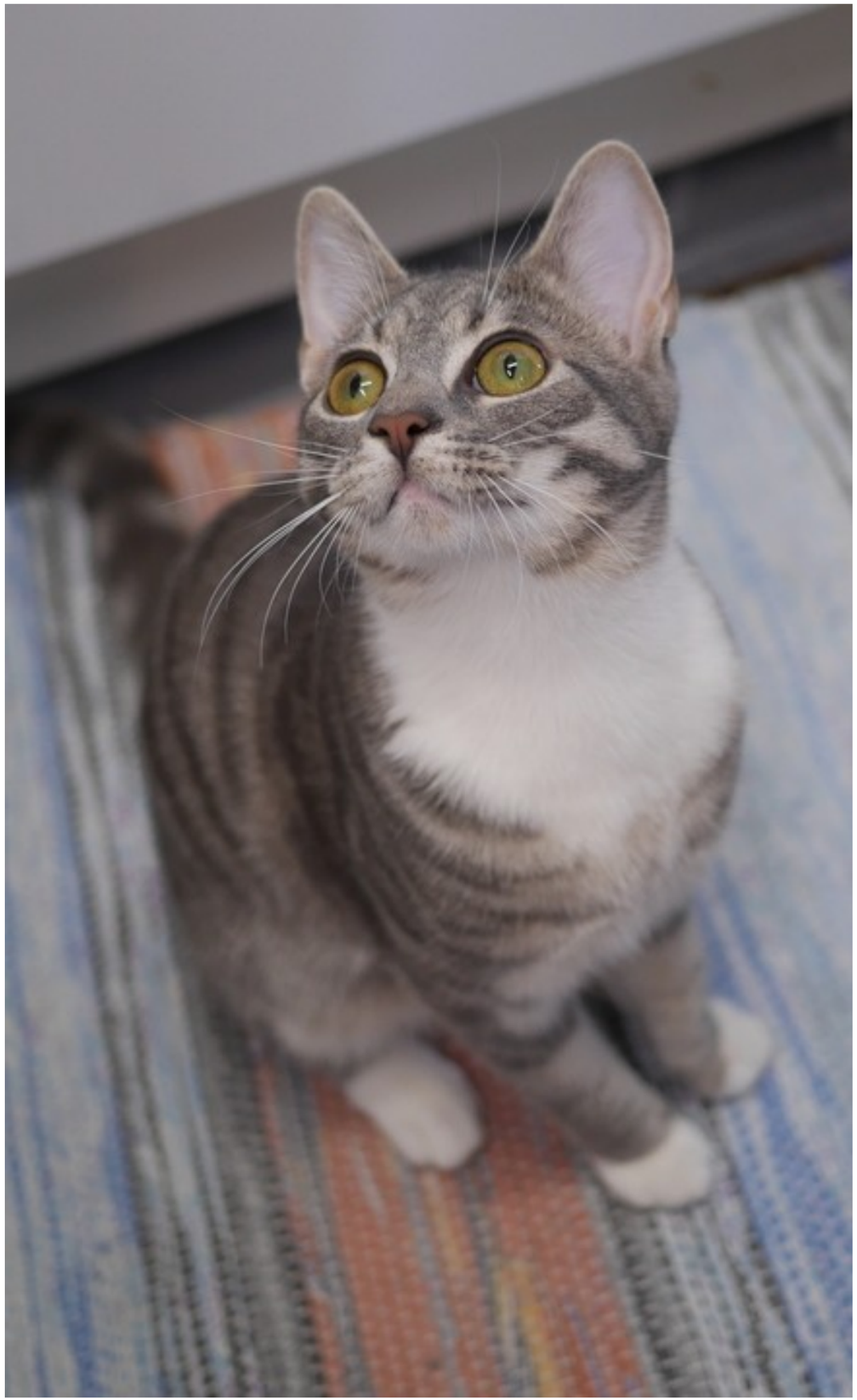
Nordic e-Infrastructure Collaboration

NordForsk

Nordic e-Infrastructure Collaboration

Overview

- What is Ganeti
- What is it good for
- How does it work
- NDGF usage



NordForsk

Nordic e-Infrastructure Collaboration

What is Ganeti

- A software stack for managing virtual machines
  - Like VMware or OpenStack or libvirt or ...
  - Supporting Xen or KVM hypervisors
  - Handles
    - Storage: volume creation and assignment
    - OS installation and customization
    - Networking
    - Startup, shutdown, live migration, failover of instances
  - Written in Python and Haskell
  - Aimed for ease of use and fast and simple error recovery after physical failures on commodity hardware

NordForsk

Nordic e-Infrastructure Collaboration

What is Ganeti

- Mainly developed by Google for their own use
  - Handles VMs for corporate network (office servers, remote desktops etc), not production services (what non-employees see)
- Outside Google
  - Debian
  - NDGF-T1
  - Lufthansa
  - Etc
- Maintained by Google with significant external contributions

NordForsk

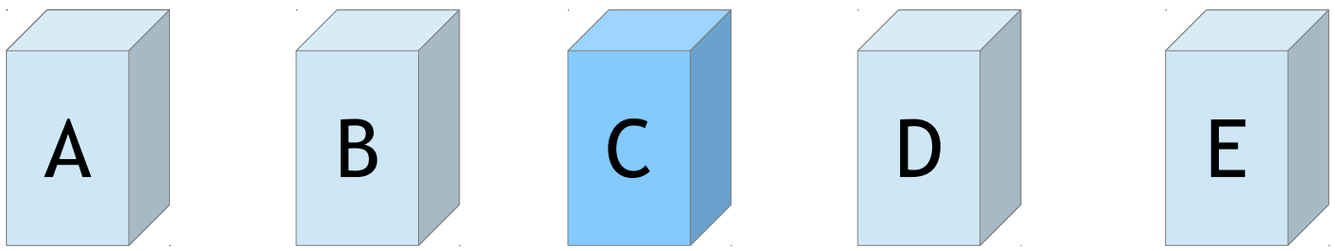
Nordic e-Infrastructure Collaboration

What is Ganeti good at

- Running highly available services on a small set of hardware
  - DRBD or external reliable block devices (CEPH, Enterprise storage)
  - Live migrations in case of impending hardware failure
    - Or reboot into new kernel security upgrade on the hardnode
  - Failover handled automatically in case of sudden hardware failure
  - No external dependencies beyond networking
    - Well, if you use external storage...
    - But no extra servers or services needed
  - Typical reasonable cluster size, 3 – 50 hardnodes
    - Multiple clusters integrate well though in admin tools

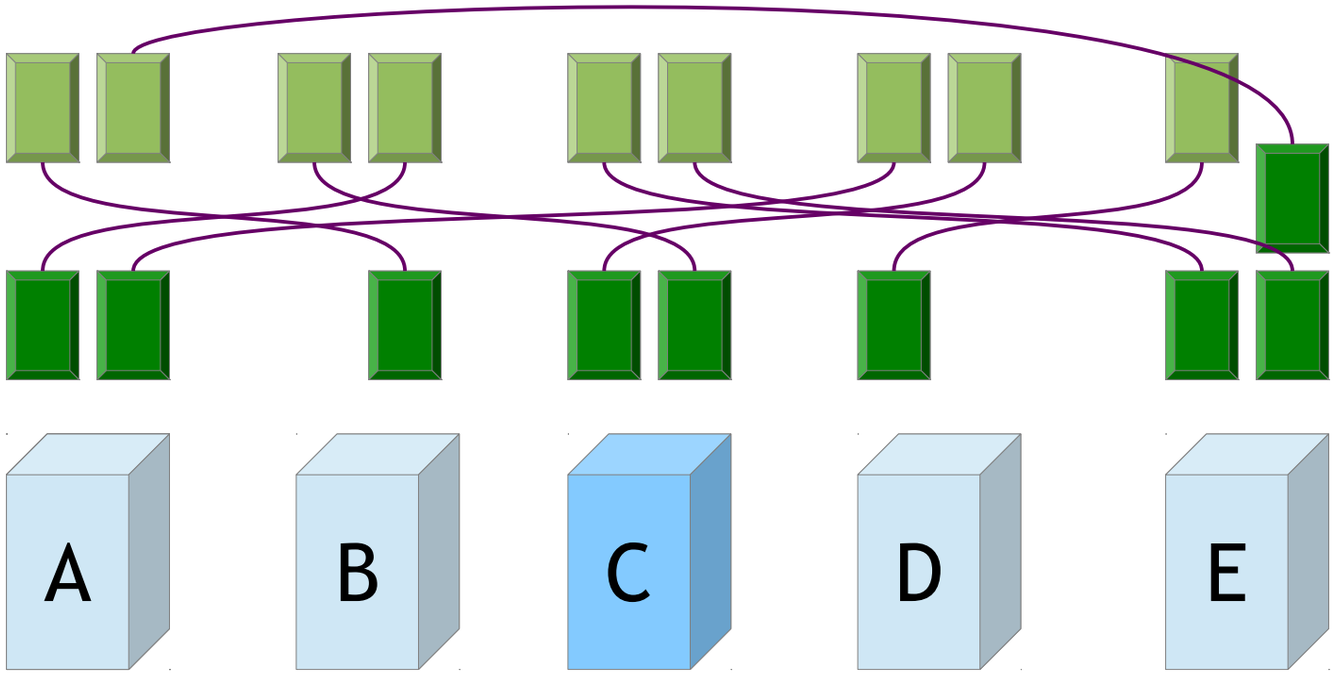
How does Ganeti work

- gnt-cluster init ...
  - Creates a cluster of ganeti nodes
  - We'll assume DRBD for storage, as at NDGF
  - One member is a master node
    - Others can take over with master-failover if they can get quorum



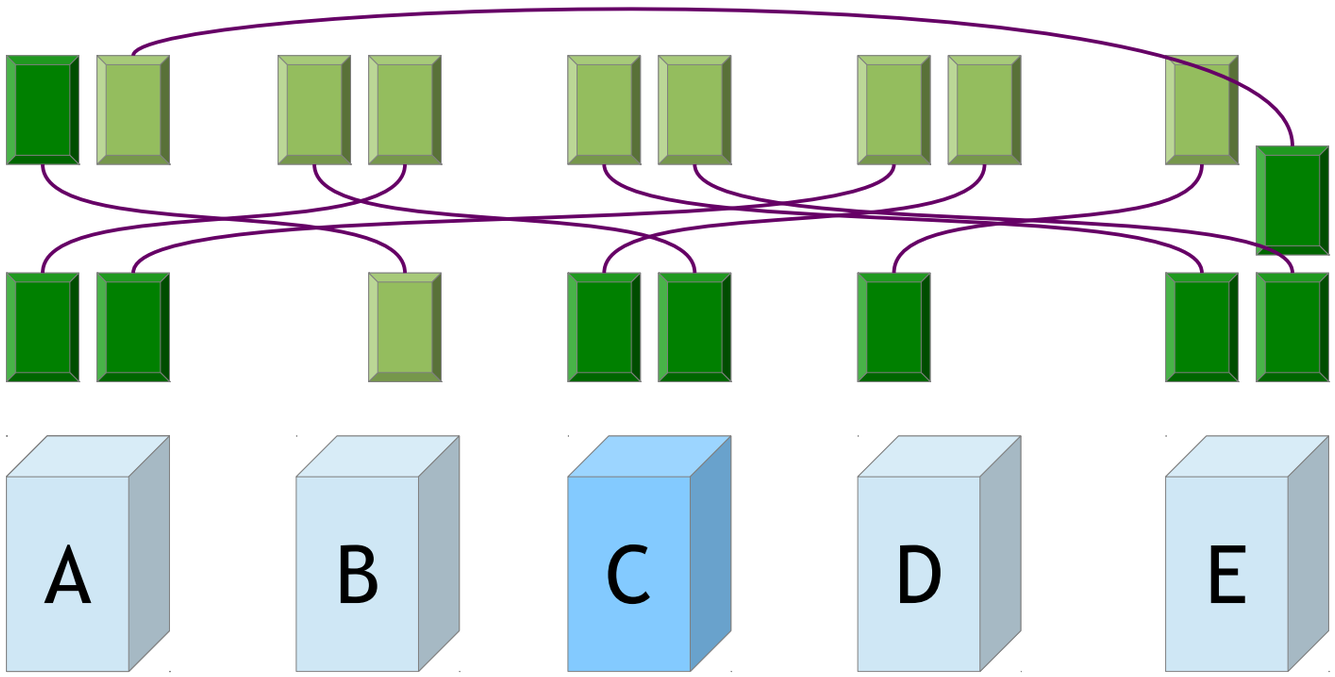
How does Ganeti work

- gnt-instance add
  - Creates VMs, with OS install scripts (image, debootstrap, pxe)
  - Each VM has a secondary location (DRBD mirror, sync)



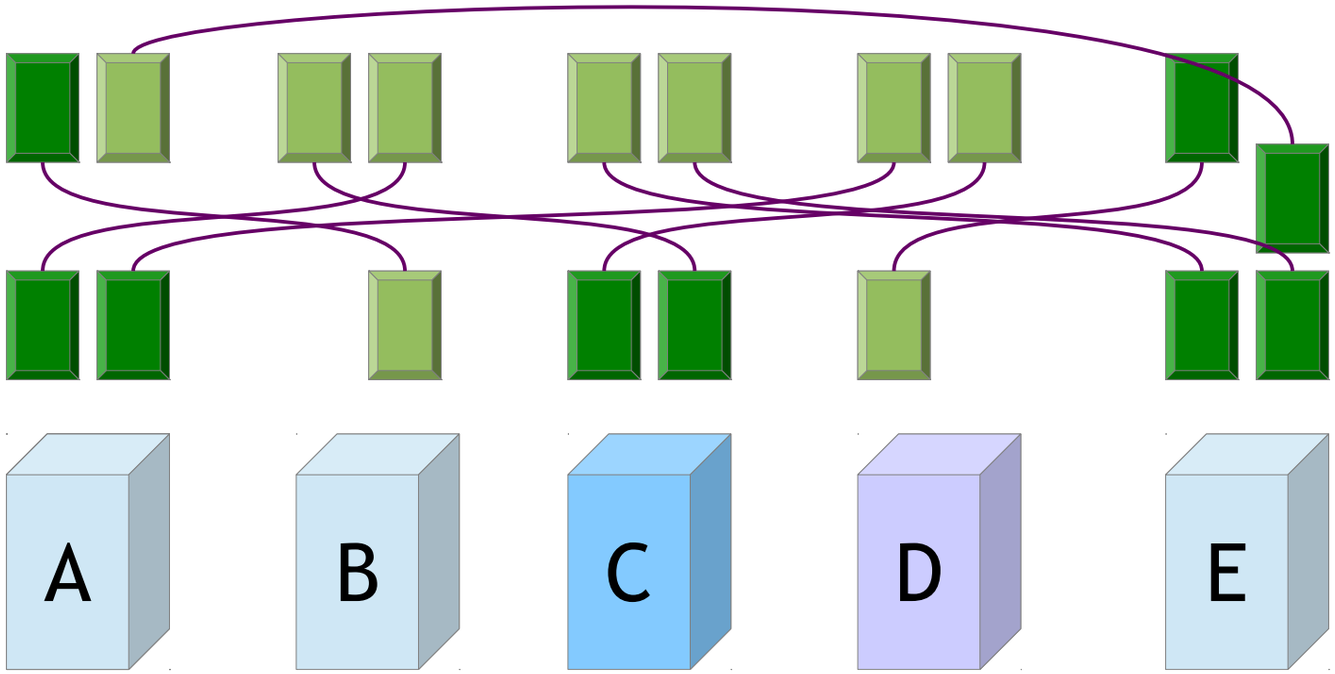
How does Ganeti work

- gnt-instance migrate
  - No noticable service impact from live migration, <1s network pause
    - Unless something is broken...



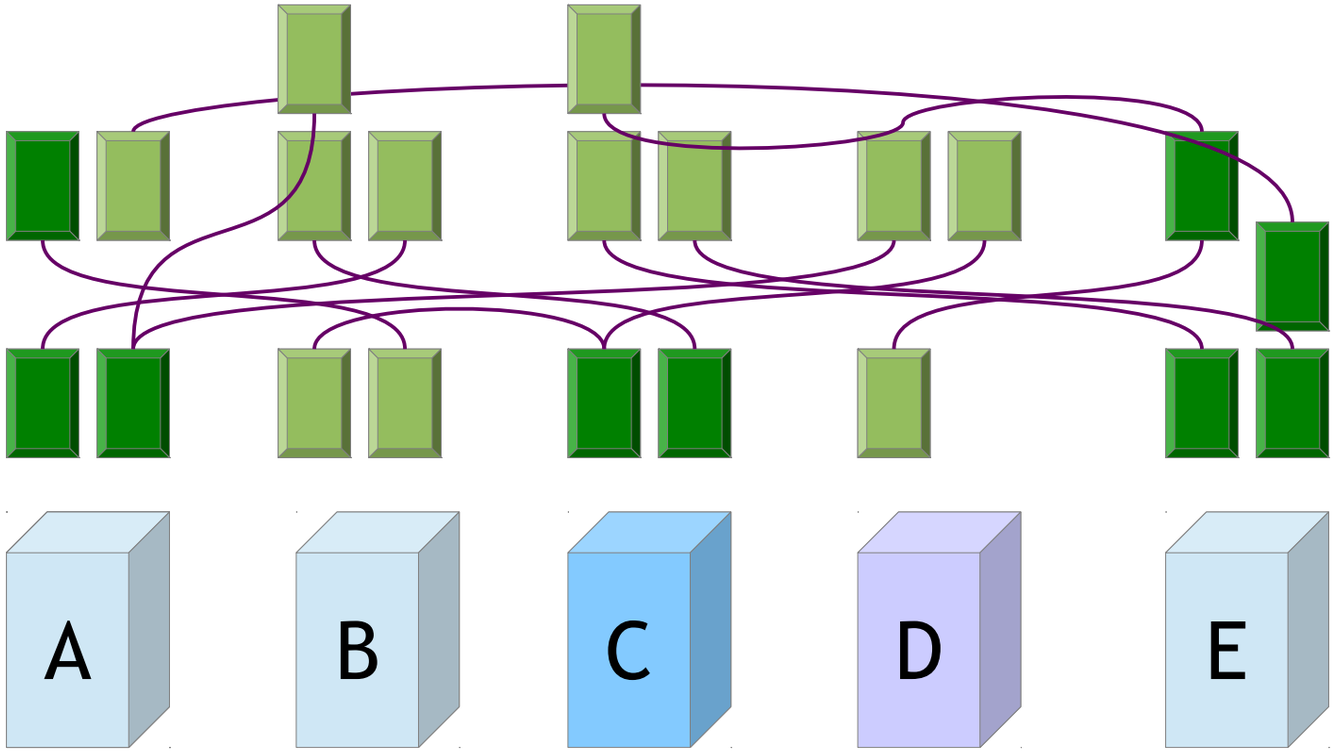
How does Ganeti work

- Full evacuation of a node
  - Removing a node for longer
  - Or wanting full redundancy all the time



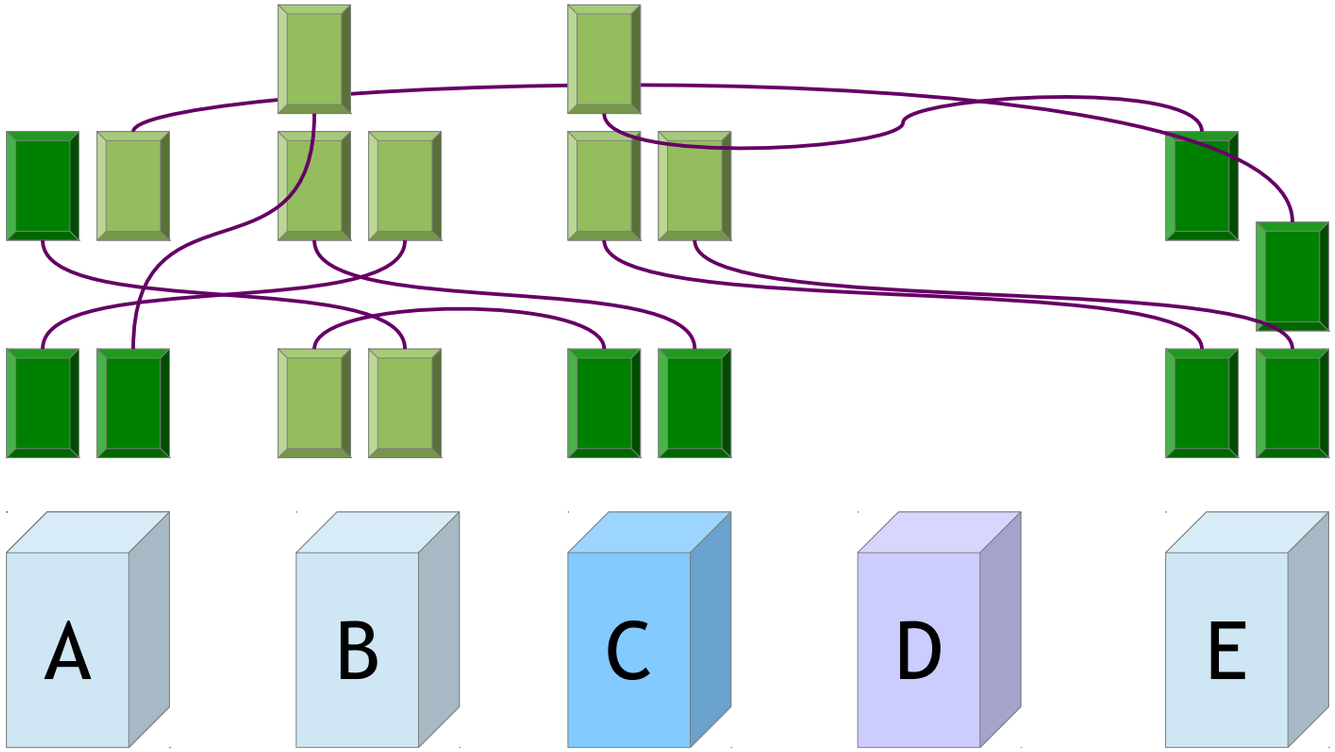
How does Ganeti work

- gnt-node evacuate to move secondary instances
  - Removing a node for longer
  - Or wanting full redundancy all the time



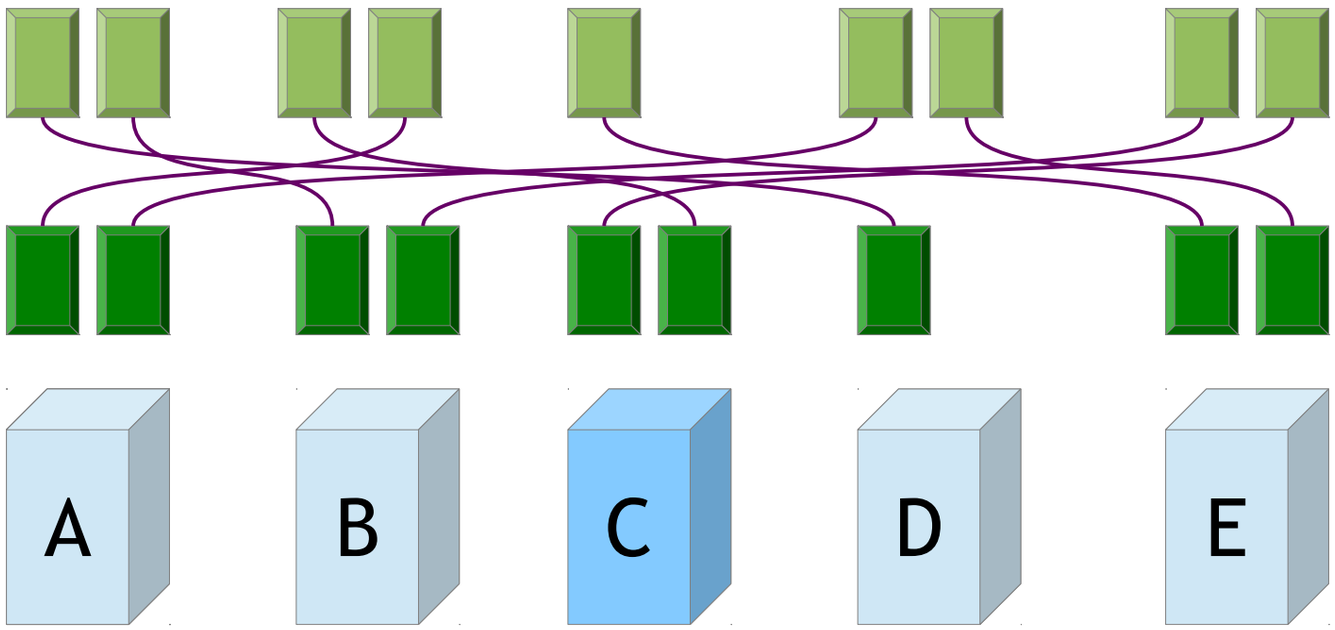
How does Ganeti work

- gnt-node evacuate to move secondary instances
  - Removing a node for longer
  - Or wanting full redundancy all the time



How does Ganeti work

- hbal -L -X
  - Rebalance to minimize a cost function based on uneven distribution



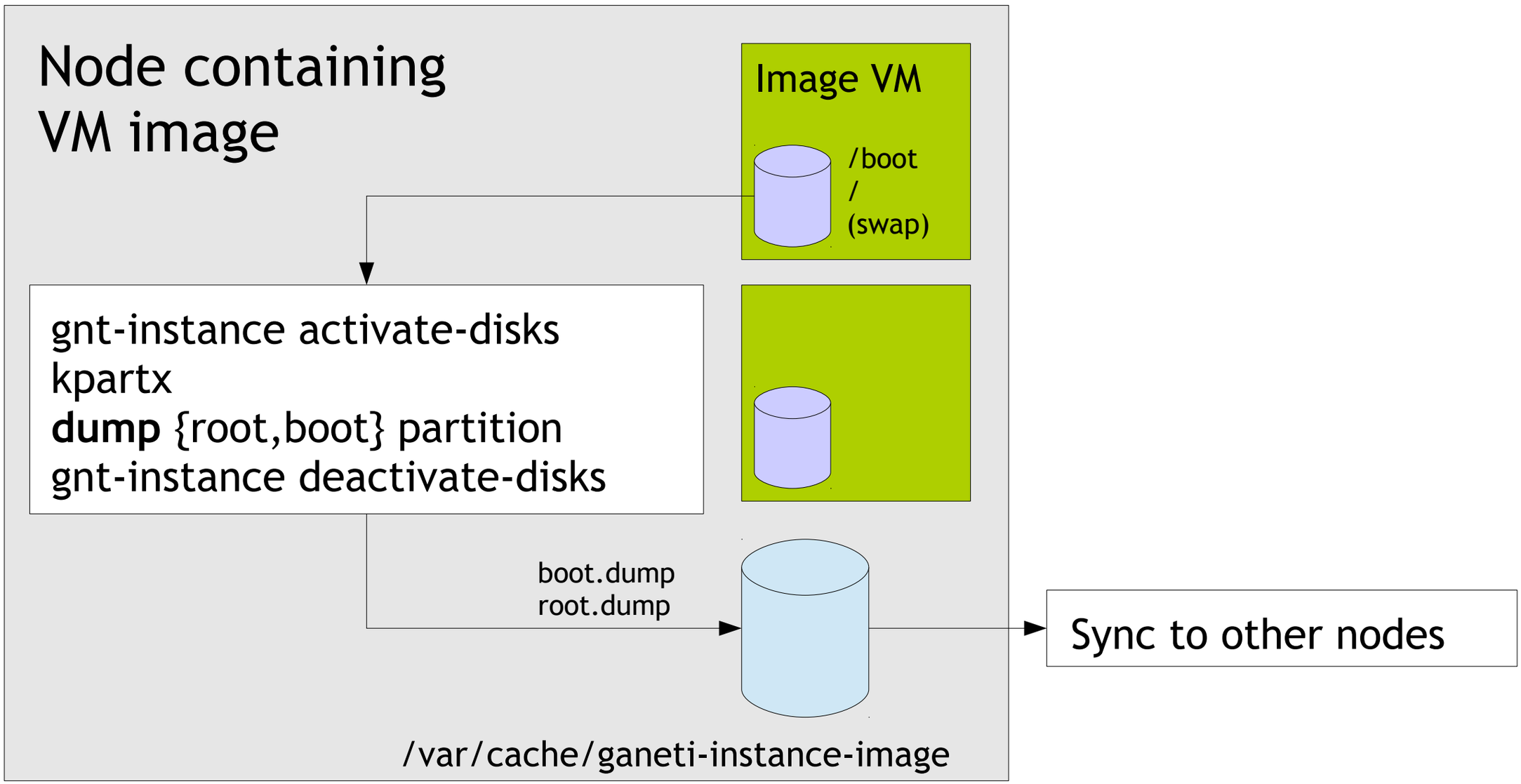
Instance Creation

- Ganeti sets up the VM with specified hardware.
- An OS Definition describes how to install the VM:
  - Image based:
    - ganeti-instance-debootstrap
    - ganeti-instance-image (← our option, next slides)
    - ganeti-os-defs
    - snf-image
  - Automated installation (PXE booting, i.e. FAI, kickstart).
  - Manual using CD medium/image and VNC or serial console, e.g. for initial image creation.

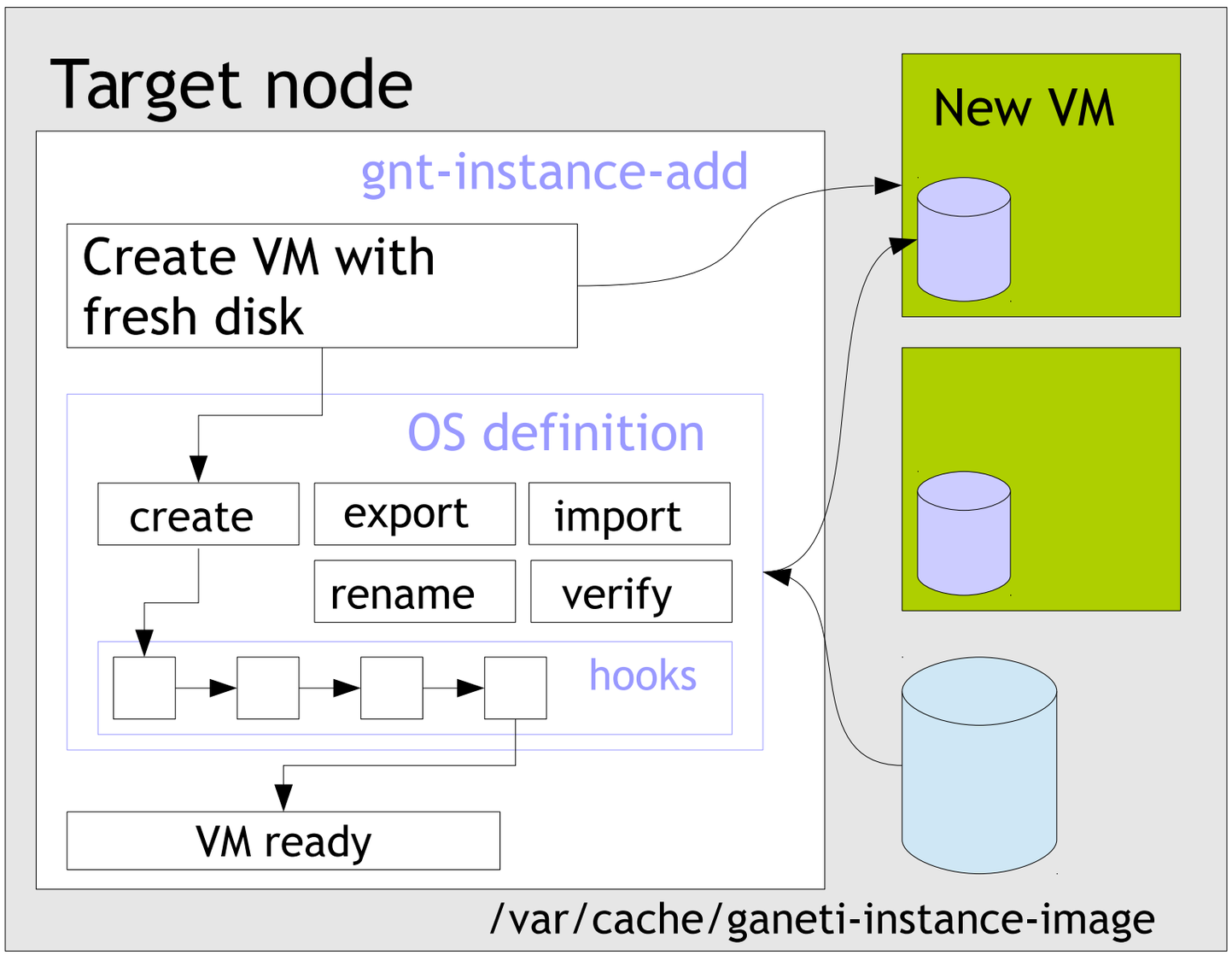
NordForsk

Nordic e-Infrastructure Collaboration

Instance Creation



Instance Creation



High availability

- Ganeti is geared towards accepting loss of any hardware compontent
  - If you have common points of failure (racks, switches, etc), tagging node groups so that primary and secondary are put separately
  - A watcher component will automatically reconnect storage (after a reboot), failover VMs if primary node goes down, etc
  - gnt-cluster verify checks that everything is OK, including N+1 for resources
- Separate internal fast network for replication and migration strongly encouraged

NordForsk

Nordic e-Infrastructure Collaboration

How does Ganeti work

- Networking
  - Bridging, easy default and gives VMs proper IPs
  - Host routing
  - Arbitrarily advanced SDN plugins
- Other storage
  - Native CEPH RBD support
  - Enterprise storage via plugins
    - Volume creation, exporting to the right nodes, release, etc
- Very featureful APIs
  - For self-service portals, provisioning frameworks, etc

NordForsk

Nordic e-Infrastructure Collaboration

NDGF Ganeti use

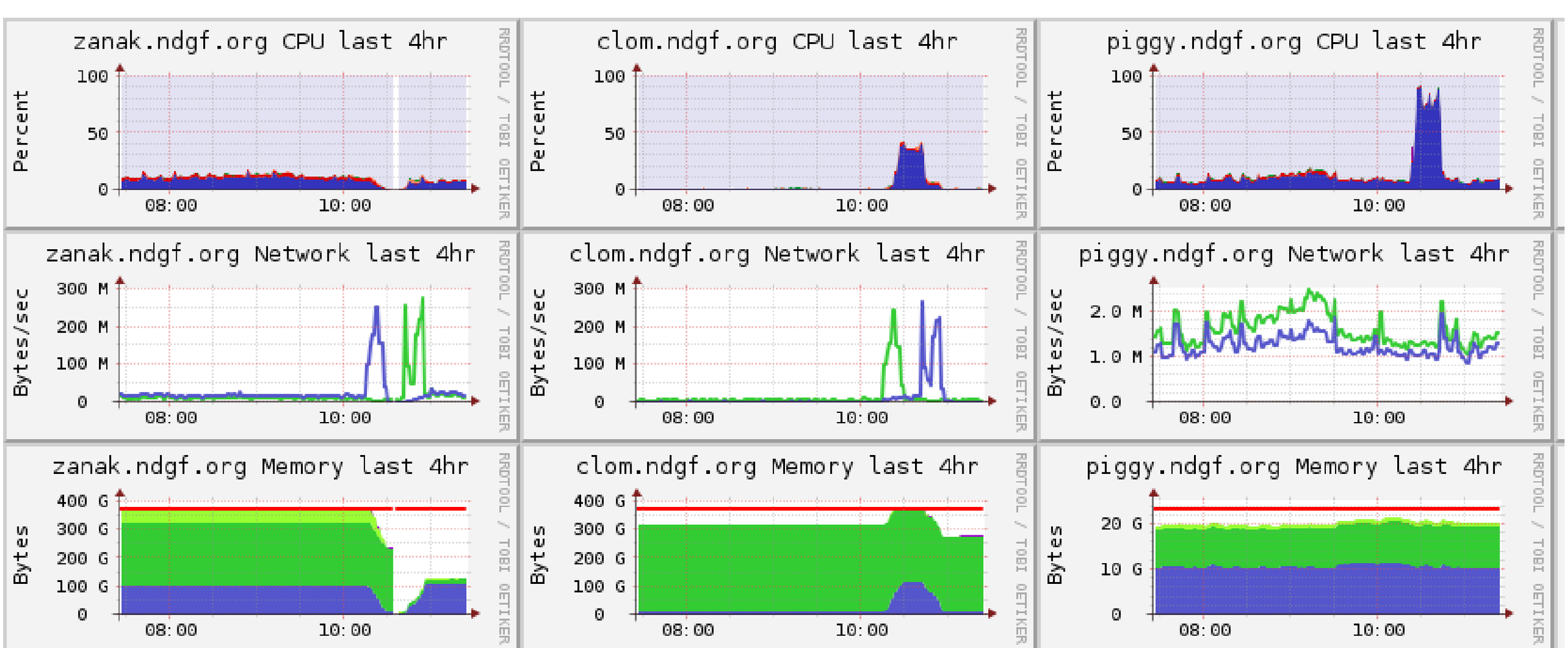
- Production
  - Two servers with direct 10Gbit/s connection
  - Running Ganeti for dCache head nodes, nagios, accoutning, DNS hidden master, monitoring
  - Also running postgresql on hardware in parallel
    - Main concern was adding latency to sync writes
  - Two servers = less automatic failover, but still taking advantage of live migrations to minimze downtime length
  - One big headnode plus two doors
  - Everything can run on one node in case of catastrophy

NDGF Ganeti use

- “Level2 services”
  - Non-critical but useful services
    - No facility guarantees outside office hours, etc
  - As I mentioned in site report, repurposed condensed pool nodes
  - Running even elasticsearch in redundant VMs
    - In part to keep Ganeti knowledge sharp
    - Very seldom a need to add or remove nodes from the production cluster, or rebalance VMs
  - 10GBase-T internal network, 1GBase-T external

NDGF Ganeti issue

- One strangeness with KVM live migrations
  - Only affects one VM, but it is the central dCache node...
  - Anyone else seen this with KVM live migrations?



NordForsk

Nordic e-Infrastructure Collaboration

Questions?