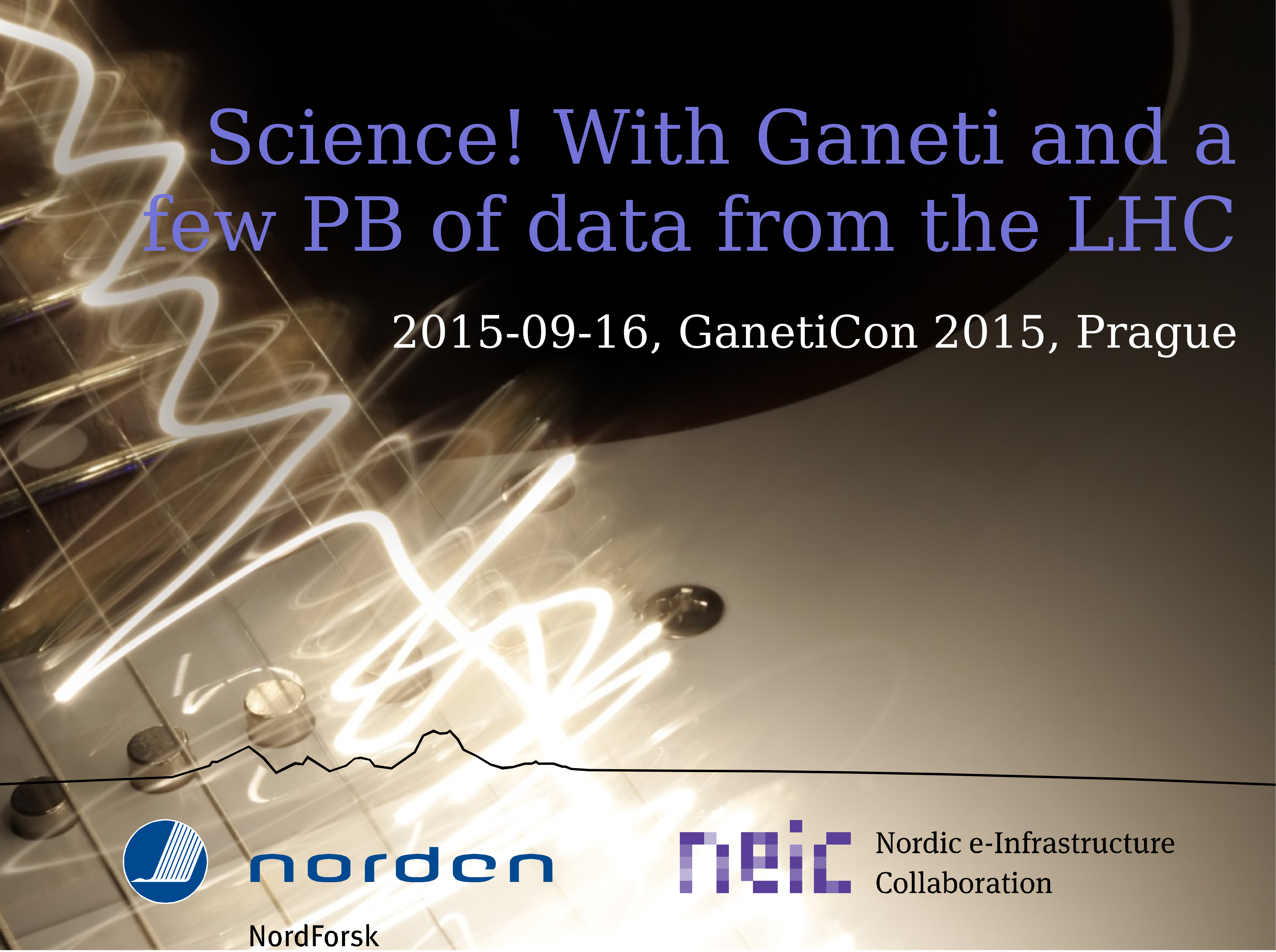


Science! With Ganeti and a few PB of data from the LHC

2015-09-16, GanetiCon 2015, Prague

norden

NordForsk

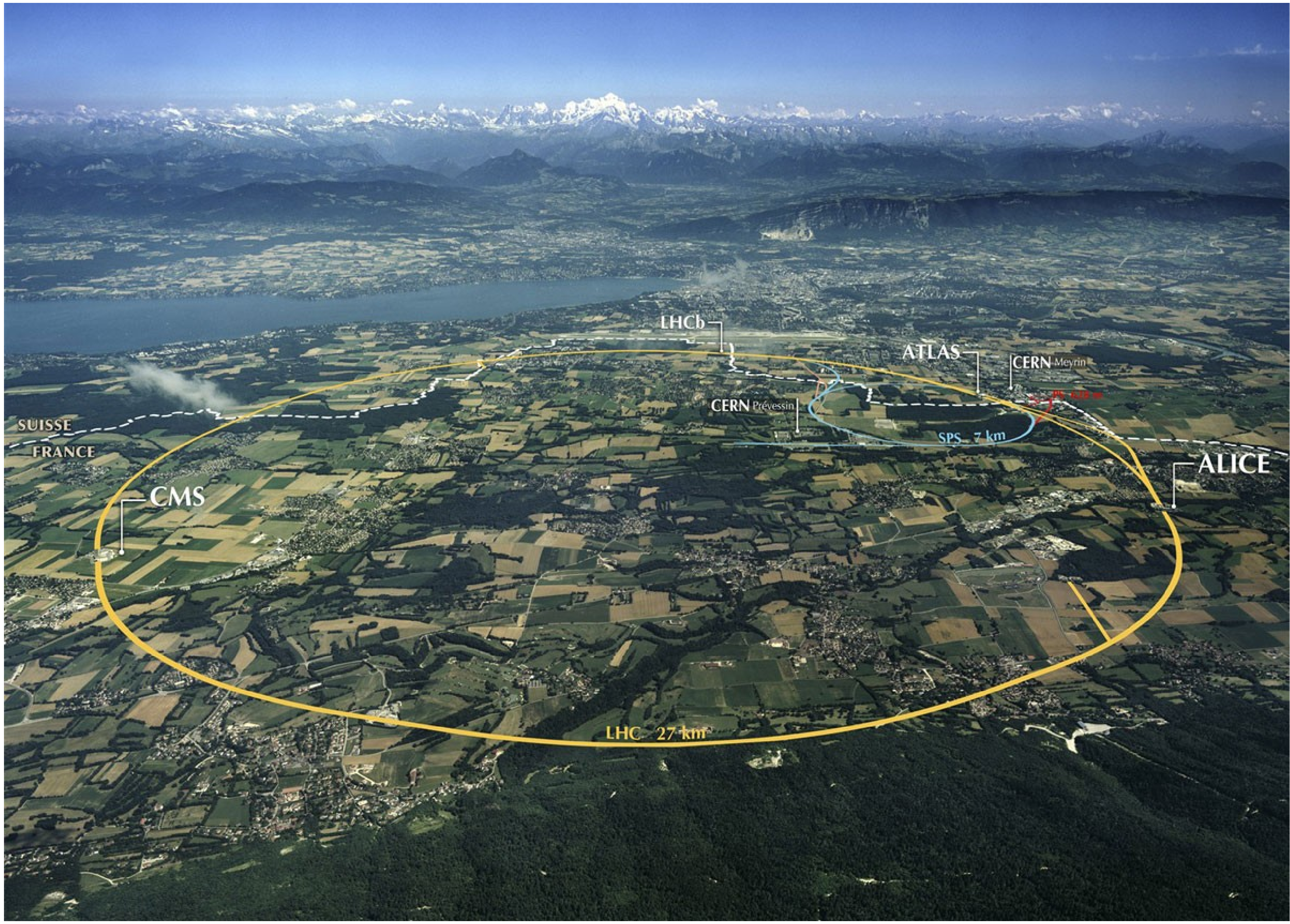
neic Nordic e-Infrastructure Collaboration

Overview

- Overview
- The LHC
- The WLCG
- NeIC
- The Site
- Ganeti

What is the Large Hadron Collider?

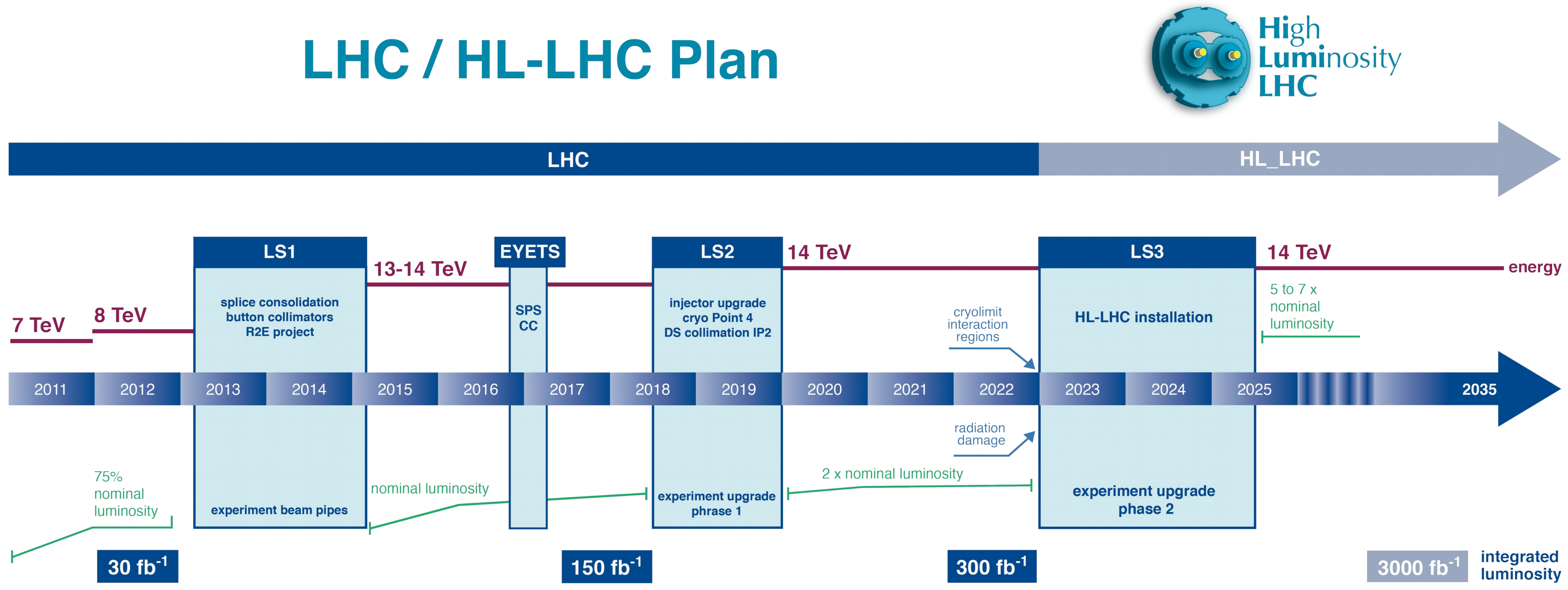
- The LHC is a gigant collider: phsyics experiement infrastructure
- Four big experiments
- Big as in:
  40x20 – 10x20m
  5000 – 15000t
  10-100TBytes/s



What is the Large Hadron Collider?

- From my perspective, the LHC is a very expensive machine that turns physics into data
- Now data is interesting
- To me, as a storage and computing challenge
- To my users, somehing to turn into papers and gold

And they have upgrade plans

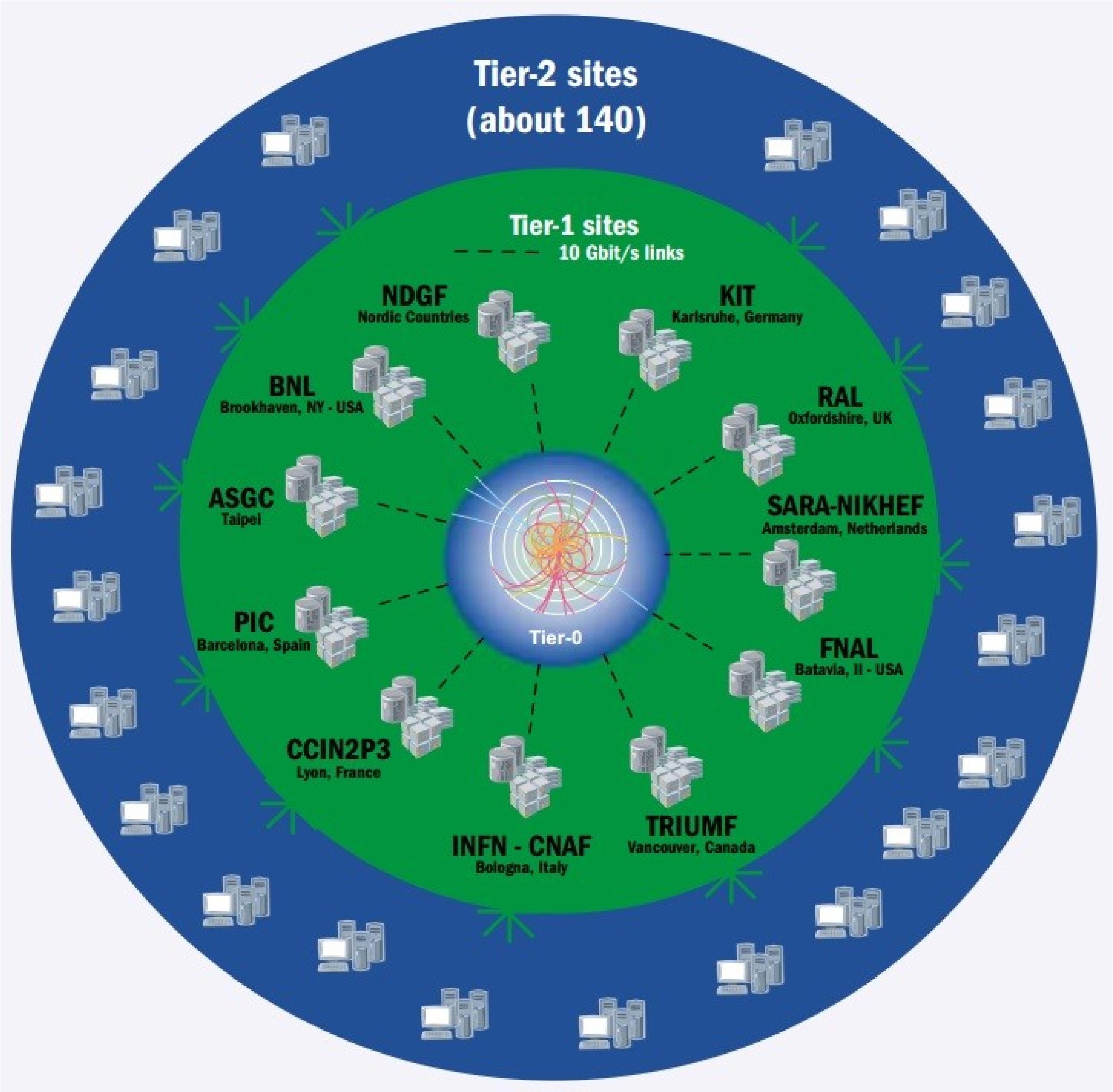


Data challenge

- Collisions happen at 20-40MHz
- And produce about 1-2 MB each
- 20-80TB/s is a bit much for the computing budget, so enter the Trigger
- Trigger triggers on “interesting” events
  - So most can be thrown away
  - Reduces the data rate by a factor of 1000-10000
- Which still is a pretty challenging rate to save and make available for analysis

World-wide LHC Computing Grid, WLCG

- Data gets distributed around the world
- Analysis and other jobs go to where the data is
- Tier-1 sites range 5-50PB online storage and about as much tape
- Typical IO rate is around 2-20GByte/s

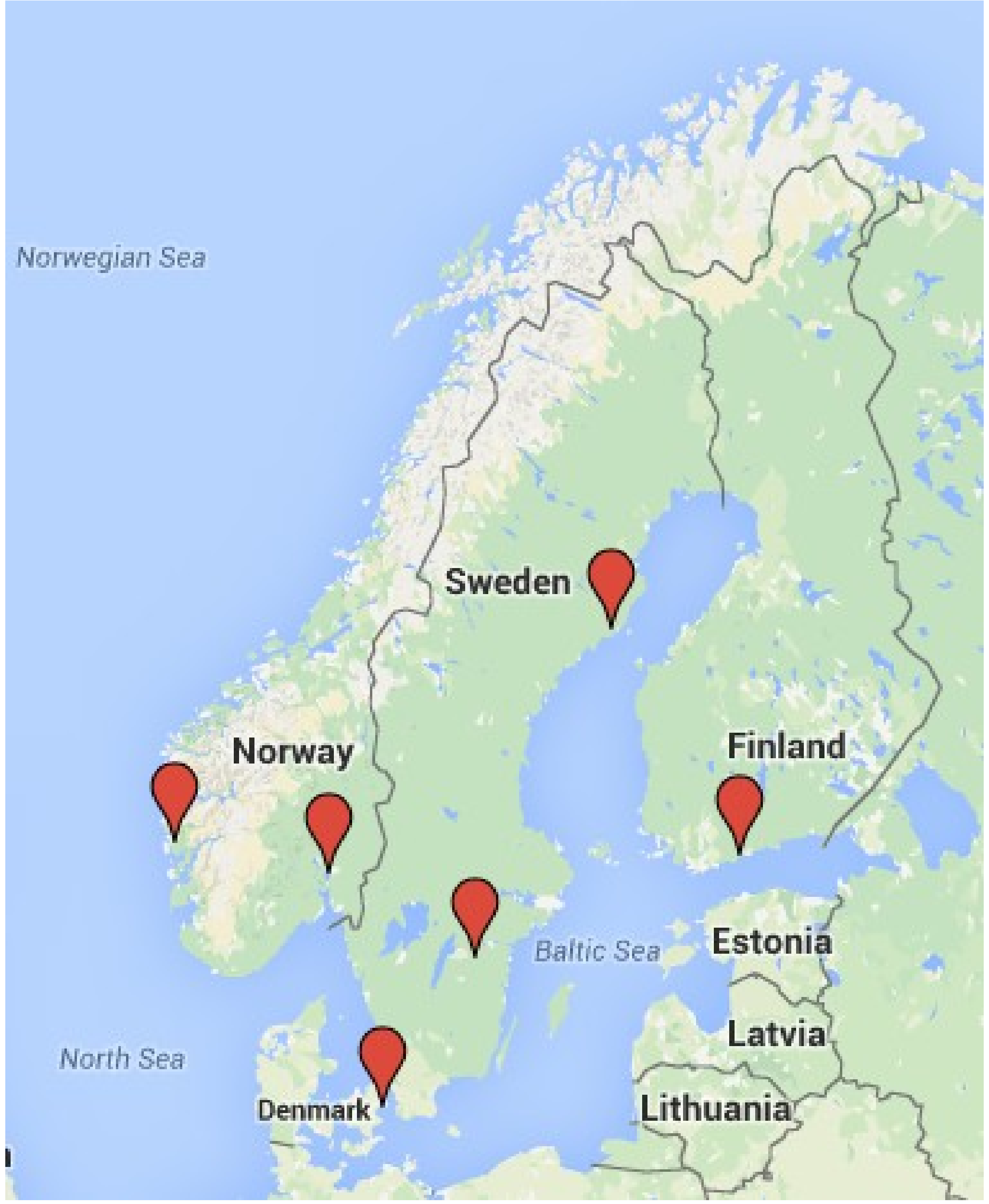


Nordic e-Infrastructure Collaboration

- Distribited organsiation based in the Nordic countries
- Based on academic HPC for scientific computing
- Developing and operating services for science
  - Pooling competence, sharing resources, etc
- About 50 people, half that in FTE
- Hosted by NordForsk
  - NordForsk is an organisation under the Nordic Council of Ministers that provides funding for and facilitates Nordic cooperation on research and research infrastructure

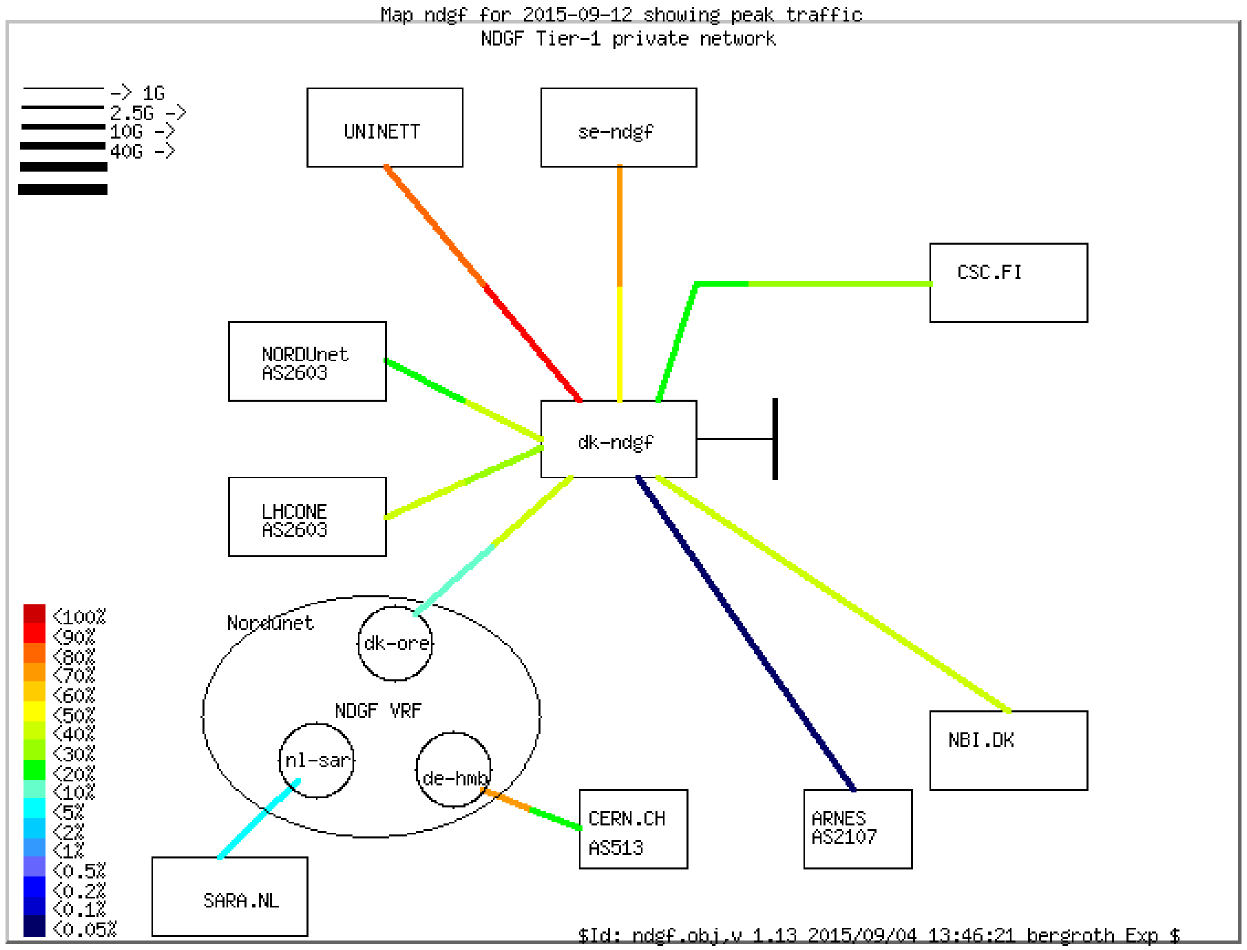
NDGF, the Tier-1 site run by NeIC (and me!)

- A distributed Nordic site
- Storage and computing spread over 6 academic HPC sites
- Using a distributed storage software called dCache
- And grid computing software, mostly ARC
- 2 full-time developers
- 6 half-time operations
- 2 FTE exp support and mgmt



The interesting parts

- dCache consists of
  - Storage pools (distributed)
  - Namespace (postgresql)
  - Admin nodes
  - Doors (network protocols)
- The last two run in Ganeti
- Together with monitoring, accounting, DNS, etc
- Close to our network SPOF
- A few hours downtime is OK
  - Longer outage or data loss: :(



Ganeti hardware and setup

- Two HP DL380 with 384 GB ram and 25 HDDs
  - 3-way mirroring for data availability
- Postgresql on hardware on the same machines
  - Hot standby on the other server
- 16 VMs of various size and OS
  - Mostly ubuntu 14.04 but some centos 5/6/7
  - Biggest 24G ram, 240G disk, 12 CPUs
  - Adding new ones as needed
- KVM and DRBD and bridged networking

Ganeti setup

- Using whatever version is in the Ubuntu 14.04 main repo (2.9.3)
- ganeti-instance-image with a few base Oses
  - Ubuntu 14.04, CentOS 5, CentOS6, CentOS7
  - Some software is annoying to port and the WLCG crowd has standardized on Scientific Linux, so RHEL
  - Most VMs run Ubuntu though
- Was originally looking at debootstrap, but it seemed to take a lot of fiddling
  - Also, RHEL-derivatives and their lack of rhbootstrap

Other Ganeti in my vicinity

- HPC2N, Umeå University
  - One of the 6 HPC sites and my home institute
  - Also two servers with live migration for random service VMs that it would be nice to have decent availability on
  - Running very ancient version of Ganeti
- Academic Computer Club, Umeå University
  - Recent install with latest version from Ubuntu 14.04 PPA (2.12.4)
  - Only deployment in my surroundings with n>2 nodes
  - Looking to run both services and have it available as a playground to test new technology and get experience
    - For instance, been hearing nice things about CEPH?
    - Computer club members self service portal with v6-only private VMs?
- Both of these run KVM+DRBD and PXE/FAI

Ganeti experiences

- Ganeti itself is nice and pretty easy to set up from scratch
  - Newer versions better than older
- OS creation is an order of magnitude more tricky
  - Collegues confused by creating custom debootstraps
  - Confused by images and bootloaders (grub1, then run magic script post-instatiation to switch to grub2)
  - Easier when you already have a proper infrastructure
    - Like PXE-booting into FAI

Ganeti experiences

- Live migration was slow
  - Finally found that I hadn't readjusted the very conservative default limit of MB/s memory transfer
- Split brain...
  - Documentation not exactly clear on how to deal with it
  - DRBD documentation a maze of twisty terms, all different
  - Is the next slide a correct way of dealing with it?

Recovering from split brain

- If we are unlucky the drbd storage back end can suffer from split brain. For instance this can be seen when disks are labelled DEGRADED and in /proc/drbd on one node it is in state "StandAlone" and the other "WFConnection". In dmesg one can also see things like this (if one looks carefully between all the drbd chatter):
  - block drbd11: Split-Brain detected but unresolved, dropping connection!
- First, choose your victim (zanak or clom) carefully. Split brain means that drbd couldn't figure out with certainty which is the most up to date copy of the data. The victim's copy of the disk state will be clobbered with a resync from the winner. If in doubt, and it is precious, make a copy of both images first.
- First shut it down:
  - gnt-instance shutdown fax
- But activate-disks to get a /dev/drbdXX device:
  - gnt-instance activate-disks fax
- On the **victim**:
  - drbdsetup /dev/drbdXX invalidate
- Then activate-disks again to reconnect:
  - gnt-instance activate-disks fax
  - gnt-instance info fax

Questions?