Surviving (with) Debian in a scientific computer center

- `whoami`
  - Christopher Huhn, C.Huhn@gsi.de
  - I'm from:
    - GSI Darmstadt, Germany
    - IT/HPC department since 2002
  - I'm here because:
    - GSI is running Debian

About GSI

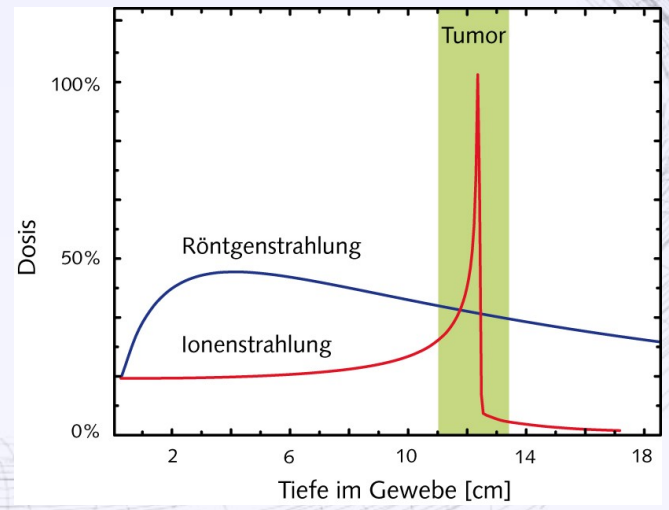
- „GSI Helmholtz Centre for Heavy Ion Research“
  - Research lab for fundamental research
  - High energy physics
- Founded 1969
- Shareholders:
  - Federal republic of Germany (90%)
  - States of Hesse, Rhineland-Palatine, Thuringia
- ~1050 employees
- 1000 - 1200 external scientists per year
- 2 heavy ion accelerators:
  - UNILAC (1975): ~10% light speed
    - Also the injector for
  - SIS18/ESR (1990): ~90% light speed

Superheavies

- Elements first “discovered” at GSI:
  - Bohrium, Bh, 107 (1996)
  - Hassium, Hs, 108 (1984)
  - Meitnerium, Mt, 109 (1982)
  - Darmstadtium, Ds, 110 (1994)
  - Roentgenium, Rg, 111 (1994)
  - Copernicium, Cn, 112 (1996)
- Currently hunting Element 119
  - Probability of production < 1 atom/month

Biophysics

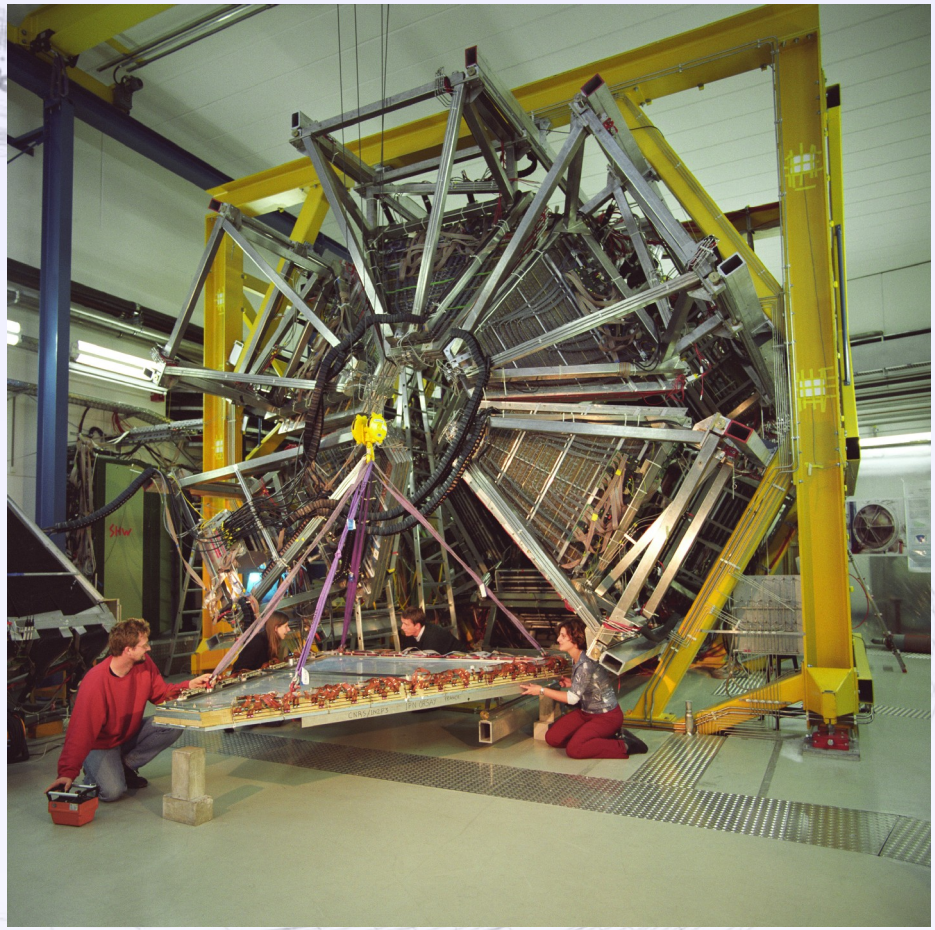
- Cancer therapy
  - Tumor irradiation with accelerated carbon ions
    - sharp Bragg peak



SIS18 Detectors

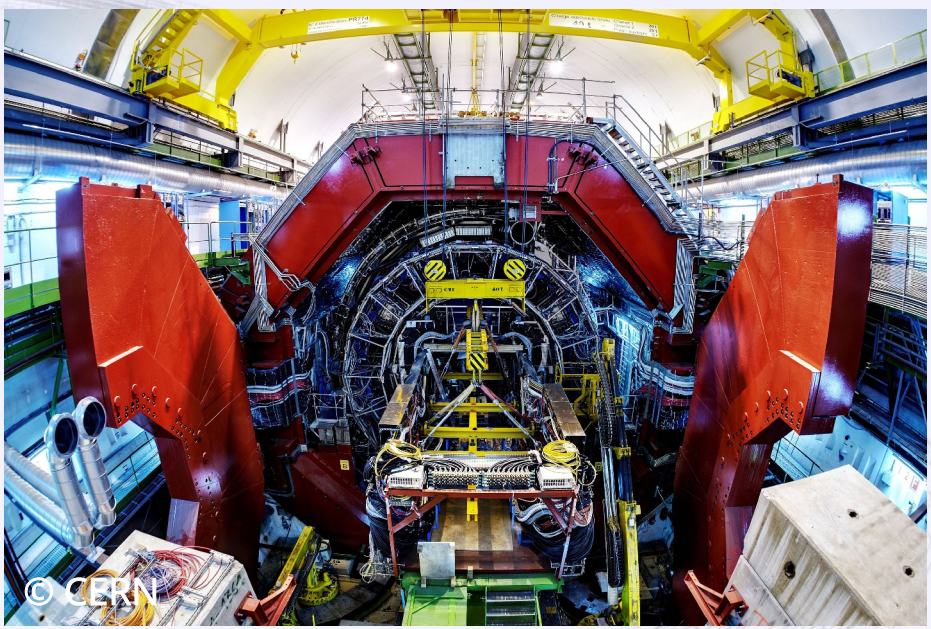
- Hades,
- Fopi and
- Land

produce respectable amounts of data to feed our HPC clusters



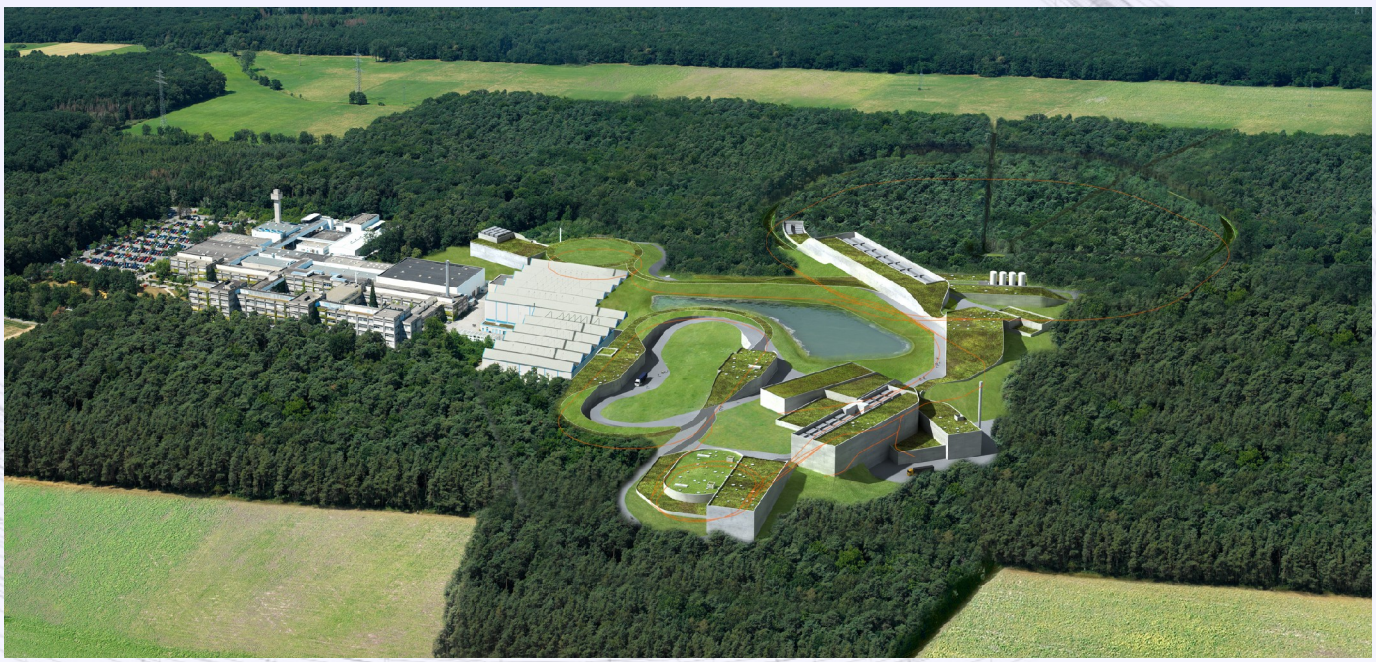
LHC participation (Alice)

- Detector development at GSI
- Now mainly computing
  - local computing
  - also Grid computing



FAIR

- PANDA: proton-antiproton collisions
- CBM: super-dense nuclear matter
- NUSTAR: Nuclear structure
- APPA: plasma and bio physics ...



GSI HPC department

GSI HPC department

- Team of 9 people:
  - 2 physicists, 2 chemists, 4 computer scientists, 1 mathematician
  - 2.5 FTE assigned to IT security
- Hived off the universal IT department last year
- We attempt to run the 'even cheaper' solution
- Computers spread over 4 computer rooms atm.

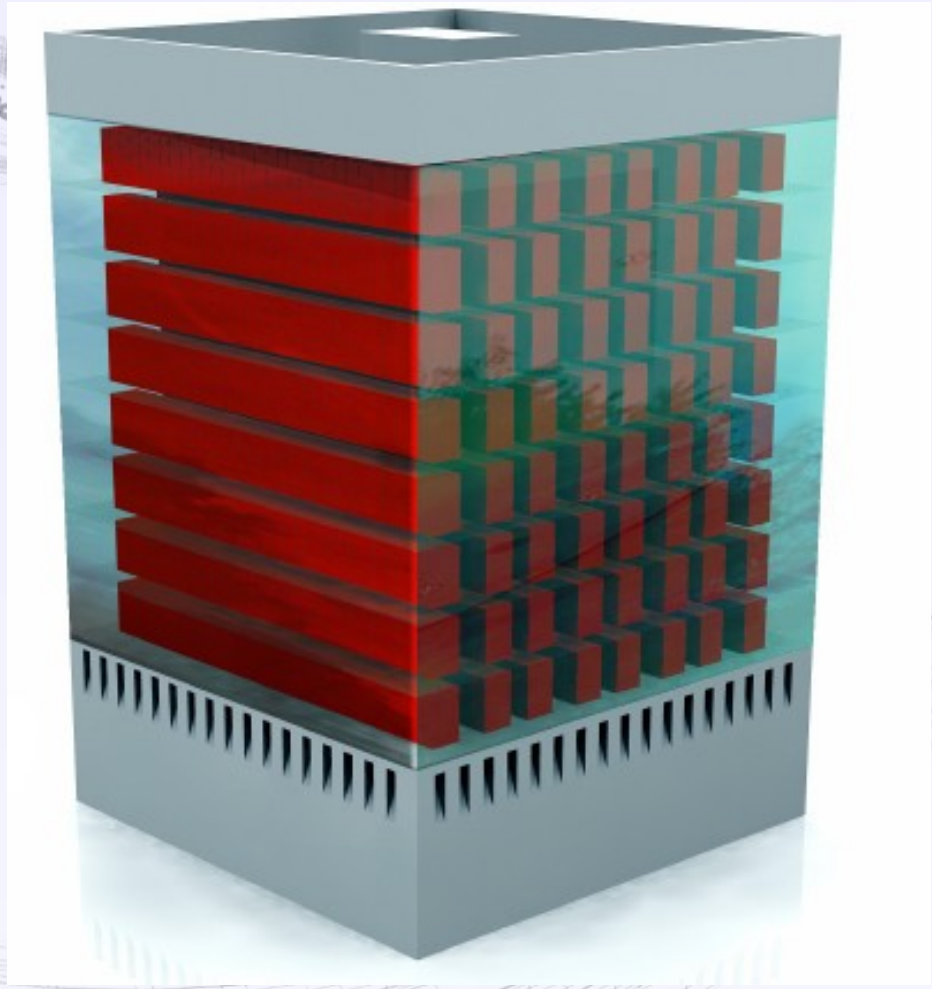
- HPC provides
  - Batch farm(s)
  - Lustre storage + ~ 50 NFS servers
  - Core IT services
    - DHCP, DNS, Radius
    - central mail hubs,
    - web servers, trouble ticket system,
    - CUPS and LprNG print services
  - Linux Desktops and central interactive login nodes
  - Collaborative Web services for experiment groups
    - Wiki, Discussion forum, SVN+Trac, Elog, Instant Messaging
- All these services run on Debian GNU/Linux

Scientific Computing at GSI

- HPC is not directly involved in scientific software development
- Experiment software has to be available in dozens of variants in parallel and not suitable for packaging
- Commercial software is a niche player
  - But freely available software without proper licensing is common
- Up to now:
  - Locally written analysis and simulation software based on ROOT
    - e.g. http://fairroot.gsi.de/
  - Mostly single threaded, independent jobs
  - IO is normally the biggest bottleneck
  - → Data is on-site and so is the computing
- Coming up: Lattice QCD
    - larger-scale MPI jobs, low latency inter-process communication and GPGPU computations
- We try to combine these on general purpose clusters

The green cube

- New data-center
- Ready for operations: 2014
  - Water-cooled racks
  - Evaporative cooling
  - 30 °C operating temp.
  - PUE < 1.1
- > 6 MW
- Capacity: ~ 1000 racks



HEP community

- Hepix conference
- Twice a year
- IT department staff from
  - CERN, SLAC, FERMI, INFN, IN2P3-CC, DESY, Diamond light source, RAL, ...
- Next meetings
  - Bejing, Fall 2012
  - Bologna, Spring 2013
- https://www.hepix.org/

Debian @ GSI

- Started around 1996
- Why Debian?
- Competitors
  - Windows compute clusters ?!
  - SuSE, RedHat
  - "Scientific" Linux, CentOS
  - MacOS
  - Ubuntu
- Why not Debian?

Debian characteristics

- Roll-your-own custom distro
  - There's always a well-supported alternative

```
# apt-get remove exim
# apt-get install [ postfix | nullmailer | … ]
```

- Debian is not a company that sells you a product
- Surprises are rare
- Software tested on many platforms
- It's ready when it's ready
  - Usual complaint: outdated software
    - People want the hippest and newest browser …
      - … while their software still only compiles with gcc 2.95 or g77
    - Nevertheless true for the Debian kernel and drivers
      - Painful to install on newer desktop and laptop hardware

Debian @ GSI

- Hardware acquisition
  - Call for tenders for server hardware request Debian Stable to be pre-installed
- (Re-)Installations done by FAI
  - Debconf, pre-seeding
  - Cfengine/Chef take over for orchestration after install
- The larger the infrastructure grows, the more distribution agnostic you (have to) become
  - Goal: be able to automatically reinstall your whole infrastructure from scratch → PaaS
  - Debian strengths become less important
  - but advantages of commercial distros also become less important
  - After all Debian is more flexible - and less expensive?

Mastering the package management

- Still one of Debian's biggest highlights
- Partial Debian mirror (`i386` and `amd64`) with `debmirror`
- Repository for custom packages with `debarchiver`
  - alien-converted proprietary RPM packages
  - back-ports
  - rebuilds with different build options e.g.
    - enabled Oracle support
    - GridEngine support enabled in OpenMPI
  - GSI meta-packages automatically built from FAI package lists
- Nightly `apt-get dist-upgrade` run via `cron-apt`
  - Upgrades normally just work
  - Also mostly true for distribution upgrades

Debian Desktops

- „subterraneously bad“?
  - Not as posh and well-integrated as MacOS or Ubuntu
- Standard desktop is KDE
  - Alternative: XFCE, no support for Gnome
- Desktop hardware from DELL remains identical for one year
- Some hundred desktops and central interactive servers (cluster head nodes ...)

Debian Desktops

- Netboot thin clients
  - Normal desktops - and fai install systems - boot from the network
  - NFS-mounting of the root partition handled by initramfs-tools
  - Shared read-only root filesystem just works thanks to Debian (Sarge release goal?)
  - Writeable `/etc` and `/var` via unionfs/aufs
    - 2 small init-scripts, no changes to initramfs required
  - `/var`, `/tmp` and data on the local hard disk
  - Images for net-boot automatically created via `fai dirinstall`
    - maintained by cron scripts
- Security updates applied directly to the master images
  - won't miss nodes that are off
- Distribution upgrades only require a reboot after some DHCP settings have been changed

Compute Cluster

- 3 Clusters:
  - LSF cluster: 190 nodes, 1500 CPU cores
    - to be phased out soon
    - Debian Lenny, Etch finally turned off just now
  - Icarus cluster: 100 nodes, 1600 cores
    - GridEngine batch scheduler
    - No shared filesystems
      - except for Lustre
    - CVMFS for software distribution
      - HTTP-based read-only filesystem
      - Meta-data collected in sqlite-DBs
      - Mounted as fuse-fs by an automounter extension
- *Prometheus* cluster: 270 nodes, 6500 cores
  - Plus 100 nodes, 2400 cores in a temporary MPI test cluster
  - Setup similar to *Icarus*
  - Infiniband-only networking
    - No Infiniband support in klibc's ipconfig
      - added by a small initramfs-tools hook script (no DHCP yet)

(SUN) Grid Engine

- 6.2u5 on Debian Squeeze
  - Almost as easy as "apt-get install gridengine"
  - MPI functionality and stability needs improvements
- Future uncertain?
  - 4 different forks
  - no clues yet which route to follow
- Evaluate slurm instead?
- Hepix-BOF: https://forge.in2p3.fr/projects/gess/wiki

Lustre

- Distributed file-system consisting of
  - one central metadata-server (MDS) and
  - many Object Storage Servers (OSS)
- Upstream concentrates on RHEL
  - SuSE also supported but client-only for 2.0 and beyond
- GSI pays Credativ for the adaption
  - http://pkg-lustre.alioth.debian.org/
  - Atm. waiting for Lustre 2.1
- Lustre client
  - External kernel module
  - Relatively easy to adapt to (older) vanilla kernels
    - No support for > 2.6.38 yet
- Lustre server
  - Kernel patch
    - Heavily relies on the ext4 code
    - Difficult to adapt to different kernels
  - We just used a debianized SuSE Kernel for Lenny
  - LLNL has Lustre running patchless on top of ZFS
    - License incompatibilities prevents ZFS integration into vanilla kernel

Lustre clusters

gsilust

- 10G Ethernet
- 135 OSSs
- 2.2 PB usable space
- 550 500GB disks
- 3500 1TB disks

hera

- 40Gb Infiniband
- 47 OSSs
- 1.4 PB usable space
- 2250 1TB hard disks
- Aggregated throughput > 100 GB/s

(The road to hell is paved with) good intentions

- Track Debian sid/testing
  - We never really managed to test testing yet
  - Be ready for the release
  - Discover bugs beforehand
  - Start Wheezy evaluation next week!
- Upgrade from oldstable to stable in time
  - Or start a Debian LTS effort?
- Improve contributing to the community
- Open up our internal wiki knowledge base

Thank you

- And
  thank you, Debian!
  - Debian knowledge helps me in being successful in a job that I like a lot!
  - Debian helps in making it *interesting* and *challenging* every day