Maxwell Cluster Overview

- Current Maxwell configuration

- Remote Graphical login

HELMHOLTZ | ASSOCIATION

DESY

What is HPC / Maxwell?

High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly.

High Performance Compute resources are usually …
- Providing peta-flops (super-computer)
- Homogenous
- Capable of rapid i/o
- Accessible only via proposal-system
- Supported by very motivated HPC experts

What is HPC / Maxwell?

High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly.

High Performance Compute resources are usually …
- Providing peta-flops (super-computer)
- Homogenous
- Capable of rapid i/o
- Accessible only via proposal-system
- Supported by very motivated HPC experts

Maxwell is the only (kind of) HPC resource at DESY
- Defined through the infiniband fabric
- Rather heterogeneous hardware
- Equipped with fast cluster file-systems (and 10GE)
- Accessible for anyone with suitable applications (or…)
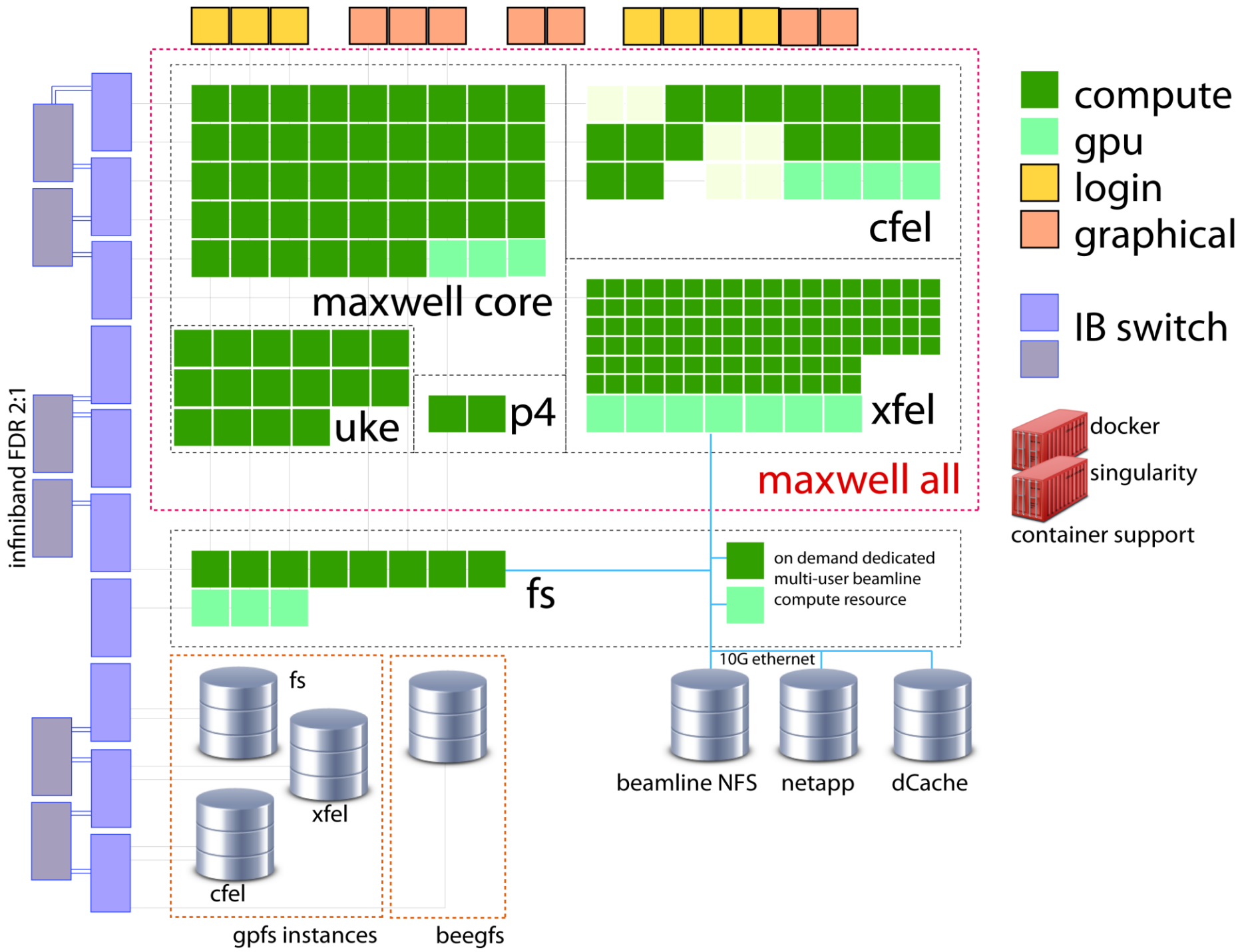- Resources are rather limited, but we try our best

Current hardware

Compute	Total	IT	FS	CFEL	XFEL	UKE	Shareable/SLURM
AMD (it-hpc)	39 (39)	16	-	-	6	17	36
Intel+GPUs	21 (13)	2(6)	4(7)	8	7	-	16
Intel	154 (32)	26(62)	8(16)	20	100	-	140
GL nodes	5 (0)	3	2	-	-	-	-
Total	219 (84)	47 (87)	14 (25)		113	17	**192** in **all** partition

(6/2017), (4/2016). See sview and/or https://confluence.desy.de/display/IS/Maxwell+Hardware for details

Storage System	Bindings	Volume	Exports
AFS	10GE / AFS	Quota per volume	global
BeeGFS	56GE IB / BeeGFS	**220** (120)TB	None
dCache	10GE / pnfs (NFS4.1)	Unlimited	DESY, global
GPFS home	56GE IB / gpfs	20GB / user	none
GPFS p3/cfel/**exfel**	56GE IB / gpfs	>1PB each	restricted
Netapp	10GE / NFS 3	30TB (3TB quota)	DESY

See Sven's talk for details

compute
gpu
login
graphical

IB switch

docker
singularity

container support

maxwell core

cfel

uke

p4

xfel

maxwell all

infiniband FDR 2:1

fs

on demand dedicated
multi-user beamline
compute resource

10G ethernet

fs

xfel

cfel

gpfs instances

beegfs

beamline NFS   netapp   dCache

Maxwell fabric
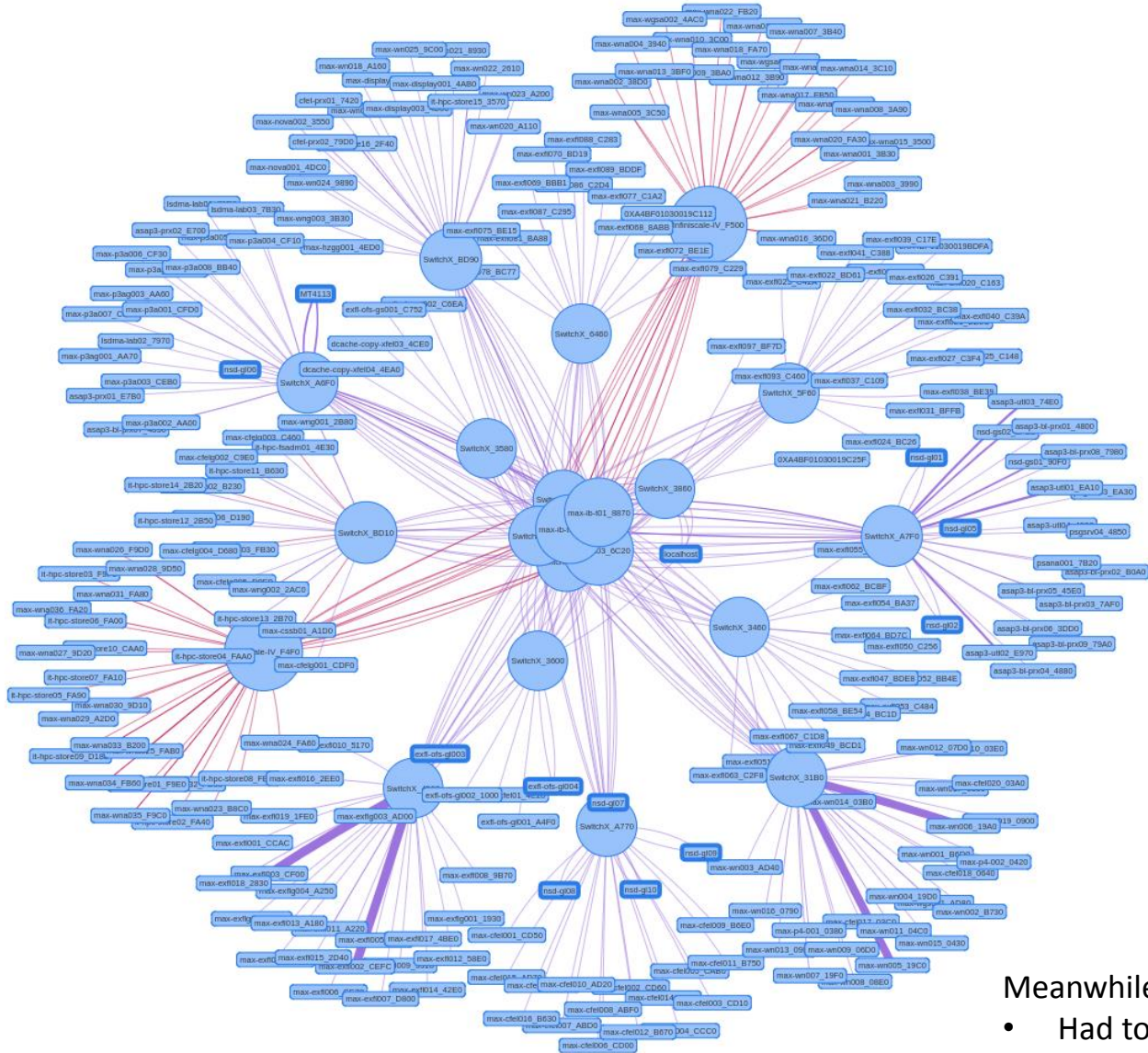
Meanwhile fairly complex setup
- Had to simplify configuration
- Made support contract to faster resolve issues
- Do lots of monitoring, but …

Monitoring & Troubleshooting

Troubles:

- *GPFS/BeeGFS is slow since 2 weeks already*
- *I have the impressions things used to be faster*
- *I frequently have problems opening files*
- *I got an error running application xyz*
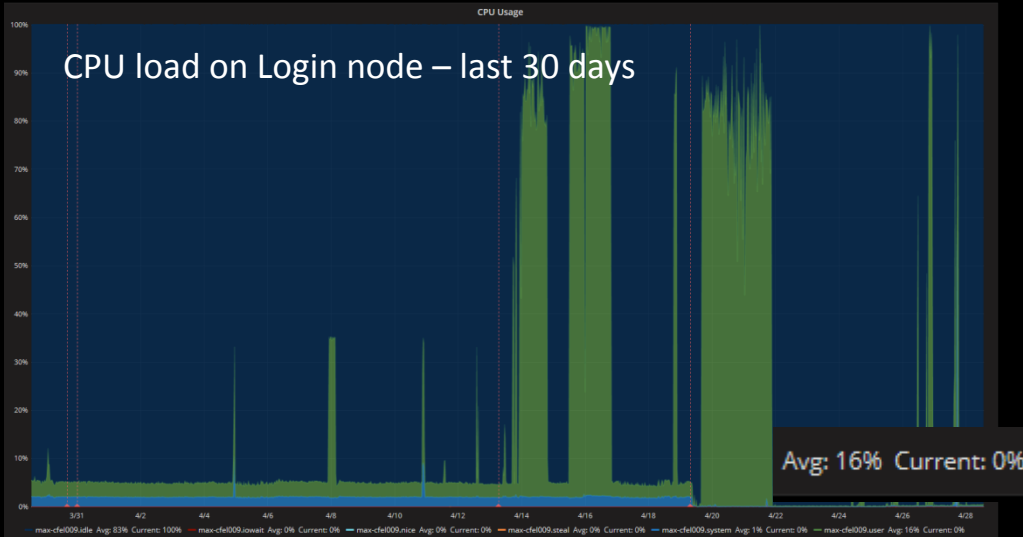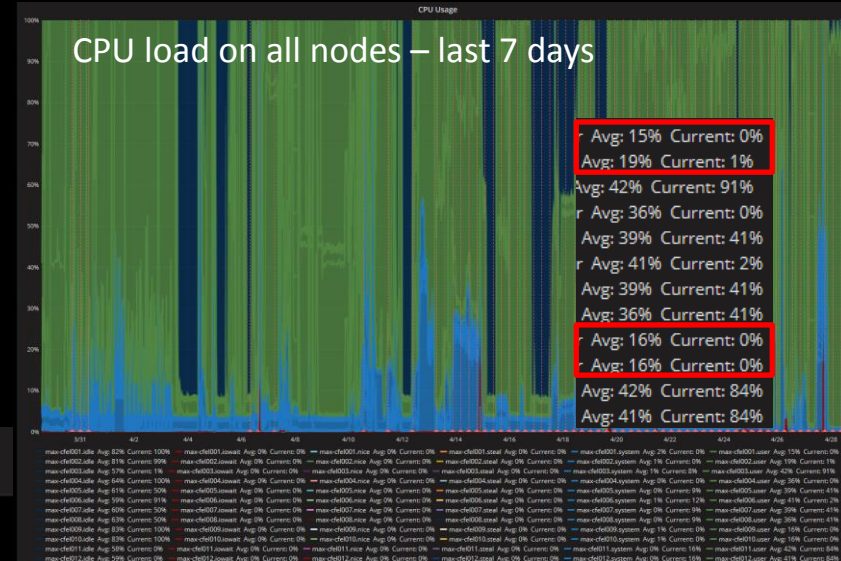
Solutions:

- None

Needed:

- Report as soon as possible, not weeks later. We need a timestamp!
- Specify at least host, application, and PATH
- Any error message
- How you connected (putty, fastx, vgl) if problems are X-related
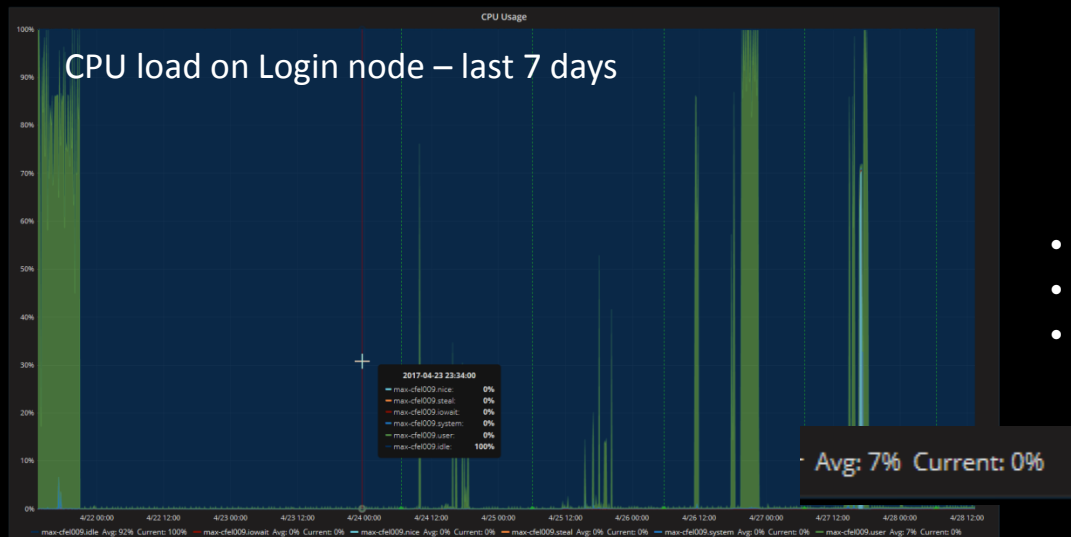- Job-Ids are very helpful

Login Nodes

CPU load on Login node – last 30 days
Avg: 16% Current: 0%

CPU load on all nodes – last 7 days
Avg: 15% Current: 0%
Avg: 19% Current: 1%
Avg: 42% Current: 91%
Avg: 36% Current: 0%
Avg: 39% Current: 41%
Avg: 41% Current: 2%
Avg: 39% Current: 41%
Avg: 36% Current: 41%
Avg: 16% Current: 0%
Avg: 16% Current: 0%
Avg: 42% Current: 84%
Avg: 41% Current: 84%

CPU load on Login node – last 7 days
Avg: 7% Current: 0%

- Mix of Login and Compute caused problems
- Switched login nodes to purely interactive nodes
- Doubt that login nodes are needed at all

Who can use Maxwell?

High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly.

Partition	Managed	Netgroup	Applications
maxwell/all	UCO maxwell.service	maxwell-users	Restricted to HPC applications. Need some information.
cfel	CFEL-DESY admins	cfel-wgs-users	Entirely up to the resource/group admins to decide who and what and where.
exfel	Exfel admins	exfel-wgs-users	
exfel-wp72		exfel-theory-users	
exfel-spb		exfel-theory-users	
efxel-th		exfel-theory-users	
p4	M admins	p4_sim	
uke	PS admins	max-uke-users	
ps (slurm)	PS admins	max-ps-users, max-psx-users	
max-fs (no slurm)	PS admins	Hasy-users, psx-users	

Depends which part of the cluster you intend to use!

How to access Maxwell?

Standard Path:

- ssh max-wgs / max-wgsa / max-cfel / max-cfelg …
- From remote: ssh tunnel to specific node
- Ok, but not too convenient.
- No support for OpenGL except via FastX, which is very CPU intensive
 - Insufficient for some applications

How to access Maxwell?

Standard Path:

- ssh max-wgs / max-wgsa / max-cfel / max-cfelg …
- From remote: ssh tunnel to specific node
- Ok, but not too convenient.
- No support for OpenGL except via FastX, which is very CPU intensive
    - Insufficient for some applications

**New graphical login will full GPU hardware acceleration**

- 3 nodes with 4 nvidia quadro M6000 (similar to TitanX)
- 2 nodes with 2 nvidia quadro M6000 dedicated for NOVA project
- Entry points are load balanced
- Session-based distribution of jobs across GPUs
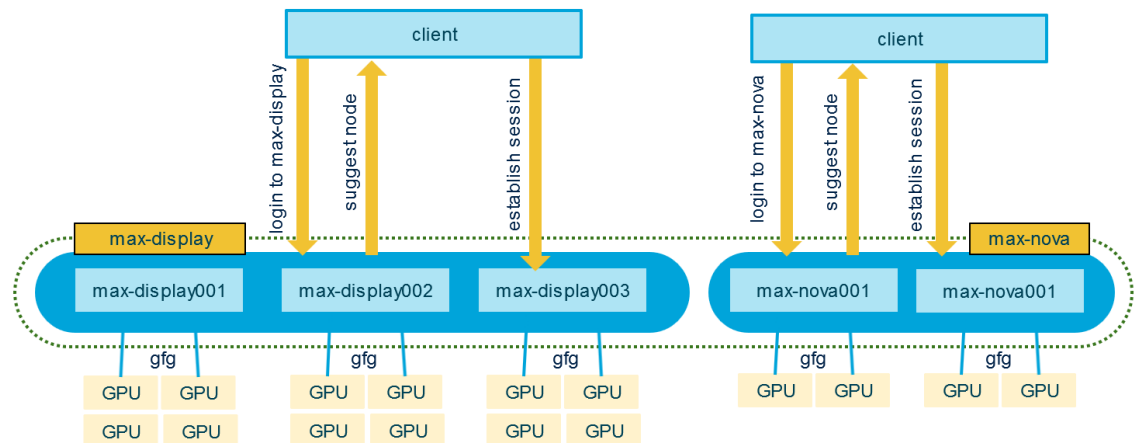
Remote Graphical logins

Essential information can be found here: https://confluence.desy.de/display/IS/Remote+Login

- Login required (don't want to advertise open ports)

- Reachable via ssh (fastx) as max-display.desy.de (max-nova.desy.de)

- Reachable via https as https://max-display.desy.de:3443

- Both ports are open world wide → no ssh tunnel through bastion

- Very convenient & fast – even from home over a slow DSL link
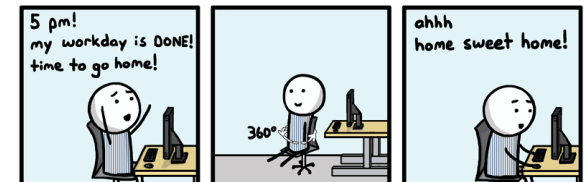
Remote Graphical logins

- Connect to other hosts inside maxwell: ssh

- Connect to GPU node, GPU configured for rendering: vglconnect
  - Use vglrun on remote node to enable vgl-gpu-rendering

- Running GL-applications on display nodes
  - That's what the resources are meant for
  - No special commands needed
  - Don't unset LD_PRELOAD
  - Some OpenGL applications try to outsmart vgl
    - warnings but usually works anyway

- Compute jobs on the rendering GPUs
  - Short tests are perfectly ok
  - Don't run lengthy jobs on the GPUs!

- Compute jobs on the CPUs
  - Perfectly ok.
  - Keep them short and small (cores/memory)

Remote Graphical logins

World wide open comes with a price

- Updates will be applied continuously

- Security updates requiring reboot will be done within 24h
  - If necessary with short notification

- Nodes might be rebooted roughly once a week
  - Save your work!

- Quick demo … if you have a maxwell/cfel/xfel-resource try it out …