DTU Compute
Department of Applied Mathematics and Computer Science

High-Performance Computing

DTU Unix Systems
&
High-Performance Computing

Using the DTU Unix systems
for your computations

Bernd Dammann

Assoc. Professor, Scientific Computing
DTU Compute, building 303B

MSc in Physics, PhD in Phys. Chemistry
work with DTU’s central HPC (and G-bar) since 2001
co-founder of DTU Compute GPUlab (2007)
teach HPC courses (02614, 02616, 41391) and related topics

DTU Computing Center (DCC)

Bernd Dammann

HPC Architect, Consultant & Scientific Lead
DTU Computing Center, building 324

DCC started in 2008
in-sourcing of operations of DTU’s central HPC (and G-bar)
HPC consulting for DTU users
HPC Competence Center (since 2013)
GPUlab activities (since 2013)

DTU Computing Center (DCC)

What are we?

- DCC is a (central) DTU unit
- we are **not** part of DTU Compute
  - ... but we have our office at DTU Compute(!)
- we provide ‘free’ access to HPC resources
- we offer HPC hosting services for departments at DTU (e.g. Compute, Chemistry, Environment, ...)
- we offer HPC consulting services for all of DTU
  - ... free of charge (up to a certain extent)
  - ... as part of larger projects

DTU Computing Center (DCC)

Who are we?

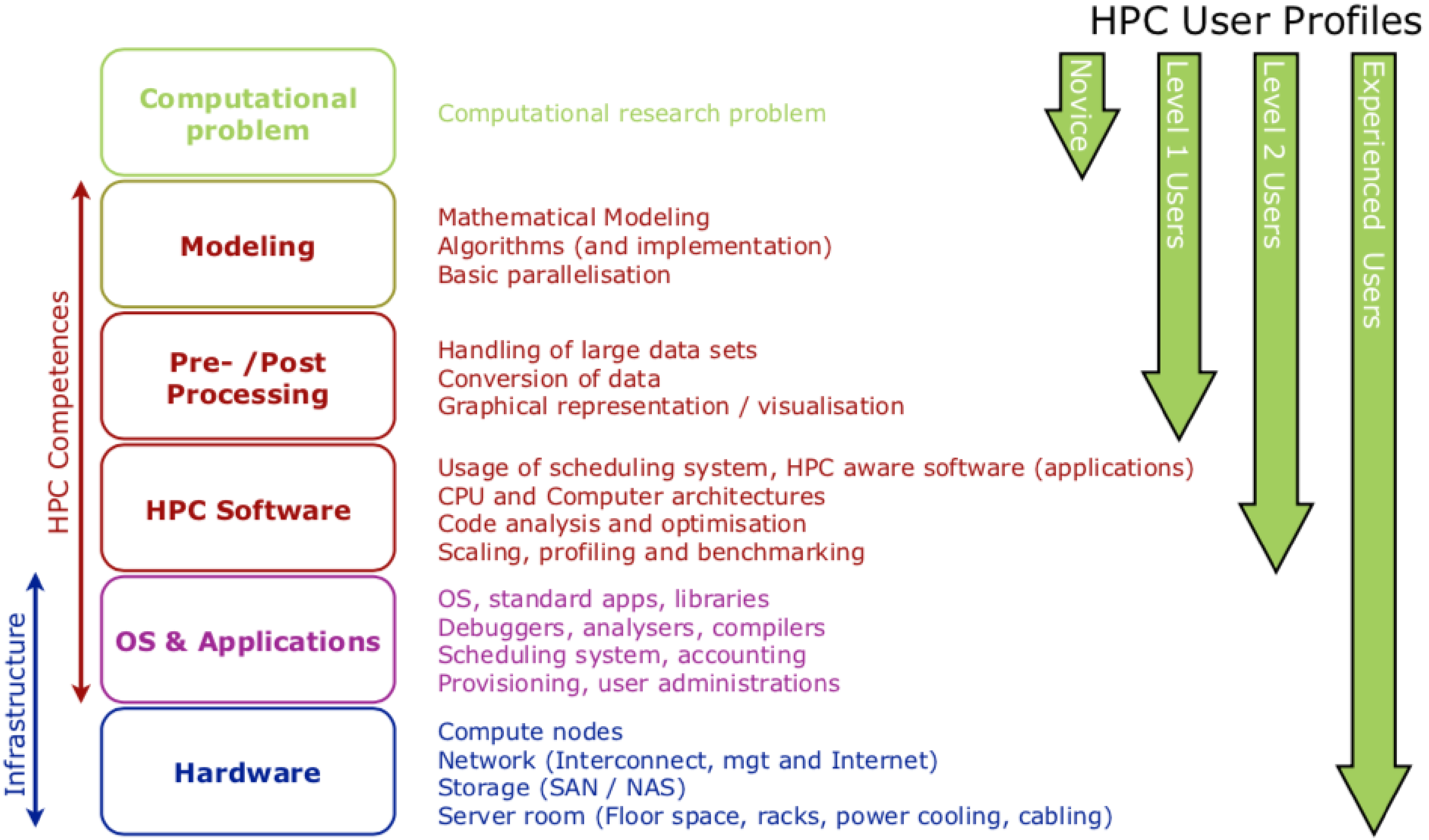
- Andrea: HPC expert
- Bernd: HPC architect, expert, scient. lead
- Hans Henrik: HPC expert (GPUs)
- Henning: team head
- Ian: system admin
- Jette: contract administrator
- Nick: HPC expert
- Pietro: project coordinator
- Sebastian: tech lead & system admin

DTU Computing Center (DCC)

Our mission: “make HPC work”

- *easy* access to HPC resources
- provide the support needed – on all levels
  - first steps for newbies
  - performance analysis and tuning
  - new architectures: testing, consulting & porting
- guide users in the HPC landscape
  - local, national and international
- if you use our services, please mention it in your publications (e-mail support@hpc.dtu.dk for details)

DTU Computing Center



HPC User Profiles

Novice

Level 1 Users

Level 2 Users

Experienced Users

HPC Competences

Infrastructure

Computational problem

Computational research problem

Modeling

Mathematical Modeling
Algorithms (and implementation)
Basic parallelisation

Pre- /Post Processing

Handling of large data sets
Conversion of data
Graphical representation / visualisation

HPC Software

Usage of scheduling system, HPC aware software (applications)
CPU and Computer architectures
Code analysis and optimisation
Scaling, profiling and benchmarking

OS & Applications

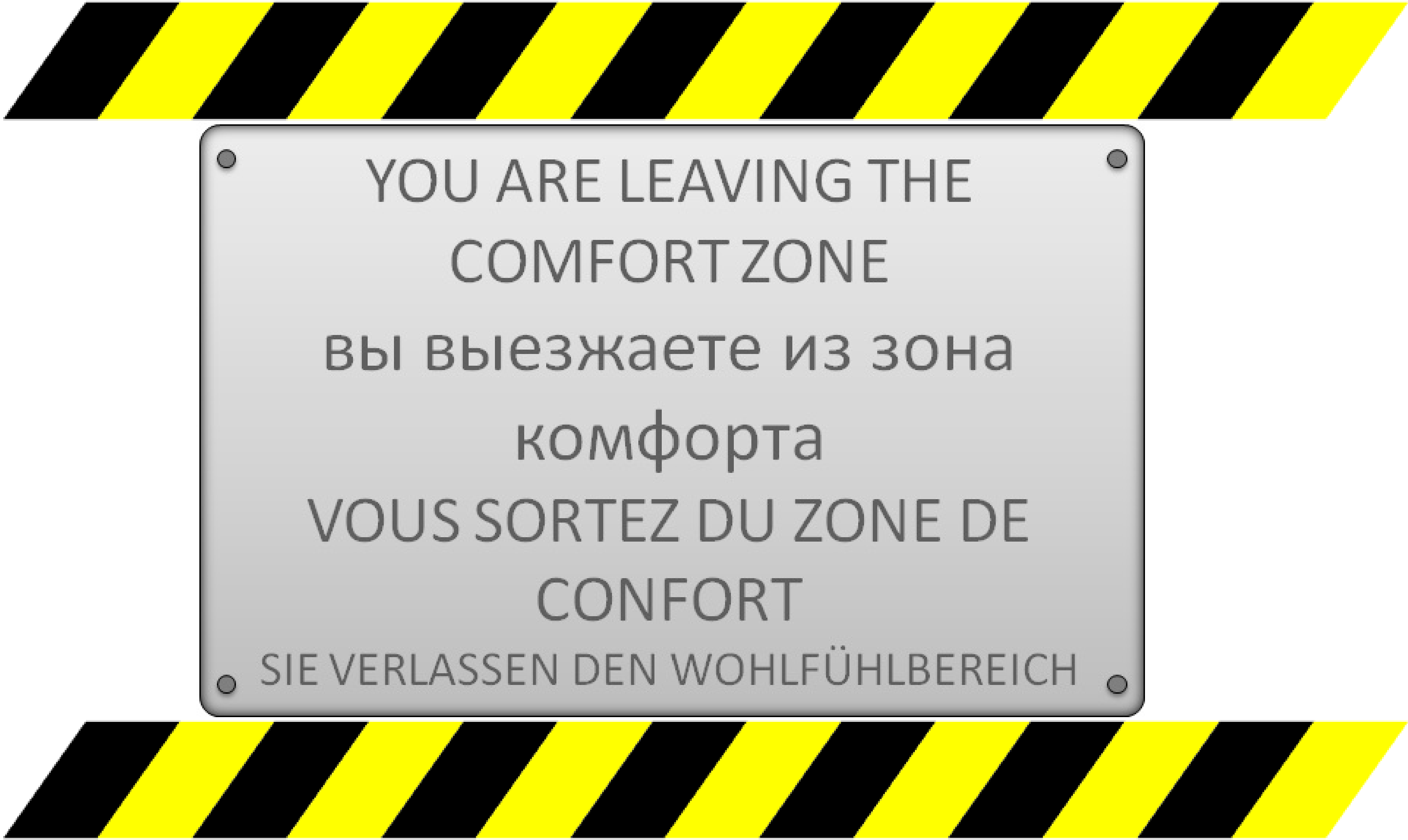
OS, standard apps, libraries
Debuggers, analysers, compilers
Scheduling system, accounting
Provisioning, user administrations

Hardware

Compute nodes
Network (Interconnect, mgt and Internet)
Storage (SAN / NAS)
Server room (Floor space, racks, power cooling, cabling)

Today's goal

- There is a world 'outside' your laptop – so let's make the first steps!


YOU ARE LEAVING THE
COMFORT ZONE
вы выезжаете из зона
комфорта
VOUS SORTEZ DU ZONE DE
CONFORT
SIE VERLASSEN DEN WOHLFÜHLBEREICH

The DTU Unix systems

a remote databar

also known as G-bar

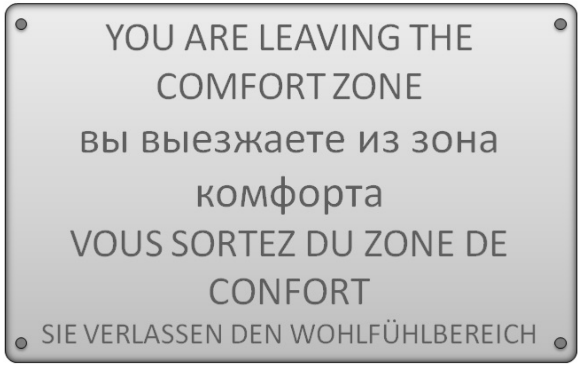
The DTU Unix systems

Access to the system

- Remote access – from everywhere:
  - ThinLinc remote desktop session:
    - download ThinLinc client from www.thinlinc.com
    - connect to thinlinc.gbar.dtu.dk
    - preferred way, if you work a lot with GUIs
    - on mobile devices: https://thinlinc.gbar.dtu.dk/
  - Secure SHell (ssh) connection (login2.gbar.dtu.dk)
    - for the command line oriented users   :-)
- On Campus:
  - ThinLinc from Windows computers

The DTU Unix systems

- But why – and when – should you use it?
- Sooner or later you ...
  - want to solve problems that don't fit your laptop any longer!
  - have to solve many problems, and you run out of time!
  - want to use a software, that doesn’t run on your PC.
- So why not start today ... ???
  - it's free!
  - trust me: *“it doesn't hurt!”*



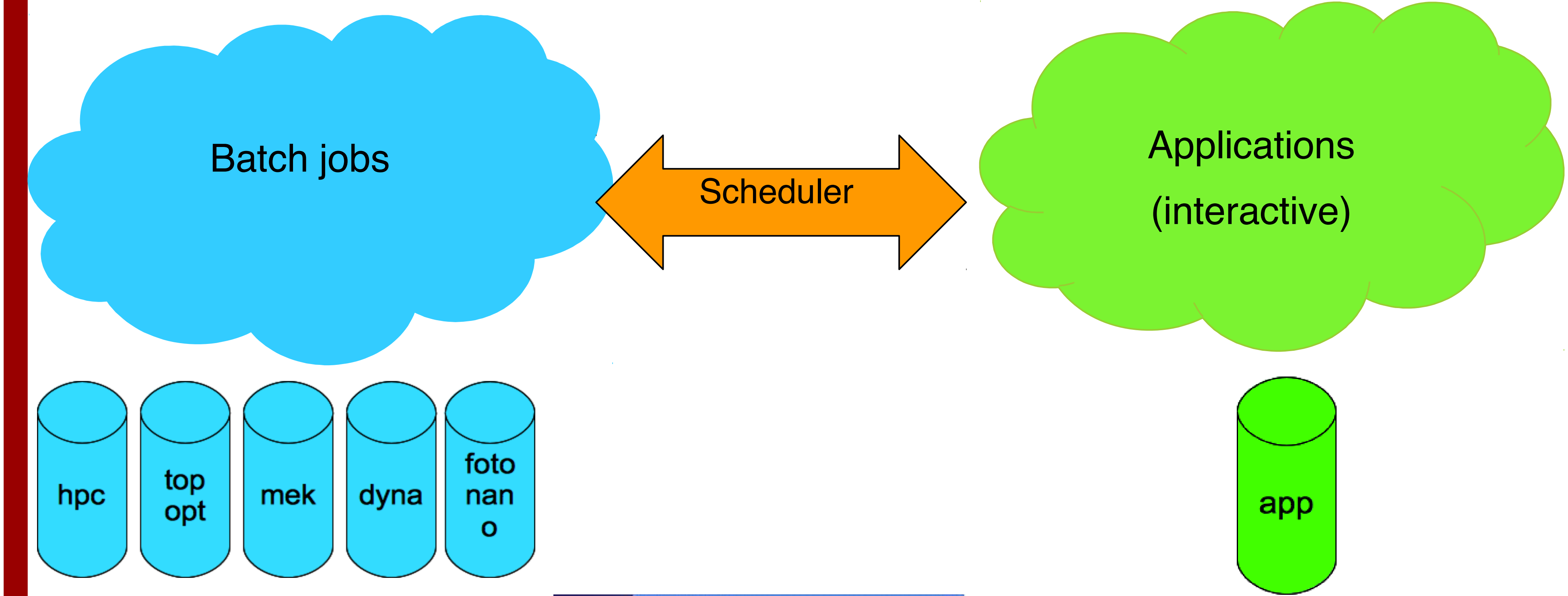
YOU ARE LEAVING THE COMFORT ZONE

вы выезжаете из зона комфорта

VOUS SORTEZ DU ZONE DE CONFORT

SIE VERLASSEN DEN WOHLFÜHLBEREICH

Our Setup



Batch jobs

Scheduler

Applications

(interactive)

hpc

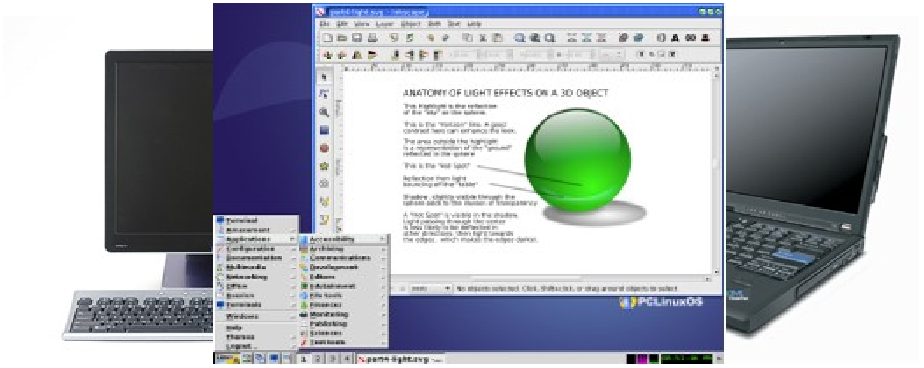
top opt

mek

dyna

foto nan o

app



The central DTU Unix systems

- Application servers – x86_64 based:
  - 6 Huawei XH620 V3 (2x Xeon E5-2660 v3 2.6 GHz)
  - 6 Dell PowerEdge FC430 (2x Xeon E5-2670 v3 2.3 GHz)
  - Scientific Linux 7.x
- Desktop servers (ThinLinc):
  - 3 servers (4x AMD Opteron 6376, 2.4 GHz)
- 10000+ users (students + employees)
  - “everybody has access”!!!

The DTU Unix systems

- HPC servers (for 'everybody'), e.g.
  - 40 IBM NeXtScale nx360 M4 (2x Xeon E5-2680v2 2.8 GHz, 128 GB memory)
  - 15 Huawei XH620 V3 (2x Xeon E5-2660v3 2.6 GHz, 128 GB memory)
  - 40 Huawei XH620 V3 (2x Xeon E5-2650v4 2.2 GHz, 256 GB memory)
  - 24 Lenovo ThinkSystem SD530 (2x Xeon Gold 6126 2.6 GHz, 192-384 GB memory)
- + “private” clusters
  - DTU Compute, DTU Nanotech, DTU Photonics, DTU Chemistry, DTU Elektro, DTU Environment, DTU Management, ...

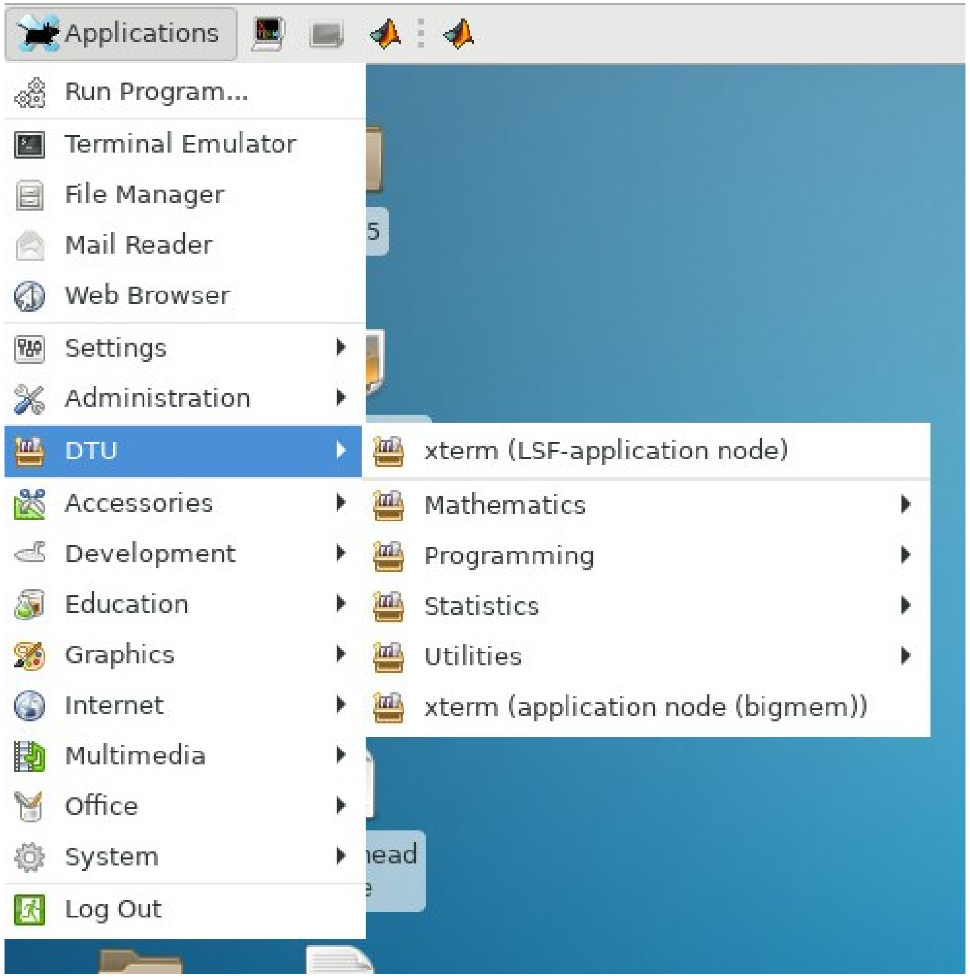
The DTU Unix systems

- HPC servers DTU Compute
  - 26 HP Proliant SL230s (2x Xeon E5-2665 2.4 GHz, 128 GB memory)
  - 7 IBM NextScale nx360 M5 (2x Xeon E5-2660v3 2.6 GHz, 128 GB memory)
  - 27 Huawei XH620 V3 (2x Xeon E5-2660v3 2.6 GHz, 128 GB memory)
  - 30 Huawei XH620 V3 (2x Xeon E5-2650v4 2.2 GHz, 256 GB memory)
  - 1 node with 1TB memory
  - nodes with NVIDIA V100 GPUs

Interactive Work

Interactive Work

- Starting applications from the desktop menu
- Most desktop apps will run on the same server as your desktop (th1, ..., th3)
- Applications from the DTU menu get dispatched on to one of the application nodes



Interactive Work

- “Code of conduct”:
  - applications from the menu run on multi-user systems – so please behave
  - there are resource limits, too:
    - CPU usage: max. 24 hours CPU time
    - memory: max. 16 GB
- If you need more than that, you have to use the batch system!

Access to your files

- To be able to work on the Unix systems, you'll need to transfer your files
- There are several ways to do that:
    - copying: WinSCP, FileZilla, scp, ...
    - mounting your Unix home-directory on your laptop
        - works while on a DTU network, e.g. eduroam
        - for more details see http://gbar.dtu.dk/faq/78-home-directory
- Best practices:
    - avoid spaces and national characters in file names
    - some applications require Unix text format

Software

What is availabe?

Installed Software

- Which software is available?
  - quick answer: a lot!
- How can I find it?
  - can be tricky ... ;-)
  - only the ‘basic’ things are in the desktop menu
  - most programs need to be started from the terminal
  - if in doubt, look in /appl (250+ packages installed)
- Which version of XYZ is installed?
  - check /appl/XYZ for subfolders
  - ... or use ‘modules ...’ (next slide)

Using modules

- modules help to organize certain Unix environment settings, e.g. PATH, MANPATH, LD_LIBRARY_PATH, etc. for different versions of the same application
- list available modules: `module avail`
- load a module: `module load python3`
- swap a version: `module swap python3/3.5.4`
- swap to default: `module swap python3`
- info: http://gbar.dtu.dk/index.php/faq/83-modules

The (new) DCC software stack

- The default module setup is
  - overwhelming
  - confusing
  - not very user friendly
- Try the new DCC software stack:
  - more structured
  - only two releases per year
  - try it: source /dtu/sw/dcc/dcc-sw.bash
  - feedback is welcome!

More Software

- What if it is not there?
  - install in your $HOME folder
  - open a ticket with support, and we will try to help
- Things we do not support:
  - containers (Docker, Singularity) – might come ...
  - Windows software

Lab time ...

THE FUTURE
BELONGS TO
THE FEW OF US
STILL WILLING
TO GET OUR
HANDS DIRTY.

Using the system

- On-line demo:
  - logging in
  - download ZIP file: https://bit.ly/2PbA4T2
    - wget -O ImageXmasWS.zip https://bit.ly/2PbA4T2
  - create folder
  - unzip downloaded file

Batch jobs

Starting point: Take a look at www.hpc.dtu.dk

Resource Managers

To handle the workload on an HPC installation, one needs a tool to manage and assign the resources: a Resource Manager – sometimes also called 'batch queue system'

- Most common systems:
  - Torque/PBS (ext. scheduler, like Maui or MOAB)
  - LSF
  - Grid Engine
  - Slurm

Resource Managers

Before submitting a job, one has to specify the resources needed, e.g.

- ❐ # of CPUs/cores
- ❐ amount of memory
- ❐ expected run time (wall-clock time)
- ❐ other resources, like disk space, GPUs, etc

This is done in a special job script and is system (RM) dependent – but very similar for all RMs.

Resource Managers

- Examples for the DTU batch system, based on
  - ~~MOAB (scheduler) and Torque (resource manager)~~
  - Spectrum LSF (setup used here!)
    - info: HPC User Guides for LSF
- Notes:
  - you need to be on the correct front-end node, either via ThinLinc or ‘ssh login2.hpc.dtu.dk’
  - you cannot submit executables directly, you have to use a job script!
  - don't expect jobs to start immediately – the scheduler has to find free resources first!

Resource Managers

The simplest job script:

```
#!/bin/bash
sleep 60
```

submit.sh

```
$ bsub < submit.sh
Job <702572> is submitted to default queue <hpc>.

$ bstat
JOBID  USER   QUEUE   JOB_NAME SLOTS STAT  START_TIME       ELAPSED
702572 gbarbd hpc     NONAME       1 RUN   Dec 13 12:17     0:00:00
$ bjobs
JOBID  USER   QUEUE   JOB_NAME SLOTS STAT  START_TIME   TIME_LEFT
702572 gbarbd hpc     NONAME       1 RUN   Dec 13 12:17 00:15:00 L
$ ls -g
total 4
-rw-r--r-- 1 gbar 1493 Dec 13 12:18 NONAME_702572.out
-rw-r--r-- 1 gbar   22 Dec 13 12:05 simple.sh
```

Resource Managers

The simplest job script – the full story:

```
#!/bin/bash
sleep 60
```

simple.sh

```
$ bsub < simple.sh
bsub info: Job has no name! Setting it to NONAME!
bsub info: Job has no wall-clock time! Setting it to 15 minutes!
bsub info: Job has no output file! Setting it to NONAME_%J.out!
bsub info: Job has no memory requirements! Setting it to 1024 MB!
bsub info:   You need to specify at least -R "rusage[mem=...]"!
Job <702608> is submitted to default queue <hpc>.
```

Resource Managers

A simple job script:

```
#!/bin/bash
#BSUB -J sleeper
#BSUB -o sleeper_%J.out
#BSUB -q hpc
#BSUB -W 2
#BSUB -R “rusage[mem=512MB]”

sleep 60
```

```
$ bsub < submit.sh
Job <702645> is submitted to queue <hpc>.
```

```
$ ls -g
total 3
-rw-r--r-- 1 gbar  121 Dec 13 12:32 submit.sh
-rw-r--r-- 1 gbar 1592 Dec 13 12:36 sleeper_702646.out
```

Resource Managers

- The output file:

```
Sender: LSF System <lsfadmin@n-62-21-20>
Subject: Job 702646: <sleeper> in cluster <dcc> Done

Job <sleeper> was submitted from host <hpclogin3> by user <gbarbd> in
cluster <dcc> at Wed Dec 13 12:34:59 2017.
Job was executed on host(s) <n-62-21-20>, in queue <hpc>, as user
<gbarbd> in cluster <dcc> at Wed Dec 13 12:34:59 2017.
</zhome/../../...> was used as the home directory.
</zhome/../../.../02614/Batch/LSF> was used as the working directory.
Started at Wed Dec 13 12:34:59 2017.
Terminated at Wed Dec 13 12:36:00 2017.
Results reported at Wed Dec 13 12:36:00 2017.

Your job looked like:

------------------------------------------------------------

# LSBATCH: User input
#!/bin/bash
#BSUB -J sleeper
#BSUB -o sleeper_%J.out
```

Resource Managers

- The output file (cont’d):
  - job summary

```
Successfully completed.

Resource usage summary:

    CPU time :                                    0.28 sec.
    Max Memory :                                  4 MB
    Average Memory :                              4.00 MB
    Total Requested Memory :                      512.00 MB
    Delta Memory :                                508.00 MB
    Max Swap :                                    -
    Max Processes :                               4
    Max Threads :                                 5
    Run time :                                    65 sec.
    Turnaround time :                             61 sec.

The output (if any) is above this job summary.
```

Resource Managers

Separating output and errors:

```
#!/bin/bash
#BSUB -J sleeper
#BSUB -o sleeper_%J.out
#BSUB -e sleeper_%J.err
#BSUB -q hpc
#BSUB -W 2 -R "rusage[mem=512MB]"

rm nonexistent.txt
echo "Just a minute ..."
sleep 60
```

```
$ bsub < submit2.sh
...
$ ls -g
total 3
-rw-r--r-- 1 gbar  184 Dec 13 13:56 submit2.sh
-rw-r--r-- 1 gbar   63 Dec 13 13:59 sleeper_702793.err
-rw-r--r-- 1 gbar 1744 Dec 13 14:00 sleeper_702793.out
```

Resource Managers

Separating output, errors – and mail summary:

```
#!/bin/bash
#BSUB -J sleeper
#BSUB -o sleeper_%J.out
#BSUB -e sleeper_%J.err
#BSUB -q hpc
#BSUB -W 2 -R "rusage[mem=512MB]"

rm nonexistent.txt
echo "Just a minute ..."
sleep 60
```

```
$ bsub -N < submit2.sh
...
$ ls -g
total 3
-rw-r--r-- 1 gbar  184 Dec 13 13:56 submit2.sh
-rw-r--r-- 1 gbar   63 Dec 13 14:04 sleeper_702814.err
-rw-r--r-- 1 gbar   18 Dec 13 14:04 sleeper_702814.out
```

send summary
at end of job

Resource Managers

A simple parallel job script:

- for shared memory (on a single node), using 4 cores

```
#!/bin/bash
#BSUB -J simple_para
#BSUB -o simple_para_%J.out
#BSUB -q hpc
#BSUB -W 2
#BSUB -R "rusage[mem=512MB]"
#BSUB -n 4
#BSUB -R "span[hosts=1]"

export OMP_NUM_THREADS=$LSB_DJOB_NUMPROC
...
```

- Note: the mem=xyzMB is per core!

Resource Managers

more options and examples:

- see http://www.hpc.dtu.dk/ under
  - LSF User Guides
    - http://www.hpc.dtu.dk/?page_id=2534
- do the lab exercises
- use 'man bsub', 'man bjobs', etc

Resource Managers

DTU Computing Center specific commands:

- bstat – shows the status of your jobs; use ‘bstat -h’ for help for other options
- classstat – shows the status of the queues, e.g. free and used cores, pending jobs, etc
- nodestat – shows the current status of all nodes (use 'nodestat hpc' for the nodes of the 'hpc' queue)
- all commands above have a ‘help option’ (-h), but no man-page!

“Big Data”: where to put my (large) datasets?

Large datasets: where & how to

- Limitations of your home folder ($HOME):
  - 30 GB quota
  - ***“slow”*** network file system
- $HOME is backed up
- there are snapshots (hourly, daily, ...) available
- not really suitable for
  - large files
  - temporary files (e.g. during computations)
  - files that change often

Large datasets: where & how to

- The /workN/$USER folders (N = 1,3):
  - a parallel filesystem
  - faster than $HOME
  - only on request
    - but we created one for everybody present today (in /work1)
- Limitations:
  - no backup – and no snapshots(!)
  - good for large files
  - avoid too many files in one folder (true for $HOME, too)
- Don’t use it as permanent storage!!!

Large datasets: where & how to

- How do I transfer my files?
  - use ‘transfer.gbar.dtu.dk’
  - fast uplink from DTU network
  - fast link to all DCC storage systems, i.e. $HOME and /workN
- Do not use transfer host for anything else!
- Do not use the login nodes for large data transfers
  - it works, but it is slow – and it slows down everybody else, too!

Large datasets: where & how to

Other possibilities:

1) access to files that are on some DTU Compute storage
   - ❒ possible via project folders under /dtu-compute/
   - ❒ needs an agreement with DTU Compute IT
2) shared filesystem(s) with other installations at DTU
   - ❒ shared storage with Niflheim (DTU Fysik)
3) bring your own money
   - ❒ ... and we'll buy storage for you

Large datasets: where & how to

Please note:

- please don’t upload data that has to follow any GDPR regulations!

Misconceptions
&
Misunderstandings

The HPC system is always faster!

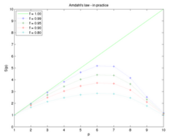
- Not always true!
- If you have a brand-new laptop, your CPU is probably one or two generations more advanced than the CPUs in the HPC setup ...
- ... but on the HPC system you have access to more resources/cores, than the usual two cores in your laptop!
- Limits: ~100 concurrent cores (batch system)
  - i.e. 100 single core jobs
  - or ten 10-core jobs, or ...

There is more memory available!

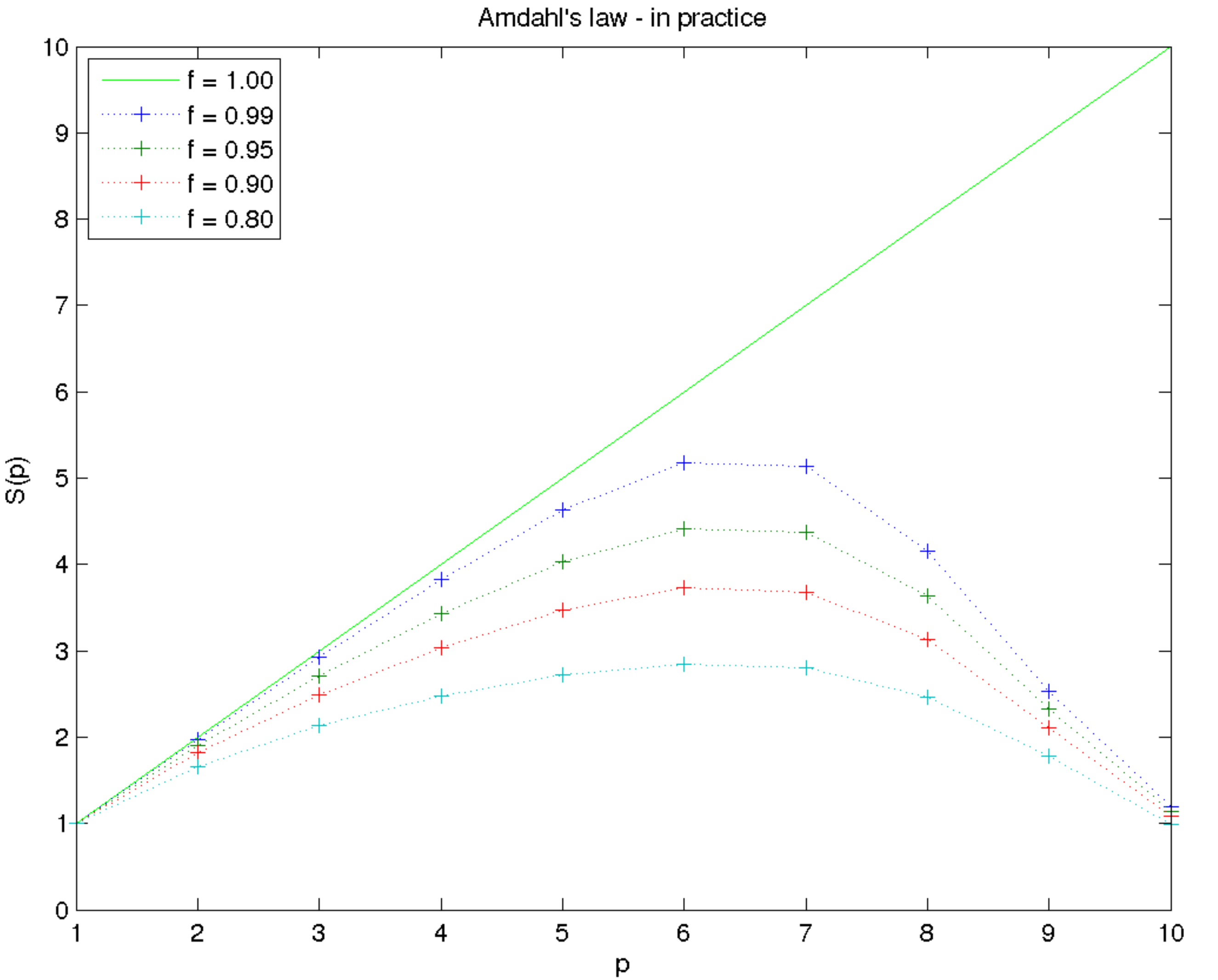
- True ... but there are limits as well!
  - we have machines with 64 GB to 256 GB memory
  - Don’t ask for the full memory – leave space for the OS, etc (rule of thumb: ~10%, e.g. 240GB on a 256GB machine)
- Always try to estimate your memory needs, and specify those needs in your job scripts.
- With LSF, we force you to specify your memory needs – and we enforce the limits!
- This helps the scheduler to dispatch to nodes, that can cope with the amount of memory needed.

The more cores – the better ...

- Mostly wrong!
- Your application has to be able to use the cores, i.e. it has to be parallel
  - requesting more than one core for a serial application is a waste of resources
- You might need to activate the parallelism, e.g. via command line options
- Remember Amdahl's law ...
- You'll probably have to run some tests to find the "optimal" number of cores



Amdahl’s law


Amdahl's law - in practice
f = 1.00
f = 0.99
f = 0.95
f = 0.90
f = 0.80
S(p)
p
0
1
2
3
4
5
6
7
8
9
10

“Somebody told me that ...”

- ... but it's probably wrong!
- In the support team, we see the same mistakes by users, made over and over again – and that is not productive! Neither for you, nor for the other users, e.g. if you block resources you do not need!
- If you are in doubt, check the web pages, or ask the support people – they are there to help, and just an e-mail away!

"None of my jobs failed ..."

- ... but still there is something wrong!
  - e.g. no output, incomplete output, etc

- LSF reports job success (DONE) or failure (EXIT)
- this information is taken from the return code of the last command in your batch script!
- Things like "echo done." or "date" will always return success – and that will hide problems!
- Conclusion: make your application the last command in the batch script!

Summary

Summary – DTU Unix Systems

- a general computing resource
- Linux based
- accessible
  - from everywhere
  - for everybody with a DTU account
- extends the compute power of your own computer/laptop
- Not to forget: your data is backed up!

Where to get help?

- Today:
  - ask us – we are here to help!
- Later:
  - check our webpages:
    - www.hpc.dtu.dk
    - www.gbar.dtu.dk
  - send e-mail to:
    - support@hpc.dtu.dk (HPC related help)
    - support@cc.dtu.dk (general help)