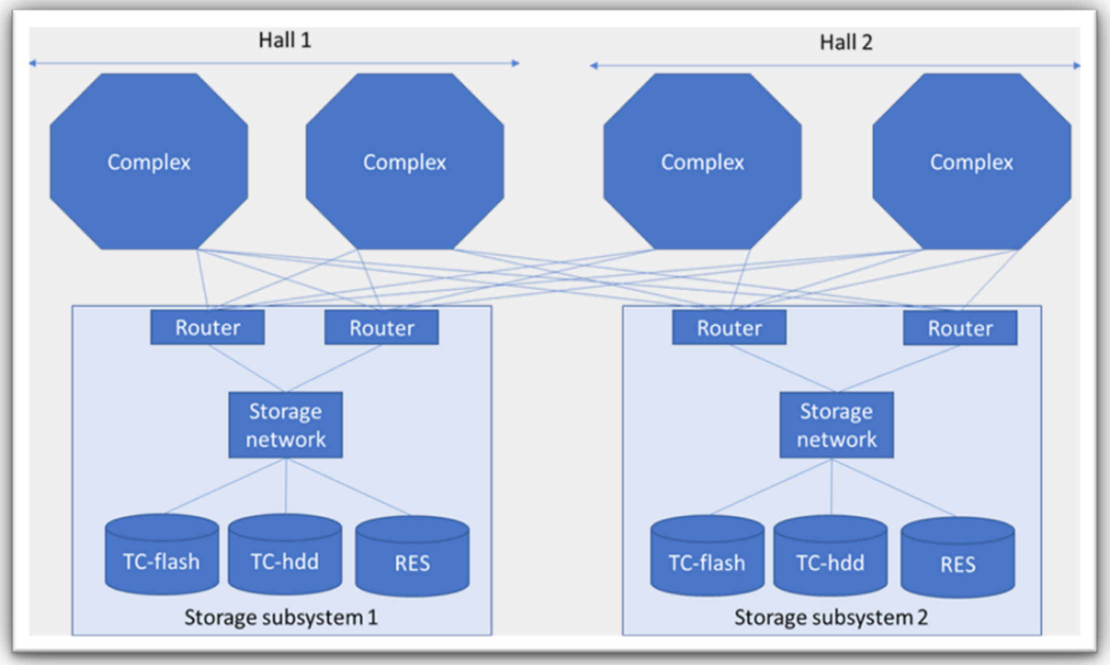
Atos Environment (and more)

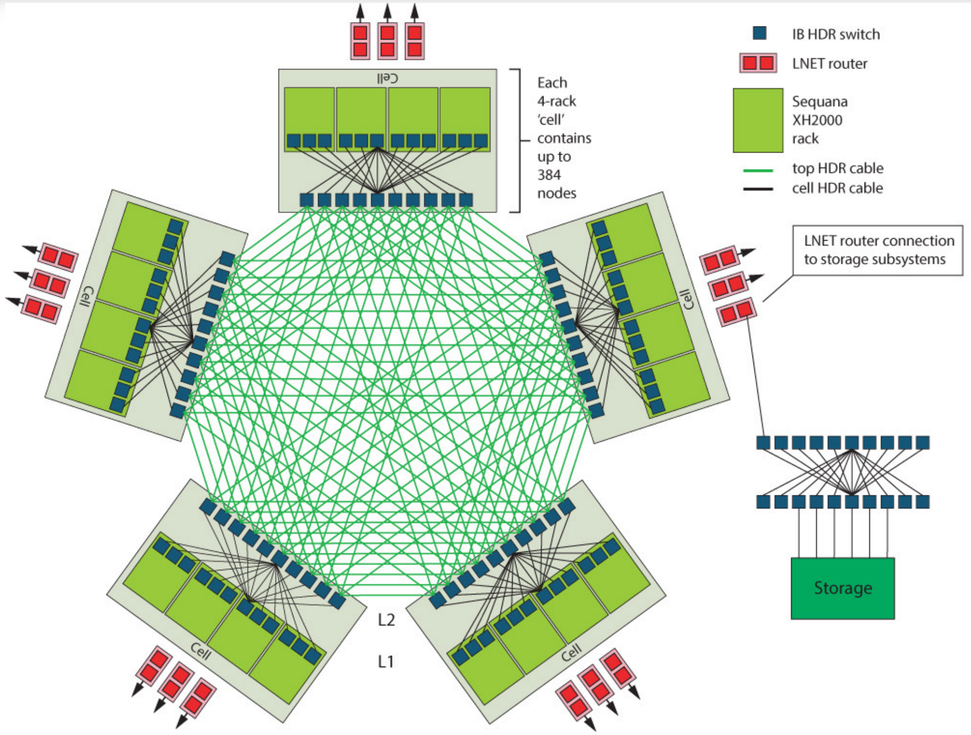
Xavier Abellan

xavier.abellan@ecmwf.int

Atos BullSequana XH2000

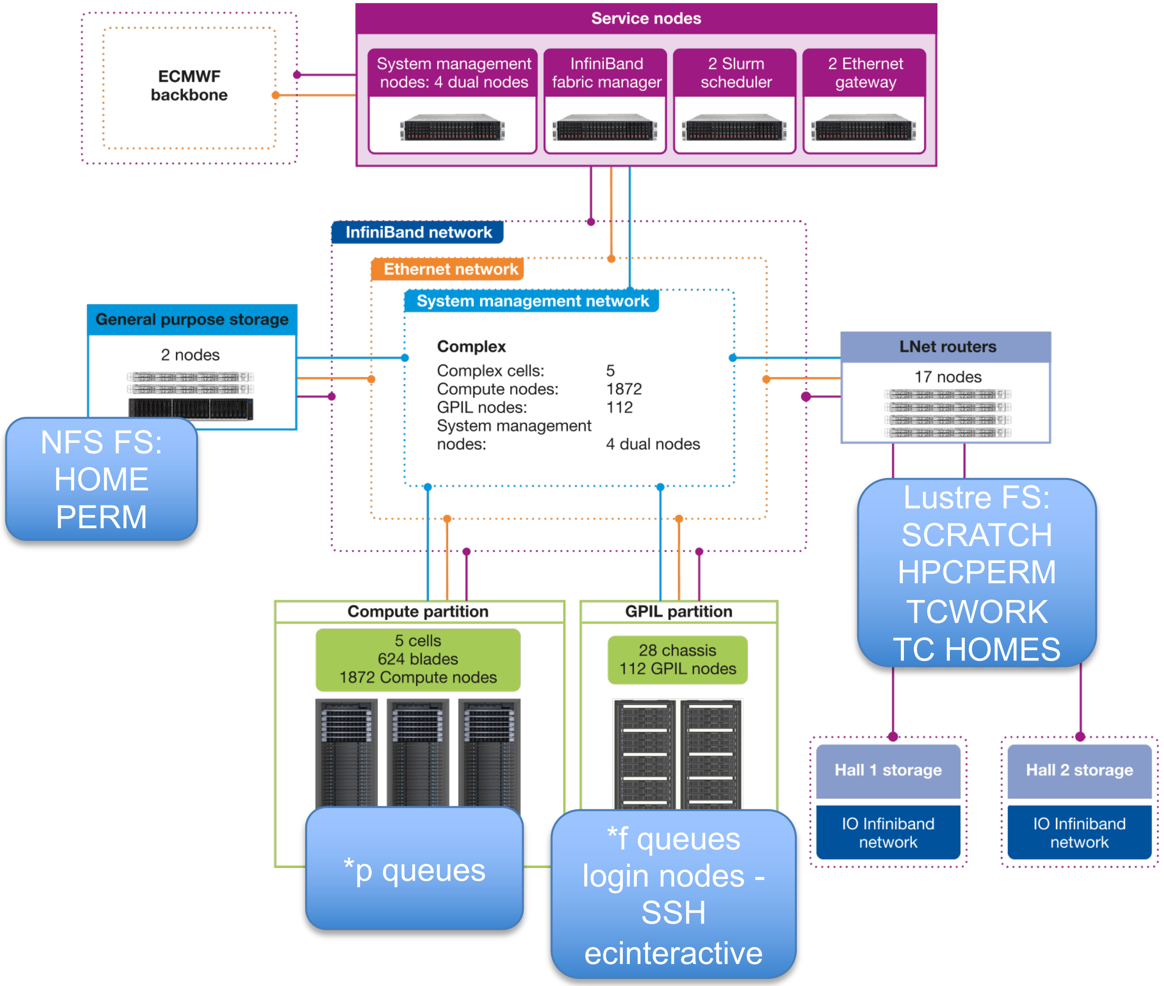
- 4 Complexes
  - Two in each hall
  - Each Complex consists of two partitions:
  - Parallel:
    - ATOS XH2000 Water cooled racks
    - Arranged in 5 “cells”, 4 racks per cell
    - IB HDR Fat Tree in each cell. Each cell connected to every other cell
    - 1920 nodes for parallel compute
    - AMD Rome 64 core processors
  - General Purpose (GPIL)
    - 112 nodes for general purpose use
      - More memory, local SSD
- One Slurm scheduler in each complex





The Atos HPCF

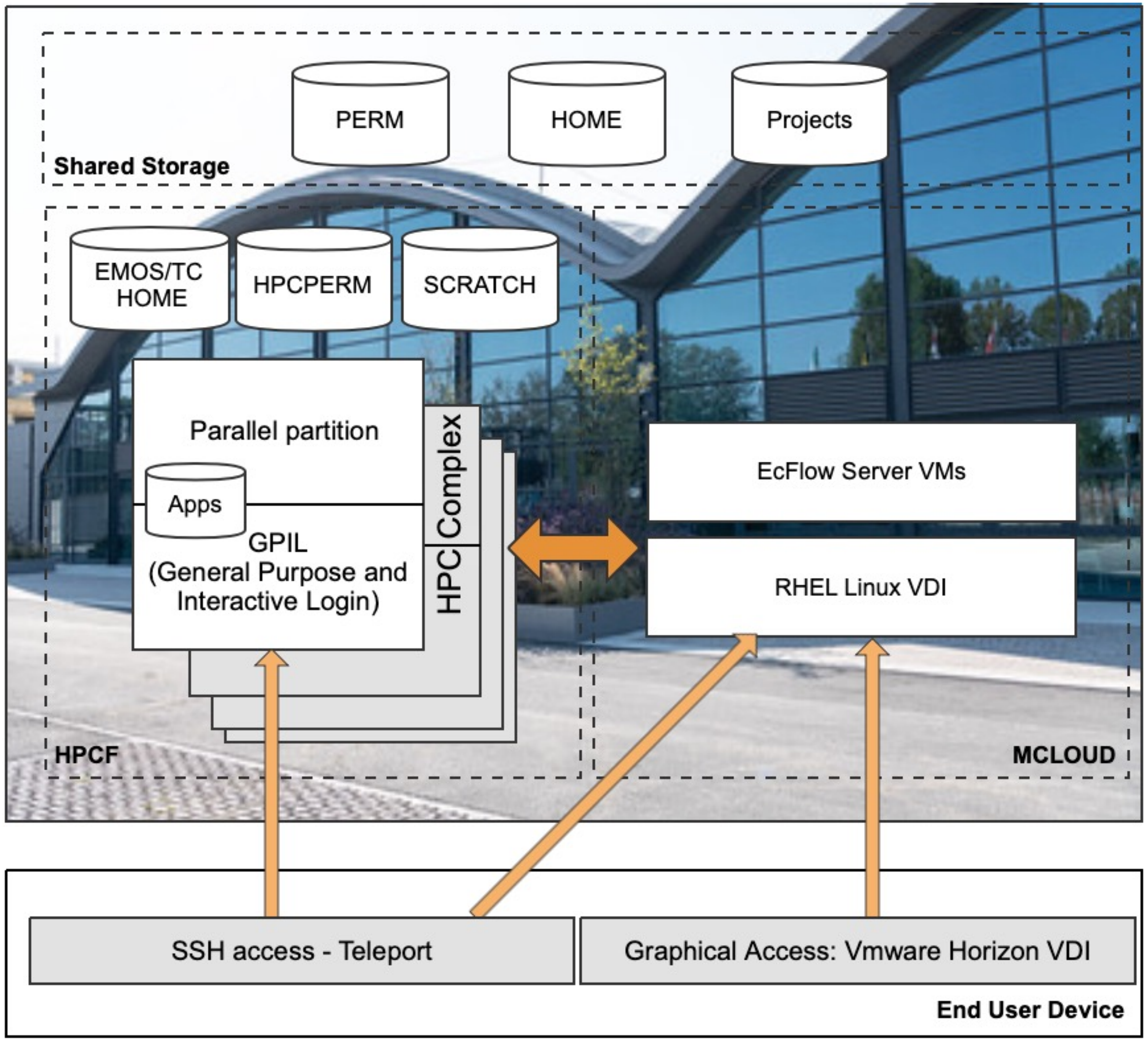
- 4 HPCF complexes: A[ABCD]
- ECS (ECGATE Class Service) “virtual” cluster
  - For users with no formal HPCF access
  - Nodes from 4 complexes
  - Same Apps and Filesystems as main complexes
  - Independent Slurm Batch system
    - Serial or very small parallel workloads
    - No SBU billing



Old vs New HPCF

	Cray	Atos
Performance factor	1	4.67
Clusters	2	4
Compute nodes	7,020	7,680
General purpose nodes	208	448
Processor type	Intel Broadwell	AMD EPYC Rome
Cores per node	36	128
Memory per node (GiB)	128	256 (compute) / 512 (general purpose)
Total cores	260,208	1,040,384
Total memory (PiB)	0.88	2.2
Parallel storage type	HDD Lustre	HDD & SSD Lustre
Total parallel storage (PB)	22	90
Total storage bandwidth	355 GB/s	2,408 GB/s

The new remote working model



The new remote working model: SSH service

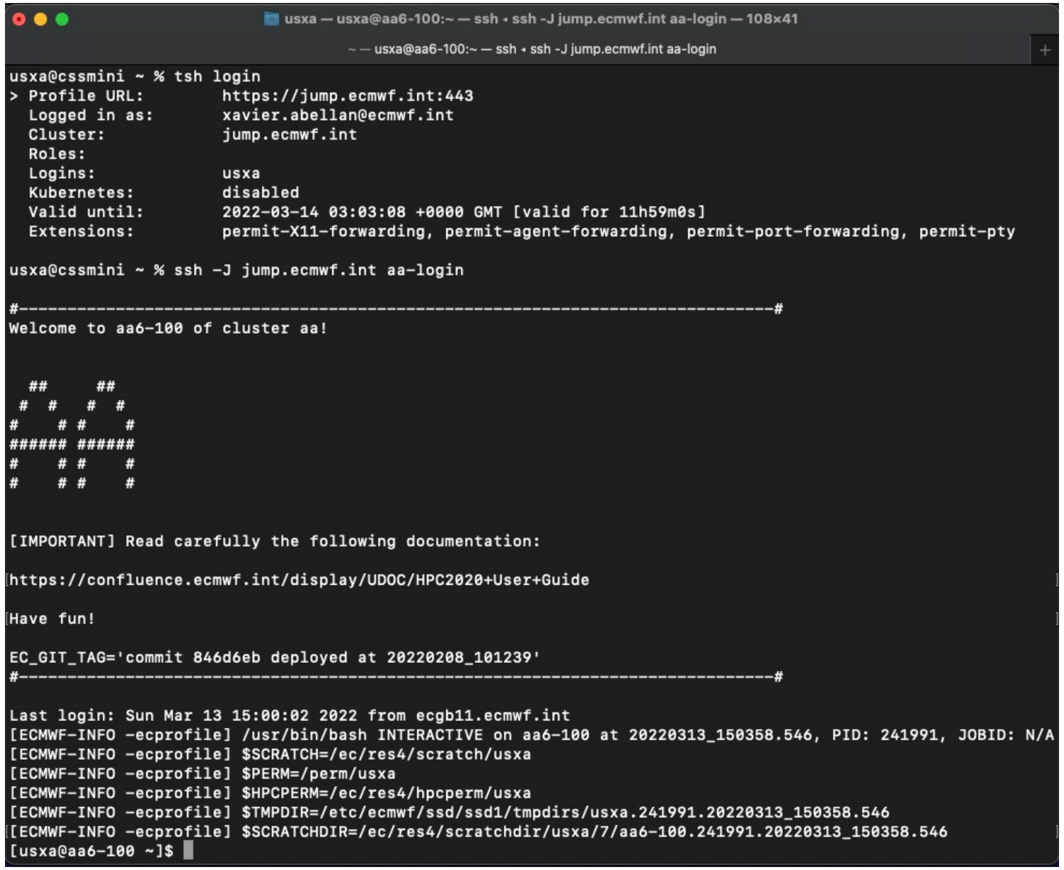
The Teleport service replaces ECACCESS SSH service, and provides:

- Single SSH hop from client systems anywhere on the internet to ECMWF servers
- Re-authentication required only every 12 hours (once per day)
- Integration with standard tools such as the OpenSSH ssh client, scp, ssh-agent and rsync
- Web-SSH interface for in-browser terminal access, with scp
- X11 and Port forwarding

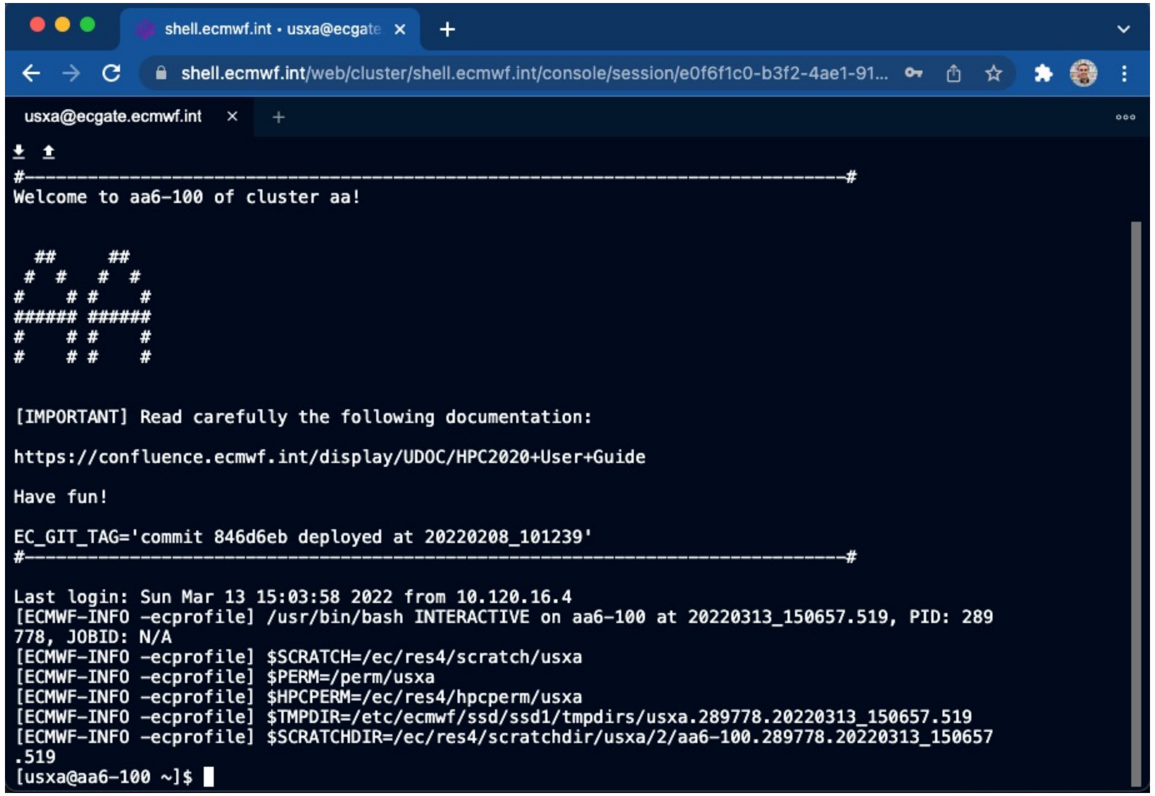
For Command line access, **tsh** client needs to be installed for the single sign-on step.

- A browser window will pop up for you to authenticate (2-factor) into ECMWF website

The new remote working model: SSH service

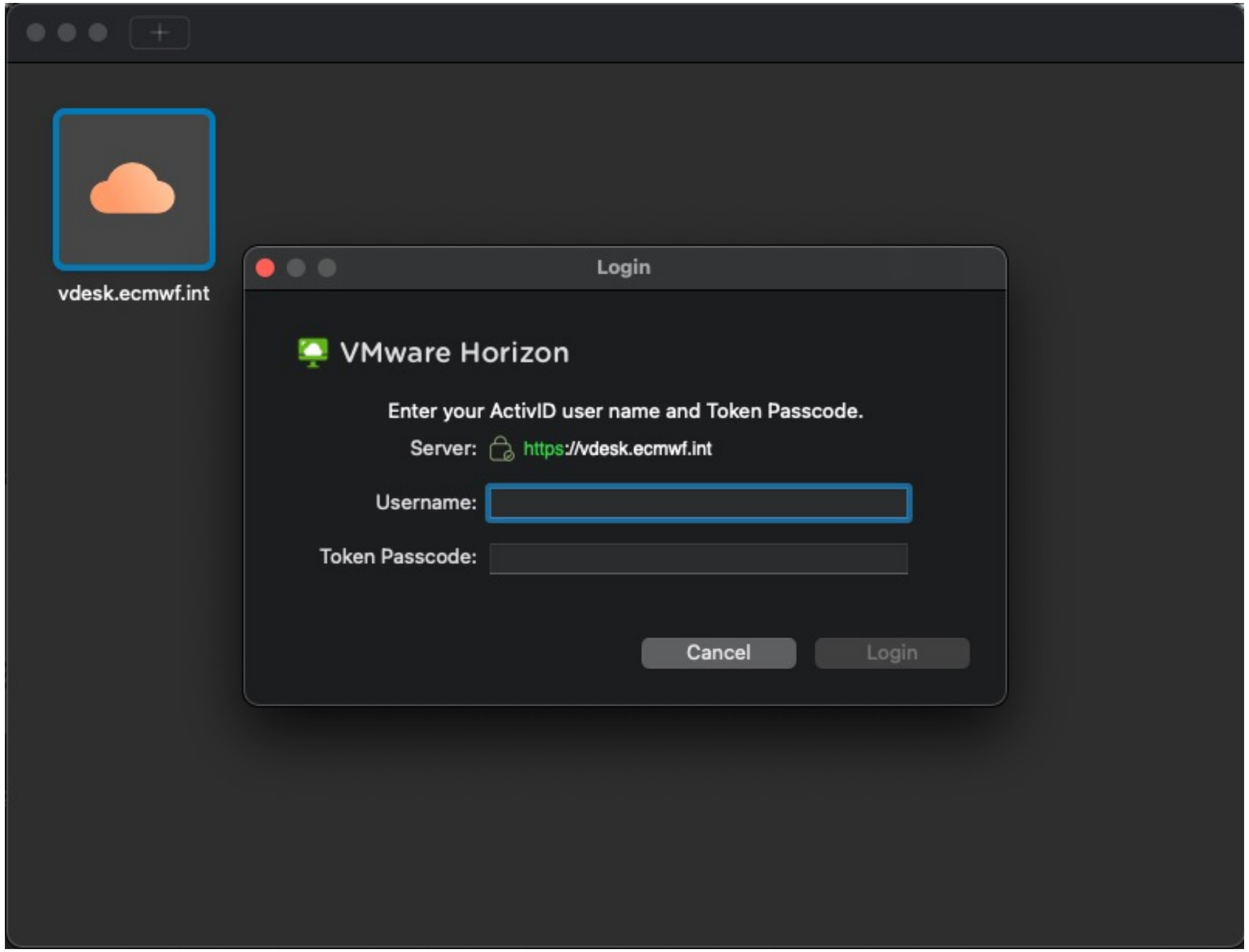

usxa — usxa@aa6-100:~ — ssh ◂ ssh -J jump.ecmwf.int aa-login — 108×41
~ — usxa@aa6-100:~ — ssh ◂ ssh -J jump.ecmwf.int aa-login

```
usxa@cssmini ~ % tsh login
> Profile URL:        https://jump.ecmwf.int:443
  Logged in as:       xavier.abellan@ecmwf.int
  Cluster:            jump.ecmwf.int
  Roles:
  Logins:             usxa
  Kubernetes:         disabled
  Valid until:        2022-03-14 03:03:08 +0000 GMT [valid for 11h59m0s]
  Extensions:         permit-X11-forwarding, permit-agent-forwarding, permit-port-forwarding, permit-pty

usxa@cssmini ~ % ssh -J jump.ecmwf.int aa-login

#-------------------------------------------------------------------------------#
Welcome to aa6-100 of cluster aa!


  ##     ##
 #  #   #  #
#    # #    #
###### ######
#    # #    #
#    # #    #


[IMPORTANT] Read carefully the following documentation:

https://confluence.ecmwf.int/display/UDOC/HPC2020+User+Guide

Have fun!

EC_GIT_TAG='commit 846d6eb deployed at 20220208_101239'
#-------------------------------------------------------------------------------#

Last login: Sun Mar 13 15:00:02 2022 from ecgb11.ecmwf.int
[ECMWF-INFO -ecprofile] /usr/bin/bash INTERACTIVE on aa6-100 at 20220313_150358.546, PID: 241991, JOBID: N/A
[ECMWF-INFO -ecprofile] $SCRATCH=/ec/res4/scratch/usxa
[ECMWF-INFO -ecprofile] $PERM=/perm/usxa
[ECMWF-INFO -ecprofile] $HPCPERM=/ec/res4/hpcperm/usxa
[ECMWF-INFO -ecprofile] $TMPDIR=/etc/ecmwf/ssd/ssd1/tmpdirs/usxa.241991.20220313_150358.546
[ECMWF-INFO -ecprofile] $SCRATCHDIR=/ec/res4/scratchdir/usxa/7/aa6-100.241991.20220313_150358.546
[usxa@aa6-100 ~]$ 
```


shell.ecmwf.int • usxa@ecgate
shell.ecmwf.int/web/cluster/shell.ecmwf.int/console/session/e0f6f1c0-b3f2-4ae1-91...
usxa@ecgate.ecmwf.int

```
#-------------------------------------------------------------------------------#
Welcome to aa6-100 of cluster aa!


  ##     ##
 #  #   #  #
#    # #    #
###### ######
#    # #    #
#    # #    #


[IMPORTANT] Read carefully the following documentation:

https://confluence.ecmwf.int/display/UDOC/HPC2020+User+Guide

Have fun!

EC_GIT_TAG='commit 846d6eb deployed at 20220208_101239'
#-------------------------------------------------------------------------------#

Last login: Sun Mar 13 15:03:58 2022 from 10.120.16.4
[ECMWF-INFO -ecprofile] /usr/bin/bash INTERACTIVE on aa6-100 at 20220313_150657.519, PID: 289
778, JOBID: N/A
[ECMWF-INFO -ecprofile] $SCRATCH=/ec/res4/scratch/usxa
[ECMWF-INFO -ecprofile] $PERM=/perm/usxa
[ECMWF-INFO -ecprofile] $HPCPERM=/ec/res4/hpcperm/usxa
[ECMWF-INFO -ecprofile] $TMPDIR=/etc/ecmwf/ssd/ssd1/tmpdirs/usxa.289778.20220313_150657.519
[ECMWF-INFO -ecprofile] $SCRATCHDIR=/ec/res4/scratchdir/usxa/2/aa6-100.289778.20220313_150657
.519
[usxa@aa6-100 ~]$ 
```

The new remote working model: VDI service

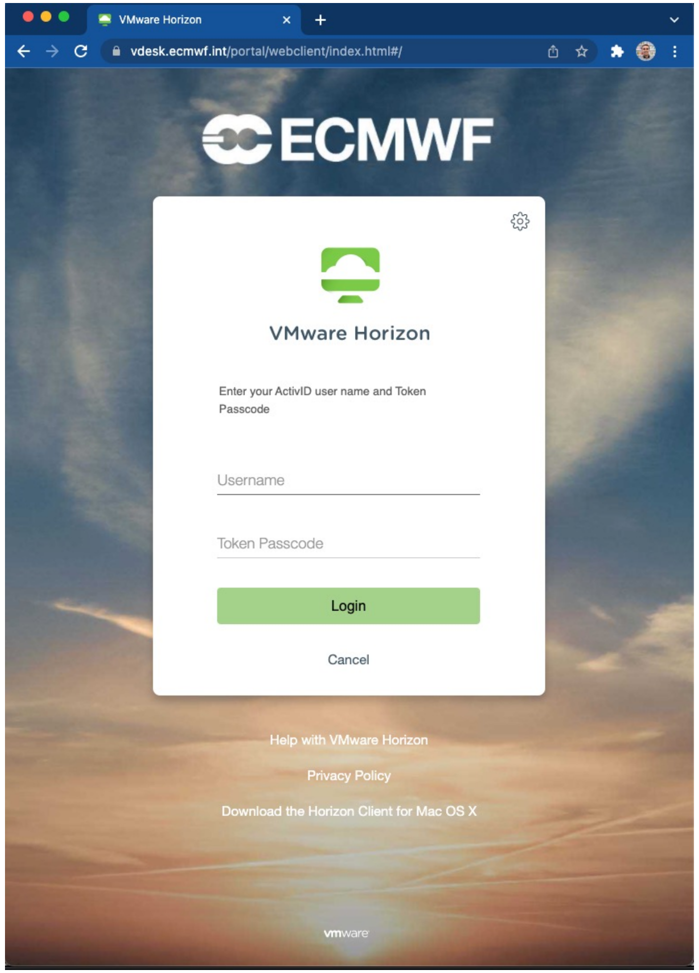


Enter your ActivID user name and Token Passcode.

Server: https://vdesk.ecmwf.int

Username:

Token Passcode:

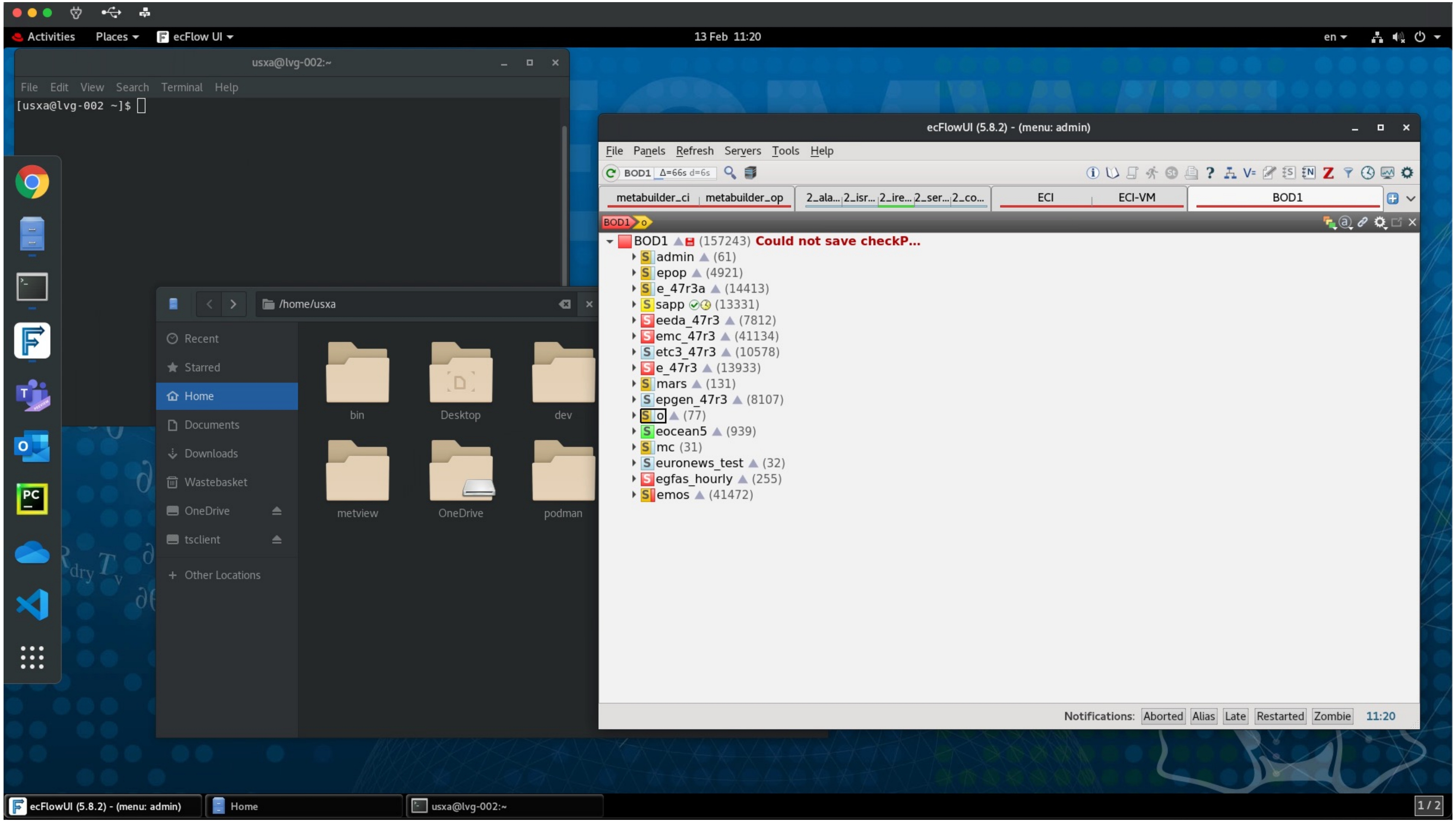


Enter your ActivID user name and Token Passcode

Username

Token Passcode

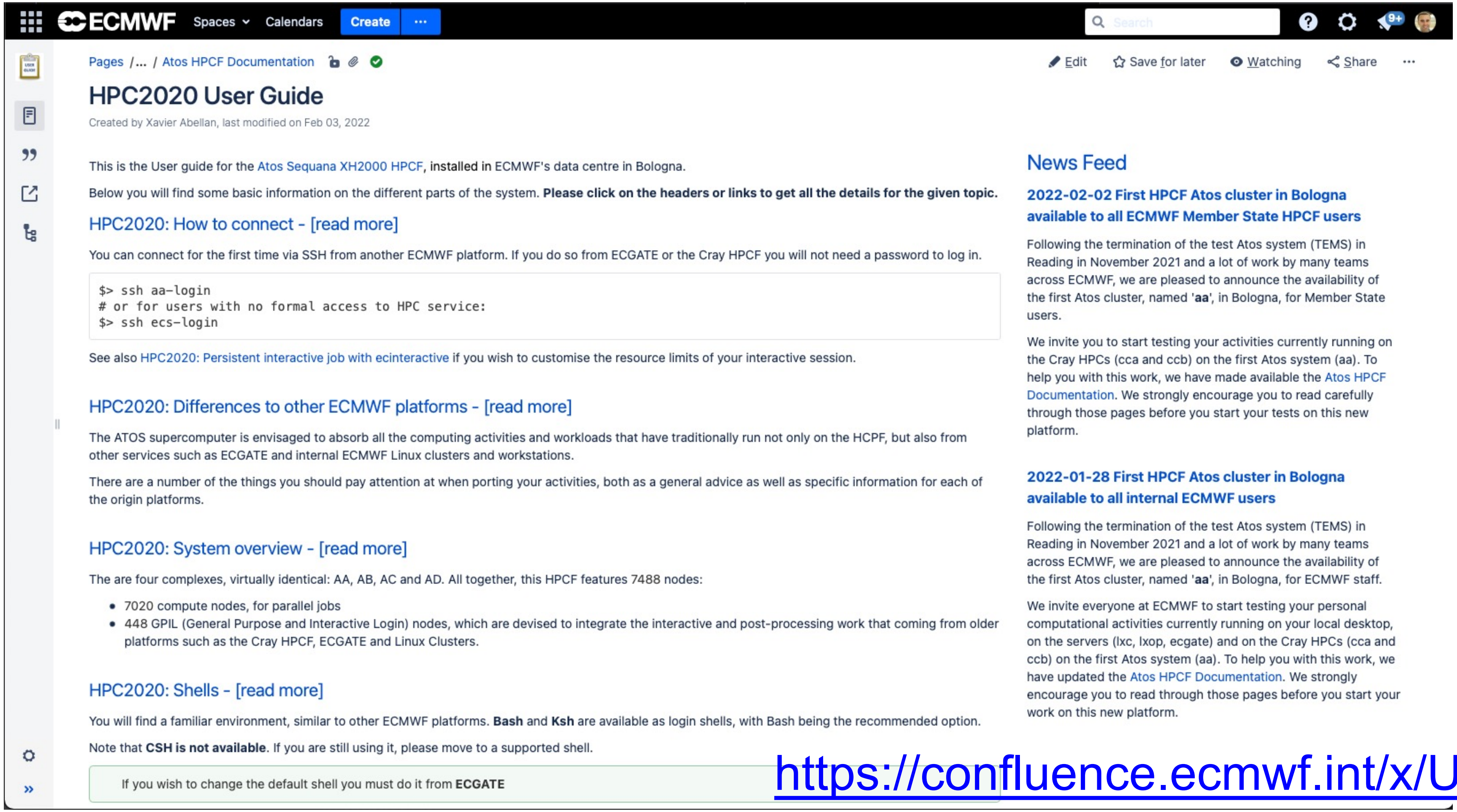
The new remote working model: VDI service



```
[usxa@lvg-002 ~]$
```

```
BOD1 (157243) Could not save checkP...
  S admin (61)
  S epop (4921)
  S e_47r3a (14413)
  S sapp (13331)
  S eeda_47r3 (7812)
  S emc_47r3 (41134)
  S etc3_47r3 (10578)
  S e_47r3 (13933)
  S mars (131)
  S epgen_47r3 (8107)
  S o (77)
  S eocean5 (939)
  S mc (31)
  S euronews_test (32)
  S egfas_hourly (255)
  S emos (41472)
```

Where to start



Pages /... / Atos HPCF Documentation

HPC2020 User Guide

Created by Xavier Abellan, last modified on Feb 03, 2022

This is the User guide for the Atos Sequana XH2000 HPCF, installed in ECMWF's data centre in Bologna.

Below you will find some basic information on the different parts of the system. **Please click on the headers or links to get all the details for the given topic.**

HPC2020: How to connect - [read more]

You can connect for the first time via SSH from another ECMWF platform. If you do so from ECGATE or the Cray HPCF you will not need a password to log in.

```
$> ssh aa-login
# or for users with no formal access to HPC service:
$> ssh ecs-login
```

See also HPC2020: Persistent interactive job with ecinteractive if you wish to customise the resource limits of your interactive session.

HPC2020: Differences to other ECMWF platforms - [read more]

The ATOS supercomputer is envisaged to absorb all the computing activities and workloads that have traditionally run not only on the HCPF, but also from other services such as ECGATE and internal ECMWF Linux clusters and workstations.

There are a number of the things you should pay attention at when porting your activities, both as a general advice as well as specific information for each of the origin platforms.

HPC2020: System overview - [read more]

The are four complexes, virtually identical: AA, AB, AC and AD. All together, this HPCF features 7488 nodes:

- 7020 compute nodes, for parallel jobs
- 448 GPIL (General Purpose and Interactive Login) nodes, which are devised to integrate the interactive and post-processing work that coming from older platforms such as the Cray HPCF, ECGATE and Linux Clusters.

HPC2020: Shells - [read more]

You will find a familiar environment, similar to other ECMWF platforms. **Bash** and **Ksh** are available as login shells, with Bash being the recommended option.

Note that **CSH is not available**. If you are still using it, please move to a supported shell.

If you wish to change the default shell you must do it from **ECGATE**

News Feed

2022-02-02 First HPCF Atos cluster in Bologna available to all ECMWF Member State HPCF users

Following the termination of the test Atos system (TEMS) in Reading in November 2021 and a lot of work by many teams across ECMWF, we are pleased to announce the availability of the first Atos cluster, named '**aa**', in Bologna, for Member State users.

We invite you to start testing your activities currently running on the Cray HPCs (cca and ccb) on the first Atos system (aa). To help you with this work, we have made available the Atos HPCF Documentation. We strongly encourage you to read carefully through those pages before you start your tests on this new platform.

2022-01-28 First HPCF Atos cluster in Bologna available to all internal ECMWF users

Following the termination of the test Atos system (TEMS) in Reading in November 2021 and a lot of work by many teams across ECMWF, we are pleased to announce the availability of the first Atos cluster, named '**aa**', in Bologna, for ECMWF staff.

We invite everyone at ECMWF to start testing your personal computational activities currently running on your local desktop, on the servers (lxc, lxop, ecgate) and on the Cray HPCs (cca and ccb) on the first Atos system (aa). To help you with this work, we have updated the Atos HPCF Documentation. We strongly encourage you to read through those pages before you start your work on this new platform.

https://confluence.ecmwf.int/x/UxhbDg

Shell and Filesystems

- No CSH support
  - You must move to bash (default for new users), or ksh
- No Cross-mounted filesystems from existing platforms
  - You must transfer what you need.
- New Flat directory structure:
  - `/home/user` instead of `/home/group/user` or `/home/ms/group/user`
- All filesystems define their corresponding environment variable:
  - $HOME, $PERM, etc
- For local temporary files, avoid using /tmp or /var/tmp: use $TMPDIR instead!
  - Automatically cleaned up at the end of session or job

Shell and Filesystems

File System	Suitable for ...	Technology	Features	Quota
HOME	permanent files, e. g. profile, utilities, sources, libraries, etc.	NFS	It is backed up. Snapshots available. Shared with VDI and ecFlow VMs **Throttled I/O bandwidth from parallel compute nodes (less performance)**	**10 GB** for Member State users
PERM	permanent files without the need for automated backups, smaller input files for serial or small processing, std output, etc.	NFS	No backup Snapshots available. Shared with VDI and ecFlow VMs **Throttled I/O bandwidth from parallel compute nodes (less performance)**	**500 GB** for Member State users
HPCPERM	permanent files without the need for automated backups, bigger input files for parallel model runs, climate files, etc.	Lustre	No backup No snapshots Only accessible from Atos HPCF No automatic deletion	**100 GB** for Member State users without HPC access **1 TB** for Member State users with HPC access
SCRATCH	all temporary (large) files. Main storage for your jobs and experiments input and output files.	Lustre	**Automatic deletion after 30 days of last access to be configured at a later stage** No backup No snapshots Only accessible from Atos HPCF	**50 TB** for Member State users with HPC access **2 TB** for users without HPC access
SCRATCHDIR	Big temporary data for an individual session or job, not as fast as TMPDIR but higher capacity. **Files accessible from all cluster.**	Lustre	**Deleted at the end of session or job** Only accessible from Atos HPCF Created per session/ job as a subdirectory in SCRATCH	part of SCRATCH quota
TMPDIR	Fast temporary data for an individual session or job, small files only. **Local to every node.**	SSD on shared (GPIL) nodes (*f QoSs)	**Deleted at the end of session or job** Created per session/ job	**3 GB** per session/job by default. Customisable **up to 40 GB** with --gres=ssdtmp:<size>G
		RAM on exclusive parallel compute nodes (*p QoSs)		no limit (maximum memory of the node)

Toolchains

- Several compiler suites available:
  - GCC: 8, 9, 10 and 11
  - Intel: 2021.4
  - AMD AOCC 3.1
  - NVIDIA HPC SDK (former PGI)
- Several MPI implementations
  - OpenMPI 4
  - Intel MPI 2021
  - HPCX OpenMPI (based on OpenMPI 4)

Environment Modules

- New module system – Lmod
  - Same basic commands plus some nice additions
  - Massive improvement in modules handling
  - Graceful failure in case of error
  - Automatic swap if module is already loaded
  - Avail and list are ”pipe-friendly”
  - And many more…
  - https://confluence.ecmwf.int/x/eA6UCg

Environment Modules

- Working with different toolchains: the **prgenv** module
  - Active toolchain loaded
  - Affects what modules are “loadable”
  - Ensures that sensitive packages are loaded with the desired “flavour” to avoid conflicts
    - Loading a different prgenv will reload all required modules automatically
  - It allows you to load secondary compilers of different family without affecting the whole stack

```
[usxa@aa6-100 ~]$ module avail prgenv


------------------------------------------------------ Global Aliases -------------------------------------------------------
   pa -> prgenv/amd    pe -> prgenv/expert    pg -> prgenv/gnu    pi -> prgenv/intel    pn -> prgenv/nvidia

------------------------------------------ /usr/local/apps/modulefiles/lmod/prgenvs -----------------------------------------
   conda/4.10.1    prgenv/amd (a)    prgenv/expert (e)    prgenv/gnu (L,D:g)    prgenv/intel (i)    prgenv/nvidia (n)
```

Environment Modules

```
[usxa@aa6-100 ~]$
```

Software Stack

- Some ECMWF software has moved to the **ecmwf-toolbox**
  - ecCodes
  - Magics
  - Metview
  - CodesUI
  - ODC
- A single **`module load ecmwf-toolbox`** to use them all!
- Run **`module help ecmwf-toolbox`** to get the details of the bundled packages and libraries
- Note that other packages still keep their standalone module:
  - ecFlow, BUFRDC…

Software Stack

- Discontinued software:
  - GRIBEX and GRIB-API: use **ecCodes**
  - SMS, ecFlow 4, ecflowview: use **ecFlow 5 (and ecflowUI)**
  - EMOSLIB
    - For interpolation, use **MARS/Metview** with **MIR** library
    - For BUFR encoding/decoding, use **ecCodes or BUFRDC**
  - Metview 3: use Metview 5

Python

- **Only Python 3 supported!**
- Traditional Python 3 available
  - + 230 extra Python modules
- Introducing **conda** for Python
  - Users can easily create their own environments to fully customise their Python experience
  - Internal conda channels available for ECMWF software
- Conda is implemented as an extra “prgenv”:
  - If loaded, it **deactivates all other modules**.
  - Avoiding conflicts between conda packages and module packages

Container support

- Docker is not supported

- You may use Singularity if you wish to run containerised workloads.
  - Rootless containers
  - Supports docker and other OCI images
  - BYOE: Bring Your Own Environment. Develop in your laptop, run in our HPCF!

```
$ module load singularity
$ singularity exec docker://ubuntu:latest cat /etc/os-release
INFO:    Converting OCI blobs to SIF format
INFO:    Starting build...
Getting image source signatures
Copying blob 345e3491a907 done
Copying blob 57671312ef6f done
Copying blob 5e9250ddb7d0 done
Copying config 7c6bc52068 done
Writing manifest to image destination
Storing signatures
2021/06/07 17:51:35  info unpack layer:
sha256:345e3491a907bb7c6f1bdddcf4a94284b8b6ddd77eb7d93f09432b17b20f2bbe
2021/06/07 17:51:36  info unpack layer:
sha256:57671312ef6fdbecf340e5fed0fb0863350cd806c92b1fdd7978adbd02afc5c3
2021/06/07 17:51:36  info unpack layer:
sha256:5e9250ddb7d0fa6d13302c7c3e6a0aa40390e42424caed1e5289077ee4054709
INFO:    Creating SIF file...
NAME="Ubuntu"
VERSION="20.04.2 LTS (Focal Fossa)"
ID=ubuntu
ID_LIKE=debian
PRETTY_NAME="Ubuntu 20.04.2 LTS"
VERSION_ID="20.04"
HOME_URL="https://www.ubuntu.com/"
SUPPORT_URL="https://help.ubuntu.com/"
BUG_REPORT_URL="https://bugs.launchpad.net/ubuntu/"
PRIVACY_POLICY_URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy"
VERSION_CODENAME=focal
UBUNTU_CODENAME=focal
```

https://confluence.ecmwf.int/x/YhhbDg

The new ecFlow service architecture - WIP

- 1 Ecflow server – 1 Virtual Machine
  1. User requests (once) access to the service.
  2. A VM is created and configured with:
     - Same HOME and PERM: best places for job standard output/error.
     - EcFlow running as a system service.
     - Troika (the ECMWF tool for submit / kill/ monitor jobs from ecFlow) - in active development.
     - No other extra software present: avoid running local tasks.
  3. User gets hostname and it’s all ready to go. Suites can be loaded/played straightaway.
     - Everyone uses the same default ecFlow port (3141).
     - No need for the user to start it the ecFlow server manually or use crontab.
     - No interference or competition with other users.

While this is finalised, you may start the ecFlow servers on the HPCF login node.

We will ask you to move to the new model once it’s ready to go.

What to expect if coming from ECGATE

- Familiar Batch System - SLURM.
- Basic commands are the same:
  - sbatch: submit a job
  - squeue: query the queues
  - scancel: cancel jobs
- Queues names are different – name scheme closer to traditional HPCF
  - nf: default queue for serial or small parallel jobs. Shared GPILs
  - np: queue for parallel jobs. Exclusive use of compute nodes.
  - ef: ECGATE-type serial work. Shared GPILs**, only on ECS**
  - el: Long queue. Shared GPILs**, only on ECS**
- Serial work merged into the “fractional” queues

What to expect if coming from Cray HPCF

- New Batch system PBS -> Slurm
  - Jobs need to be “translated”.
- Commands:

User commands	PBS	Slurm
Job cancellation	qdel <job_id>	scancel <job_id>
Job status	qscan [-u <uid>] [<job_id>]	squeue [-u <uid>] [-j <job_id>]
Job submission	qsub [<pbs_options>] <job_script>	sbatch [<sbatch_options>] <job_script>
Queue information	qstat -Q [-f] [<queue>]	sacctmgr show qos [name=<queue>]

What to expect if coming from Cray HPCF

- New Batch system PBS -> Slurm

```
#!/bin/bash
#PBS -N HelloMPI_OpenMP
#PBS -q np
#PBS -l EC_total_tasks=36
#PBS -l EC_threads_per_task=2
#PBS -l EC_hyperthreads=2

export OMP_NUM_THREADS=$EC_threads_per_task
aprun -N $EC_tasks_per_node -n $EC_total_tasks \
      -d $OMP_NUM_THREADS -j $EC_hyperthreads ./HelloMPI_OpenMP
```

```
#!/bin/bash
#SBATCH –J HelloMPI_OpenMP
#SBATCH -q np
#SBATCH -n 128
#SBATCH --cpus-per-task=2

export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK
srun ./HelloMPI_OpenMP
```

What to expect if coming from Cray HPCF

- No MOM nodes or `aprun` for parallel jobs
  - job script on parallel nodes runs on exclusively allocated node
  - `srun / mpirun / mpiexec` to be used instead of `aprun`.
- No compiler wrappers (`cc, CC, ftn…`)
  - use compilers directly (`gcc, icc…`) or use environment variables **`$CC, $CXX, $FC`**
- Flags for module-loaded libraries will not be added automatically!
  - use environment variables provided by modules

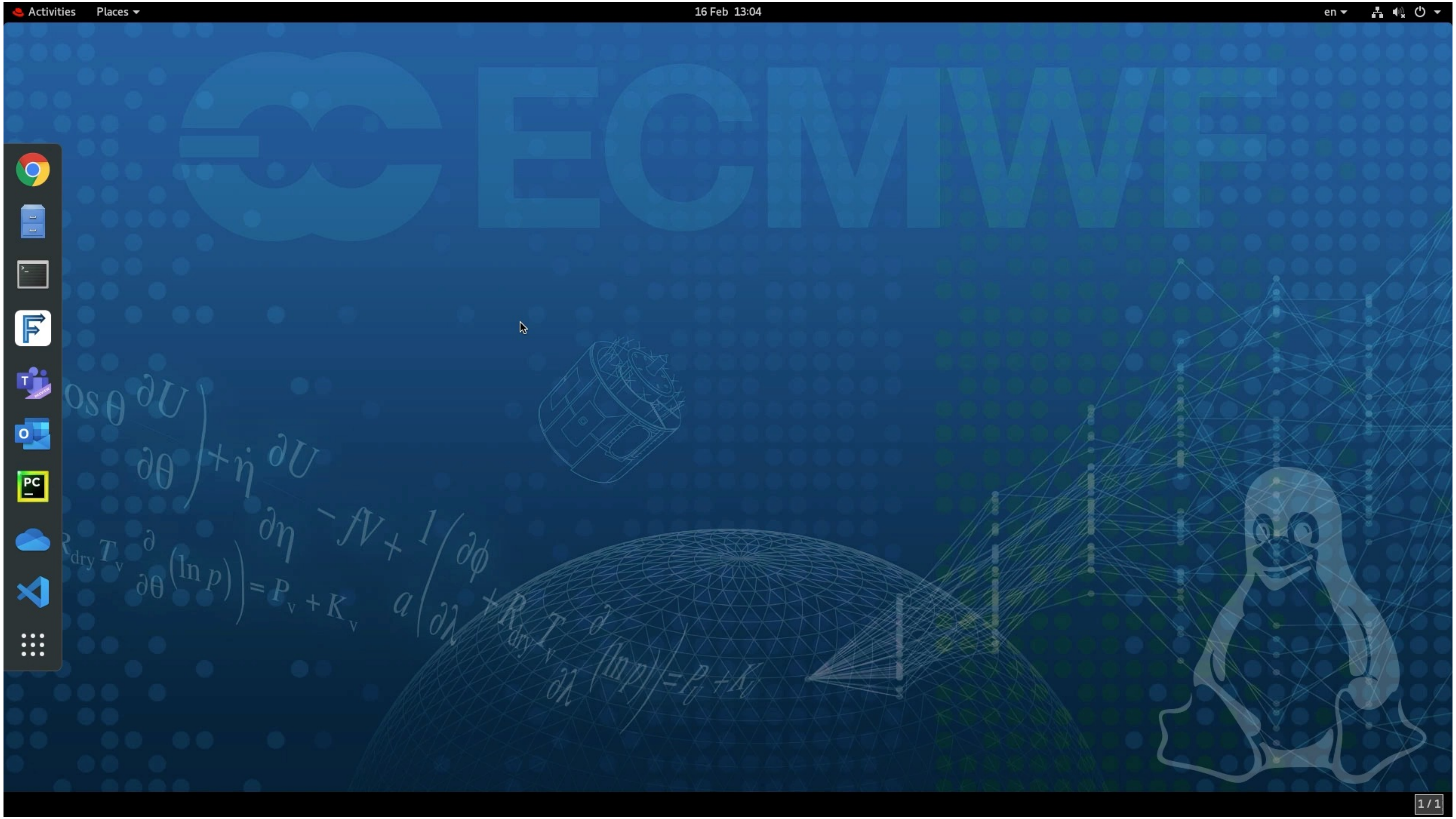
```
$ module show netcdf4 | egrep "DIR|INCLUDE|LIB"
setenv("netcdf4_DIR","/usr/local/apps/netcdf4/4.7.4/GNU/8.3")
setenv("NETCDF4_DIR","/usr/local/apps/netcdf4/4.7.4/GNU/8.3")
setenv("NETCDF4_LIB","-L/usr/local/apps/netcdf4/4.7.4/GNU/8.3/lib
       -Wl,-rpath,/usr/local/apps/netcdf4/4.7.4/GNU/8.3/lib -lnetcdff -lnetcdf_c++ -lnetcdf”)
setenv("NETCDF4_INCLUDE","-I/usr/local/apps/netcdf4/4.7.4/GNU/8.3/include")
```

Interactive sessions

- Limited resources on standard SSH sessions on main login node
- ecinteractive: For more demanding interactive workload

```
$ ecinteractive -h
Usage :  /usr/local/bin/ecinteractive [options] [--]


    -d|desktop     Submits a vnc job (default is interactive ssh job)
    -j|jupyter     Submits a jupyter job (default is interactive ssh job)
    -J|jupyters     Submits a jupyter job with HTTPS support (default is interactive ssh job)


    More Options:
    -h|help        Display this message
    -v|version     Display script version
    -p|platform    Platform (default aa. Choices: aa, ab, ac, ad, ecs)
    -u|user        ECMWF User (default usxa)
    -A|account     Project account
    -c|cpus        Number of CPUs (default 2)
    -m|memory      Requested Memory (default 8 GB)
    -s|tmpdirsize  Requested TMPDIR size (default 3 GB)
    -t|time        Wall clock limit (default 06:00:00)
    -k|kill        Cancel any running interactive job
    -q|query       Check running job
    -o|output      Output file for the interactive job (default /dev/null)
    -x
```

Interactive sessions



Take home messages

- New HPCF with x4 capacity, absorbing ECGATE service
- New ways of remote access
- Familiar ECMWF environment
  - with a few changes and improvements!

Questions?