HTC Brief Instructions
Version 31.07.2017

University of Paderborn
Paderborn Center for Parallel Computing
Warburger Str. 100, D-33098 Paderborn

http://pc2.uni-paderborn.de/
Table of Contents

1. **THE PC² COMPUTE CENTER** .............................................................................. 3

2. **THE HT-CLUSTER (HTC)** ................................................................................ 4
   2.1 FRONTEND .......................................................................................................... 4
      2.1.1 Login via ssh .................................................................................................. 4
      2.1.2 Login From a Windows Based System ......................................................... 4
      2.1.3 Customize the OpenCCS Command Line Interface ...................................... 4

2.2 FILE-SYSTEMS ..................................................................................................... 5
   2.2.1 HOME: Home Directory .................................................................................. 5
   2.2.2 PC2GROUPS: Group Directory ....................................................................... 5
   2.2.3 PC2SW: Software Directory ............................................................................ 6
   2.2.4 PC2WORK: PC² Wide Working Directory ..................................................... 6
   2.2.5 PC2SCRATCH: Cluster Local Working Directory ......................................... 6
   2.2.6 TMPDIR, CCS_TMPDIR: Node Local Directories for Jobs .......................... 6
   2.2.7 FAQs ................................................................................................................ 6

3. **CHOOSING SOFTWARE PACKAGES** ................................................................. 7
   3.1 Usage Hints ......................................................................................................... 7

4. **WORKLOAD-MANAGER (OPENCCS)** ............................................................. 8
   4.1 What is a Workload Manager? .......................................................................... 8
   4.2 Basic Job Flow ...................................................................................................... 8
   4.3 Overview of CCS Commands ............................................................................ 9
   4.4 Differences to Queueing Based Systems .......................................................... 9
   4.5 Resource Requests ............................................................................................. 9
      4.5.1 Resource Sets (Chunks) ................................................................................ 10
      4.5.2 Placement Directives .................................................................................... 10
      4.5.3 The Boot-Node ............................................................................................. 10
      4.5.4 Conversion from TORQUE .......................................................................... 11
      4.5.5 OpenMP ....................................................................................................... 11

4.6 Worker Concept .................................................................................................. 13
4.7 Batch Job Scripts ................................................................................................ 14
4.8 Job Arrays ............................................................................................................ 14

5. **JOB SUBMISSION EXAMPLES** .................................................................... 15
   5.1 Standard Applications ......................................................................................... 15

6. **ONLINE DOCUMENTATION AND CONTACT** ............................................. 16
1. The PC² Compute Center

In October 2011, the PC² moved from Fürstenallee to the building O at Pohlweg. The location provides 160m² floor space for offices and additional space for laboratories, training, and meeting rooms.

The compute center consists of five computer rooms, with 195m² floor space for HPC systems operated by the PC² and 160m² floor space for systems operated by the IMT (Zentrum für Informations- und Medientechnologien). The building infrastructure provides a redundant cooling system (pipelines to the central cooling facility of the campus and chillers on the roof) and a fault tolerant electrical system with battery backup and emergency power generator. Together with state-of-the-art security and fire protection equipment, we are able to provide reliable services and HPC systems with up to 550 kWatt of electricity consumption.

Building O. The compute center is behind the Alan Kay quote.

One of two uninterruptible power supply units (UPS)

The emergency power system (Diesel engine)
2. The HT-Cluster (HTC)

This system should be used to run node local jobs with no timeout (i.e. sequential throughput computing). The cluster has no fast interconnect and comprises a more heterogeneous set of nodes.

2.1 Frontend

The HTC has one frontends:
1. fe.htc.pc2.uni-paderborn.de

The access is via ssh.
If you are a member of the University Paderborn, use your IMT account and password, else use the credentials we have sent to you.

2.1.1 Login via ssh

```
ssh account@fe.htc.pc2.uni-paderborn.de
```

2.1.2 Login From a Windows Based System

If you are login from a system running Windows, you should ask your local administrator to install two Open-Source tools:


2.1.3 Customize the OpenCCS Command Line Interface

At the first login you should call

```
ccsgenrcfiles
```

It creates the directory .ccsrc in the user’s home directory and it creates the default configuration file (.ccsr/uirc) for the OpenCCS command line interface.
In this file one can customize default values like email address(es), format of ccsinfo output, job-name templates, etc.
2.2 File-Systems

The HTC provides two types of file systems:

1. A shared file system located on an external filer. This file system is available on all clusters and on all systems running in the University network.
2. Node local disks.

We provide environment variables which point to the different directories. They are set automatically. The following table gives an overview of the PC² file systems. They are explained in more detail below.

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Source</th>
<th>Purpose</th>
<th>Initial Quota</th>
<th>Backup</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOME</td>
<td>Filer</td>
<td>Permanent small data. Per account</td>
<td>5GB</td>
<td>Yes</td>
</tr>
<tr>
<td>PC2GROUPS</td>
<td>Filer</td>
<td>Group related data. Created on request</td>
<td>200G</td>
<td>Yes</td>
</tr>
<tr>
<td>PC2SW</td>
<td>Filer / Cluster</td>
<td>Pre-installed software</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PC2SCRATCH</td>
<td>Cluster</td>
<td>Temporary working data. Per account</td>
<td>1TB</td>
<td>No</td>
</tr>
<tr>
<td>PC2WORK</td>
<td>Filer</td>
<td>Temporary working data. Per account</td>
<td>1TB</td>
<td>No</td>
</tr>
<tr>
<td>TMPDIR</td>
<td>Node local</td>
<td>Temporary, created by the WLM</td>
<td>-</td>
<td>No</td>
</tr>
<tr>
<td>CCS_TMPDIR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2.1 HOME: Home Directory

Files in this directory are accessible on all PC² systems and on all systems running in the University of Paderborn. This filesystem is restrictively quoted regarding file space and number of inodes. I/O-intensive applications should not use HOME. For this purpose we provide PC2WORK and PC2SCRATCH. HOME is backed up in regular intervals.

2.2.2 PC2GROUPS: Group Directory

To allow working group specific collaboration, we can provide a central directory. Files in this directory are accessible on all PC² systems and on all systems running in the University of Paderborn. This filesystem is restrictively quoted regarding file space. If you need such a directory, please contact us. PC2GROUPS is backed up in regular intervals.
2.2.3  **PC2SW: Software Directory**

This directory contains all pre-installed software for a system. It may be cluster local and/or PC² wide.

2.2.4  **PC2WORK: PC² Wide Working Directory**

For each user account we provide a directory on the Filer which is accessible on all PC² systems and on all systems running in the University of Paderborn. It may be used for storage of temporary files. There will be no backup service in this directory.

2.2.5  **PC2SCRATCH: Cluster Local Working Directory**

On the HTC, $PC2SCRATCH is identical to $PC2WORK.

2.2.6  **TMPDIR, CCS_TMPDIR: Node Local Directories for Jobs**

When a job is allocated via the Work Load Manager (WLM), a temporary directory for the job is created on each affected node. It will be deleted by the WLM, after the job has terminated.

2.2.7  **FAQs**

For more information about the provided file systems, refer to our Wiki.

https://wikis.uni-paderborn.de/pc2doc/index.php/PC2-FileSystems - FAQs
3. Choosing Software Packages

The PC\textsuperscript{2} provides a wide range of development software (compilers, libraries, debuggers, profilers, etc.) as well as specific scientific applications. Many of the available programs require certain environment variables to be set or changed, e.g. \texttt{PATH}, \texttt{LD_LIBRARY\_PATH}, \texttt{MANPATH}. The environment modules package is employed to access or switch between various, sometimes conflicting versions of software. It provides the means to change the environment dynamically by loading, switching or unloading specific software modules. The module installation is customized to automatically resolve dependencies between modules.

The following table shows the most used commands.

<table>
<thead>
<tr>
<th>Command</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>module avail</td>
<td>list the available software packages</td>
</tr>
<tr>
<td>module list</td>
<td>show the loaded modules</td>
</tr>
<tr>
<td>module add</td>
<td>load a module</td>
</tr>
<tr>
<td>module del</td>
<td>unload a module</td>
</tr>
<tr>
<td>module display</td>
<td>show what the module does</td>
</tr>
</tbody>
</table>

A module can either be identified by its name only or its name together with a specific version, for example \texttt{intel} or \texttt{intel/11.1}, respectively. When only the name is specified, the default version, as shown by \texttt{module avail}, is loaded.

3.1 Usage Hints

1. Loading a module changes the environment of the shell in which it is loaded.

2. If other already loaded modules conflict with the module to load, an according error message is displayed and no further changes are made.

3. If other modules are required by the module to load, these are loaded recursively and an according message is displayed.

4. Removing a module reverts the changes in the environment of the shell.

5. If other modules depend on the module to remove, these are removed recursively and an according message is displayed.

For more information about \textit{Modules}, please read the man page or refer to the Modules Project home page: \url{http://modules.sourceforge.net/}
4. Workload-Manager (OpenCCS)

On OCuLUS, we use OpenCCS as Workload Management System (WLM) to control the usage of the compute nodes. Interaction with the system is done by using CCS commands, e.g., ccsalloc, ccsinfo, or ccskill. You will find the OpenCCS User-Manual in $PC2SW/doc.

4.1 What is a Workload Manager?

A compute cluster is a collection of computers and other resources (networks, storage systems, license servers, GPUs, and so forth). Using a WLM typically reduces technical administration of resources while offering a uniform view to users. A WLM abstracts away many of the details involved with running and managing jobs, allowing higher resource utilization. For example, users typically only need to specify the constraints of a job and do not need to know the individual machine names of each host on which they are running. Using a WLM, clusters can execute jobs simultaneously. Clusters often comprise four different components:

(1) **Master Node**: A dedicated node hosting the WLM server components.
(2) **Frontend Nodes**: They provide an entry point to the system for users to manage their workload. Users are able to compile, submit and track their jobs. Additionally, other job-specific pre- and post-processing is possible.
(3) **Compute Nodes**: They execute submitted jobs, controlled by the WLM. A cluster may comprise hundreds or thousands of compute nodes.
(4) **Resources**: Resources can include high-speed networks, storage systems, GPUs, compute nodes, software licenses, and so forth. A list of allocatable resources can be printed by ccsinfo –a or ccsinfo –a --classes

4.2 Basic Job Flow

The life cycle of a job can be divided into four stages: creation, submission, execution, and finalization.

(1) **Creation**: Typically, a submit script is written to hold all of the parameters of a job. These parameters could include how long a job should run (walltime), what resources are necessary to run, and what to execute. The following is an example submit file:

```sh
#!/usr/bin/sh

#CCS --N localBlast
#CCS --res=rset=1:ncpus=2:mem=10g
#CCS -t 10d
#CCS --M user@my.organization.com
#CCS -mea
cd $PC2SCRATCH
myBlast -i -v
```

This submit script specifies the name of the job (localBlast), that it needs two cores an 10 GiByte on a single node (rset=1:ncpus=2:mem=10g), that it will run for at most
10 days, and that the WLM should email "user@my.organization.com" when the job exits or aborts. Additionally, the user specifies where and what to execute.

(2) **Submission**: A job is submitted with the `ccsalloc` command. Once submitted, the policies set by the administration of the site dictate the priority of the job and therefore, when it will start executing.

(3) **Execution**: Jobs often spend most of their lifecycle executing. While a job is running, its status can be queried with `ccsinfo`.

(4) **Finalization**: When a job completed, information about the job is available via `ccsinfo` for about 30 minutes.

### 4.3 Overview of CCS commands

The OpenCCS user interface provides seven commands:

1. `ccsalloc` Submit a job, a job-array, or reserve resources
2. `ccsalter` Alter jobs or reservations
3. `ccsbind` Reconnect to an interactive job
4. `ccsinfo` About nodes, schedule, users, jobs, limits, resources, …
5. `ccskill` Kill jobs and reservations
6. `ccsmsg` Send messages to the job’s output files
7. `ccssignal` Send signals to jobs

### 4.4 Differences to Queueing Based Systems

CCS has no explicit queues. Things like limits, privileges, default values, etc. are assigned to projects and / or single users. The assignment of resources is based on the resource requests and the related limits. Limits may be time dependent.

```
ccsinfo -l
```

shows your limits.

### 4.5 Resource Requests

Running a job requires to specify the resources the job wants to use. If you do not specify resources, CCS will take default values, normally one core and the associated amount of memory.

CCS uses a resource specification syntax which is similar to that of PBSPro (but not Torque). The syntax is:

```
--res=<resource[=value][,resource[=value],...]
```

resource may be

- a name of an allocatable resource,
- a resource set specification,
- a placement specification.
4.5.1 Resource Sets (Chunks)
A resource set (also named chunk) specifies a set of resources that has to be allocated as a unit on a single node. Chunks cannot be split across nodes. Resource sets are specified using the keyword "rset".

Syntax: rset=[N:\]chunk+[\N:\]chunk...

If N is not specified, it is set to 1.

A chunk comprises one or more resource=value statements separated by a colon. Examples:

- ncpus=2:mem=10g:hostname=Host1
- ncpus=27:vmem=20g:arch=linux+4:ncpus=15:mem=128g

ccsinfo –a shows a summary of the available resources.
ccsinfo –a --classes shows the resource classes.

4.5.2 Placement Directives
One may specify how the chunks should be placed on the nodes.

Syntax: place=[arrangement][:sharing]

- arrangement is one of free, pack, or scatter.
- sharing is one of excl or shared.

Default is free:shared. All keywords are described in the following table.

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free</td>
<td>No restriction</td>
</tr>
<tr>
<td>Pack</td>
<td>All chunks must be placed on one node only</td>
</tr>
<tr>
<td>Scatter</td>
<td>Only one chunk per node</td>
</tr>
<tr>
<td>Exclusive</td>
<td>Only this job may use the node</td>
</tr>
<tr>
<td>Shared</td>
<td>This chunk may share the node with other chunks</td>
</tr>
</tbody>
</table>

4.5.3 The Boot-Node
At allocating a partition one of the nodes becomes the boot node. This node controls the jobs started by the user (e.g. starts the job, holds the connection to the UI). The boot node is the first node in the environment variable CCS NODES and in all mapping infos. OpenCCS assigns the boot node to one of the nodes which satisfy the last specified chunk.

Example: --res=rset=ncpus=240:phi=1:arch=MIC+4:ncpus=8:mem=32g.

The boot node will be one of (4:ncpus=8:mem=32g)
4.5.4 Conversion from TORQUE

4.5.4.1 Rules

Torque: \(-\text{lnodes} = X\)
CCS: 

\(--\text{res} = \text{rset} = X; \text{ncpus} = 1, \text{place} = \text{scatter}\)

Torque: \(\text{ncpus} = X \text{ or } \text{cpp} = X\)
CCS: 

\(\text{ncpus} = X\)

Torque: \(\text{ppn} = P\)
CCS: 

\(\text{mpiprocs} = P\)

Torque: \(-\text{lnodes} = N; \text{ppn} = P\)
CCS: 

\(--\text{res} = \text{rset} = N; \text{ncpus} = P; \text{mpiprocs} = P, \text{place} = \text{scatter}\)

4.5.4.2 Examples

1. Request 5 cores and 10GiByte memory on a single node
   Torque: \(-\text{ncpus} = 5, \text{mem} = 10g\)
   CCS: 

   \(--\text{res} = \text{rset} = 1; \text{ncpus} = 5; \text{mem} = 10g\)

2. Submit the same job as a job-array
   Torque: \(-t 1-100 -\text{ncpus} = 5, \text{mem} = 10g\)
   CCS: 

   \(--J 1-100 --\text{res} = \text{rset} = 1; \text{ncpus} = 5; \text{mem} = 10g\)
   NOTE: Refer to the OpenCCS User-Manual for more information on job-arrays.

3. Request one node with a certain property
   Torque: \(-\text{lnodes} = 1; \text{property}\)
   CCS: 

   \(--\text{res} = \text{rset} = 1; \text{ncpus} = 1; \text{property} = \text{true}\)

4. Request 2 cores on each of four nodes with a given property
   Torque: \(-\text{lnodes} = 4; \text{property}; \text{ncpus} = 2\)
   CCS: 

   \(--\text{res} = \text{rset} = 4; \text{ncpus} = 2; \text{property} = \text{true}, \text{place} = \text{scatter}\)

5. Request 1 node of type "fserver" and 15 nodes which are not of type "fserver"
   Torque: \(-\text{lnodes} = 1; \text{fserver} + 15; \text{noserver}\)
   CCS: 

   \(--\text{res} = \text{rset} = 1; \text{ncpus} = 1; \text{fserver} = \text{true}\)
   \(--\text{res} = \text{rset} = 15; \text{ncpus} = 1; \text{fserver} = \text{false}\)
   \(--\text{res} = \text{place} = \text{scatter}\)

4.5.5 OpenMP

OpenCCS supports OpenMP applications by setting the OMP_NUM_THREADS variable in the job’s environment, based on the resource request of the job. The OpenMP run-time picks up the value of OMP_NUM_THREADS and creates threads appropriately.
• OpenCCS sets the value of OMP_NUM_THREADS based on the last chunk of the rset statement.
• If you request ompthreads in the last chunk, OpenCCS sets the environment variable to the value of ompthreads.
• If you do not request ompthreads in the last chunk, then OMP_NUM_THREADS is set to the value of the ncpus resource of that chunk.
• If you do not request either ncpus or ompthreads for the last chunk of the select statement, then OMP_NUM_THREADS is set to 1.

4.5.5.1 Examples
1. Submit an OpenMP job as a single chunk, for a two-core, two-thread job requiring 10GiByte of memory:
   --res=rset=1:ncpus=2:mem=10g
2. The same job but 4threads
   --res=rset=1:ncpus=2:mem=10g:ompthreads=4

4.5.5.2 Fewer Threads than Cores
You might be running an OpenMP application on a host and wish to run fewer threads than the number of cores requested. This might be because the threads need exclusive access to shared resources in a multi-core processor system, such as to a cache shared between cores, or to the memory shared between cores.
You want one chunk, with 16 cores and eight threads:
   --res=rset=1:ncpus=16:ompthreads=8

4.5.5.3 More Threads than Cores
You might be running an OpenMP application on a host and wish to run more threads than the number of cores requested, perhaps because each thread is I/O bound.
You want one chunk, with eight cores and 16 threads:
   --res=rset=1:ncpus=8:ompthreads=16

NOTE: Your job may be killed by OpenCCS due to the ncpus limit if your job uses more than the requested number of ncpus. Refer to the OpenCCS User Manual chapter „Limit Enforcement“ for more information.
4.6 Worker Concept

Cluster systems comprise nodes with full operating system capabilities and software packages like debuggers, performance analyzers, numerical libraries, and runtime environments. Often these software packages require specific pre- and post-processing. OpenCCS supports this with the so-called worker concept. Workers are tools to start jobs under specific runtime environments. They hide specific procedures (e.g., starting of a daemon or setting of environment variables) and provide a convenient way to start and control programs.

ccsinfo --worker shows a list of the available workers.

ccsinfo --whelp=WORKER gives a worker specific help.

For example, we provide workers for different MPI releases (e.g., Intel-MPI, OpenMPI, MVAPICH). All of them load the default module of the related MPI and start mpirun with appropriate arguments. They also ensure that the needed environment is available on all affected nodes. Additional mpirun arguments may be added. They are forwarded to mpirun.
4.7 Batch job scripts

Job scripts are executable (shell) scripts, which contain resource requirements as well as the actual commands to be executed.

Example job script

```bash
#!/usr/bin/sh

#CCS --res=rset=5:ncpus=4:mem=10g
#CCS --res=matlab=5
##CCS --res=place=scatter:excl
#CCS -t 1h
my_preproc
ccsworker g09 -- my_gaussianjob
```

Arguments given to ccsalloc override script directives.

You will find example scripts in $PC2SW/examples. Copy them to your working directory and adapt them to your requirements.

4.8 Job Arrays

Job arrays may be used to group closely related work into a set so that you can submit, query, modify, and display the set as a unit. Job arrays are useful whenever you want to run the same program over and over on different input files.

```bash
#!/usr/bin/sh

#CCS --res=rset=5:ncpus=4:mem=10g
#CCS --res=matlab=5
##CCS --res=place=scatter:excl
#CCS -t 1h
#CCS -j \"1-100:2,200-500:50\"
my_preproc
ccsworker g09 -- my_gaussianjob.$(CCS_ARRAY_INDEX)
```

You will find an example script in $PC2SW/examples/arrayJob.sh. Copy it to your working directory and adapt it to your requirements.

For detailed information on how to submit and manage job arrays, refer to the OpenCCS User-Manual.
5. Job Submission Examples
The standard way of submitting jobs to OpenCCS is using the command ccsalloc. For a submission, OpenCCS needs this information:

- The resources to allocate (refer to chapter 4.3).
  - The job will be killed by CCS if it tries to use more resources than requested.
- How long the job will run.
  - The job will be killed by CCS if the runtime is exceeded.
  - Using ccsalter one can request to increase the runtime.
- The corresponding executable.
  - This may be a job script or an executable. Arguments may be added.
    E.g.: ccsalloc job.sh \-x 1 \-y 3

If you are familiar with PBSPro or Torque, you will find several similar directives. Refer to the ccsalloc man page or the OpenCCS User-Manual.

5.1 Standard Applications
Refer to $PC2SW/examples. There are a lot of example script.
E.g., Abaqus, Gaussian, Matlab, Orca, Turbomole, ...
6. Online Documentation and Contact

![Fig. 2: The PC² Wiki](image)

We use a Wiki to provide up to date information about systems. Refer to [https://wikis.uni-paderborn.de/pc2doc](https://wikis.uni-paderborn.de/pc2doc).

In the Wiki we document relevant hardware and software changes, inform about system status (e.g., maintenance or failures), describe the system architecture, available software packages and available file systems. Further we provide HowTos, FAQs and give hints for tuning application performance.

In `$PC2SW/examples` you will find example scripts for submitting jobs.

In `$PC2SW/doc`, you will find manuals of the workload-manager and this document.

Use the man pages. E.g., `man ccsalloc` or `man module`.

In case you need assistance send an Email to `pc2-support@upb.de`