

Git at the PI

Computing at the PI

HTCondor and the D0 cluster

Blake Leverington, with notes from Alexey Zhelezov





The PI Git Repository

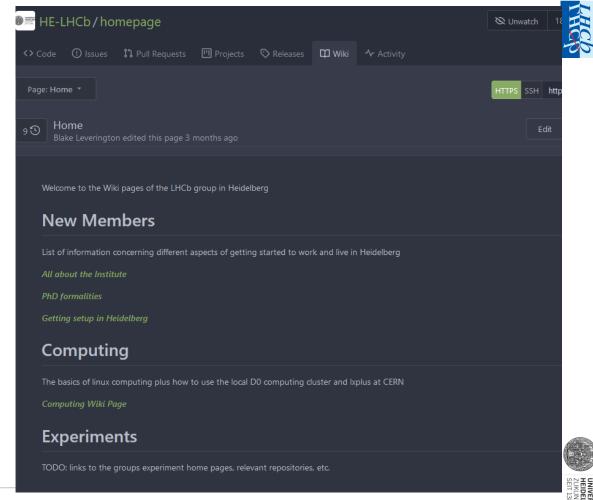
- our DI gradantiala
- Login with your PI credentials.
- Not everyone has, or will maintain CERN credentials
- Not everyone is an LHCb member
- Thesis repository and archive: https://git.physi.uni-heidelberg.de/lhcb_theses
 - Not publicly accessible. Only internal.
 - Archive your thesis pdf and relevant project work here, for future use by the group.
 - Please upload the figures in a folder, and/or figure data and macros to make them (.C).
 - If you keep some files private (i.e. not archived here), that's fine.
 - · How many bachelor, masters and phd projects have disappeared?
- <u>https://git.physi.uni-heidelberg.de/HIT-PAT</u>
- <u>https://git.physi.uni-heidelberg.de/HE-LHCb</u>
- Etc..





The PI Wiki

- <u>https://git.physi.uni-</u> <u>heidelberg.de/HE-</u> <u>LHCB/homepage/wiki</u>
- Easy to edit and create useful pages of information
- Computing How-to's,
- new joiner information
- Best when used and maintained by everyone
 - Contribute your knowledge
 - What did you wish you knew when you started?



The computing page:



- <u>https://git.physi.uni-heidelberg.de/HE-LHCb/homepage/wiki/computing</u>
- Some links with useful computing information for beginners in HEP
 - HEP Essentials: Linux, bash scripting, python, jupyter, git, etc.
 - LHCb Starter kits
 - A lot of the LHCb Software Twiki pages are old and outdated.
- How to use the local computing cluster "DO"
- Building the LHCb Software stack
- Please add more...
 - Like Git how-to's, etc



Available systems

Hich

- 4 Interactive nodes with Ubuntu environment:
 - <u>Lhcba2:</u> Dual AMD EPYC 7713 64-Core Processors (256 CPU threads), 3.7 GHz, 1 TB RAM
 - <u>Lhcba1:</u> Dual AMD EPYC 7352 24-Core Processors (96 CPU threads), 3.2 GHz, 264 GB RAM
 - <u>dOnew</u>: Dual AMD Opteron(tm) 6344 6-core Processor (24 CPU threads), 2.6 GHz, 132 GB RAM
 - delta, Dual Quad-Core AMD Opteron(tm) Processor 2360 SE (8 CPU threads), 2.5 GHz, 16 GB RAM
 - <u>Lhcbi1 (special purpose)</u>: Dual socket Intel(R) Xeon(R) Silver 4214 CPU (48 CPU threads), 2.20GHz , 96 GB RAM –
 - Don't use unless you know you need to use this specific machine
- singularity support under Ubuntu (lhcba2, lhcba1, dOnew, delta, batch system)
- batch system (190 slots HTCondor)

\$ ssh -X -p32 leverington@lhcba2.physi.uni-heidelberg.de

Ports:

24 - host Ubuntu

30 - CentOS7

32 - ALMA Linux 9 (RHEL9)





Storage

- /scratch on some local machines (eg. build stacks here)
- /work
 - Mounted on sigma0 (network connected via NFS)
 - Relatively high performance 22 TB.
 - daily backups. Please do not use this space for data sets greater then O(100 GB).
- /auto/data
 - Mounted on dOnew (network connected via GlusterFS)
 - Slow distributed 260TB storage for big files.
 - No backups.
- /home
 - Mounted on rho0 (network connected) via NFS
 - Do not use for anything more than documents.
 - Hosts your e-mail folder, etc.
 - Backed up

\$ mkdir -p /work/<yourusername> #make your own work directory if it doesn't exist





Play nice.



Pay attention to resource consumption! Use a top (only Ubuntu -p24) and/or top

1000 Mbps network connection \rightarrow read/write data is the first bottleneck seen usually

No network capacity left means no good connection for your colleagues

| ATOP | - lhcb | a1 | | | | | | 2022 | /05/05 | 14:48:55 | | | | | | | | | | | | | | | 10s | elapsed |
|-------|--------|---------|------|------------|-------------|--------|-------|---------|----------------|----------|-------|---------|----------|---------|--------|-------|-------|--------|--------|------|--------|--------|--------|--------|----------|---------|
| PRC | sys | 18.47s | user | 11.13s | | | | | #proc | 1272 | #trun | 4 | #tslpi | 2301 | #tslpu | 1 | #zomb | ie 0 | clones | s 55 | | | 1 | | #exit | 42 |
| CPU į | sys | 178% | user | 112% | irq | 20% | ĺ | | idle | 9212% | wait | 80% | steal | 0% | guest | 0% | İ | | ipc no | | cycl u | | curf 1 | .60GHz | curscal | . ?% |
| CPL į | avg1 | 3.04 j | | Í | avg5 | 2.24 | avg15 | 1.89 | | Í | | | csw 1 | .096133 | ĺ | | intr | 258259 | | | | | numcpu | 96 | | |
| MEM | tot | 251.8G | free | 75.7G | cache | 155.6G | dirty | 1.4M | buff | 341.6M | slab | 4.0G | slrec | 2.5G | shmem | 33.6M | shrss | 3.0M | shswp | 0.0M | vmbal | 0.0M | hptot | 0.0M | hpuse | 0.0M |
| SWP | tot | 46.5G | free | 46.5G | | Í | i í | | | i | | | i | | i | | i | | i İ | | vmcom | 74.2G | i i | | vmlim 1 | .72.4G |
| PSI į | cs | 0/0/0 | | i | ms | 0/0/0 | mf | 0/0/0 | | i | is | 7/4/2 | if | 7/4/2 | İ | | i | | i | | i | | i | | | |
| MDD j | | md1 | busy | 0% | | | read | Θ | write | 53 | KiB/r | Θ | KiB/w | 4 | ĺ | | MBr/s | 0.0 | MBw/s | Θ.Θ | avq | 0.00 | Í | | avio 0. | 30 ms |
| DSK j | | nvme0n1 | busy | <u>0</u> % | | Í | read | Θ | write | 25 | KiB/r | Θ | KiB/w | 9 | İ | | MBr/s | 0.0 | MBw/s | 0.0 | avq | 0.03 | i | | avio 1. | 28 ms |
| DSK j | | nvme1n1 | busy | 0% | | | read | Θ | write | 25 | KiB/r | Θ | KiB/w | | ĺ | | MBr/s | 0.0 | MBw/s | 0.0 | avq | 0.03 | Í | | avio 1. | 28 ms |
| NFM | /π | nt/raid | srv | sigma0 | read | 0K | write | 915.5M | nread | 230K | nwrit | 963.6M | dread | 0K | dwrit | 0K | | | mread | 0K | mwrit | 963.7M | i | | | |
| NFC | rpc | 124090 | read | - 0 | | | write | 123988 | retxmi | t 0 | autre | f 124e3 | İ | | İ | | i i | | İ | | Í | | İ | | | |
| NET | trans | port | tcpi | 162701 | tcpo | 719471 | udpi | 179 | udpo | 179 | tcpao | 173 | tcppo | 172 | ĺ | | tcprs | Θ | tcpie | Θ | tcpor | 1 | udpnp | Θ | udpie | 0 |
| 10.1 | HELWO | | τμι | 102120 | | | ιμυ | 0/113 | up in w | | uettv | 103153 | | | ĺ | | | | | | icmpi | Θ | | | icmpo | 0 |
| IET | enp68 | ISO 81% | pcki | 200953 | pcko | 717554 | sp 10 | 00 Mbps | si 7 | 4 Mbps | so 8 | 14 Mbps | | | coll | Θ | mlti | | erri | Θ | erro | Θ | drpi | 33 | drpo | 0 |
| | | _ | Pent | 1000 | peiro | 1000 | | 0 11000 | | | | | | | coll | Θ | mlti | Θ | erri | Θ | | Θ | drpi | Θ | drpo | 0 |
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P | ID | SYSC | PU | USRC | CPU | V | GROW | RGF | ROW | RUI | D | | EUID | | ST | EX | (C | THR | | S | CPUN | NR . | ACPU | | CMD | 1/7 |
| 30446 | 97 | 0.1 | 3s | 8.5 | 56 s | -46 | 9.0M | -39 | .8M | cas | pary | | caspary | | | | | 8 | | S | 1 | 10 | 87% | | python | |
| 29435 | 82 | 0.0 | 2s | 1.5 | 51s | 135 | 536K | 1954 | 14K | her | rero | | herrero | | | | | 71 | | S | 2 | 25 | 15% | | node | |
| 30861 | 47 | 1.2 | 1s | 0.6 |)0s | | 0K | | 0K | lev | ering | | levering | | | | | 1 | | D | 6 | 51 | 12% | | dd | |
| 30560 | 77 | 0.8 | 5s | 0.0 | 00s | | 0K | | 0K | roo | t | | root | | | | - | 1 | | I | 5 | 54 | 9% | | kworker/ | u196:0 |



7

Disk Space

HICP

(this seems to be our bottle-kneck most times)

- There are 30-40 people using the resources and only 22 TB on /work
- Easy to fill these days with 500 GB of data sets per person

\$ du -h --summarize /work/clangenb/ #check the size of your directory

\$ /home/bleverin/software/gdu_linux_amd64 -m 8 /work #check the size of all directories



8

Compression

Fast and efficient lossless compression with zstd:

Hich

\$ zstd -8 -T8 -rm -r --output-dir-mirror -o /work/compressedfolder/*.zst
/work/sourcefolder/*.mdf

.root files already have some significant compression.

<u>https://root.cern/manual/io/#compression-and-performance</u>

ROOT offers <u>several options</u>, such as LZMA with very high compression ratio, or LZ4 with very high decompression throughput, or ZSTD with a good compromise in performance. The default compression for <u>RNTuple</u> is ZSTD level 5; for everything else it's zlib with compression level 1. Algorithm and compression level can be selected using <u>TFile::SetCompressionAlgorithm()</u> and <u>TFile::SetCompressionLevel()</u>, respectively, at the time data is written.



Compression

(2017)

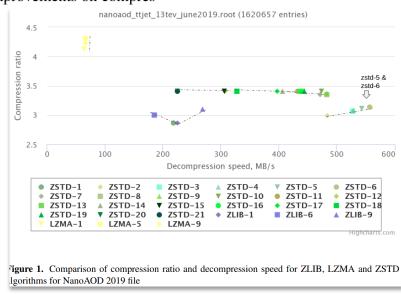
Three years ago, Facebook [7] open-sourced ZSTD, widely used in its software projects. It is largely supported by the community and enhanced by ZSTD authors, who released a variety of advanced capabilities, such as improved decompression speed and better compression ratios.

The initial promise of ZSTD was that it allows users to replace their existing data compression implementation, such as ZLIB, for one with significant improvements on compres-

sion speed, compression ratio, and decompression speed. [6]

The ratio here is for a particular standard file.

In LHCb, global .mdf can compress by 5 times! SciFi NZS by 20 times.



https://www.epj-conferences.org/articles/epjconf/pdf/2020/21/epjconf_chep2020_02017.pdf





UNIVERSITÄT



Best Practices at the PI (to be discussed)



- use lhcba2 as your interactive login machine.
- Compile your development code and store histogram files and results on /work. Consider using /scratch to build standard stacks, lbconda, etc.
 - /work for things that need to be backed up. Weeks of work lost if something happens.
- Put large data sets on /auto/data
 - things that can be downloaded from elsewhere or rerun in a couple days
- Try not to stream data from cern and run multi-core process on lhcba1/2 for extended periods (>2 minutes)
 - Fine for 1 person, doesn't scale beyond a couple of people on our system.
 - Consider using the/a batch system and /auto/data, keep results on /work, if you are running data/computationally intensive process.
 - Can also be run interactively on the batch system
- Full LHCb analyses should be sent to the GRID via Ganga (lhcba1/2 or lxplus)
- If in doubt, ask Alexey how to best run your data.



Environment Setup



#See what is already defined in your environment: leverington@lhcba2:/work/leverington/lhcb \$ env

#Basic environment which will give you access to the python-based Lb-XXXX env-tools: leverington@lhcba2:/work/leverington/lhcb

\$ /cvmfs/lhcb.cern.ch/lib/LbEnv

#Check which platforms are available on the server you logged in to: leverington@lhcba2:/work/leverington/lhcb \$ lb-describe-platform

#set the environment to point to an LCG software stack
leverington@lhcba2:/work/leverington/lhcb
\$ source /cvmfs/sft.cern.ch/lcg/views/LCG_105a/x86_64-el9-gcc12-opt/setup.sh



CVMFS



- CernVM File System (CernVM-FS)
 - global delivery of scientific software, containers, and auxiliary data
 - read-only filesystem for those who access it
 - designed to be scalable and reliable, with known deployments involving hundreds of millions of files and many tens of thousands of clients.
 - https://cvmfs-contrib.github.io/cvmfs-tutorial-2021/01 introduction/
- Exactly the same software version and compiled the same way at CFRN and here at the PL



15-Apr-25

LCG Software Stack in CVMFS



The LCG Software Stacks in a nutshell

Contents

- ~450 packages
 - contrib: gcc, clang, ...
 - projects: ROOT, Geant4, COOL/CORAL, DD4hep, VecGeom, ...
 - externals: external, non-purely Python external packages
 - pyexternals: purely Python external packages
 - generators: physics generators

Main platforms/compilers

- Baseline stable CERN linux distros: SLC6, CERN-CentOS7 , RHEL9 (Alma Linux 9)
 - Also Ubuntu LTS, MacOsX
- Latest compilers: gcc8, gcc9, clang10 , gcc11, gcc12, gcc13, etc

Python3

- Baseline
 - Debug builds

LCG Software Stacks, pre-GDB on Software Deployment, 5 May 2020, CERN virtual



0

leverington@lhcba1:/work/leverington/lhcb

\$ source /cvmfs/sft.cern.ch/lcg/views/LCG_101/x86_64-centos7-gcc8-opt/setup.sh leverington@lhcba1:/work/leverington/lhcb

\$ which python

/cvmfs/sft.cern.ch/lcg/views/LCG_101/x86_64-centos7-gcc8-opt/bin/python

leverington@lhcba1:/work/leverington/lhcb

\$ python --version

Python 3.9.6

leverington@lhcba1:/work/leverington/lhcb

\$ gcc -v

gcc version 8.3.0 (GCC)

leverington@lhcba1:/work/leverington/lhcb

\$ which root

\$ which root

```
alias root='root -l'
```

/cvmfs/sft.cern.ch/lcg/views/LCG_101/x86_64-centos7-gcc8-opt/bin/root

leverington@lhcba1:/work/leverington/lhcb

\$ root --version

ROOT Version: 6.24/06

Personal experience: not all LCG versions will work with your code. Depends on compilation options a lot.

LCG_105a seems to be bug free and works with current LHCb software stack.



NIVERSITÄT EIDELBERG UKUNFT EIT 1386

AHC C



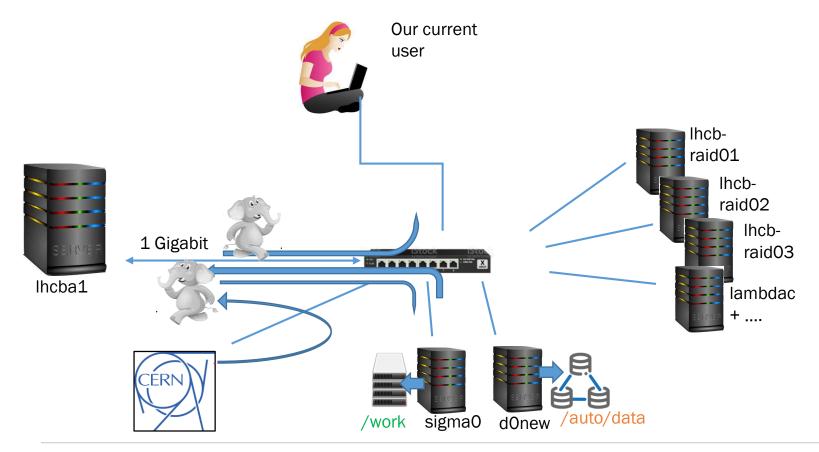
HTCondor and Batch Jobs





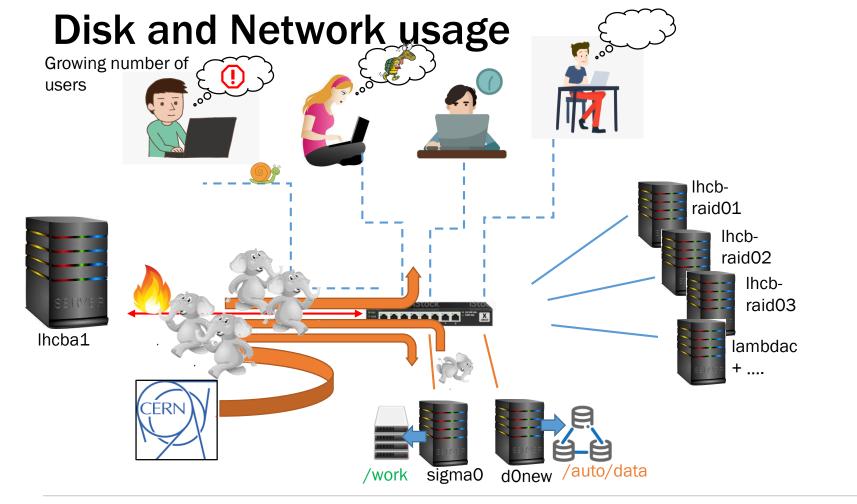
Disk and Network usage







UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

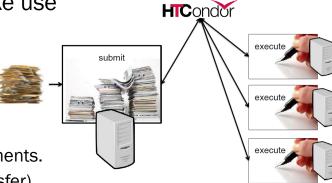


JNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386



Batch submission via HTCondor

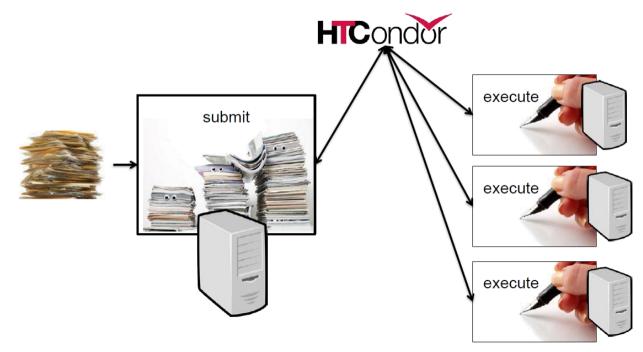
- Split a computational job into multiple jobs to make use of computing resources and your time.
- The job (program) must be executable from the command line (not interactive).
 - You can specify input files, and output files, and pass arguments.
 - We used a shared filesystem in Heidelberg (no need to transfer)
- Will run in a "sandbox". Doesn't know your environment variables unless you pass them or set them up explicitly.









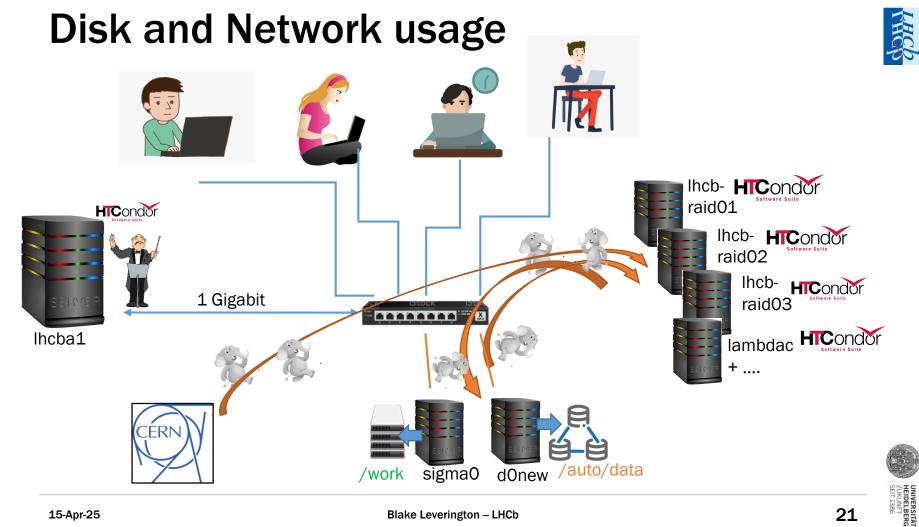


Use case example:

- 1000 events LHCb detector simulation (Gauss) takes ~150 hours on 1 CPU.
 - split it into 20 different 50 event jobs and run on 20 CPUs (each finishes in 8 hours)
 - I can specify when each job is to start (spread them out in time, or run at night to be nice.)
- Many configurable options. <u>https://htcondor.readthedocs.io/en/latest/users-manual/</u>









Preparing a job submission



- Two files to be prepared:
- 1. The submission file: (.sub)
 - Contains all the parameters that HTCondor needs to run your executable on the available slots of the computing pool
- 2. The executable: typically a bash script
 - Sets up your environment, folder structure
 - (maybe) takes the arguments passed from the submission file to execute the program in a different way
 - i.e. random number seed, input files, etc.





The job submission file:

\$ condor_submit examplejob.sub

| - | n under local CentOS7 container, on modern servers only What to run and arguments, can (and probably always | | | | | | | | | |
|--|---|--|--|--|--|--|--|--|--|--|
| should be) a script | | | | | | | | | | |
| executable = ./j | | | | | | | | | | |
| 0 | lusterId) \$(ProcId) | | | | | | | | | |
| # \$(ClusterId) is the number assigned to the group of sub-jobs | | | | | | | | | | |
| # \$(ProcId) is the number assigned to each sub-job | | | | | | | | | | |
| # Safe option, so files are not transfer when not required (our servers have access # to the same storage, /auto/work, /auto/data, /home). | | | | | | | | | | |
| <pre>should_transfer_files = IF_</pre> | NEEDED | | | | | | | | | |
| | | | | | | | | | | |
| # Without variable part, ou | tput files will be the same for all jobs. # That can be confusing. | | | | | | | | | |
| output = out | put/\$(ClusterId)/job.\$(ProcId).out | | | | | | | | | |
| error = out | put/\$(ClusterId)/job.\$(ProcId).err | | | | | | | | | |
| log = out | put/\$(ClusterId)/job.log | | | | | | | | | |
| # It is possible apply current environment variables in jobs, with the following # flag. But in general when particular environment is required, safe option # is directly call the script which setup correct environment (lb-run, etc). In LHCb case # that can also set optimal platform. | | | | | | | | | | |
| #getenv = Tr | ue | | | | | | | | | |
| # The following flag by itself set 2 following options. It is set by default # when submiting from CentOS7 container, but setting it | | | | | | | | | | |
| explicitly produce | | | | | | | | | | |
| # desired result from Ubuntu as well (when submited from hosts supporting it). | | | | | | | | | | |
| +FromVIRTC = "CentOS7" | | | | | | | | | | |
| # the number of jobs queue 10 | https://htcondor.readthedocs.io/en/latest/users-manual/submitting-a-job.html | | | | | | | | | |
| 15 4 05 | | | | | | | | | | |

VERSITÄT



The executable:



- Typically a bash script, but not necessarily. Should be executable on its own.
- <u>https://git.physi.uni-heidelberg.de/HE-LHCb/homepage/wiki/computing_shell_scripts</u>

#!/bin/bash #examplejob.sh

#setup the job folders HOMEFOLDER=/work/leverington/lhcb/examplejob #\\$1 is the first argument passed to a bash script (job ClusterID in this case) mkdir -p \$HOMEFOLDER/output/\${1}/\${2} cd \$HOMEFOLDER/output/\${1}/\${2}

#setup a simple LCG environment
#source /cvmfs/sft.cern.ch/lcg/views/LCG_101/x86_64-centos7-gcc8-opt/setup.sh

#delay each executable so that they do not all start and stop at the same instant. Be nice to the network. sleeptime=\$(echo "\$2 * 30" | bc) sleep \$sleeptime

#our executable that we really wanted to perform (something simple to try, write a 1 GB file). Use full paths. dd if=/dev/zero of=test.file bs=64M count=16 oflag=dsync



Run htcondor in interactive mode

Run your executable from a condor node. Check for errors.

leverington@lhcba1:/work/leverington/lhcb/examplejob/output \$ condor_submit --interactive Submitting job(s). 1 job(s) submitted to cluster 729. Welcome to slot1@delta.physi.uni-heidelberg.de! /w You will be logged out after 7200 seconds of inactivity. ROOT version: bash: /bin/root-config: No such file or directory GCC version: gcc (Ubuntu 7.4.0-1ubuntu1~18.04) 7.4.0 Copyright (C) 2017 Free Software Foundation, Inc. This is free software; see the source for copying conditions. There is NO warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. leverington@delta:/var/lib/condor/execute/dir_3124256 \$ /work/leverington/lhcb/examplejob/examplejob.sh (standard_in) 1: syntax error sleep: missing operand Try 'sleep --help' for more information. 1+0 records in 1+0 records out 67108864 bytes (67 MB, 64 MiB) copied, 0.723605 s, 92.7 MB/s





Run and Monitor your jobs

leverington@lhcba1:/work/leverington/lhcb/examplejob

\$ condor_submit example.sub Submitting job(s)..... 10 job(s) submitted to cluster 724.

leverington@lhcba1:/work/leverington/lhcb/examplejob \$ condor_q

-- Schedd: Ihcba1 : <147.142.19.151:9618?... @ 05/05/22 16:59:59 OWNER BATCH NAME SUBMITTED DONE RUN IDLE TOTAL JOB IDS leverington ID: 724 5/5 16:59 10 10 724.0-9

Total for query: 10 jobs; 0 completed, 0 removed, 0 idle, 10 running, 0 held, 0 suspended Total for leverington: 10 jobs; 0 completed, 0 removed, 0 idle, 10 running, 0 held, 0 suspended Total for all users: 13 jobs; 0 completed, 0 removed, 3 idle, 10 running, 0 held, 0 suspended

leverington@lhcba1:/work/leverington/lhcb/examplejob \$ condor_wait -status output/job.724.log







NIVERSITÄT EIDELBERG JKUNFT EIT 1386

Common errors

leverington@lhcba1:/work/leverington/lhcb/examplejob
\$ condor_q

-- Schedd: lhcba1 : <147.142.19.151:9618?... @ 05/06/22 09:42:28 OWNER BATCH_NAME SUBMITTED DONE RUN IDLE HOLD TOTAL JOB_IDS leverington ID: 725 5/6 09:42 _ _ _ 10 10 725.0-9

Total for query: 10 jobs: 0 completed, 0 removed, 0 idle, 0 running, 10 held, 0 suspended Total for leverington: 10 jobs; 0 completed, 0 removed, 0 idle, 0 running, 10 held, 0 suspended Total for all users: 13 jobs; 0 completed, 0 removed, 3 idle, 0 running, 10 held, 0 suspended

Cause: my error, stdout/stderr output and logfiles were to be written to a folder that doesn't exist

| leverington@lhcba1:/work/leverington/lhcb/examplejob \$ condor_q -hold | |
|--|--------|
| Schedd: lhcba1 : <147.142.19.151:9618? @ 05/06/22 09:50:54 ID OWNER HELD_SINCE HOLD_REASON 726.0 leverington 5/6 09:49 Error from slot1@lhcb-raid02: Failed to open '/auto/work/leverin gton/lhcb/examplejob, outputs/job.726.0.out' as standard output: No such file or directory (errno 2) 726.1 leverington 5/6 09:49 Error from slot2@lhcb-raid02: Failed to open '/auto/work/leverin gton/lhcb/examplejob/outputs/job.726.1.out' as standard output: No such file or directory (errno 2) | |
| <pre>leverington@lhcba1:/work/leverington/lhcb/examplejob \$ condor_rm 725 All jobs in cluster 725 have been marked for removal ngton - LHCb</pre> Check that one or two small jobs run before submitting anything large and time consuming. | - 2 |



Held Jobs



- HTCondor will put your job on hold if there's something YOU need to fix.
- A job that goes on hold is interrupted (all progress is lost) and kept from running

again, but remains in the queue in the "H" state.



Job attributes can be edited while jobs are in the queue using: condor_qedit [U/C/J] Attribute Value

\$ condor_qedit 128.0 RequestMemory 3072 Set attribute "RequestMemory".

If a job has been fixed and can run again, release it with: $condor_release [U/C/J]$

- Job has used more memory than requested
- Incorrect path to files that need to be transferred
- Badly formatted bash scripts (have Windows instead of Unix line endings)
- · Submit directory is over quota
- The admin has put your job on hold



Disk and Network usage.

Hich

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

A series of jobs all started and finished at once.



Summary



- The resources are there to be used, so use them.
 - But be aware of what you are using and be nice.
 - Share them or make a plan for the best time.
- Mattermost channel: sign up with PI email or CERN email.
 - Probably ask any questions here first.
 - <u>https://mattermost.web.cern.ch/heidelberglhcb/channels/lhcbi1-comrades</u>
- Alexey Zhelezov is the person to contact when you have problems or requests.
 - He is maintaining the servers.
- Network and disk monitoring from the terminal (previous plots) <u>https://github.com/tenox7/ttyplot</u>



15-Apr-25



Backup







Disk and Network usage

Hich

scp many files from CERN (~40 MBytes/s)

