Informative session and best practices

DIPC 2019/03/13

Diego Lasa, DIPC (@dilasgoi)

Materials

http://dipc.ehu.es/cc/computing_resources

Outline

- Storage solutions.
- Good practices regarding the use of the storage solutions.
- Data transfers and connecting to the HPC systems.
- Fine-tuning your job specification.
- What is coming next?
- Getting help.

Home directories (A.K.A. /dipc)

Home directories (A.K.A. /dipc)

- 22TB permanent storage solution. It should be used to store your most important information of permanent nature.

Home directories (A.K.A. /dipc)

- 22TB permanent storage solution. It should be used to store your most important information of permanent nature.
- Your home directories are mounted under `/dipc` directory accross all the HPC systems and they are always accesible from: AC computers (`ac.sw.ehu.es`, `ac-01.sw.ehu.es`, `ac-0X.sw.ehu.es`), login nodes of a cluster (`ponto.sw.ehu.es`, `atlas.sw.ehu.es`).

Home directories (A.K.A. /dipc)

- 22TB permanent storage solution. It should be used to store your most important information of permanent nature.
- Your home directories are mounted under `/dipc` directory accross all the HPC systems and they are always accesible from: AC computers (`ac.sw.ehu.es`, `ac-01.sw.ehu.es`, `ac-0X.sw.ehu.es`), login nodes of a cluster (`ponto.sw.ehu.es`, `atlas.sw.ehu.es`).
- Home directories are synchronized everyday.

Home directories (A.K.A. /dipc)

- 22TB permanent storage solution. It should be used to store your most important information of permanent nature.
- Your home directories are mounted under `/dipc` directory accross all the HPC systems and they are always accesible from: AC computers (`ac.sw.ehu.es`, `ac-01.sw.ehu.es`, `ac-0X.sw.ehu.es`), login nodes of a cluster (`ponto.sw.ehu.es`, `atlas.sw.ehu.es`).
- Home directories are synchronized everyday.
- Full and incremental backups on tape. You can recover your home directory's state if is no more than 2 months old.

Home directories: What is going on?

Home directories: What is going on?

- /dipc got filled up 4 weeks ago.

Home directories: What is going on?

- /dipc got filled up 4 weeks ago.
- It is partially our fault because we got rid of user quotas (previously set to 600GB).

Home directories: What is going on?

- /dipc got filled up 4 weeks ago.
- It is partially our fault because we got rid of user quotas (previously set to 600GB).
- The system is full and due to the lack of space + malfunction, data cannot be removed.

Home directories: What is going on?

- /dipc got filled up 4 weeks ago.
- It is partially our fault because we got rid of user quotas (previously set to 600GB).
- The system is full and due to the lack of space + malfunction, data cannot be removed.
- Since it is a corporative non-free solution we had hardware support contracted.

Home directories: What is going on?

- /dipc got filled up 4 weeks ago.
- It is partially our fault because we got rid of user quotas (previously set to 600GB).
- The system is full and due to the lack of space + malfunction, data cannot be removed.
- Since it is a corporative non-free solution we had hardware support contracted.
- But turns out the problem is software related.

Home directories: What now?

Home directories: What now?

- Support provider does not guarantee the integrity of the data.

Home directories: What now?

- Support provider does not guarantee the integrity of the data.
- We performed an additional backup that was completed last week.

Home directories: What now?

- Support provider does not guarantee the integrity of the data.
- We performed an additional backup that was completed last week.
- Support provider is taking care of the system right now.

Scratch filesystems: Atlas

Scratch filesystems: Atlas

- 6 node BeeGFS filesystem. (Parallel and distributed)

Scratch filesystems: Atlas

- 6 node BeeGFS filesystem. (Parallel and distributed)
- 88 TB.

Scratch filesystems: Atlas

- 6 node BeeGFS filesystem. (Parallel and distributed)
- 88 TB.
- Low-latency and high bandwidth FDR Infiniband connection. RDMA procotol.

Scratch filesystems: What is the /scratch filesystem for?

Scratch filesystems: What is the /scratch filesystem for?

- `/scratch` filesystems are shared high performance storage solutions.

Scratch filesystems: What is the /scratch filesystem for?

- `/scratch` filesystems are shared high performance storage solutions.
- Provide access to large amounts of disk for short periods of time at much higher speed than `/dipc`.

Scratch filesystems: What is the /scratch filesystem for?

- `/scratch` filesystems are shared high performance storage solutions.
- Provide access to large amounts of disk for short periods of time at much higher speed than `/dipc`.
- It is meant to be used as the work space for jobs.

Scratch filesystems: What /scratch filesystems are not for?

- `/scratch` filesystems are not meant to be used permanent storage solution.

Scratch filesystems: What /scratch filesystems are not for?

- `/scratch` filesystems are not meant to be used permanent storage solution.
- You should use it only to submit jobs and redirect all your I/O.

Scratch filesystems: What /scratch filesystems are not for?

- `/scratch` filesystems are not meant to be used permanent storage solution.
- You should use it only to submit jobs and redirect all your I/O.
- Once the jobs is finished, you should clear the `/scratch`.

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)
- Total number of files older than 30 days: 12869847 (77%)

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)
- Total number of files older than 30 days: 12869847 (77%)
- Total number of files older than 90 days : 10101103 (60%)

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)
- Total number of files older than 30 days: 12869847 (77%)
- Total number of files older than 90 days : 10101103 (60%)
- Total number of files older than 180 days : 7527168 (45%)

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)
- Total number of files older than 30 days: 12869847 (77%)
- Total number of files older than 90 days : 10101103 (60%)
- Total number of files older than 180 days : 7527168 (45%)
- Total number of files older than 360 days : 4869983 (29%) (24% of occupation)

Scratch filesystems: Used as permanent storage solution

Researchers are using (have always been using) `/scratch` as a permanent storage solution.

- Total number of files in the /scratch : 16791329 (100%)
- Total number of files older than 30 days: 12869847 (77%)
- Total number of files older than 90 days : 10101103 (60%)
- Total number of files older than 180 days : 7527168 (45%)
- Total number of files older than 360 days : 4869983 (29%) (24% of occupation)

When the occupancy goes above 80% the BeeGFS filesystem shows a performance degradation that affects all users.

Scratch filesystems: Number of files per directory

Scratch filesystems: Number of files per directory

- The same applies with large numbers of small files, since the BeeGFS filesystem is not behaving ideally when dealing with high volumes of small files.

Scratch filesystems: Number of files per directory

- The same applies with large numbers of small files, since the BeeGFS filesystem is not behaving ideally when dealing with high volumes of small files.
- The performance of the `/scratch` filesystem has been deeply degraded 3 times over the lifetime of Atlas: Apr. 2018, Nov. 2018, and Feb. 2019. The three events were related to the fact that some researchers had tens and hundreds of thousands of files under their directories.

Scratch filesystems: Number of files per directory

- The same applies with large numbers of small files, since the BeeGFS filesystem is not behaving ideally when dealing with high volumes of small files.
- The performance of the `/scratch` filesystem has been deeply degraded 3 times over the lifetime of Atlas: Apr. 2018, Nov. 2018, and Feb. 2019. The three events were related to the fact that some researchers had tens and hundreds of thousands of files under their directories.
- Avoid storing more than a few thousand of files (<3000) per directory.

Scratch filesystems: More users

Scratch filesystems: More users

- The number of accounts allowed to use Atlas: 173

Scratch filesystems: More users

- The number of accounts allowed to use Atlas: 173
- The number of accounts is increasing: 37 new accounts in the last 5 months (15 in 2019).

Scratch filesystems: More users

- The number of accounts allowed to use Atlas: 173
- The number of accounts is increasing: 37 new accounts in the last 5 months (15 in 2019).
- That is a increase of a 22%.

Scratch filesystems: More users

- The number of accounts allowed to use Atlas: 173
- The number of accounts is increasing: 37 new accounts in the last 5 months (15 in 2019).
- That is a increase of a 22%.
- However the capacity of the `/scratch` remains constant.

Scratch filesystem: How can we deal with this?

Scratch filesystem: How can we deal with this?

- Quota enforcement: disk space (brought down to 1.5TB from 3TB and affected 10% of the users), number of files.

Scratch filesystem: How can we deal with this?

- Quota enforcement: disk space (brought down to 1.5TB from 3TB and affected 10% of the users), number of files.
- Periodical deletion of files older than X days. (Not considering this right now)

Scratch filesystem: How can we deal with this?

- Quota enforcement: disk space (brought down to 1.5TB from 3TB and affected 10% of the users), number of files.
- Periodical deletion of files older than X days. (Not considering this right now)
- Update the permanent storage solution.

Scratch filesystem: How can we deal with this?

- Quota enforcement: disk space (brought down to 1.5TB from 3TB and affected 10% of the users), number of files.
- Periodical deletion of files older than X days. (Not considering this right now)
- Update the permanent storage solution.
- However, if the permanent storage is not enough to hold your data, then you should move it to your local machine or any local device.

Scratch filesystems: Quota and disk space usage

- To check your quota and disk space usage on Atlas while you are in the login nodes type:

```
getquota
```

Scratch filesystems: Quota and disk space usage

- To check your quota and disk space usage on Atlas while you are in the login nodes type:

```
getquota
```

- There are no restrictions on Ponto in this regard.

Scratch filesystems: What if I need to temporarily keep many files?

- Distribute your files accross multiple directories.
- Archive your files using tools like `tar`.

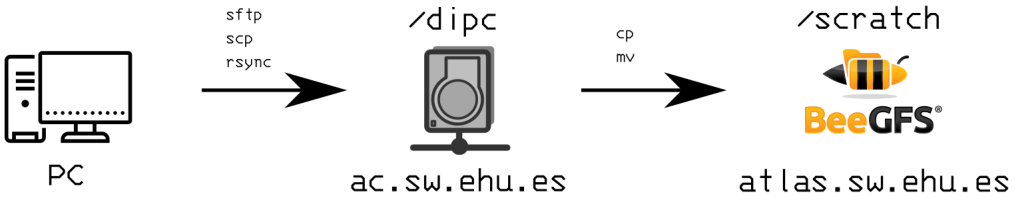
Archive a bunch of files:

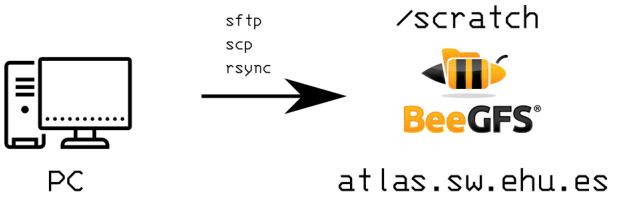
```
tar -cvf archive.tar file1 file2 [..] fileN
```

Unarchive an archive file:

```
tar -xvf archive.tar
```

Data transfer





Fine-tuning job specifications

One of the biggest concerns is:

- How much memory should I request for my job?

Fine-tuning job specifications

One of the biggest concerns is:

- How much memory should I request for my job?
- How much cputime?

Fine-tuning job specifications

One of the biggest concerns is:

- How much memory should I request for my job?
- How much cputime?
- How many nodes/cores?

Fine-tuning job specifications

There is not one clear answer to that questions.

If you do not know the amount of resources in terms of memory or time your jobs are going to need, you should overestimate this values in the first runs and tweak those values up as you learn how jobs behave

Fine-tuning job specifications: learning how jobs behave

Once a jobs finishes we can query the system to get information about the resources the job consumed.

- `tracejob`

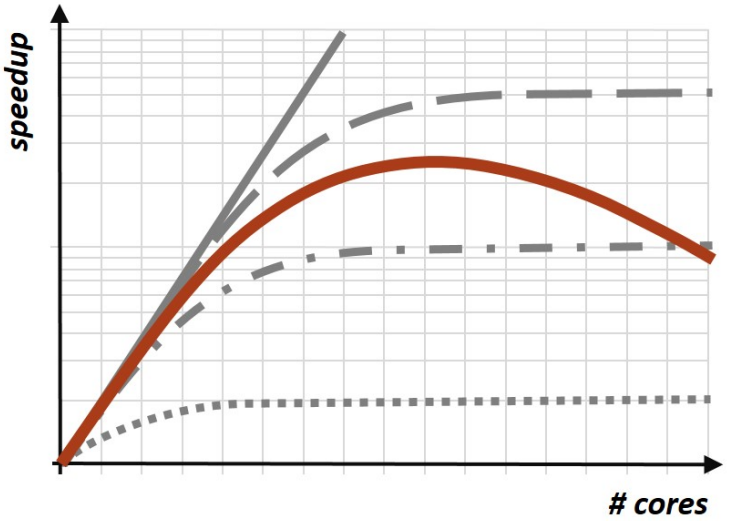
Usage:

`tracejob -n X <job_id>`

- DCRAB

http://dipc.ehu.es/cc/computing_resources/apps/dcrab.html

Fine-tuning job specifications: Number of cores



What you can expect in the near future: CEPH filesystem

What you can expect in the near future: CEPH filesystem

New CEPH disk server as a permanent storage solution:

- Accessible from all the HPC systems and desktop computers.

What you can expect in the near future: CEPH filesystem

New CEPH disk server as a permanent storage solution:

- Accessible from all the HPC systems and desktop computers.
- 800 TB of net space.

What you can expect in the near future: CEPH filesystem

New CEPH disk server as a permanent storage solution:

- Accessible from all the HPC systems and desktop computers.
- 800 TB of net space.
- Fully backed up.

What you can expect in the near future: CEPH filesystem

New CEPH disk server as a permanent storage solution:

- Accessible from all the HPC systems and desktop computers.
- 800 TB of net space.
- Fully backed up.
- Fast connection: 40 Gbps Ethernet.

What you can expect in the near future: ATLAS-EDR

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.
- 20 of them featuring 2 x GPGPUs (Tesla P40).

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.
- 20 of them featuring 2 x GPGPUs (Tesla P40).
- Independent login nodes for login and compilation.

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.
- 20 of them featuring 2 x GPGPUs (Tesla P40).
- Independent login nodes for login and compilation.
- New software servers.

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.
- 20 of them featuring 2 x GPGPUs (Tesla P40).
- Independent login nodes for login and compilation.
- New software servers.
- New 200 TB /scratch filesystem shared between old and new nodes.

What you can expect in the near future: ATLAS-EDR

New upgrade of Atlas:

- 37 new computing nodes: 48 cores, 64-384 GB RAM.
- 20 of them featuring 2 x GPGPUs (Tesla P40).
- Independent login nodes for login and compilation.
- New software servers.
- New 200 TB /scratch filesystem shared between old and new nodes.
- New nodes will also be able to access the old 88 TB /scratch solution.

Migrating from TORQUE/Maui to SLURM

Migrating from TORQUE/Maui to SLURM

- Maui is not being updated since 2012.
- It is not prepared to manage accelerators such as GPGPUs.

Migrating from TORQUE/Maui to SLURM

- Maui is not being updated since 2012.
- It is not prepared to manage accelerators such as GPGPUs.
- SLURM has great community and support from developers.
- Becoming more and more a standard in the HPC industry.

Migrating from TORQUE/Maui to SLURM

- Maui is not being updated since 2012.
- It is not prepared to manage accelerators such as GPGPUs.
- SLURM has great community and support from developers.
- Becoming more and more a standard in the HPC industry.
- First and exclusively on Atlas-EDR.
- We will gradually add nodes from Atlas-FDR (A.K.A Atlas).

Walltime limitation for jobs

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.
- People abuse the system asking for to much time. Therefore, backfilling policies do not kick in.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.
- People abuse the system asking for to much time. Therefore, backfilling policies do not kick in.
- As a consequence, less jobs executed per unit of time.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.
- People abuse the system asking for to much time. Therefore, backfilling policies do not kick in.
- As a consequence, less jobs executed per unit of time.
- A queue labelled "long" without time limitations. However, it will be thoroughly monitored.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.
- People abuse the system asking for to much time. Therefore, backfilling policies do not kick in.
- As a consequence, less jobs executed per unit of time.
- A queue labelled "long" without time limitations. However, it will be thoroughly monitored.
- Do not panic! Job priority does not depend on the queue, but it is primarily based on a FairShare factor that measures the past use of the HPC platform made by a user compared to the overall past use of the same HPC platform made by all users.

Walltime limitation for jobs

- Creation of very-fast, fast, slow, very-slow queues with time limitations.
- Still deciding what these limits will be.
- People abuse the system asking for to much time. Therefore, backfilling policies do not kick in.
- As a consequence, less jobs executed per unit of time.
- A queue labelled "long" without time limitations. However, it will be thoroughly monitored.
- Do not panic! Job priority does not depend on the queue, but it is primarily based on a FairShare factor that measures the past use of the HPC platform made by a user compared to the overall past use of the same HPC platform made by all users.
- This FairShare factor halves every X days, so if a user stops using the system their future jobs will have more priority than the ones sent in the present.

Why are we introducing time limits?

Why are we introducing time limits?

- So backfill policies kick in and the job throughput is increased.

Why are we introducing time limits?

- So backfill policies kick in and the job throughput is increased.
- So people do not monopolize the resources.

What is 'backfilling'?

Where you can help and assistance?

You can find the contact information of the DIPC-CC staff here:

http://dipc.ehu.es/cc/computing_resources/general/staff.html

You can find information relative to the use of the HPC Systems here:

http://dipc.ehu.es/cc/computing_resources

You can reach us by:

- Email (preferred)
- Phone
- In person at our offices located in Building 3

Thank you!