Science & Technology Facilities Council

Parallel Processing using the LOTUS cluster

Alison Pamment / Cristina del Cano Novales

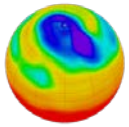
JASMIN/CEMS Workshop

February 2015

Overview

- Parallelising data analysis
- LOTUS HPC Cluster
- Job submission on LOTUS
- Re-factoring code for efficiency
- Future of parallel data analysis

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

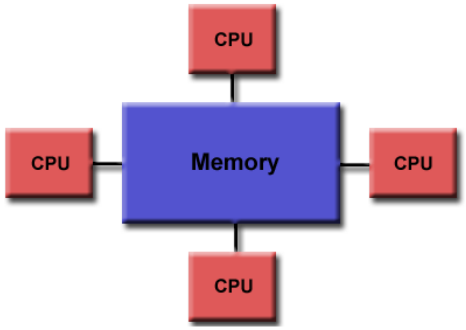
Parallelising Data Analysis

Science & Technology Facilities Council

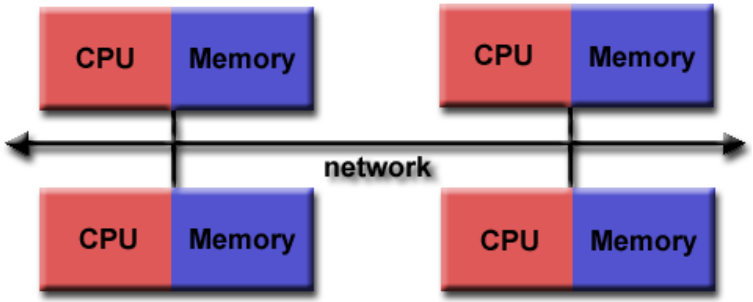
Processing big data: the issues

- Parallel processing in the Environmental Sciences has historically focussed on **highly-parallel models.**
- Data analysis was typically run **sequentially** because:
  - It was a smaller problem
  - It didn't have parallel resources available
  - The software/scientists were not equipped to work in parallel
- But now we generate enormous datasets (e.g. UPSCALE – around 300Tb) means that:
  - Processing big data **requires** a parallel approach, but
  - Platforms, tools, and programmers are becoming better equipped

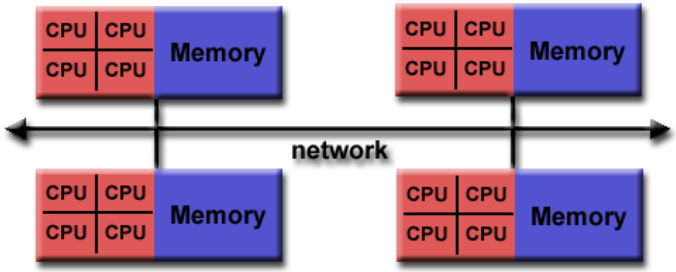
The traditional view of parallel processing



shared memory



distributed memory



shared distributed memory

Parallel processing for data analysis

- Data analysis tools do **not** (typically) do parallelisation automatically.
- But parallelisation is normally achievable at a small price.
- A lot can be done with:
  - Batch processing
  - Decomposition of large jobs into smaller jobs
  - Understanding tools and schedulers

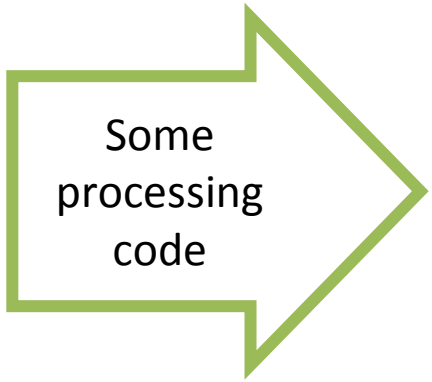
We will look at these and show examples.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Simple parallelism by hand (1)

Long list (100,000) of text files: each file contains the text from a whole scientific paper.



A text file: listing all lines in all papers that match the word "Einstein"

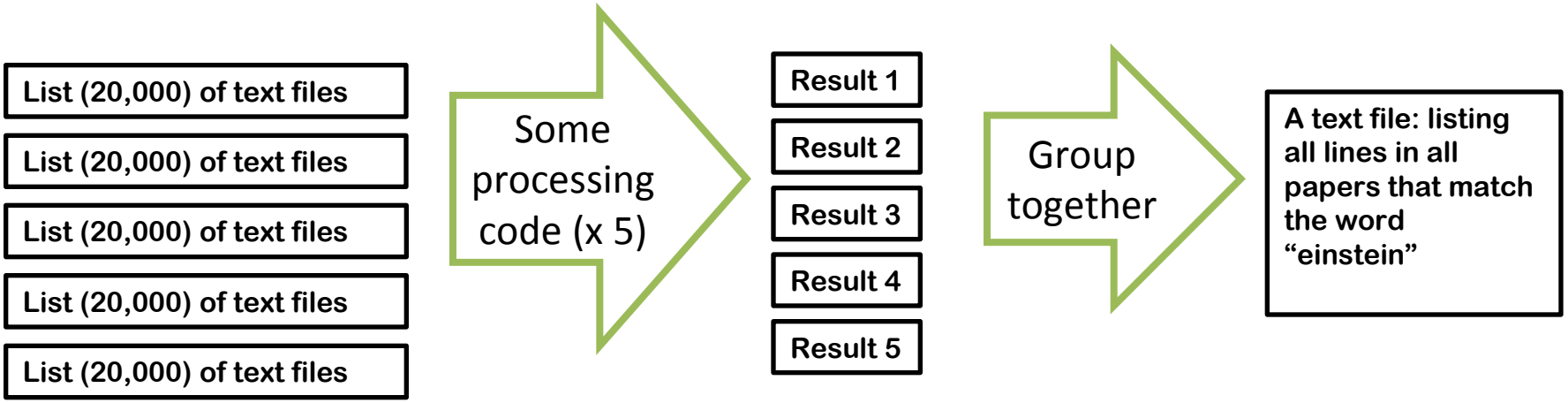
grep_for_einstein.sh

```
#!/bin/bash
input_file=$1
while read FILENAME; do
       grep Einstein $FILENAME >> ${input_file}_result.txt
done <  $input_file
```

Science & Technology Facilities Council

Simple parallelism by hand (2)

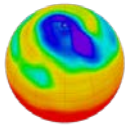
- A simple re-factoring splits the job into five parts:




```
$ split -l 20000 -d list_of_files.txt   # Writes to "x00", "x01", ..., "x04"
$ for i in x??; do grep_for_einstein.sh $i & done
$ cat *_result.txt > output.txt
```

"&" means "run in background". So all will run in parallel.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

LOTUS HPC Cluster

Science & Technology Facilities Council

LOTUS

- JASMIN Scientific Analysis Servers are limited in resources – LOTUS is bigger!
- High performance computing facility
- Cluster of both physical and virtual machines
- High speed access to Panasas storage
  - BADC and NEODC Archive
  - Group workspaces
  - Users' home directories
  - Scratch area (read/write)
  - /tmp

http://www.ceda.ac.uk/help/users-guide/lotus/

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

LOTUS Hardware

Model	Processor	Cores	Memory
194 x Viglen HX525T2i	Intel Xeon E5-2650 v2 “Ivy Bridge”	16	128GB
14 x Viglen HX545T4i	Intel Xeon E5-2650 v2 “Ivy Bridge”	16	512GB
6 x Dell R620	Intel Xeon E5-2660 “Sandy Bridge”	16	128GB
8 x Dell R610	Intel Xeon X5690 “Westmere”	12	48GB
3 x Dell R610	Intel Xeon X5675 “Westmere”	12	96GB
1 x Dell R815	AMD Opteron	48	256GB

- 226 physical hosts
- 3556 cores
- Intel/AMD processors
- 17 large memory hosts
- Mix of generations/specs

Science & Technology Facilities Council

LOTUS Resources

- RedHat Enterprise Linux 6 OS
- IBM Platform LSF batch scheduler
- Platform MPI
- Intel and PGI compilers (available through environment modules)
- Jasmin Analysis Platform software
  - All LOTUS nodes have the same software packages available as the JASMIN Scientific Analysis servers. So you can:
    - Develop code on the generic Analysis Servers
    - Run in batch mode via LOTUS

LOTUS Directory Layout

- /home/users
  - Users' home directory
  - Read/write
  - Shared with other JASMIN machines
    - i.e. jasmin-login1.ceda.ac.uk, jasmin-xfer1.ceda.ac.uk, jasmin-sci1/2.ceda.ac.uk
    - Data can be moved between hosts without login into lotus.jc.rl.ac.uk
  - 10GB Quota
    - Current usage can be checked with the command:

      `pan_quota`
  - Home directory backed up on a nightly basis

LOTUS Directory Layout

- /badc and /neodc
  - BADC and NEODC archives
  - Read only
  - Access control by UNIX groups
- /group_workspaces/jasmin/*
  - Group workspaces
  - Read/write
  - Access control by UNIX groups

Directory Layout

- /work/scratch
  - Shared across the whole cluster:
    - Parallel jobs can access the same files during execution
  - Stored in Panasas
  - 16TB quota, but shared between all users!
  - Users should create subdirectories for their jobs (i.e. /work/scratch/jsmith)
- /tmp
  - Local directories, one per node
  - Store temporary data that only needs to be read by the local process
  - Job should delete any files in /tmp when completed

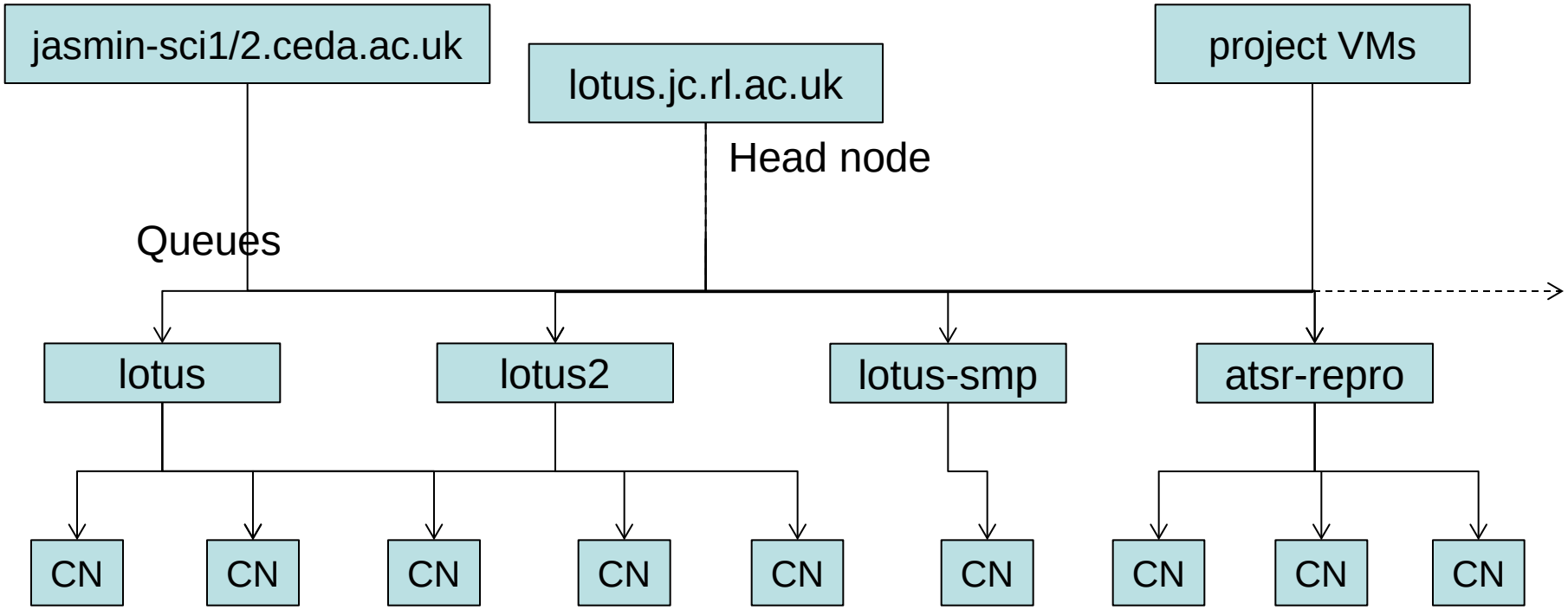
Data in these directories is temporary and may be arbitrarily removed at any point once the job has finished running.

All important data should be saved in Group Workspaces.

LOTUS Queues

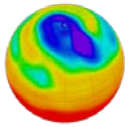
- Queues: Groups of hosts with the same policies
- Many queues available in LOTUS – only a few are relevant
  - lotus: Default queue - **mixed generations/cores/memory**
  - lotus2: Currently closed to users
  - lotus-smp:
    - 1 node with 4x12 cores AMD 2.6GHZ 256GB RAM
  - project queues …

LOTUS Configuration



Compute nodes (physical hosts or virtual machines)

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



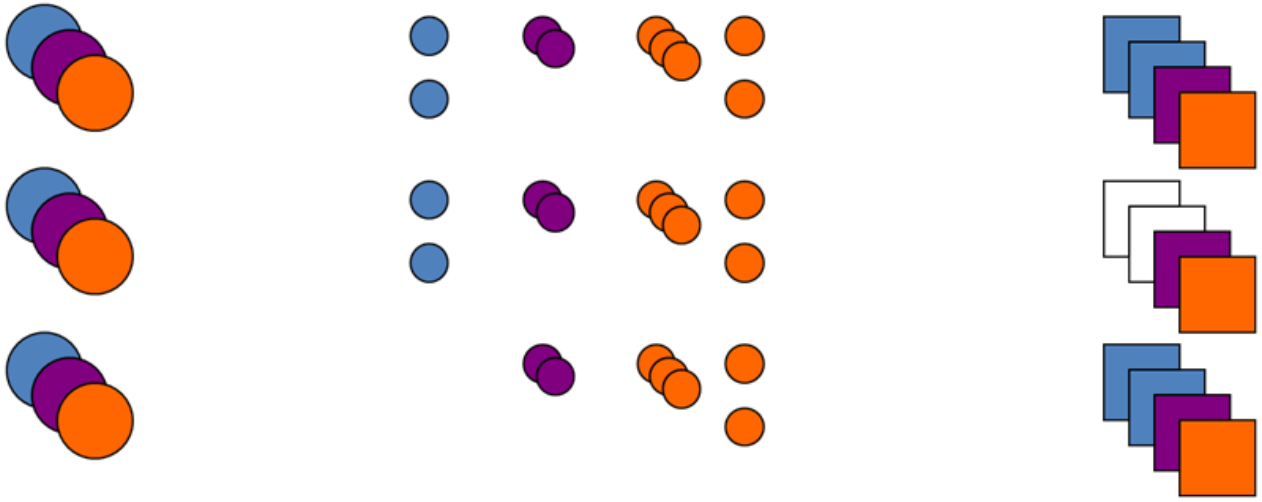
Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Job Submission

Science & Technology Facilities Council

Job Submission

- Jobs are submitted using the LSF scheduler
- Resources are allocated as they become available
- Fair share of resources between users



Job Submission

- Jobs submitted with the `bsub` command

```
[cdelcano@lotus ~]$ bsub -q lotus -o test.out < hostname_test.sh
        Job <71880> is submitted to queue <lotus>.
```

- Check status of job(s) with `bjobs` command

```
[cdelcano@lotus ~]$ bjobs
JOBID   USER    STAT  QUEUE   FROM_HOST  EXEC_HOST JOB_NAME SUBMIT_TIME
71880  cdelcan  PEND  lotus  lotus.jc.rl   */hostname Mar 18 16:26
```

- Different flags can be used when submitting jobs:
  - **-q** lotus (which queue to run on)
  - **-n** 36 (number of processors to run on)
  - **-W** 00:30 (predicted time in hours and minutes)
  - **-x** (exclusive node use, to avoid sharing with other jobs)
  - **-o** %J.log (name of file for output of job)
  - **-e** %J.err (name of file for error log)

Job Arrays

- Group of jobs with the same executable and resource requirements
- Using different input files
- Can be submitted, controlled and monitored as a single unit or as individual jobs or groups of jobs
- Can be dependent on another job (or job array)

Job Arrays

- Submitting a job array:

```
[cdelcano@lotus array]$ bsub -J "myArray[1-2]" -o output.%J.%I -e
error.%I -i input.%I /home/users/cdelcano/array/array_test
Job <231562> is submitted to default queue <lotus>.
```

- Checking status of a job array:

```
[cdelcano@lotus array]$ bjobs -A
JOBID    ARRAY_SPEC  OWNER   NJOBS PEND DONE  RUN EXIT SSUSP USUSP
231562   myArray[ cdelcano     2    0    2    0    0     0     0
```

MPI

Some background: What is MPI?

- **MPI** stands for ***Message Passing Interface***.
- Explaining MPI is beyond our scope.
- MPI provides a standard specification that enables message passing between different nodes of a parallel computer system.
- It follows a distributed memory model.

So it maps nicely to use on the LOTUS cluster...and lots of existing parallel code uses MPI (as implemented in C, Fortran, Python etc).

MPI on LOTUS

- The LOTUS User Guide provides detailed instructions on how to:
  - Load MPI modules
  - Compile C/Fortran code to use MPI
  - Log on to interactive nodes for compilation
  - Submitting jobs using MPI
- There is not time to cover this in detail, please see:

  http://www.jasmin.ac.uk/how-to-use-jasmin/lotus-documentation/

Re-factoring the code:
Efficiency gains through re-factoring (1)

- Major gains can be made by changing the order and structure of your code
- Problems with your code might include:
  - Code will not run because of memory requirements
  - Code runs sequentially and takes a long time
  - Code does run but falls over because of resource limits.
- In some cases you can create loops that can be scripted as separate processes (or JUG tasks) allowing you to submit them in parallel

Efficiency gains through re-factoring (2)

Here is a real-world example:

The Problem: Trying to run the NCO tool "ncea" to calculate an average from a large dataset. It will not run!

Why? The "ncea" command reports this...and then exits:

- "unable to allocate 7932598800 bytes" (which is about 8 Gbytes)

Possible solutions:

1. Data files hold multiple variables: Operate on one at a time:

ncea **-v vosaline** means/199[45678]/*y01T.nc -o test.nc

2. Reduce the number of files (i.e. years) processed each time:

ncea means/**199[45]**/*y01T.nc -o test.nc

The future of parallel data analysis

- Analysing Big Data is a challenge! Software needs to adapt and scientists need to be able to adapt their code to keep up!

Number of files	3,222,967
Number of datasets	54,274
Archive Volume (TB)	1,483
Models with data published	64
Models with documentation published in archive	38
Experiments	108
Modelling centres	32
Data Nodes	22

CMIP5 Status (early 2013)

The future of parallel data analysis

We are likely to see more:

- Parallel I/O in software libraries;
- Web processing services that do the parallel analysis remotely;
- Analysis Platforms (like JASMIN) that allow scientists to run code next to the data;
- Learning to write parallel code now is likely to be of great benefit in future;

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Further Information

JASMIN Analysis Platform (software packages):

http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/AnalysisPlatform/Packages

LOTUS Overview:

http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/LOTUS

LOTUS User Guide:

http://www.ceda.ac.uk/help/users-guide/lotus/

Jug:

http://pythonhosted.org/Jug/

Parallel processing:

https://computing.llnl.gov/tutorials/parallel_comp/

Science & Technology Facilities Council