Digital Institute

# Processing data intensive Matlab jobs through Condor

Stephen McGough

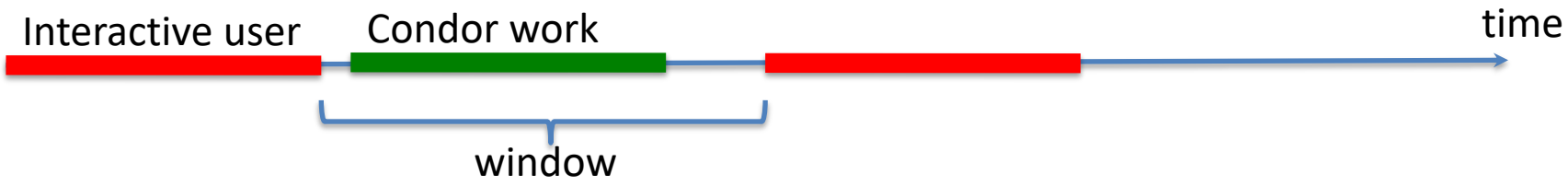
Fanar M. Abed

Newcastle University

Digital Institute

# Outline

- Condor provides a powerful job execution environment
- Problems come with
  - lack of check-pointing in Windows
  - Unknown time window for execution




- Like many Universities Newcastle has a large (cycle stealing) Condor cluster of student computers (Windows Based, some Linux)
  - No Shared file space

Newcastle University

Digital Institute

# Aims

- Run large number of ‘jobs’ against single data set
  - Jobs are independent
- Stage data as few times as possible
  - Reduce network use
- Don’t change user’s code
  - Allows users to adapt to their other situations
- Waste little computation time

Newcastle University

Digital Institute

# Problem with the Condor Model

- You can use Condor to process many jobs
  - Scatter / Gather approach
    - Requires modification of code
  - Remote access to local files
    - Requires re-compiling of user's code
  - Process a set of jobs together
    - Processing can be lost if job is evicted before finishing
    - Could process more if computer is still idle after job
- Want a model which keeps on processing till either all jobs complete or evicted
  - But results not lost

Newcastle University

Digital Institute

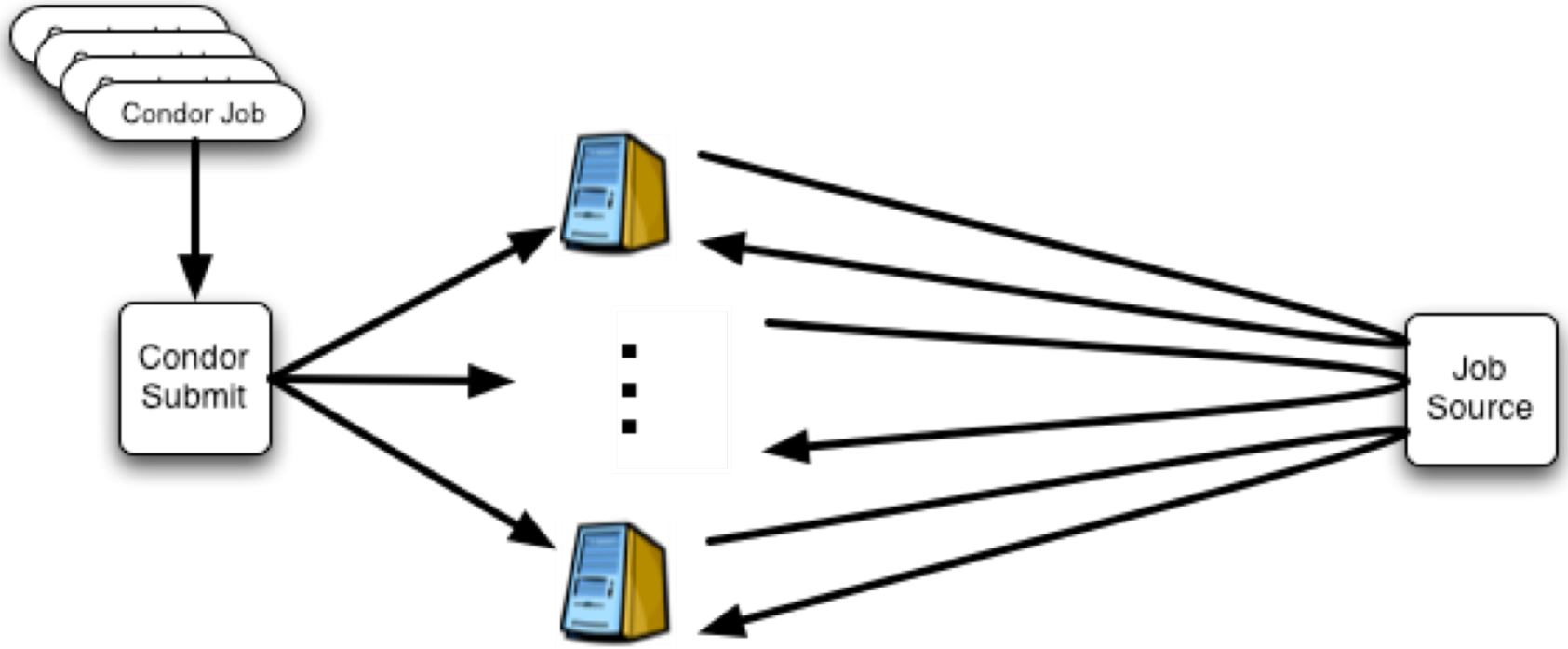
# We need a Pull Model

- Conventionally Condor works as a Push Model
  - Condor pushes jobs to workers
  - Once the job completes the worker is given a new (potentially different) job to perform
- With a pull model the worker requests a new (sub) job
  - Part of the same job set so can maintain the same data
  - So jobs can be made arbitrarily small without wasting data transfer

Newcastle University

Digital Institute

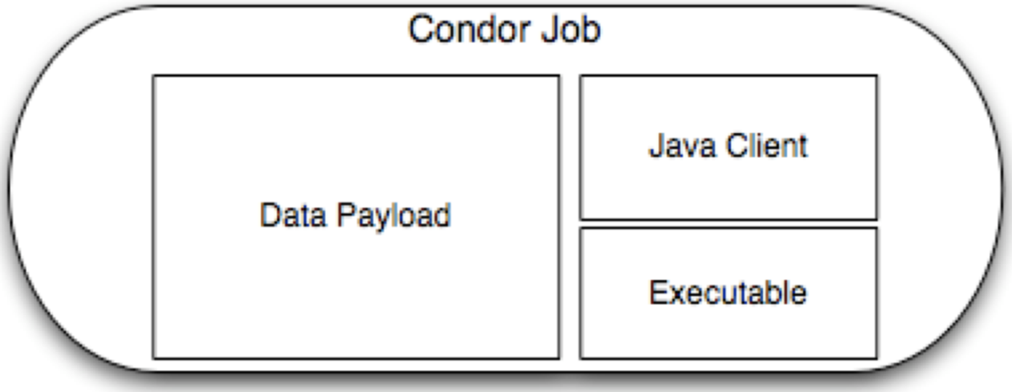
# General Architecture

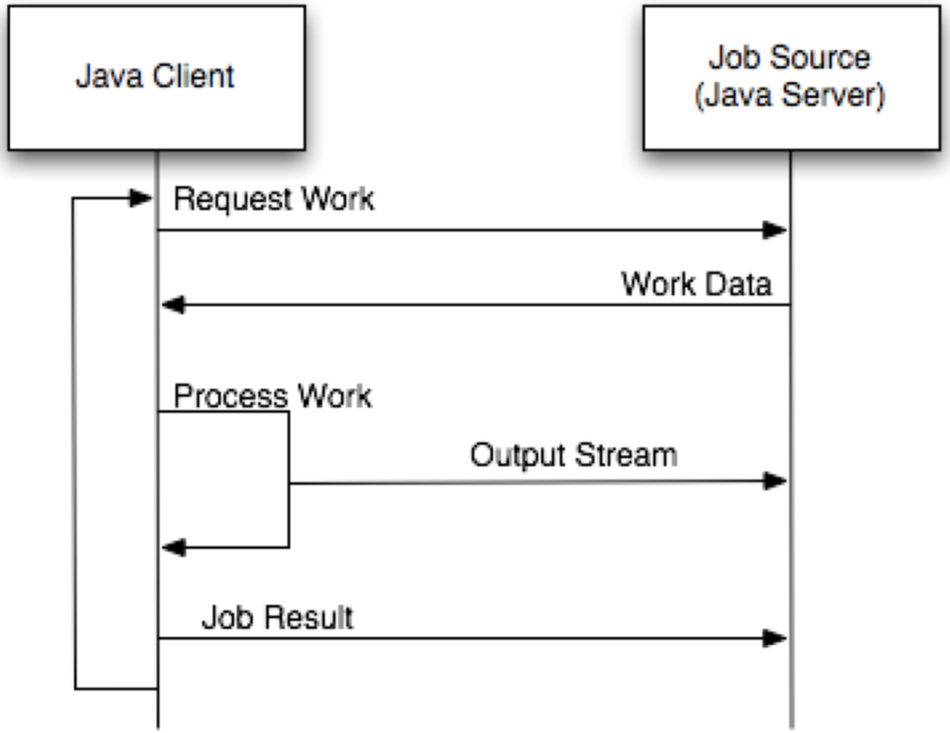



Newcastle University

# Condor Pull Job

Digital Institute



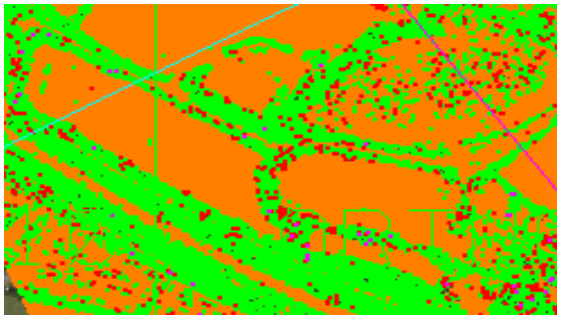




Newcastle University

# E.g.: Lidar (Light Detection And Ranging)

Digital Institute

- An active remote sensing technology which can acquire highly accurate 3D data of the earth's surface along with radiometric information.
- The new generation systems can supply significantly more physical information
  - Strips contain millions of points that needs to be processed to obtain the required information
  - Hence larger data files (~500MB – 4GB)
- This information can help to distinguish between different earth surface features (e.g. vegetation, buildings, roads. etc.) according to their reflectivity and roughness.



Newcastle University

Digital Institute

# Lidar With Condor

- Due to high accuracy requirements, a novel, rigorous Gaussian pulse detection method has been developed in-house (at Newcastle University) and successfully applied for full waveform point cloud post processing
  - Each point is dealt with separately in a matter of seconds
  - Time complexity comes from the number of points
  - Lends itself well to high-throughput computing (Condor)

Newcastle University

Digital Institute

- Reduced run time from Months to Days
  - Average run reduced time from three months to four days

- “The benefit of adopting Condor in my PhD project is to process all overlapped flightlines for two dense project datasets. Both geometric and physical output in the overlapped regions will be used to facilitate ongoing research into lidar point cloud segmentation. This could provide me more data to investigate different land cover regions within the study sites which I couldn’t apply before even with high performance workstation” – Fanar M. Abed.

Newcastle University

Digital Institute

# Conclusion

- Pull Model helps with
  - Processing Large data sets in the absence of
    - Shared file space
    - Condor Compliable Code
    - A *NIX based cluster
  - Using free time on worker nodes more dynamically
  - Removing lost work due to job eviction

Newcastle University