# Use of Archer for Particle Physics

**Andrew Washbrook**
University of Edinburgh

Physics at Extreme Scales, Edinburgh
15th April 2014

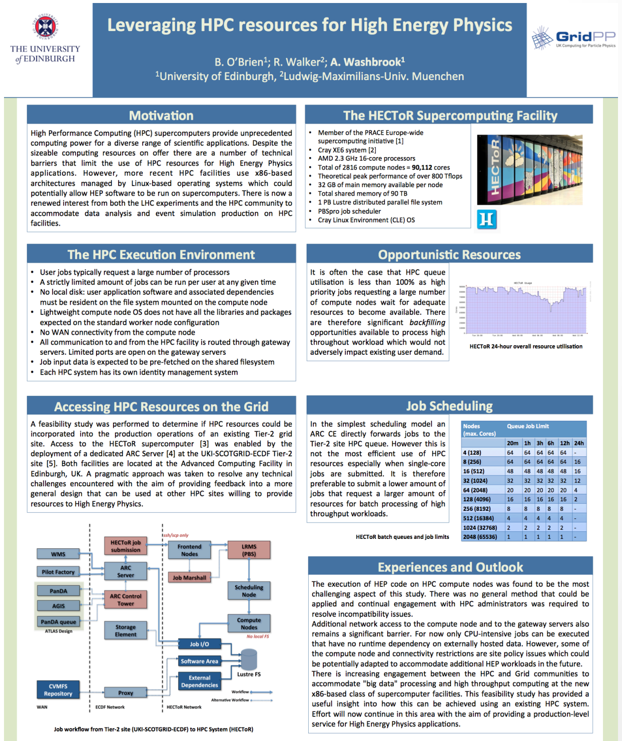
# Motivation

- The Edinburgh Particle Physics group had access to a share of resources on HECToR facility - and now on Archer

**CHEP 2013 conference note:**
**Leveraging HPC resources for High Energy Physics**
[http://indico.cern.ch/event/214784/session/9/contribution/438](http://indico.cern.ch/event/214784/session/9/contribution/438)

- A feasibility study was performed to determine how an example HPC resource could be incorporated into a WLCG Tier-2 Grid site hosted at the same facility

- Now would like to move from HECToR feasibility studies a production-level service using Archer
- We have recently been looking at the issues and challenges in using the new Archer HPC facility for HEP-EX



THE UNIVERSITY of EDINBURGH

## Leveraging HPC resources for High Energy Physics

B. O'Brien[1]; R. Walker[2]; **A. Washbrook[1]**
[1]University of Edinburgh, [2]Ludwig-Maximilians-Univ. Muenchen

GridPP UK Computing for Particle Physics

### Motivation

High Performance Computing (HPC) supercomputers provide unprecedented computing power for a diverse range of scientific applications. Despite the sizeable computing resources on offer there are a number of technical barriers that limit the use of HPC resources for High Energy Physics applications. However, more recent HPC facilities use x86-based architectures managed by Linux-based operating systems which could potentially allow HEP software to be run on supercomputers. There is now a renewed interest from both the LHC experiments and the HPC community to accommodate data analysis and event simulation production on HPC facilities.

### The HECToR Supercomputing Facility

- Member of the PRACE Europe-wide supercomputing initiative [1]
- Cray XE6 system [2]
- AMD 2.3 GHz 16-core processors
- Total of 2816 compute nodes = **90,112** cores
- Theoretical peak performance of over 800 Tflops
- 32 GB of main memory available per node
- Total shared memory of 90 TB
- 1 PB Lustre distributed parallel file system
- PBSpro job scheduler
- Cray Linux Environment (CLE) OS

### The HPC Execution Environment

- User jobs typically request a large number of processors
- A strictly limited amount of jobs can be run per user at any given time
- No local disk: user application software and associated dependencies must be resident on the file system mounted on the compute node
- Lightweight compute node OS does not have all the libraries and packages expected on the standard worker node configuration
- No WAN connectivity from the compute node
- All communication to and from the HPC facility is routed through gateway servers. Limited ports are open on the gateway servers
- Job input data is expected to be pre-fetched on the shared filesystem
- Each HPC system has its own identity management system

### Opportunistic Resources

It is often the case that HPC queue utilisation is less than 100% as high priority jobs requesting a large number of compute nodes wait for adequate resources to become available. There are therefore significant *backfilling* opportunities available to process high throughout workload which would not adversely impact existing user demand.

HECToR 24-hour overall resource utilisation

### Accessing HPC Resources on the Grid

A feasibility study was performed to determine if HPC resources could be incorporated into the production operations of an existing Tier-2 grid site. Access to the HECToR supercomputer [3] was enabled by the deployment of a dedicated ARC Server [4] at the UKI-SCOTGRID-ECDF Tier-2 site [5]. Both facilities are located at the Advanced Computing Facility in Edinburgh, UK. A pragmatic approach was taken to resolve any technical challenges encountered with the aim of providing feedback into a more general design that can be used at other HPC sites willing to provide resources to High Energy Physics.

Job workflow from Tier-2 site (UKI-SCOTGRID-ECDF) to HPC System (HECToR)

### Job Scheduling

In the simplest scheduling model an ARC CE directly forwards jobs to the Tier-2 site HPC queue. However this is not the most efficient use of HPC resources especially when single-core jobs are submitted. It is therefore preferable to submit a lower amount of jobs that request a larger amount of resources for batch processing of high throughput workloads.

| Nodes (max. Cores) | Queue Job Limit | | | | | |
|---|---|---|---|---|---|---|
| | 20m | 1h | 3h | 6h | 12h | 24h |
| 4 (128) | 64 | 64 | 64 | 64 | 64 | - |
| 8 (256) | 64 | 64 | 64 | 64 | 64 | 16 |
| 16 (512) | 48 | 48 | 48 | 48 | 48 | 16 |
| 32 (1024) | 32 | 32 | 32 | 32 | 32 | 12 |
| 64 (2048) | 20 | 20 | 20 | 20 | 20 | 4 |
| 128 (4096) | 16 | 16 | 16 | 16 | 16 | 2 |
| 256 (8192) | 8 | 8 | 8 | 8 | 8 | - |
| 512 (16384) | 4 | 4 | 4 | 4 | 4 | - |
| 1024 (32768) | 2 | 2 | 2 | 2 | 2 | - |
| 2048 (65536) | 1 | 1 | 1 | 1 | 1 | - |

HECToR batch queues and job limits

### Experiences and Outlook

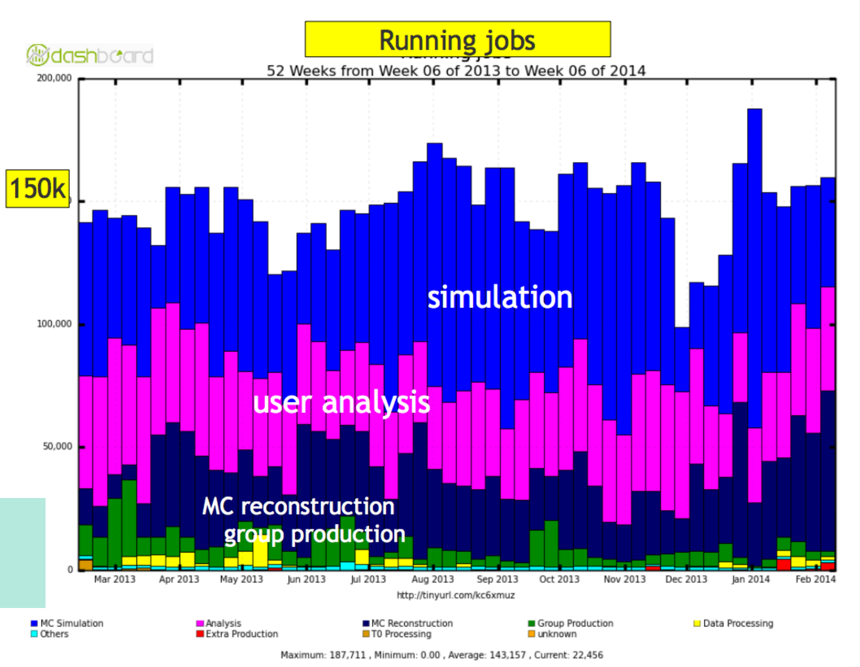
The execution of HEP code on HPC compute nodes was found to be the most challenging aspect of this study. There was no general method that could be applied and continual engagement with HPC administrators was required to resolve incompatibility issues.

Additional network access to the compute node and to the gateway servers also remains a significant barrier. For now only CPU-intensive jobs can be executed that have no runtime dependency on externally hosted data. However, some of the compute node and connectivity restrictions are site policy issues which could be potentially adapted to accommodate additional HEP workloads in the future.

There is increasing engagement between the HPC and Grid communities to accommodate "big data" processing and high throughput computing at the new x86-based class of supercomputer facilities. This feasibility study has provided a useful insight into how this can be achieved using an existing HPC system. Effort will now continue in this area with the aim of providing a production-level service for High Energy Physics applications.

# ATLAS Computing Context

- The ATLAS experiment at the LHC:
  - Processes and manages more than 130 PB of data
  - Uses more than 150k CPUs distributed across 100 computing centres managed by central workload management system (PanDA)



**ATLAS Dashboard 2013/14 Running Jobs**

- Run 2 (2015-2018) data processing will require a **lot** more computing and storage resources
- Can HPC and Leadership Class Facilities help with the increased demand?

- Several HPC sites in Europe and the US are working with ATLAS including:
  - Mira
  - Titan
  - Stampede
  - Hydra, RZG Munich
  - ***Archer***

# HPC vs. High Throughput Computing

The following restrictions apply for HPC usage compared with traditional high throughput computing methods:

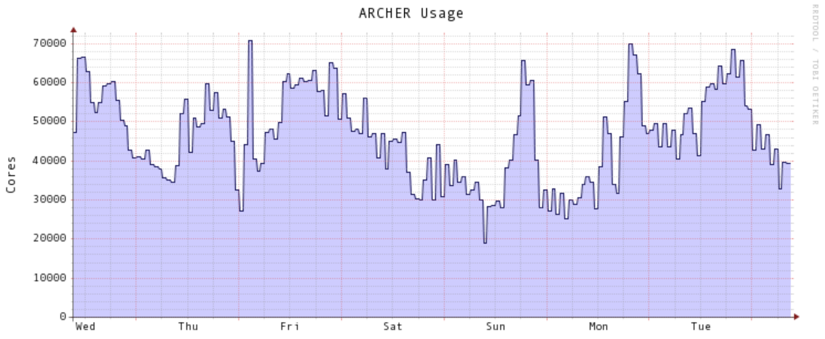
- Network access to and from HPC compute nodes more restrictive than Grid worker nodes
  - No WAN connectivity available
- OS deployed on compute nodes (CLE) is more lightweight than WN OS
  - Optimise code execution by limiting the number of interruptions to compute processes
  - Standard software libraries and packages not available
- No local disk on compute node
  - All job data is expected to reside on the shared filesystem
  - Not designed to cater for applications handling large input data sets and sustained I/O calls during job execution
- Separate identity management policy cannot be coupled to the federated systems we use on the Grid
  - All jobs submitted through my local account (for now)

- Restrictions are mostly driven by HPC user expectations rather than by strict technical barriers
- Exploring where adjustments to system configuration can be potentially adapted to accommodate ATLAS workloads

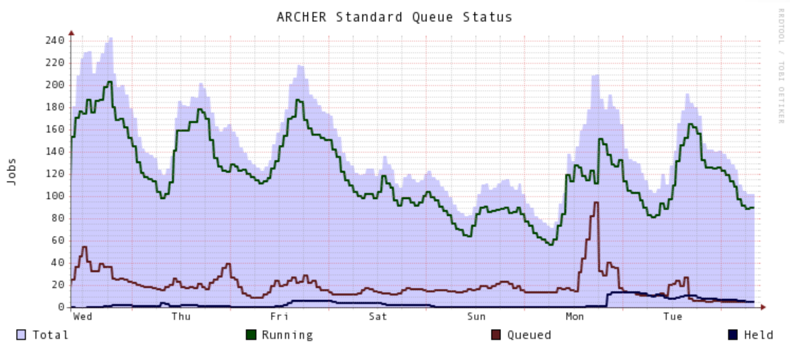
# Job Submission and Scheduling

- Deployed Grid Middleware services at our existing Tier-2 site (ECDF) to enable jobs from the Grid to be routed to Archer

- ATLAS software not currently suited for MPI-type jobs **but** can efficiently process *multi-core* workloads
- Submit single HPC job can steer hundreds of *wholenode* jobs
- Other options explored include offloading critical sections of workload well suited for HPC resources
- Job resource request size can be adapted to queue conditions
- Backfilling could generate slots for HEP-EX use without loss of service to other HPC users



Archer Utilisation (1 week view)



Archer Queue Status (1 week view)

# Outlook

- Aim to provide a production level HPC service in concert with local HEP-EX Tier-2 operations at ECDF in Edinburgh
- Currently investing time and effort into a building a robust setup and to resolve compatibility issues
- Previous experience with running LHC software at a shared cluster facility (ECDF) is proving useful
- Novel solutions will be required to fit the computing environment expectations from ATLAS and other HEP experiments
- Incorporating ideas and solutions from other HPC facilities in the US and Europe rather than working in isolation
- Edinburgh and Archer are well placed to contribute in this area