Flipping the priority: effects of prioritising HTC jobs on energy consumption in a multi-use cluster

Matthew Forshaw[1] A. Stephen McGough[2]

[1]School of Computing Science
Newcastle University, UK

[2]School of Engineering and Computing Sciences,
Durham University, UK

Presentation for ENERGY-SIM 2015

Outline

Background

- High throughput computing
  - Large computational tasks - which can be broken down into short chunks - *'Jobs'*
  - *Well suited to 'embarrassingly parallel' workloads*
  - *Resilient architecture*
  - *All attempts are made to make sure each job completes*
  - *Despite job interruptions due to:*
    - *Hardware and software failures*
    - *'Multi-use' cluster - interactive users*

- Volunteer Computing (e.g. HTCondor, BOINC)
  - Leverage spare capacity on existing infrastructure
  - Resource owners choose who has priority
    - Normally in situations of contention, computers are relinquished, e.g. termination, suspension
  - Leads to detrimental impact on HTC jobs, which must then be re-run elsewhere

Background

- Energy consumption of IT faces increasing scrutiny
- Newcastle University has a strong desire to reduce energy consumption and CO_2 emissions
- Newcastle University's ICT is responsible for 18% of the total electricity bill and the *desktop estate* represents 37% of electricity cost (approx. £320,000)

- Here we relax some of the common computer management policies used in large organisations
- In doing so, can we improve performance and energy consumption?

Motivation: Moving users

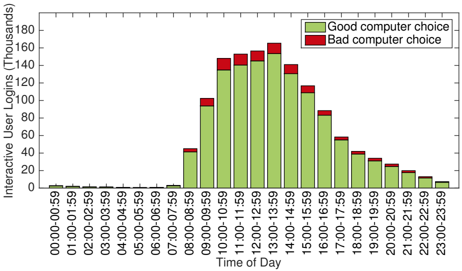
Is it always sensible to terminate a job when a user arrives?

Motivation: Moving users

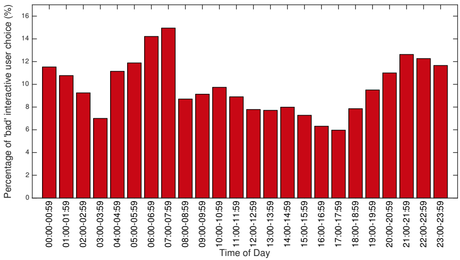
Is it always sensible to terminate a job when a user arrives?

Newcastle University HTCondor System

Interactive user logins by hour

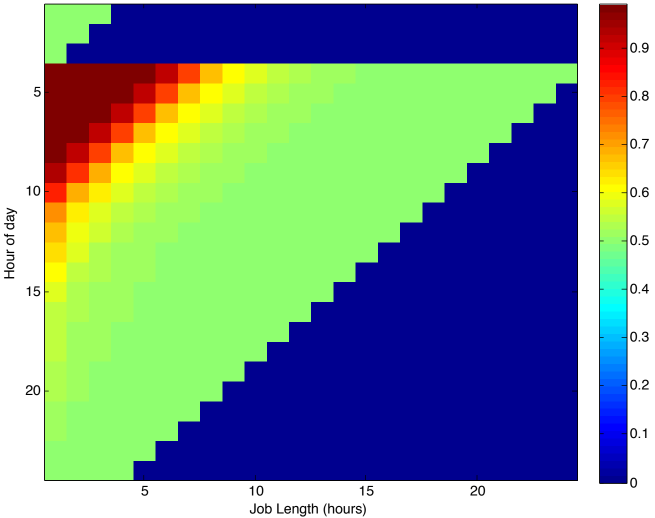


Percentage of 'bad' users per hour

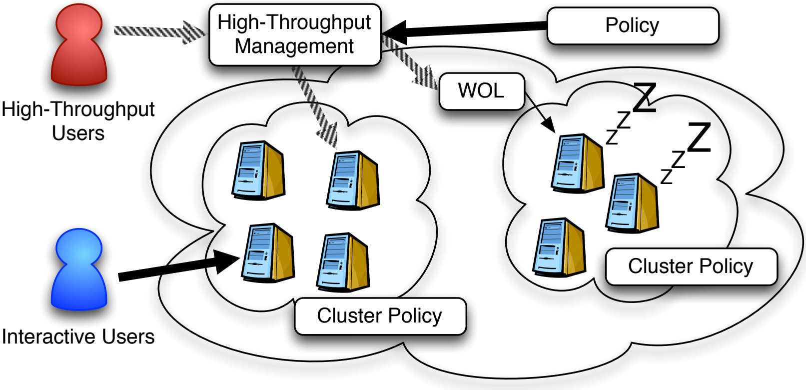


Motivation: Moving nightly reboots

Probability job will complete per hour



Newcastle University HTCondor System



- ~1,400 machines in 35 clusters
  - Opening times
  - Location
  - Availability
  - Four-year procurement cycle
  - Computational power
  - Energy efficiency
- Nightly reboot between 3-5am for maintenance and updates

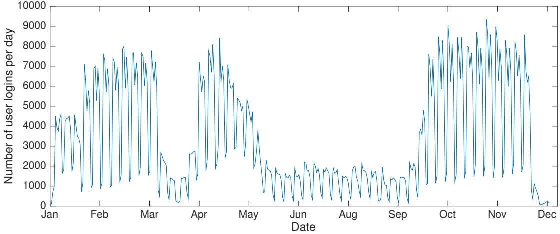
Trace-Driven Simulation

- Developed a trace-driven simulation for evaluation of different policy sets
- Trace logs from a twelve month period from Newcastle University's HTCondor system
  - Interactive user activity
    - Log in timestamp, log out timestamp, Computer name
  - HTCondor Job submissions
    - Submission time, job duration, memory footprint, resource requirements, ...

Type	Cores	Speed	Power Consumption		
			Active	*Idle*	*Sleep*
Normal	2	~3Ghz	57W	40W	2W
High End	4	~3Ghz	114W	67W	3W
Legacy	2	~2Ghz	100-180W	50-80W	4W

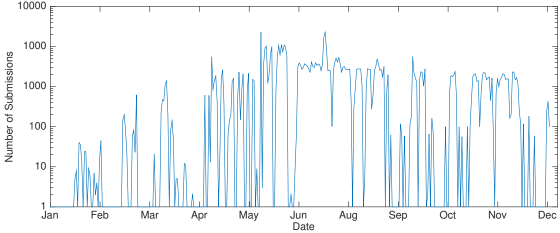
Newcastle University HTCondor System

Interactive User Trace



- 1,229,820 user sessions, 39,610 unique users

HTCondor Workload Trace



- 561,851 jobs, ~53 years of work

Policies: Reboot Policies

- RB1
  - Machines reboot according to cluster management policies enacted in 2010, between 3-5am
- RB2
  - Machines reboot when cluster closes for the night
  - Machines within 24 hour clusters reboot at midnight
- RB3(n,r)
  - An extension of RB2; if an HTC job is currently running on a machine, reboot is deferred until n minutes before the cluster reopens.
  - We introduce a random component in the reboot scheduling η, where η is uniformly distributed on $[-r, r]$
- RB4
  - Newcastle University default power saving scripts
  - Active machines are polled ever 10 minutes and are suspended if there is no user present, and the CPU is idle
  - Computers scheduled to reboot randomly between 01:00-06:59

Policies: User allocation policies

- U1: Exact
  - Users arrive to the computer specified in our trace data for 2010
- U2: Random
  - Users are allocated to their original computer choice if this computer is not currently occupied with a job or interactive user
  - Alternatively, an idle or sleeping computer is selected at random
- U3(n)
  - Users are allocated to their original computer choice if this computer is idle, sleeping, or has an HTC job with a runtime less than n minutes
- U4
  - An extension of Policy U3, allowing users to be reassigned to other clusters within the same physical location

Policies: Computer power management

- P1: Computers are permanently awake
- P2: Computers are on during cluster opening times or sleeping otherwise with no ability to wake up
- P3(n): Computers sleep after n minutes of inactivity with no wakeup for high-throughput jobs
- P4(n): Computers sleep after n minutes of inactivity with HTC being made aware of their availability
  - Allows the HTC system to wake computers when required

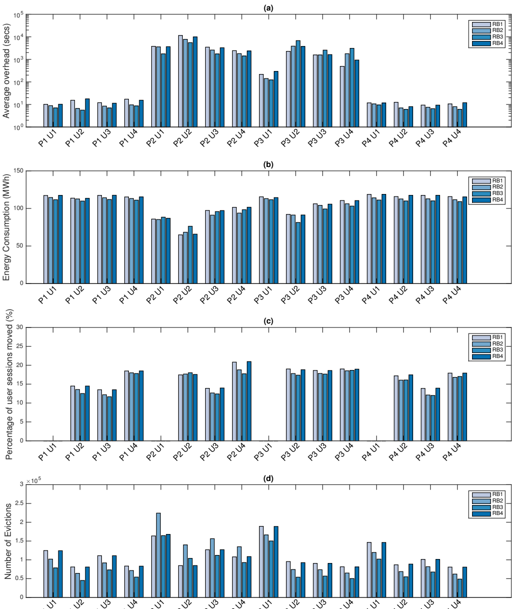
Results - TBC

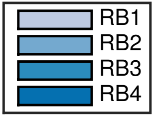
Average overhead (s)

Energy Consumption (MWh)

% of user sessions moved

Number of evictions





Conclusions

- We have explored, through trace-driven simulation, the impact of relaxing commonly adopted policies governing the operation of volunteer HTC clusters
- Potential for significant improvements of performance on energy consumption
  - ˜20-74% reduction in overheads incurred by HTC jobs
  - ˜12.4% reductoin in energy consumption
- Communication among campus cluster operators and HTC system managers is essential
- Future Work: Operating policies for HTC systems which reconcile the different (often opposing) demands of the cluster owner, HTC submitter, and interactive user

matthew.forshaw@newcastle.ac.uk stephen.mcgough@durham.ac.uk