Batch Farm Evolution

James Adams

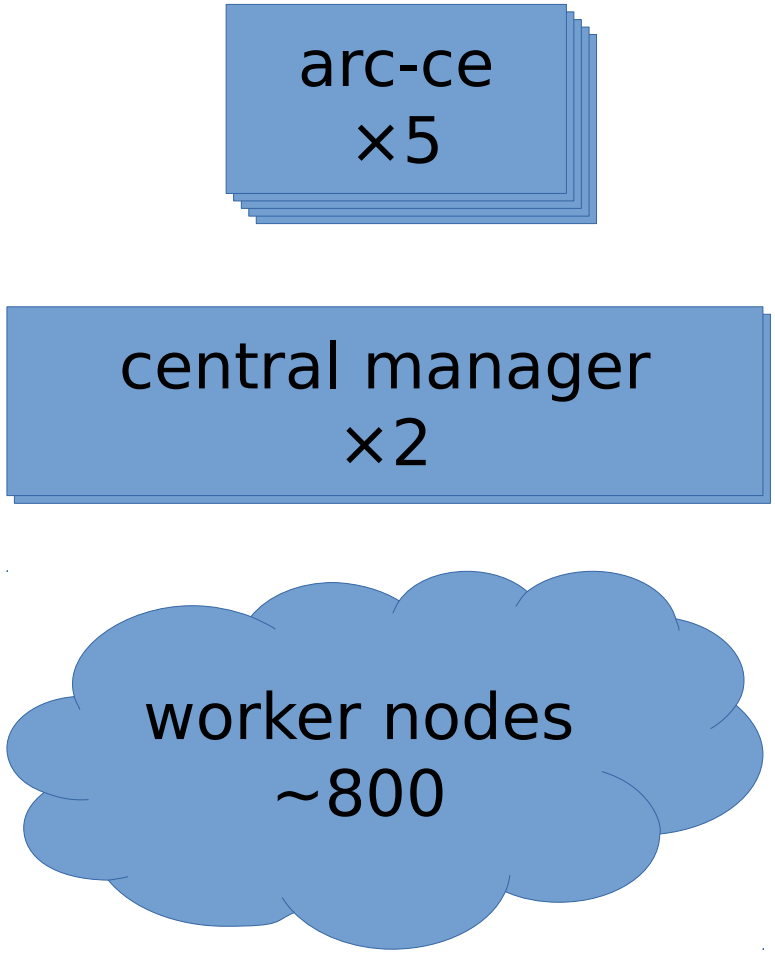
<obligatory pokemon picture goes here>

Overview

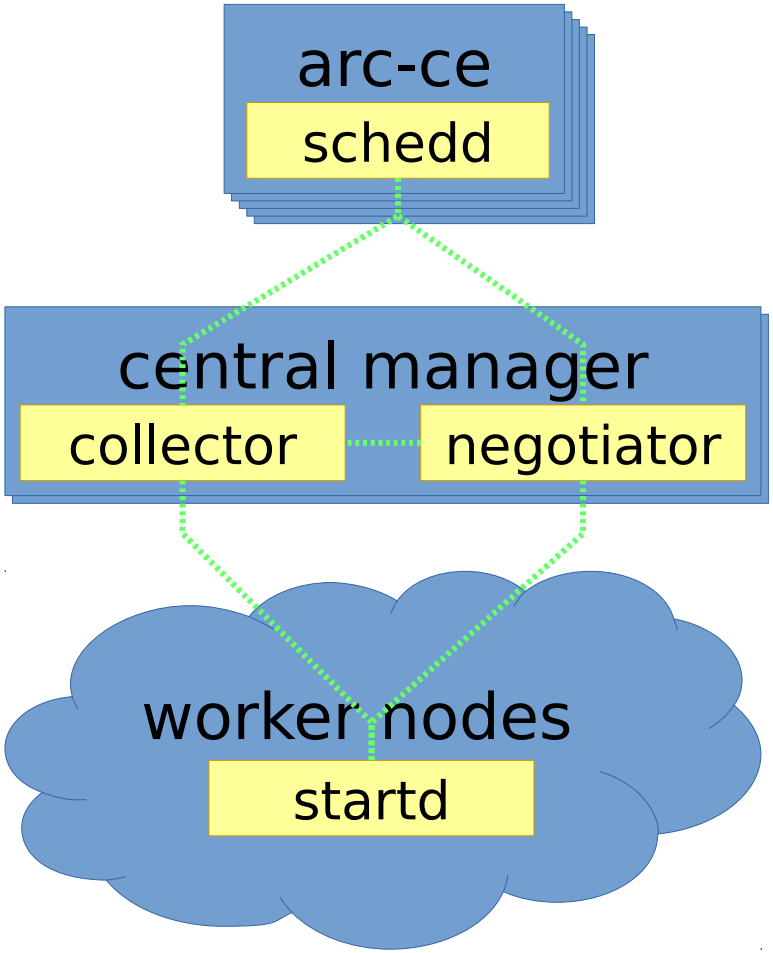
- Current state
- Recent events
- Future thoughts and plans

Disclaimer: I have only recently become reacquainted with the batch farm, many thanks to John Kelly and Catalin Condurache for bringing me up to speed.

Batch Farm



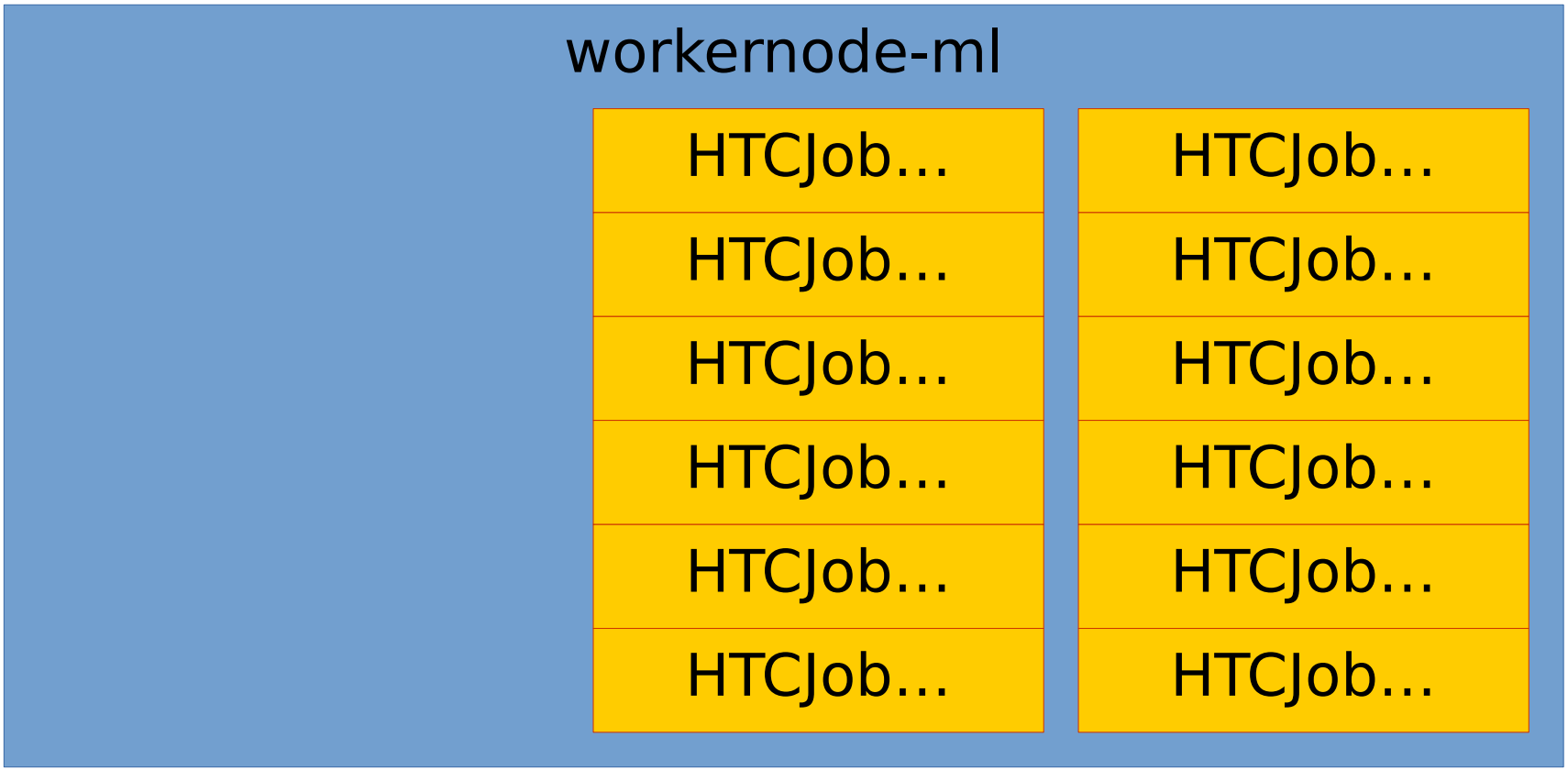
HTCondor Batch Farm



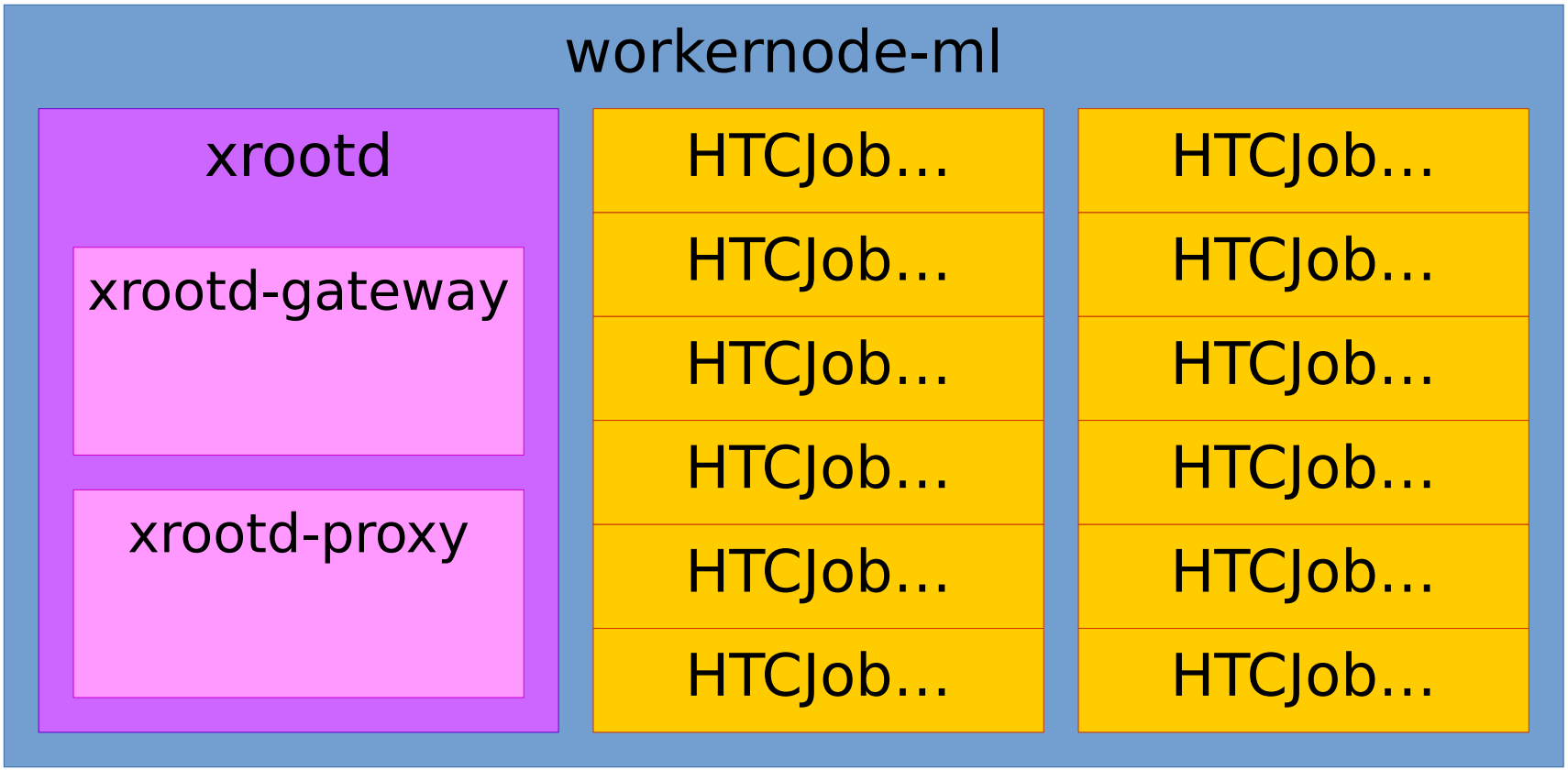
Worker nodes

- All 800+ run:
  - SL7.5 user-land
  - Kernel 4.17.11 or later (kernel-ml from elrepo)
  - HTCondor 8.6.9
  - Docker 18.03.0-ce
- Both CentOS 6 and 7 user-lands provided for jobs.
  - ARC-CE wraps jobs in docker container
  - Based on VO submitting to *EL7* “queue”
  - Migration therefore under VO control
    - Roughly 18% jobs are using EL7

Inside a Worker Node



Inside a Worker Node



Historical Aside

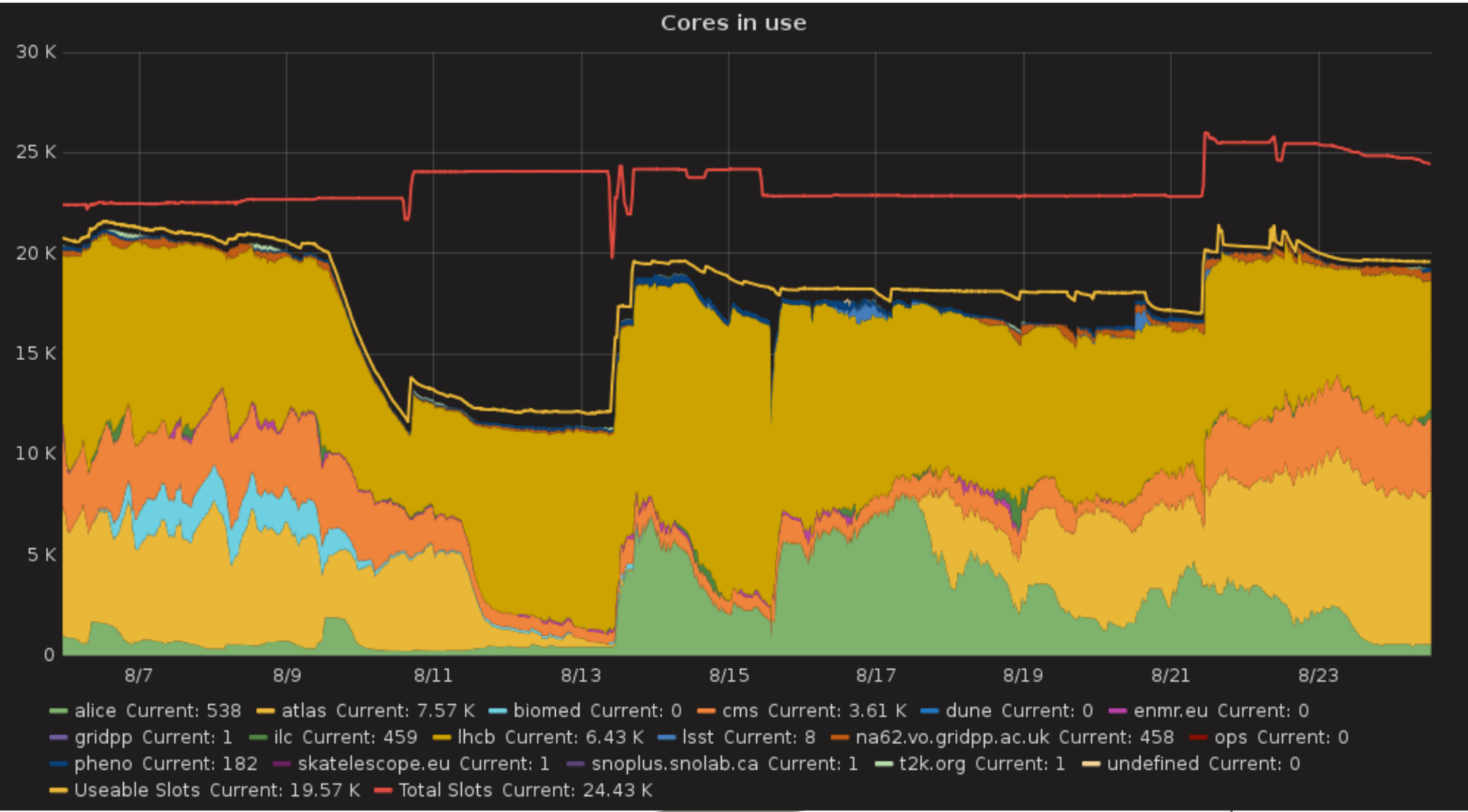
- Went through many changes under Andrew Lahiff
  - Not all of these well understood by team
  - Using Docker for everything
  - Gateways for echo on Wns
  - Magic
- Became somewhat unloved over the last year
  - Some things stopped working
  - Others replaced by upstream functionality in HTCondor
- Has now been made my problem

Recent News

- 2017 XMA nodes in production
  - 56 nodes each with 56 slots (additional 3136 slots)
- Resurrected local cloud-bursting capability
  - Additional ~1500 slots
  - Still semi-manual (for now)
- Identified number of nodes with serious problems
  - Took decision to fix as a big-bang change

Recent Problems

- Some nodes still running stock kernel
  - Fixed — Nodes re-installed
- Typo in mainline kernel script
  - Fixed
- Some file-systems created with incorrect flags
  - Fixed — Nodes re-installed
- Some nodes running older Docker version
  - Fixed — Upgraded
- Network configuration fighting authconfig over NIS
  - Fixed — Removed NIS
- Attempting to configure incorrect/non-existent network interfaces
  - Fixed — Configuration now renames interfaces (update reality to match config)
- Efficient reboot script cron being called without args
  - Fixed
- Docker using 172.17.0.0/16 — routed on campus
  - Mitigated
- Telegraf version too old for mainline kernel
  - Fixed
- Firmware bugs in half of 2015 nodes causing instability/hourly reboots
  - Ongoing, fixed code must be deployed manually


Cores in use
30 K
25 K
20 K
15 K
10 K
5 K
0
8/7
8/9
8/11
8/13
8/15
8/17
8/19
8/21
8/23
alice Current: 538
atlas Current: 7.57 K
biomed Current: 0
cms Current: 3.61 K
dune Current: 0
enmr.eu Current: 0
gridpp Current: 1
ilc Current: 459
lhcb Current: 6.43 K
lsst Current: 8
na62.vo.gridpp.ac.uk Current: 458
ops Current: 0
pheno Current: 182
skatelescope.eu Current: 1
snoplus.snolab.ca Current: 1
t2k.org Current: 1
undefined Current: 0
Useable Slots Current: 19.57 K
Total Slots Current: 24.43 K

Future

- SL7 migration
  - “Encourage” VOs to migrate (before 30th Nov 2020)
- SL8 migration
  - As soon as it appears!
- Establish routine cycle of well-tested upgrades
  - HTCondor
  - Docker
  - Kernel
- Evaluate migration to HTCondor Ces
  - Dependent on porting our magic

Future

- Period of stability
  - If nothing else, prove that everything is working
- Full automation of routine operations
  - Upgrades and rolling-reboots
  - Cloud bursting
- Work with HTCondor team to upstream our magic
  - Pre-emptable backfilling has already made it
  - Efficient multi-core draining
  - Discussions next week!

Future

- Try to improve operations and availability
  - All while using less effort
- Redesign processes to maintain deployed capacity
  - Stay above pledge!
  - VOs signing off monthly on their used capacity
- Consider some static partitioning
  - Multicore only nodes?
  - WLCG only nodes?
  - Stupidly big memory nodes?
- Find somewhere to stuff our extra DIMMs

IRIS

- For HTC batch work the answer is HTCondor
  - Users will either submit directly or via CEs
- Tier 1 unwilling to support more (or bespoke) systems
  - Less effort in the short term
  - Better in the long run to teach people to use one batch system well than many badly
- However… SCD is trying to work more closely
  - More collaboration between HPC and HTC teams?
  - HPC at RAL is migrating to SLURM
  - …?

Questions?