CERN

CERN Batch system

ben.dylan.jones@cern.ch

Batch Overview

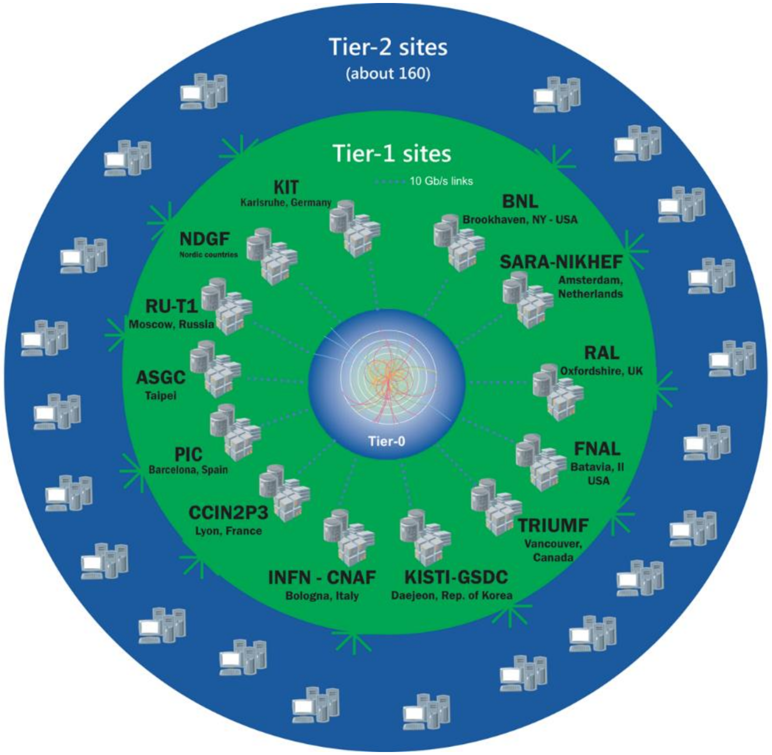
- CERN Batch system to process CPU intensive workload ensuring fairshare among various user groups
- Maximize utilization, throughput, efficiency
- Split of Grid or “local” submissions
- 110K cores
  - Mostly VM
  - 16 core or 8 core VMs
- 650K jobs finish a day

Worldwide LHC Computing Grid

TIER-0 (CERN):
data recording, reconstruction and distribution

TIER-1:
permanent storage, re-processing, analysis

TIER-2:
Simulation, end-user analysis



nearly 170 sites, 40 countries

~350’000 cores

500 PB of storage

> 2 million jobs/day

10-100 Gb links

Not just the LHC…

Local v Grid

- Roughly equal numbers of jobs submitted via each method
  - Helps smooth utilization
- Grid submission use X509 certificates, submitted via experiment workload managers to Compute Elements
- Local submission typically directly from users, using kerberos auth on shell services
- Local jobs typically less predictable workload

LSF to HTCondor

- Proprietary vs Open
- Scale
  - LSF has 5K host limit
  - Can scale but only by splitting up instances
  - Central master for queries
  - Some divergence of feature set from “high throughput computing”
- HTCondor community
  - Great support from both HTCondor core team and others in WLCG
  - So far for us, CMS global pool pushing scale

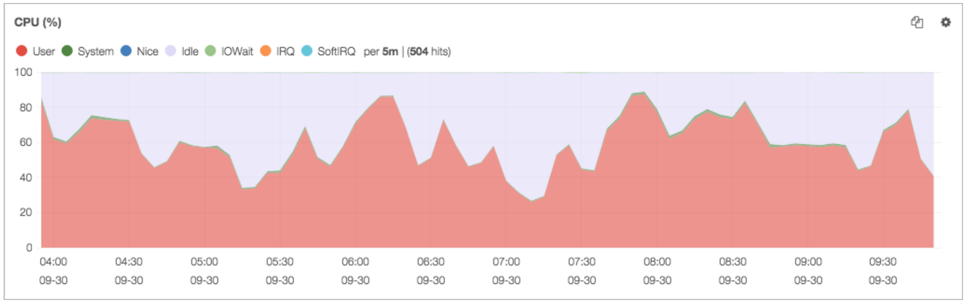
Batch Machine Size

- LSF:
  - 15 slot (16 core), 30gb RAM. Newer machines have SSD. Hyperthreaded (outside ATLAS-T0)
- HTCondor:
  - 8 core, 2gb / core advertised (-5% hv tax). Hyperthreaded
- New hw arriving with 40HT cores & 128GB RAM, making 10 core VMs for HTCondor
- External Cloud has till now been 4 core
  - 8 core in future to make things a bit more consistent

The “Kilo-1” configuration

- NUMA + Pinning
  - 1-to-1 vs. 1-to-N no difference
- 2MB huge pages
  - 1GB slightly better
- EPT on
  - EPT off still better in HS06

VM sizes (cores)	Before	After
4x 8	7.8%	3.3% (batch WN)
2x 16	16%	4.6% (batch WN)
1x 24	20%	5.0% (batch WN)
1x 32	20.4%	3-6% (bare SLC6 … batch WN)



ATLAS T0 host with batch VM running the new config: throughput for recon jobs 20% higher!

OpenStack Kilo will fully support our desired configuration!

Arne Wiebalck et al. – CPU Optimizations of Compute Resources in the CERN Cloud 9

Multicore / memory

- Normal practice is slots of 1 core / 2gb ram / 20gb scratch disk
- ATLAS T0 require more memory & no HT
- Multicore requirement is 8 core, again memory scaled, but increases job memory efficiency
  - Draining / defragging via HTCondor (not LSF)
- Newer hardware more memory per slot (~3gb)

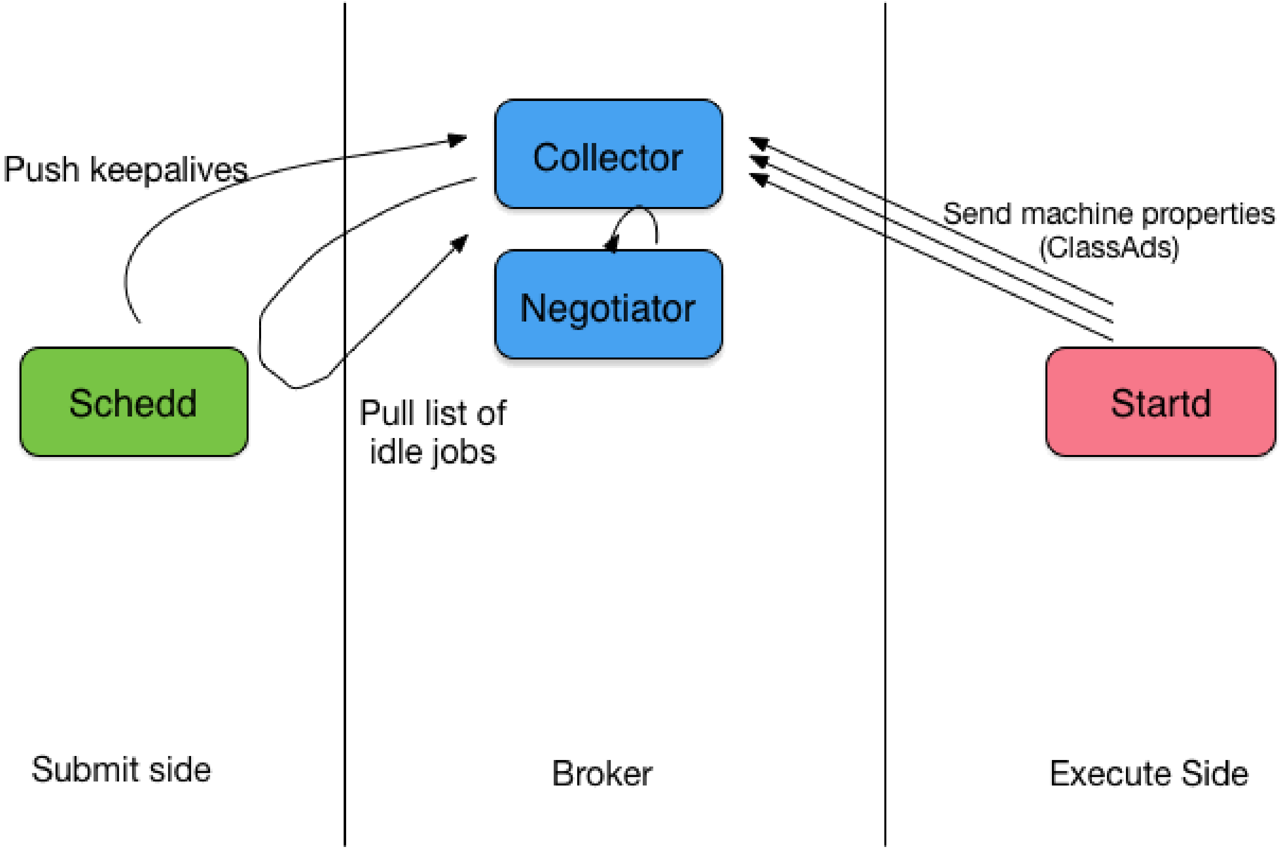
HPC

- A number of HPC facilities being deployed or expanded.
- HTCondor support for larger MPI is patchy
  - UW themselves don’t use HTCondor for HPC…
- Larger MPI jobs will run on dedicated Linux HPC cluster using SLURM
- Backfill submitted via HTCondor

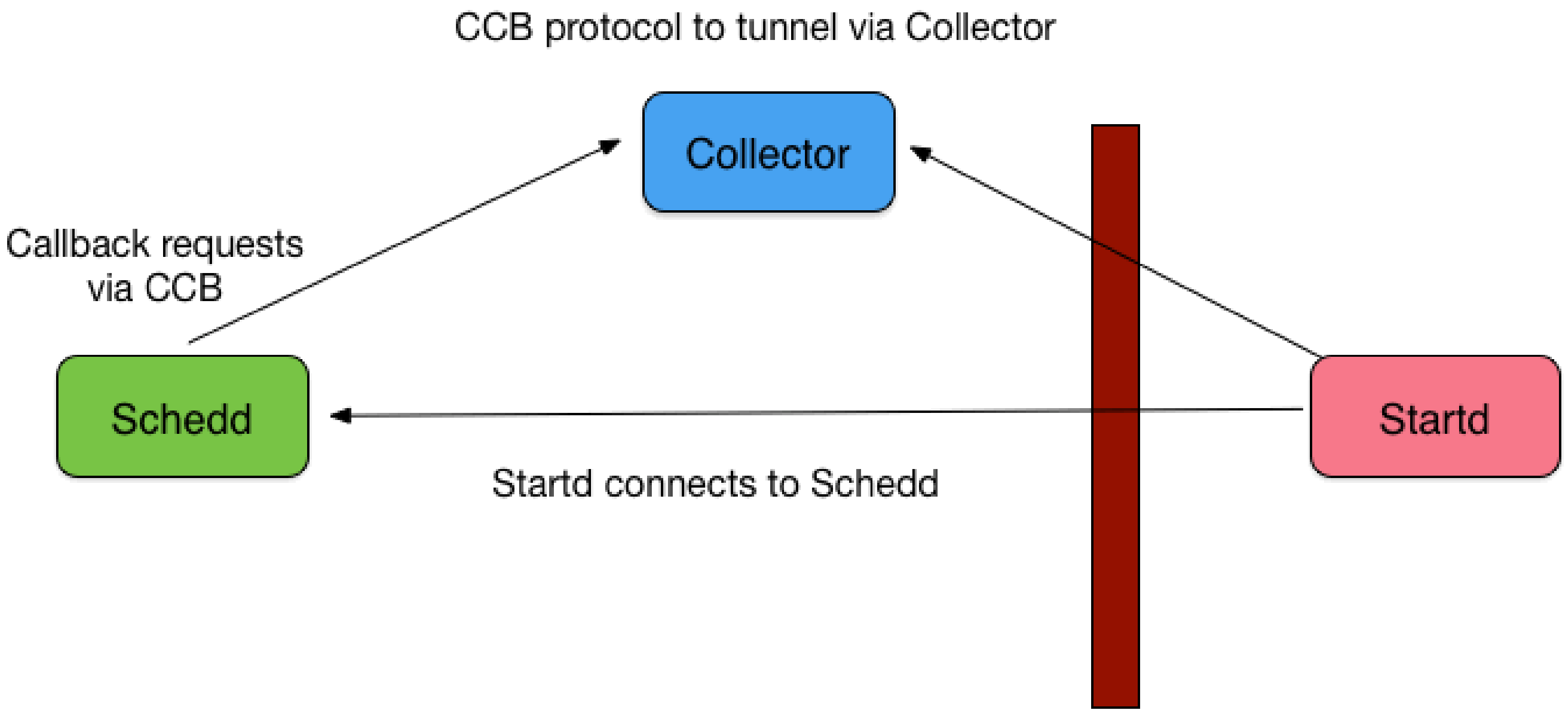
Cloud

- Addition of Cloud resources to general batch pool
- Can we manage external resources seamlessly in terms of provisioning, tools, presentation to customers?
- Activities with SoftLayer, T-Systems, and in future with HNSciCloud

HTCondor communication



Communication via firewall


CCB protocol to tunnel via Collector
Collector
Callback requests via CCB
Schedd
Startd
Startd connects to Schedd

Condor job routes

- HTCondor-CE feature
- Defaults to set default datacentre, HEPSPEC or cores of undefined machines
- Routes have helped partition public cloud whilst maintaining single point of submission

```
 [
     TargetUniverse = 5;
     name = "External_Cloud";
     set_Requirements = (XBatch =?= True);
     set_WantExternalCloud = True;
     Requirements = (TARGET.WantExternalCloud =?= True)
|| (TARGET.queue =?= "WantExternalCloud") ||
(TARGET.queue =?= "externalcloud");
   ]
```

Toolset

Monitoring	Grafana		
Orchestration	mcollective	RunDeck	Hacky ssh loops
Configuration	Puppet / Foreman		
Personalization	cloud-init		Provisioning scripts
Provisioning	Terraform	Packer	Nova Scripts

Future

- Complete migration to HTCondor
- Exploit container tech for payloads
  - Helps with external cloud to avoid messing around with images
  - HTCondor can manage containers
- Investigate bare-metal deployment
  - Container management via HTCondor available
  - Needs OpenStack Ironic
- More cloud activity
- Efficiency, Packing, HA

Questions?