Information System Evolution and Machine/Job Features

Andrew McNab
University of Manchester,
GridPP, and LHCb

Overview

- Information Systems Evolution TF
  - “Replacing BDII?”
- Machine/Job Features TF
  - “Sites communicating to jobs”
- Overview of links between task forces
  - “How do the task forces fit together?”

Information Systems Evolution task force

Information System Evolution Task Force

- Aims to simplify information system (currently BDII) and better align it with experiments'/sites' needs
- Quickly became clear that BDII is overkill and has lots of unused and/or inaccurate info
- Request from sites to remove Site BDII (and maybe local Top BDIIs)
- Three places to publish info considered: service's LDAP or HTTP server, GOCDB, or CRIC
- Maria Alandes Pradillo gave a detailed status report at the July GDB

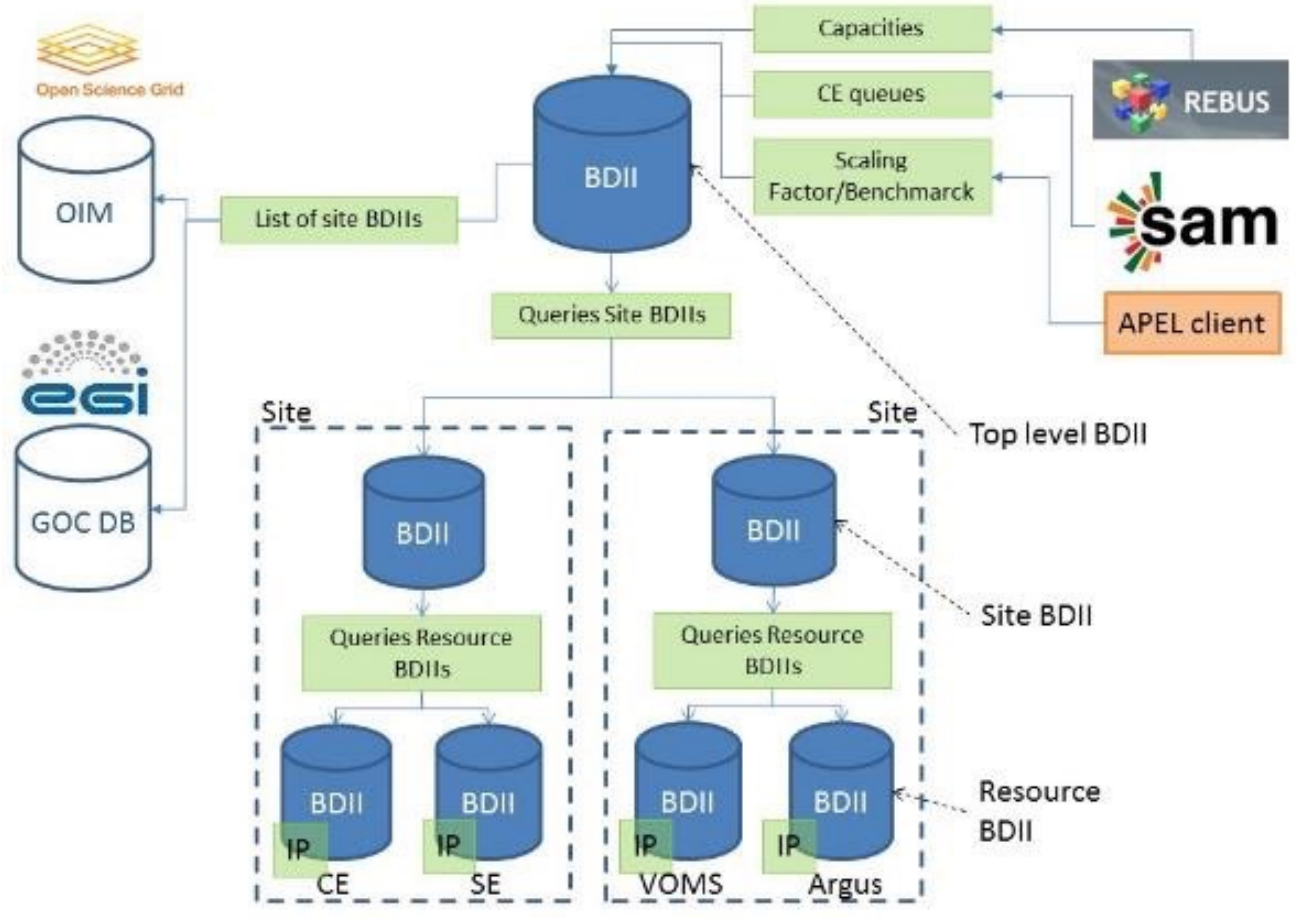
LHC VOs dependencies on the IS (I)

	Attributes vs Information Sources	Resource BDII	Site BDII	Top BDII	GOCDB	OIM	MyOSG	REBUS	Manual modifications?
ALICE	Status of CEs	D/C (CREAM)	D/C (ARC)						NO
	Number of Waiting Jobs								
	Number of Running Jobs								
ATLAS	List of services and associated information: • SEs • CEs • PerfSonars				D/C	D/C			YES
	Queue name	D		C					
	MaxCPUTime	D		C					
	MaxWallClockTime	D		C					
	List of sites and associated information				D/C	D/C			
	Site properties (Lat, Long)	D		C					
	Site downtimes				D/C	D/C			NO
	HS06	D (EGI)					D (OSG)	C	YES
	Logical CPUs	D						C	
CMS	List of CEs	D		C					NO
	Queue name	D		C					YES
	MaxCPUTime	D		C					
	MaxWallClockTime	D		C					
	Logical CPUs	D		C					
	Site downtimes				D/C	D/C			NO
LHCb	List of CEs	D		C					NO
	MaxCPUTime	D		C					
	CPUScalingReference	D		C					
	Site properties (Lat, Long)	D		C					
	Site downtimes				D/C				

	Dynamic: it changes very frequently
	Static or semi-static information: it changes very rarely
D	Defined: where the information is defined by the sites
C	Consumed: where the information is consumed by the experiments

From Maria Alandes Pradillo's July GDB talk

Current WLCG Information System

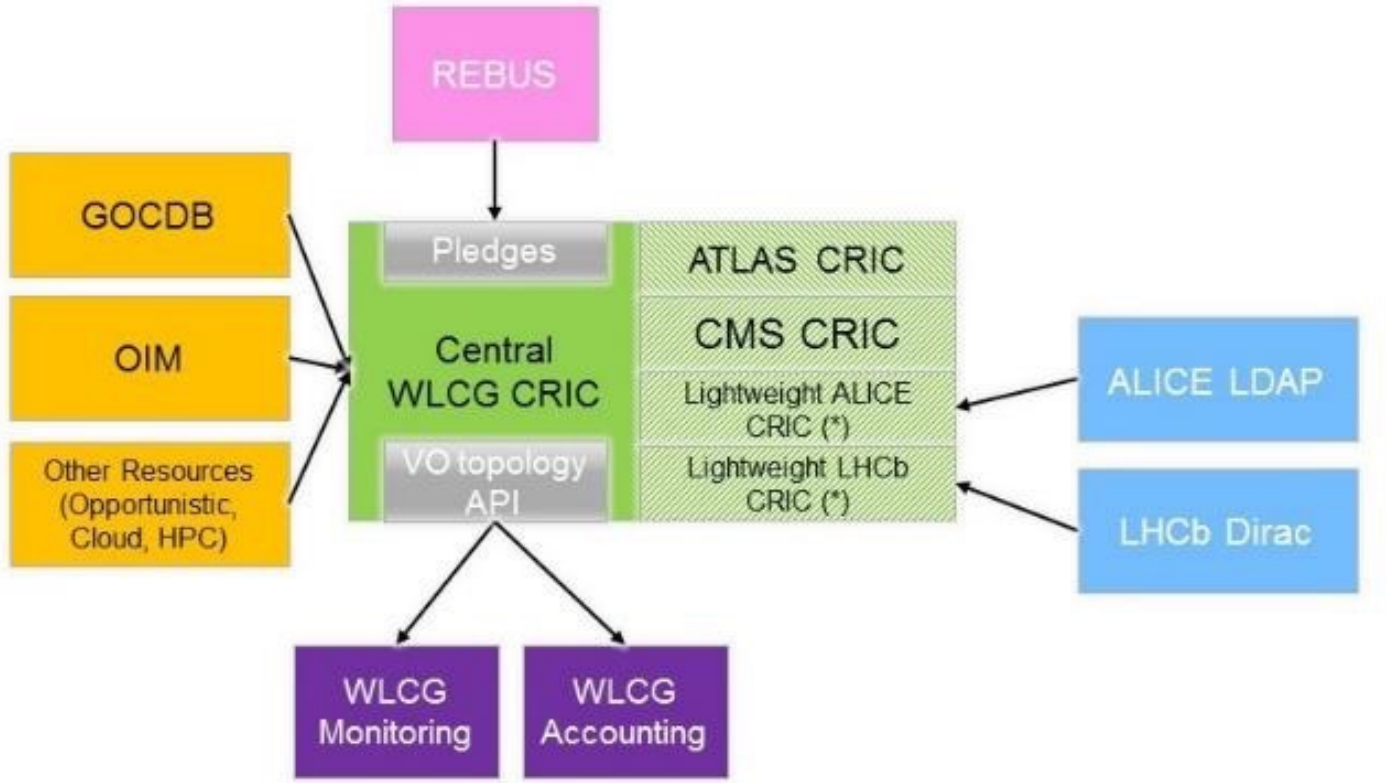


From Maria Alandes Pradillo's July GDB talk

CRIC

- ATLAS proposed that the other experiments use a generalised version of its AGIS system
  - CMS and ATLAS now developing “CRIC”
  - Quite like GOCDB, but run by experiment(s)
- ALICE and LHCb have broadly same response to this
  - May use CRIC if information better quality than alternatives, but otherwise don’t feel need for it
  - Existing plan for site services to publish information to extended GOCDB should be sufficient

Future WLCG IS



(*) Maintained by WLCG to store very simple experiment topology information (i.e. experiment names)

Implications of removing BDII

- Use CRIC for these three in the future?
  - BDII used by REBUS for capacities
  - By APEL client for finding endpoints and by site APEL for getting CE's HS06
  - SAM/ETF get CE names from BDII
- EGI assumes/requires BDII
  - Concerned about WLCG sites dropping BDII
  - But GOCDB is an EGI service too so maybe ok

Machine/Job Features task force

Machine/Job Features Task Force

- MJF is a common API that jobs can use to discover the parameters of their environment
  - eg wall clock time limit
- Otherwise requires a patchwork of environment variables and command call-outs
  - Different for each batch system: qstat etc
  - Not available in VM-based environments
- So N experiments have to write implementations for M batch systems (N x M)
  - With MJF, goes more like N + M

Key / values

- HSF-TN-2016-02 has the full list with definitions
- Sites should supply them if they know the value (eg HS06)
- Values can typically be discovered from batch system, with OS values as a fall-back
- shutdowntime allows sites to declare a cut-off when draining

$MACHINEFEATURES

total_cpu
hs06
shutdowntime
grace_secs

$JOBFEATURES

allocated_cpu
hs06_job
shutdowntime_job
grace_secs_job
jobstart_secs
job_id
wall_limit_secs
cpu_limit_secs
max_rss_bytes
max_swap_bytes
scratch_limit_bytes

Transport mechanisms

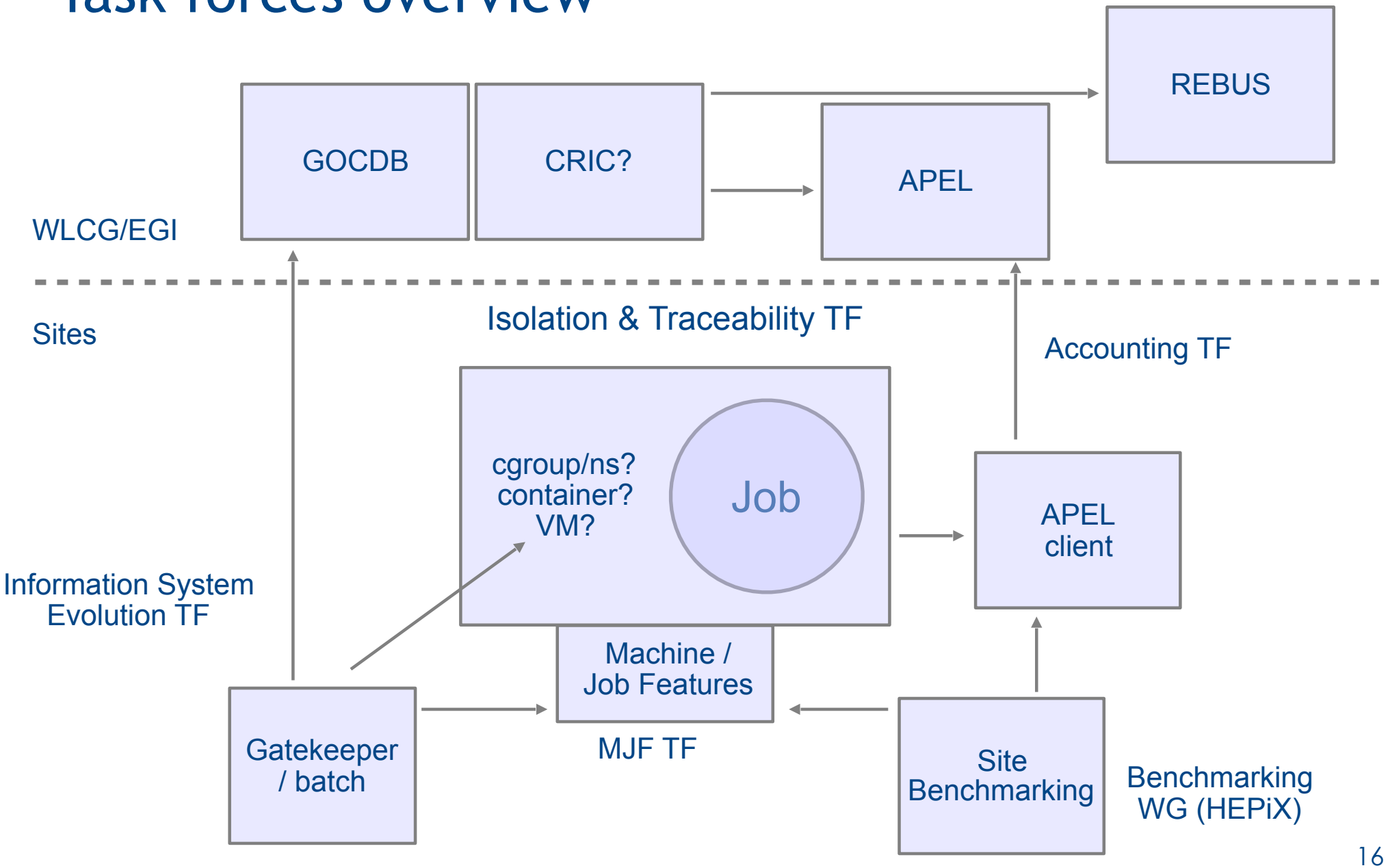
- Jobs expect $MACHINEFEATURES and $JOBFEATURES to point to “directories” containing key/value pairs
  - File name is key; content is value
- Simple cases: $MACHINEFEATURES=/etc/machinefeatures
- For worker nodes, usually local directories
- For VMs though, “directory” is a URL on a web server populated by the VM lifecycle manager
  - EC2/OS metadata keys to discover URLs

Implementations

- Vac/Vcycle supply MJF directories to their VMs
- PBS/Torque and HTCondor scripts exist in GitLab and as RPMs
  - Common code where possible; same ideas
  - Tested/running at Manchester, PIC, Cambridge, Liverpool (thanks!)
- Grid Engine scripts in production at GridKa
- More volunteer sites still needed
- See https://twiki.cern.ch/twiki/bin/view/LCG/MachineJobFeaturesImplementations

How do the task forces fit together?

Task forces overview



Common definitions key to points of overlap

InfoSys(proposed): "GLUE2ExecutionEnvironmentLogicalCPUs: the number of single-threaded benchmark instances run when benchmarking the Execution Environment, corresponding to the number of processors which may be allocated to jobs. Typically this is the number of processors seen by the operating system on one Worker Node (that is the number of "processor :" lines in /proc/cpuinfo on Linux), but potentially set to more or less than this for performance reasons. This value corresponds to the total number of processors which may be reported to APEL by jobs running in parallel in this Execution Environment, found by adding the values of the "Processor" keys in all of their accounting records."

MJF: "$JOBFEATURES/allocated_cpu: number of processors allocated to the current job"

APEL: "Processors: number of processors"

InfoSys(proposed): "GLUE2BenchmarkValue: the average benchmark when a single-process benchmark instance is run for each processor which may be allocated to jobs. Typically the number of processors which may be allocated corresponds to the number seen by the operating system on the worker node (that is the number of "processor :" lines in /proc/cpuinfo on Linux), but potentially set to more or less than this for performance reasons. This should be equal to theServiceLevel in the APEL accounting record of a single-processor job, where the APEL "Processors" key will have the value 1"

MJF: "$MACHINEFEATURES/hs06: Total HS06 rating of the full machine in its current setup. HS06 is measured following the HEPiX recommendations, with HS06 benchmarks run in parallel, one for each processor which may be allocated to jobs."

APEL: "ServiceLevel: Value of either HepSpec2000 or SpecInt2000"

Summary

- Significant changes being proposed to information system
  - Especially getting rid of BDII dependency?
- Machine/Job Features being deployed now
- Lots of connections between task forces
- Lots of scope for site people to get involved if interested …