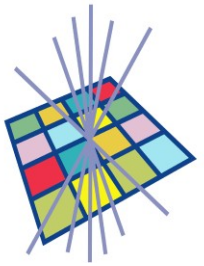
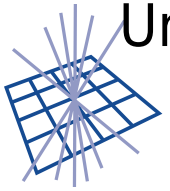


ScotGrid:
Providing an Effective Distributed Tier-2 in the LHC Era

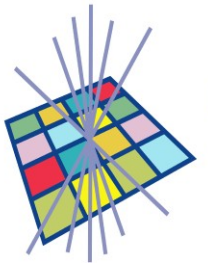
Sam Skipsey

David Ambrose-Griffith, Greig Cowan, Mike Kenyon, Orlando Richards
Phil Roffe, Graeme Stewart

Universities of Glasgow, Edinburgh and Durham

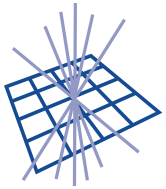


Itinerary



ScotGrid
Scottish Grid Service

- Fabric management and infrastructure consideration.
- "Special cases" – Durham + ECDF
  - 2/3 is just a large minority?
- Experiment Production
  - A solved problem?
- User analysis optimisation.
- "Tier 2.5" - supporting local users locally.

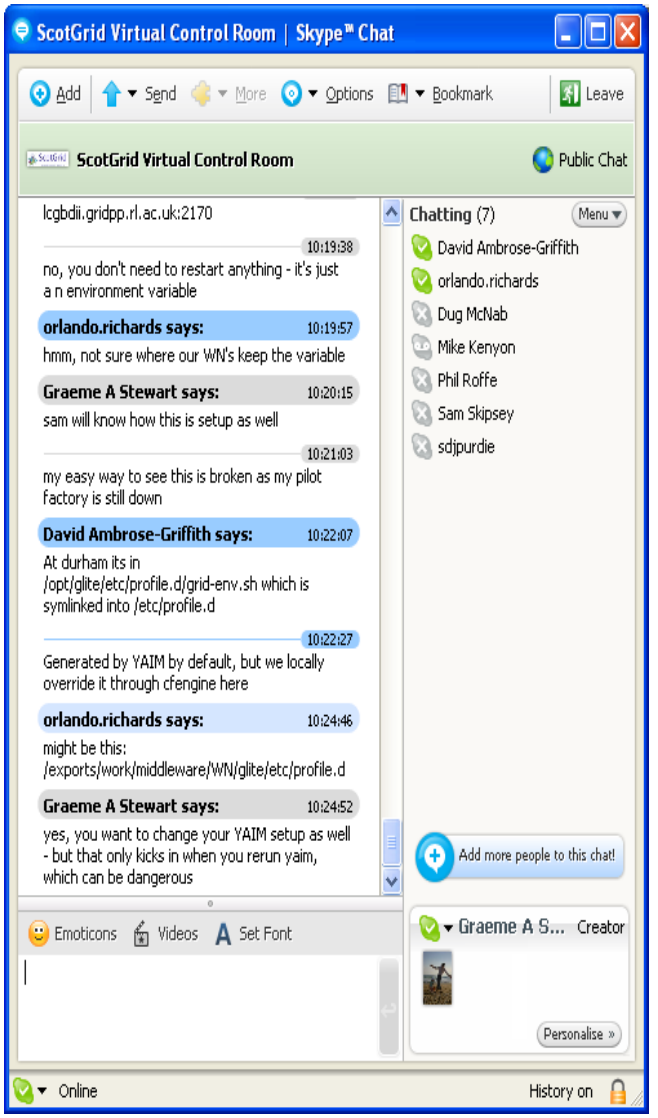


GridPP
UK Computing for Particle Physics

Management and infrastructure

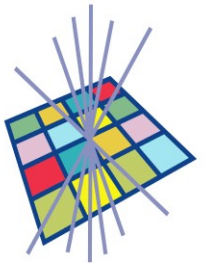
ScotGrid
Scottish Grid Service

- Cfengine
- Ganglia
- Centralised nagios alerting.
- Shared login across Edinburgh, Glasgow, Durham
- "Virtual control room"
  - Extremely useful for instant discussion and problem resolution



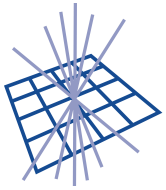
GridPP
UK Computing for Particle Physics

“Local central” Services



ScotGrid
Scottish Grid Service

- For our own reliability, and to balance load, ScotGrid runs its own:
  - WMS
  - Top-level BDII
  - VOMS server
- May add others, all at Glasgow
- Pro: more control, more choice for other sites
- Con: more management overhead, complexity

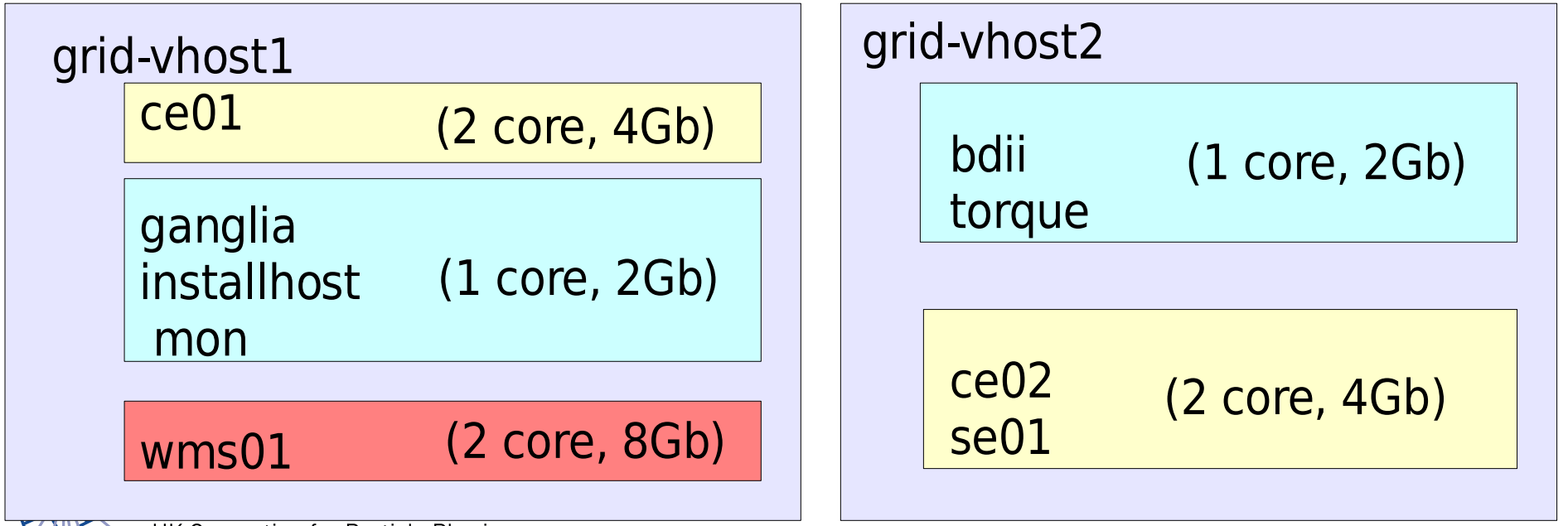


GridPP
UK Computing for Particle Physics

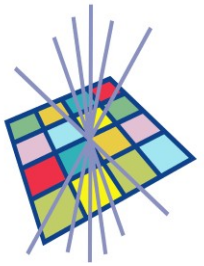
Durham – VMs abound

ScotGrid
Scottish Grid Service

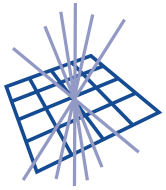
- All front-end service nodes run as VMWare-hosted virtual machines.
  - 2 Physical nodes – 8 cores, 16Gb each (+ UI)
  - Separate NFS server + master node



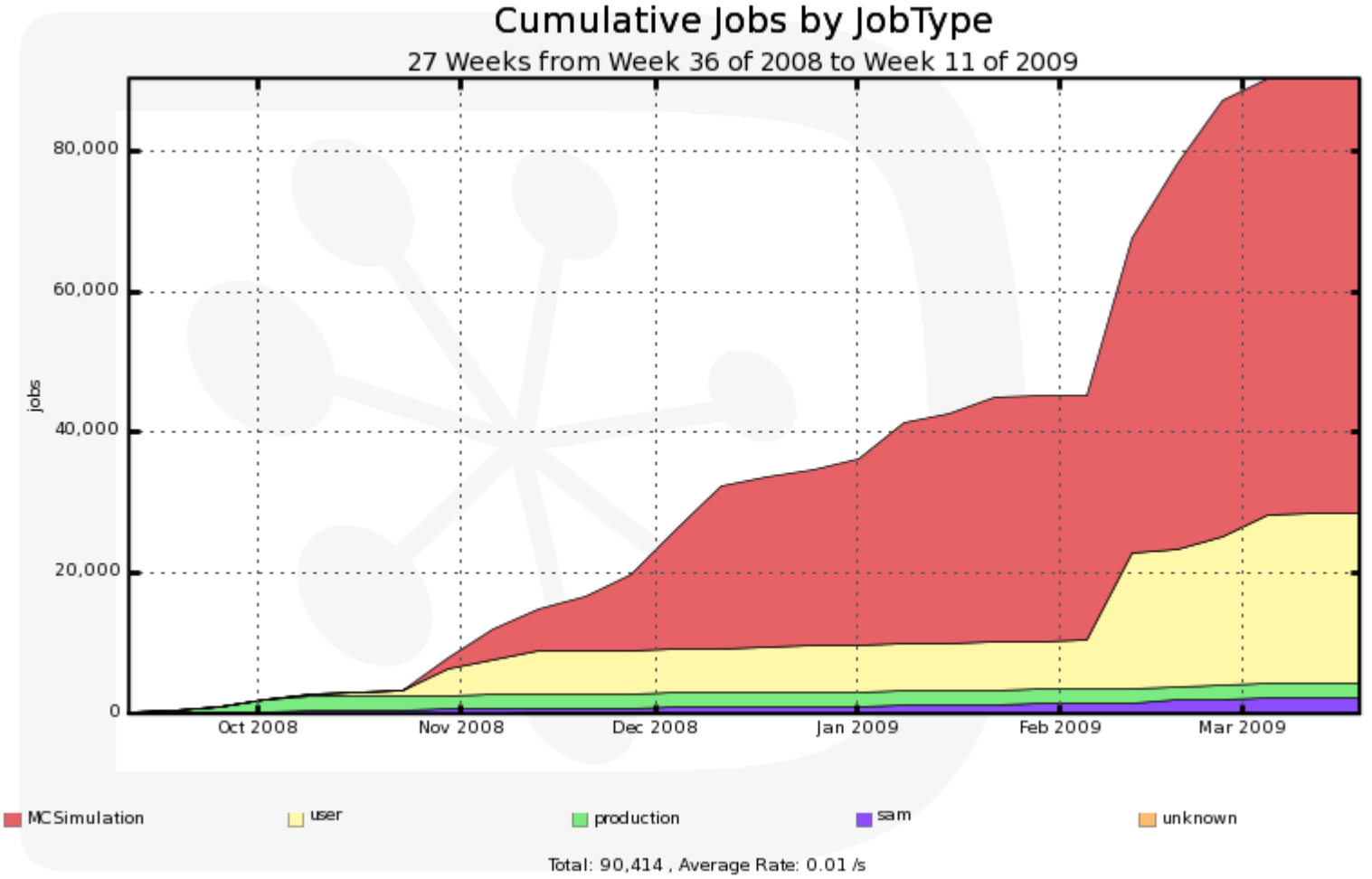
ECDF – special cases



- Central university resource – Grid share must play nicely with local jobs (and is not a majority user).
- VMEM limit special-cases.
- WN-TAR install – everything Grid sandboxed.
- Some handcrafting of “default” SGE jobmanager to support nonstandard queue configuration.



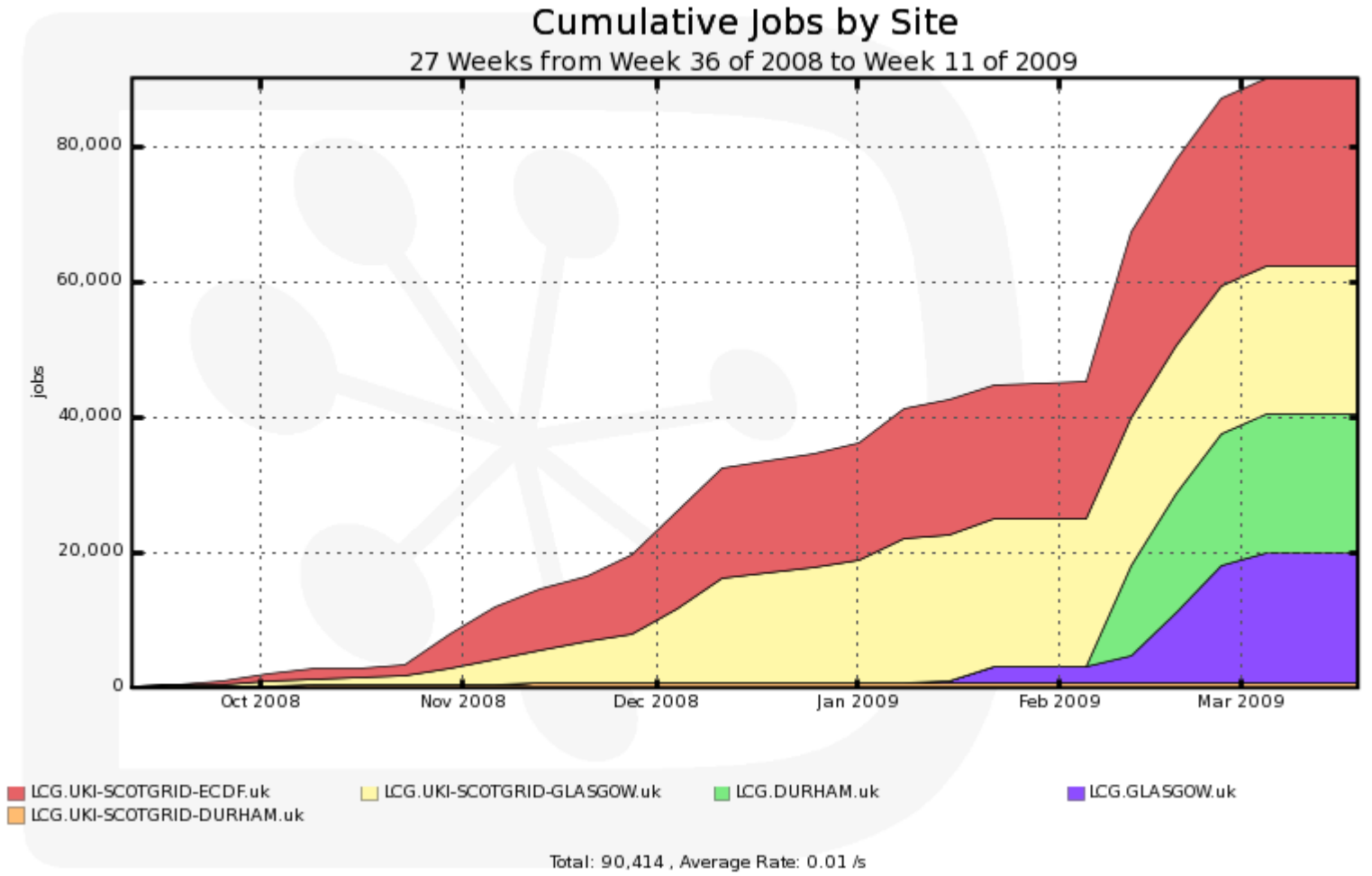
LHCb Jobtype changes



ScotGrid
Scottish Grid Service

GridPP
UK Computing for Particle Physics

LHCb usage across sites



ScotGrid
Scottish Grid Service

GridPP
UK Computing for Particle Physics

ATLAS Production

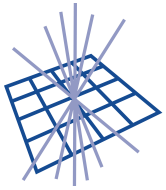
ScotGrid
Scottish Grid Service

- Glasgow is top UK ATLAS Tier-2
  - Storage and fabric easily cope with production load with 2000 running jobs
- Production at Edinburgh is significant
- Durham should improve
  - Sometimes no jobs is hard to diagnose

site	success	failure	success (walltime)	failure (walltime)	efficiency	efficiency (walltime)
RAL-LCG2	245787	74734	2830137434	820364357	76.7%	77.5%
UKI-SCOTGRID-GLASGOW	87376	16461	1859086649	70004125	84.1%	96.4%
UKI-LT2-QMUL	36035	4306	1034180037	55335363	89.3%	94.9%
UKI-LT2-RHUL	23813	1049	518662190	18846141	95.8%	96.5%
UKI-NORTHGRID-LANCS-HEP	13598	5548	348812689	47601881	71%	88%
UKI-NORTHGRID-MAN-HEP	17638	1471	740648861	15169211	92.3%	98%
UKI-NORTHGRID-LIV-HEP	13570	1835	425224890	21649453	88.1%	95.2%
UKI-NORTHGRID-SHEF-HEP	13081	203	266302242	3149928	98.5%	98.8%
UKI-SCOTGRID-ECDF	9103	1421	135756242	7462699	86.5%	94.8%
UKI-SOUTHGRID-OX-HEP	7140	2638	191695448	26819988	73%	87.7%
UKI-SOUTHGRID-CAM-HEP	8483	514	155769938	5044015	94.3%	96.9%
UKI-LT2-IC-HEP	6045	279	134262189	3776010	95.6%	97.3%
UKI-SOUTHGRID-RALPP	4260	155	113936553	7742406	96.5%	93.6%
UKI-LT2-Brunel	1798	264	56890157	1965882	87.2%	96.7%
UKI-SCOTGRID-DURHAM	1427	113	504188	21155	92.7%	96%
UKI-SOUTHGRID-BHAM-HEP	735	23	452986	172312	97%	72.4%
UKI-LT2-UCL-HEP	0	205	0	0	0%	-
total	*489889*	*111219*	*8.812322693e+09*	*1105124926*	*81.5%*	*88.9%*

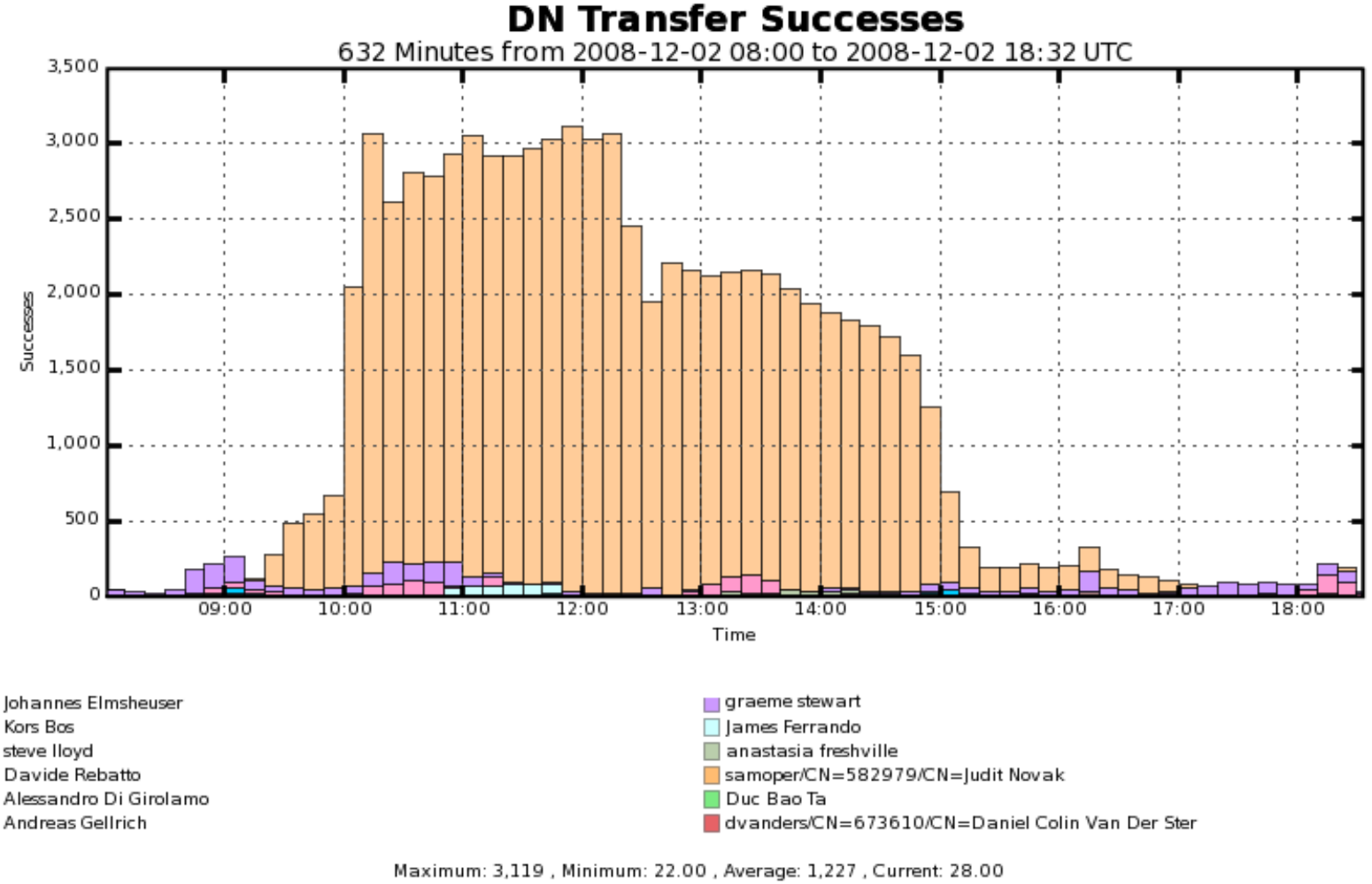
User Analysis

- Even in 2008 we had not seen significant user analysis activity on the Tier-2
- We were 'early adopters' of the ATLAS Hammercloud test framework to prepare the clusters for the challenges of user analysis
- This framework sends a large number of realistic muon analysis jobs to clusters through ganga
  - Key point is the access of very large numbers of small AOD files with non-sequential access



GridPP
UK Computing for Particle Physics

HC038 – before optimisation

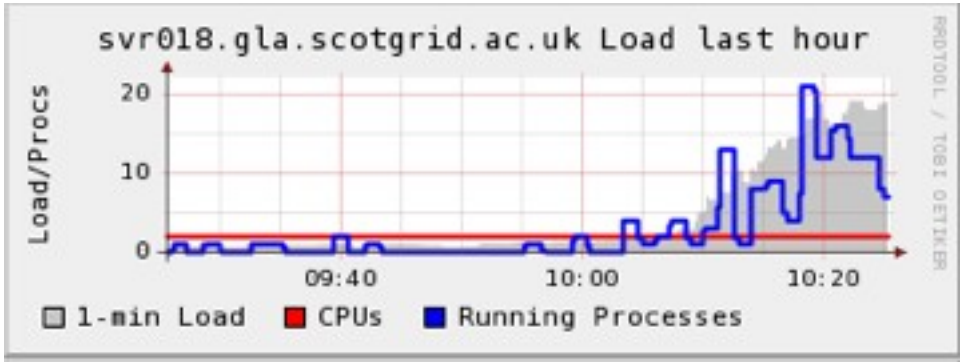


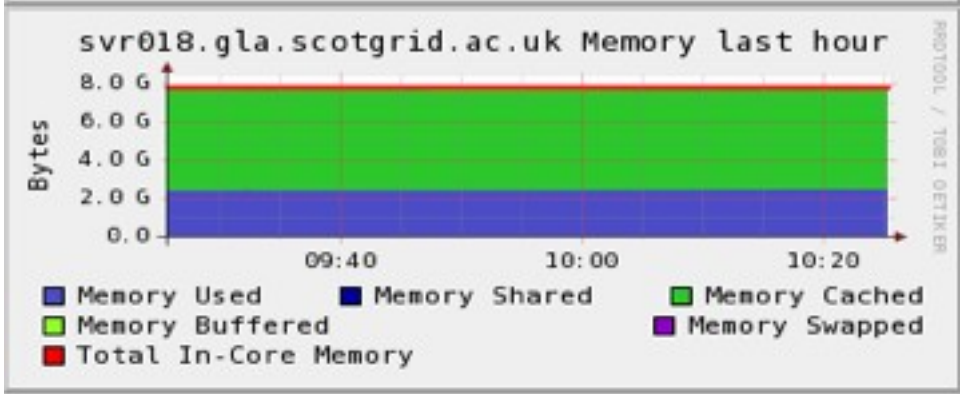
ScotGrid
Scottish Grid Service

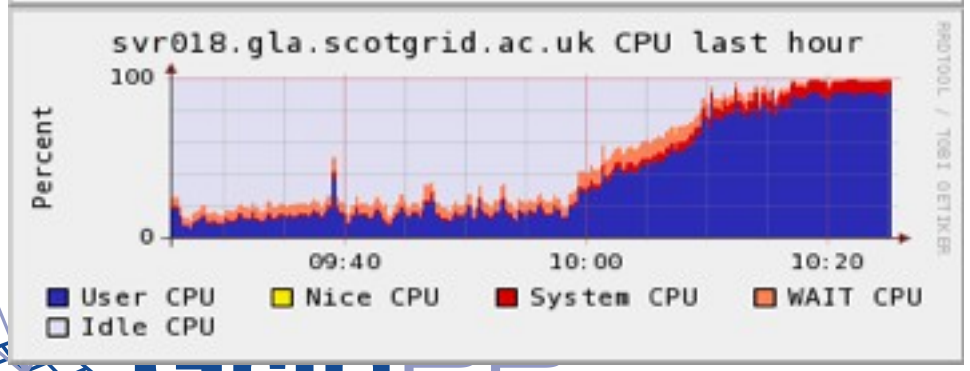
GridPP
UK Computing for Particle Physics

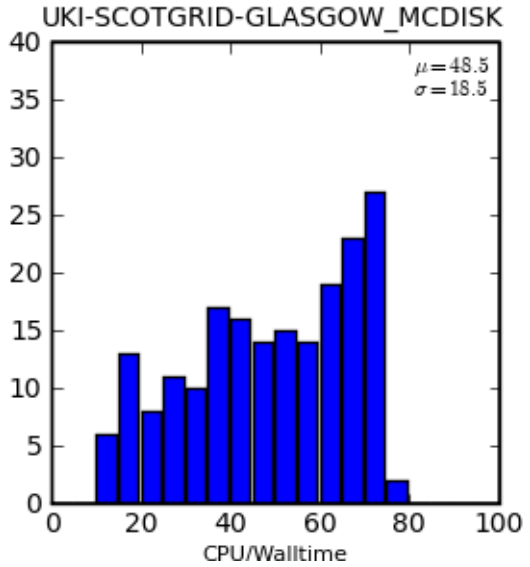
HC038 – load on services

ScotGrid
Scottish Grid Service



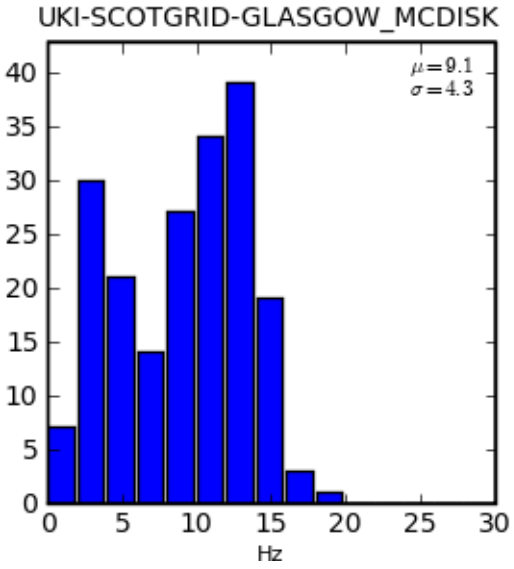






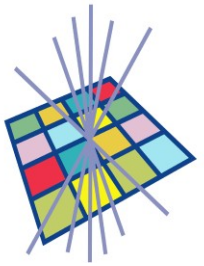
Terrible event rate -

DPM head just can't cope

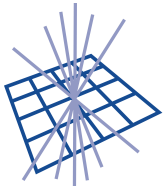


GridPP
UK Computing for Particle Physics

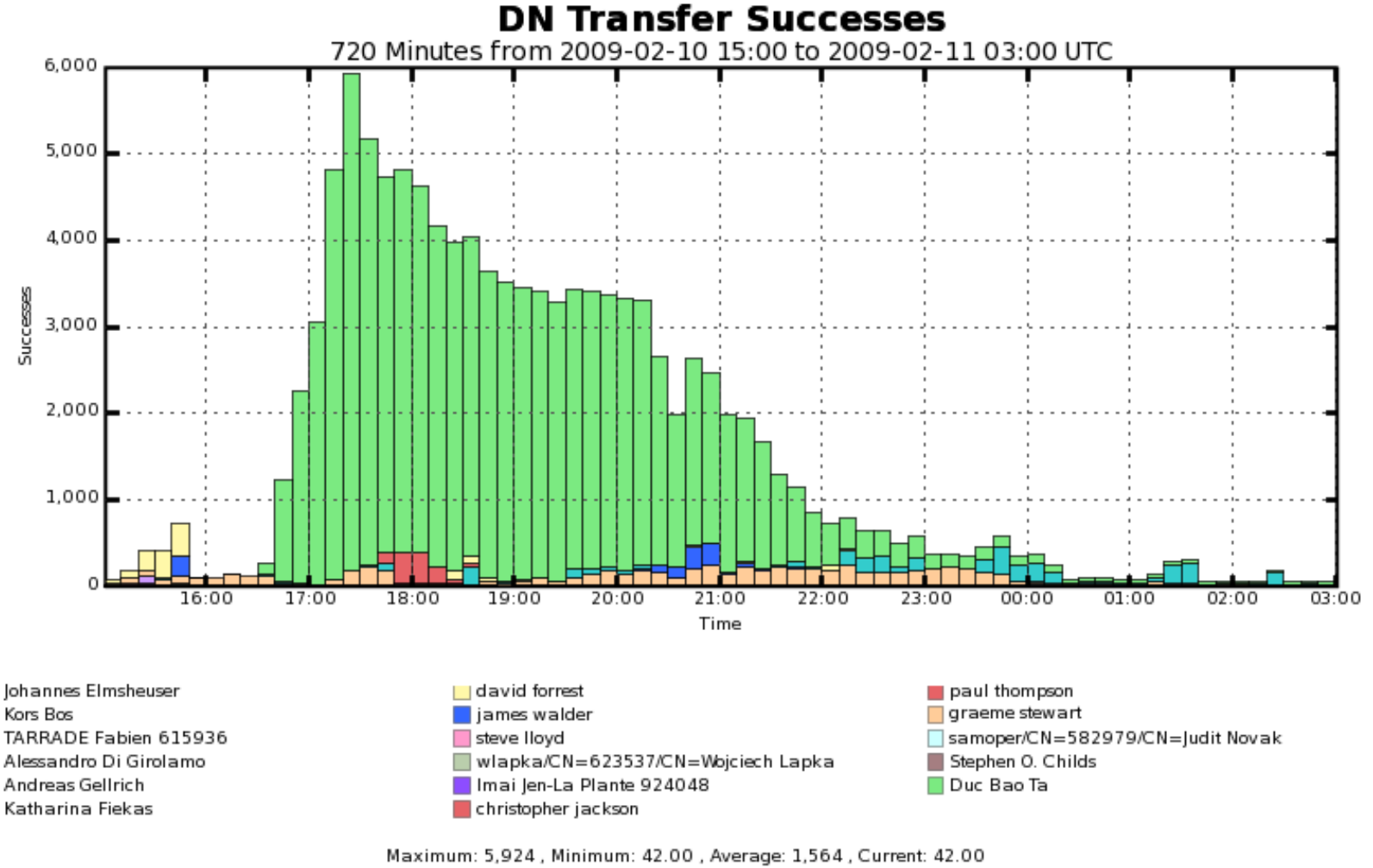
Splitting services



- DPM is clearly CPU bound
  - Load on dpm daemon was > 100% (no iowait)
- DPM services are CPU bound
  - Split these onto new node (high CPU, poor disk)
- MySQL back-end is probably IO-bound
  - Leave on “old” node, which has fast disk



HC135 – after split

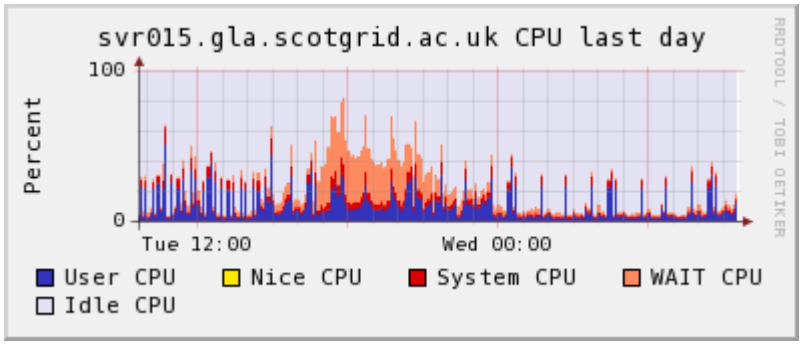


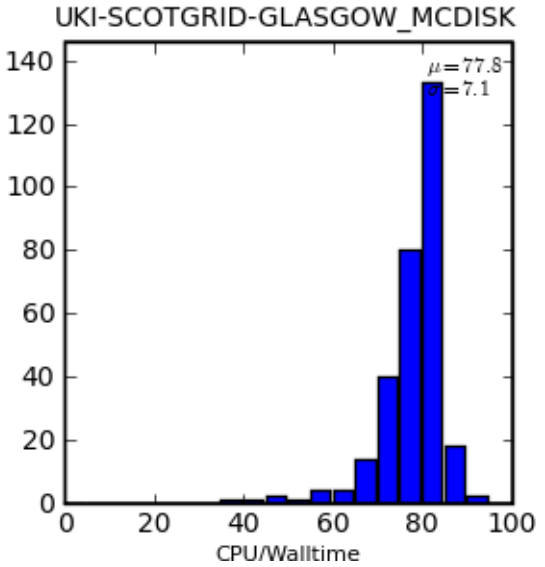
ScotGrid
Scottish Grid Service

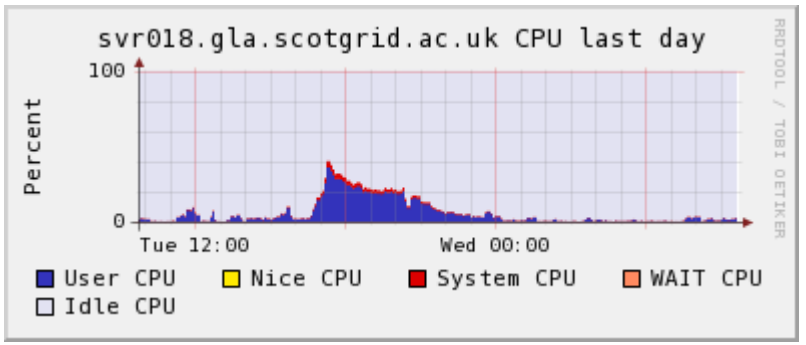
GridPP
UK Computing for Particle Physics

HC135 – load and stats

ScotGrid
Scottish Grid Service

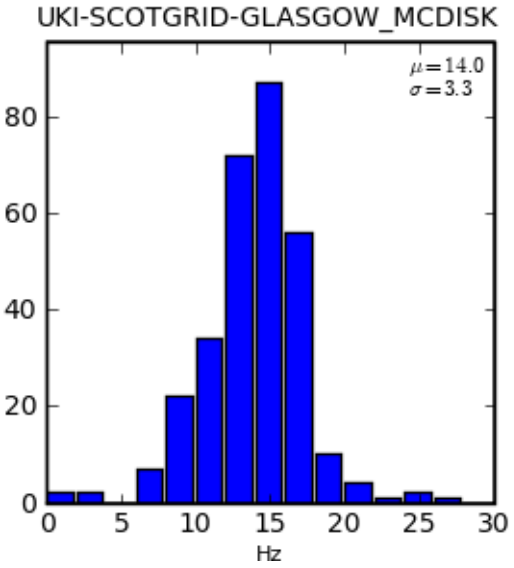






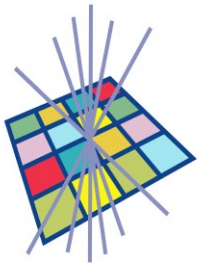
Significantly better rate + efficiency.

Limiting IOwait on MySQL node.



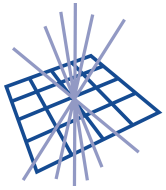
GridPP
UK Computing for Particle Physics

MySQL indexing



ScotGrid
Scottish Grid Service

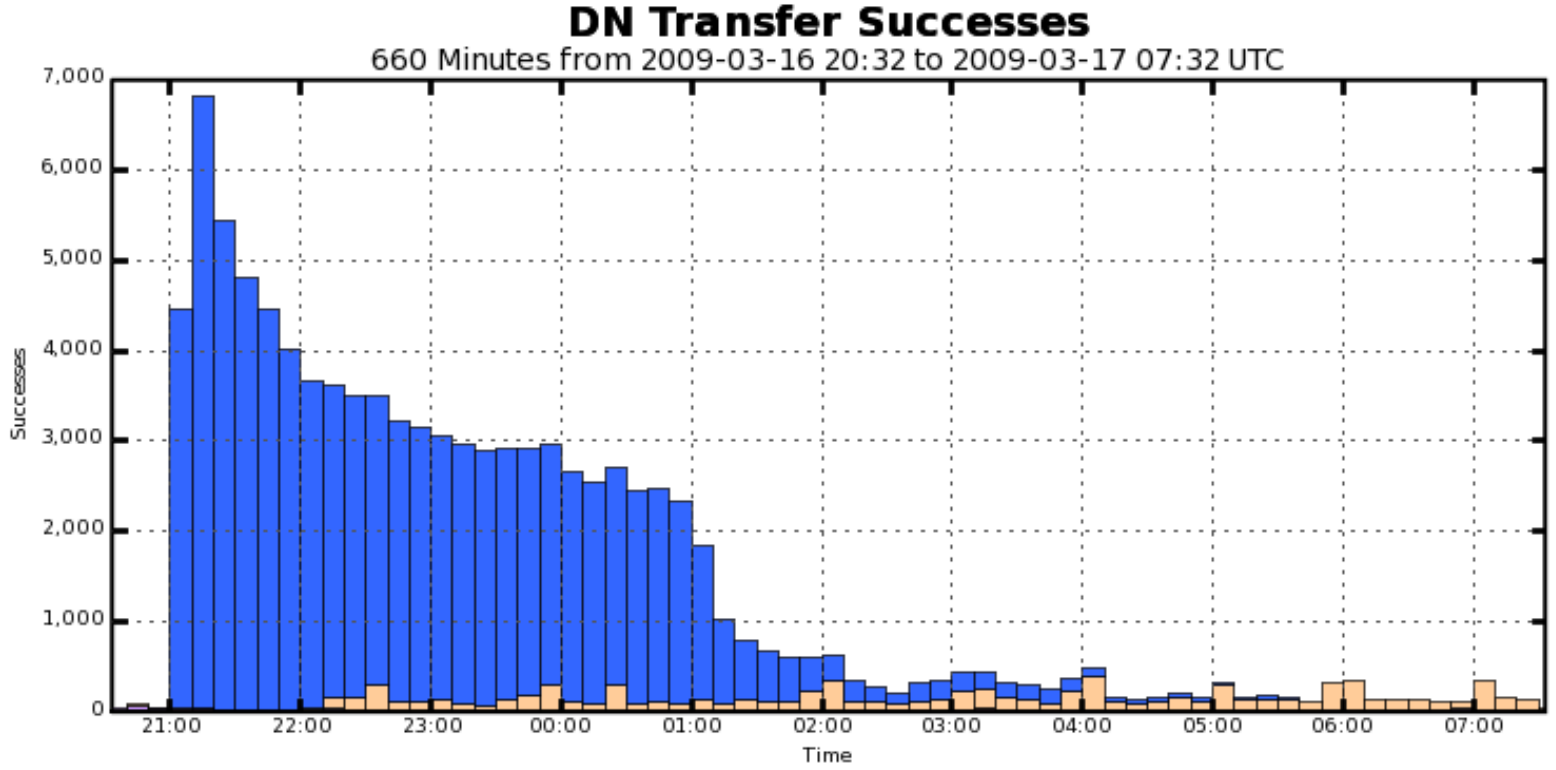
- Enable slow-queries log.
- Three common slow queries on unindexed columns in dpm_db tables.
- Add indices:
  - create index pfn_lifetime on dpm_get_filereq (pfn(255), lifetime);
  - create index status_idx on dpm_put_filereq(status);
  - create index stime_idx on dpm_req(stime);



GridPP
UK Computing for Particle Physics

HC193 – after optimisation

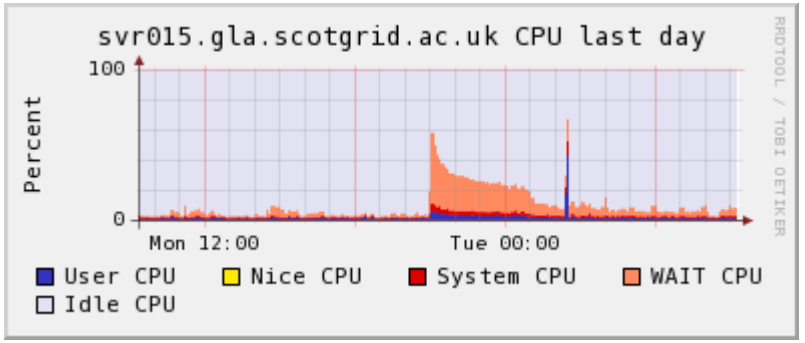
ScotGrid
Scottish Grid Service

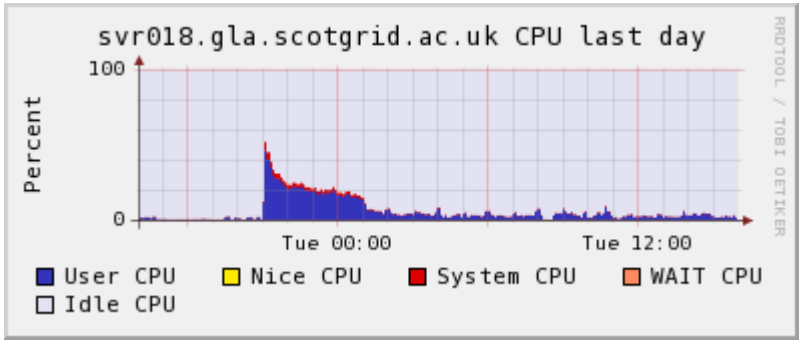


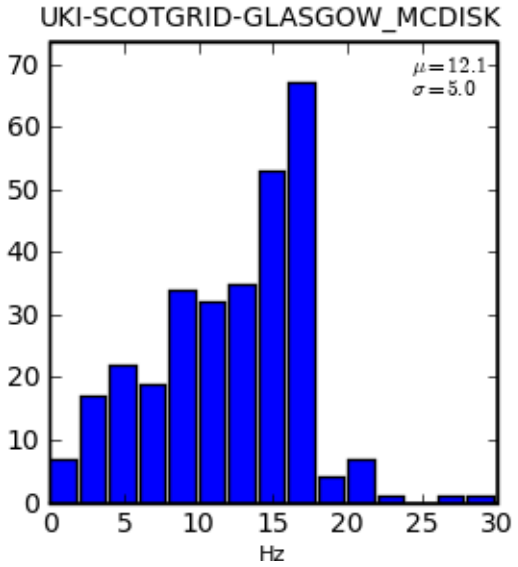
GridPP
UK Computing for Particle Physics

HC193 – load and stats

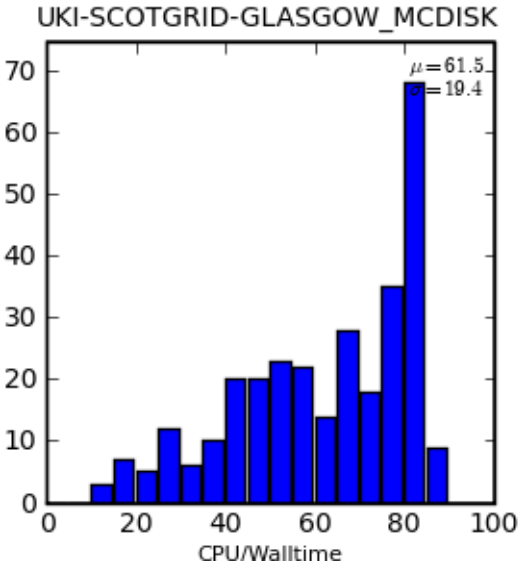
ScotGrid
Scottish Grid Service





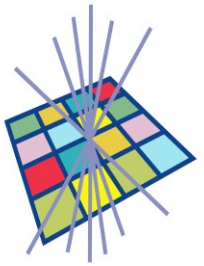


Efficiency per job lower – but we are actually processing more concurrent jobs here (cluster was empty before test)

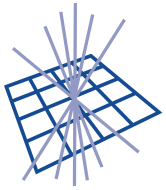


GridPP
UK Computing for Particle Physics

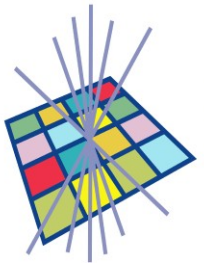
Next steps



- non-rfio access
  - dpm-xrootd, “pure” xrootd
  - (Eliminates some of the authentication overhead of rfio + gsiftp)
- Network infrastructure improvements
  - disk servers are each Gigabit each
  - channel bonding? 10Gigabit? Infiniband?

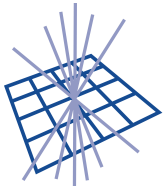


Conclusions



ScotGrid
Scottish Grid Service

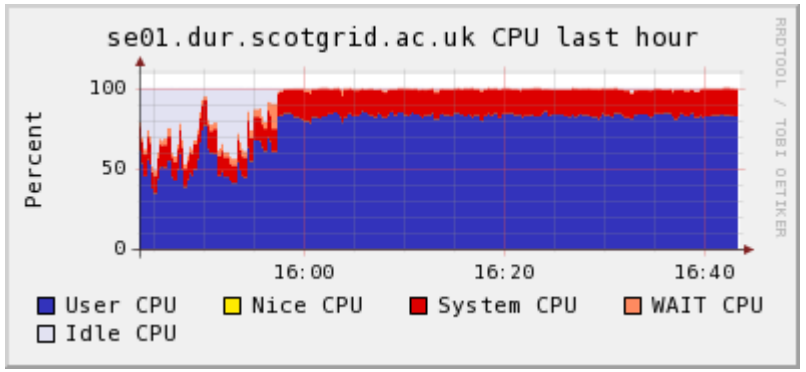
- Communication is essential!
- Be prepared to be flexible.
- Local copies of “central” services
  - Split load
  - But add overhead.

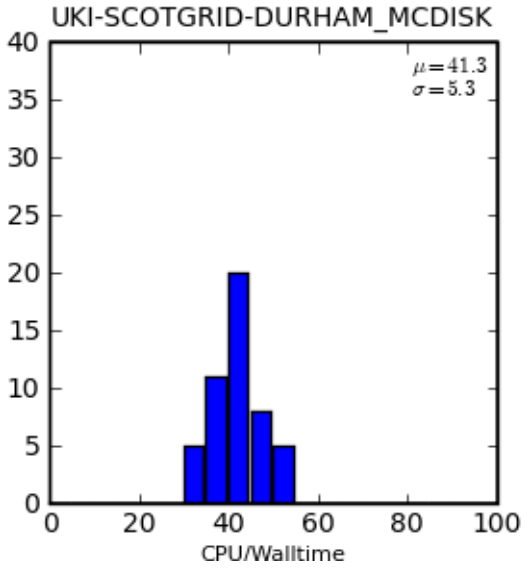


GridPP
UK Computing for Particle Physics

Durham vs HammerCloud

ScotGrid
Scottish Grid Service



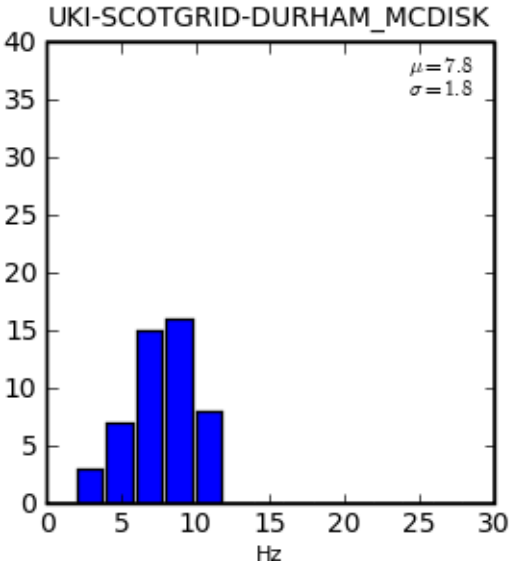


This is where we came in...

DPM is clearly CPU bound.

Virtual Machine, so assign more cores

Possibly even dynamically?



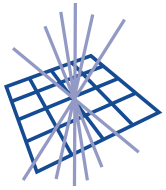
GridPP
UK Computing for Particle Physics

Glasgow Tier 2.5

Interface local (PPE) resources with Grid.

Local-level access for data, job preparation

ScotGrid's DPM ATLAS datasets available via RFIO, XROOT

gLite 3.1 UI available on all Linux desktops?

Removes need for local users to map to generic pool username (e.g. gla048 -> sskipsey).