Multi User Pilot Jobs update

GDB 2010-03-24

Maarten Litmaath

CERN

Multi-user pilot jobs working group

- 71 members, others welcome
  - http://groups.google.ch/group/wlcg-tf-pilot-jobs
- Intermediate summary Wiki
  - https://wlcg-tf.hep.ac.uk/wiki/Multi_User_Pilot_Jobs
    - To view that page the browser needs an IGTF cert loaded
- Questionnaire sent to T0/1/2 representatives on Jan 11-13
  - Vast majority of responses received by March 1
    - Some regions missing or incomplete
      - Triumf, ASGC
  - Further activity in the WG → recommendations

Summary Wiki (1)

- 1. Introduction
  - What are pilot jobs?
  - Single- vs. multi-user pilots
  - What is glexec?

- 2. Boundary conditions
  - Mainly JSPG policies
    - http://www.jspg.org/wiki/JSPG_Docs
    - Some adjustments could turn out to be desirable

- 3. Benefits of pilot jobs compared to "classic" jobs
  - Also single- vs. multi-user pilots

Summary Wiki (2)

- 4. Issues for efficient/correct scheduling of pilot jobs
  - A single class of pilot jobs may not be a panacea
- 5. Drawbacks of multi-user pilots
  - Mainly issues surrounding glexec
- 6. Multi-user pilot jobs with identity change
  - Pro: complete separation of users
  - Con: setuid complications
- 7. Multi-user pilot jobs without identity change
  - Pro: no setuid complications
  - Con: incomplete separation of users

Summary Wiki (3)

- 8. Legal considerations
  - Some sites may have more constraints than others

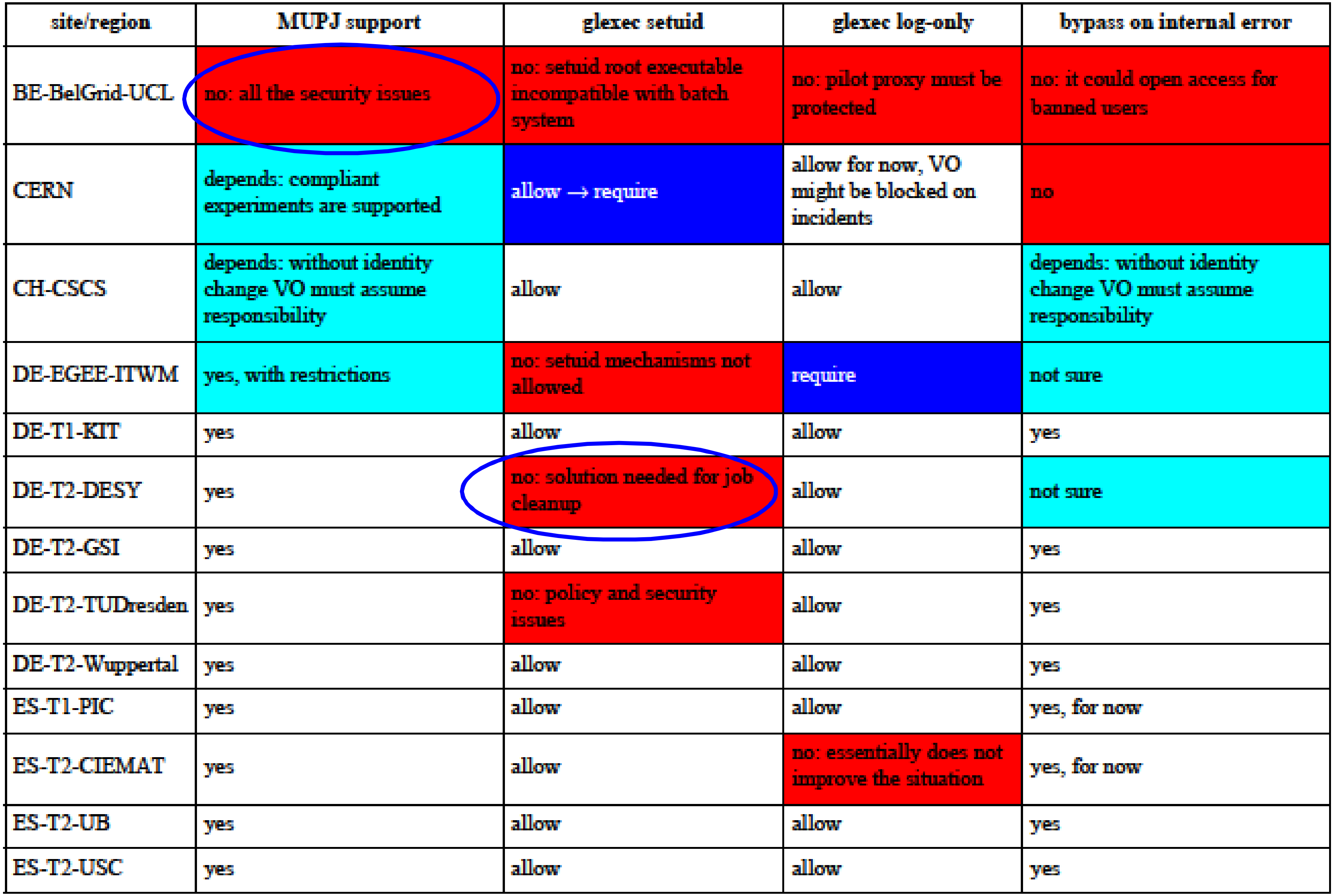
- 9. Virtual machines
  - Will simplify matters

Questionnaire

1. Does your site policy allow the use of multi-user pilot jobs by the LHC experiments you support? (no/depends/yes)
   - If no, why?
2. Does your site policy support the use of glexec in setuid mode? (no/allow/require)
   - If no, why?
3. Does your site policy support the use of glexec in log-only mode? (no/allow/require)
   - If no, why?
4. When glexec returns an internal error (e.g. SCAS/Argus/GUMS temporarily unavailable), does your site policy allow the pilot to continue and run the payload itself? (no/depends/yes)
   - If depends, on what?

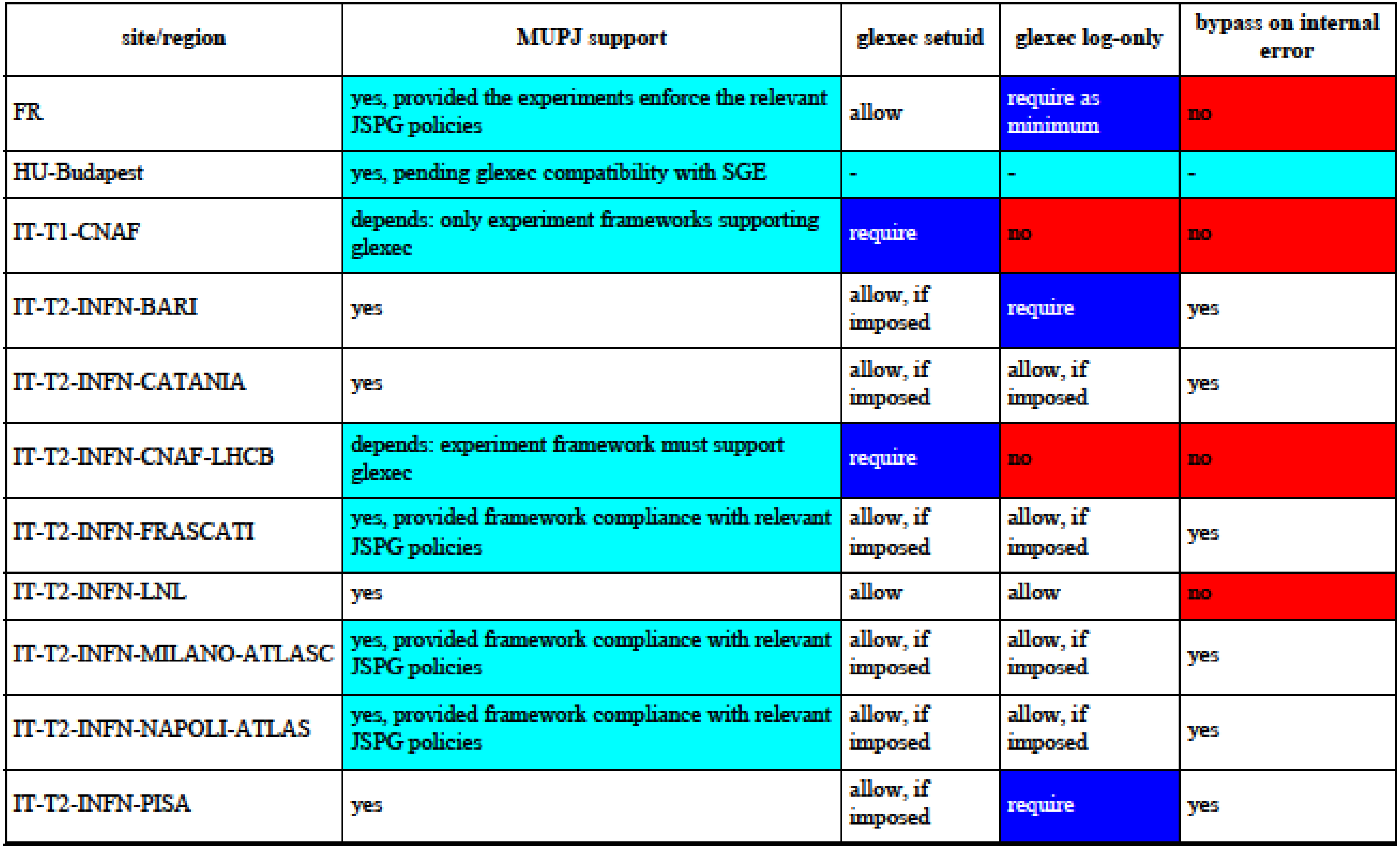
- Results: http://litmaath.web.cern.ch/litmaath/MUPJ-quest.html

Questionnaire responses (1/6)



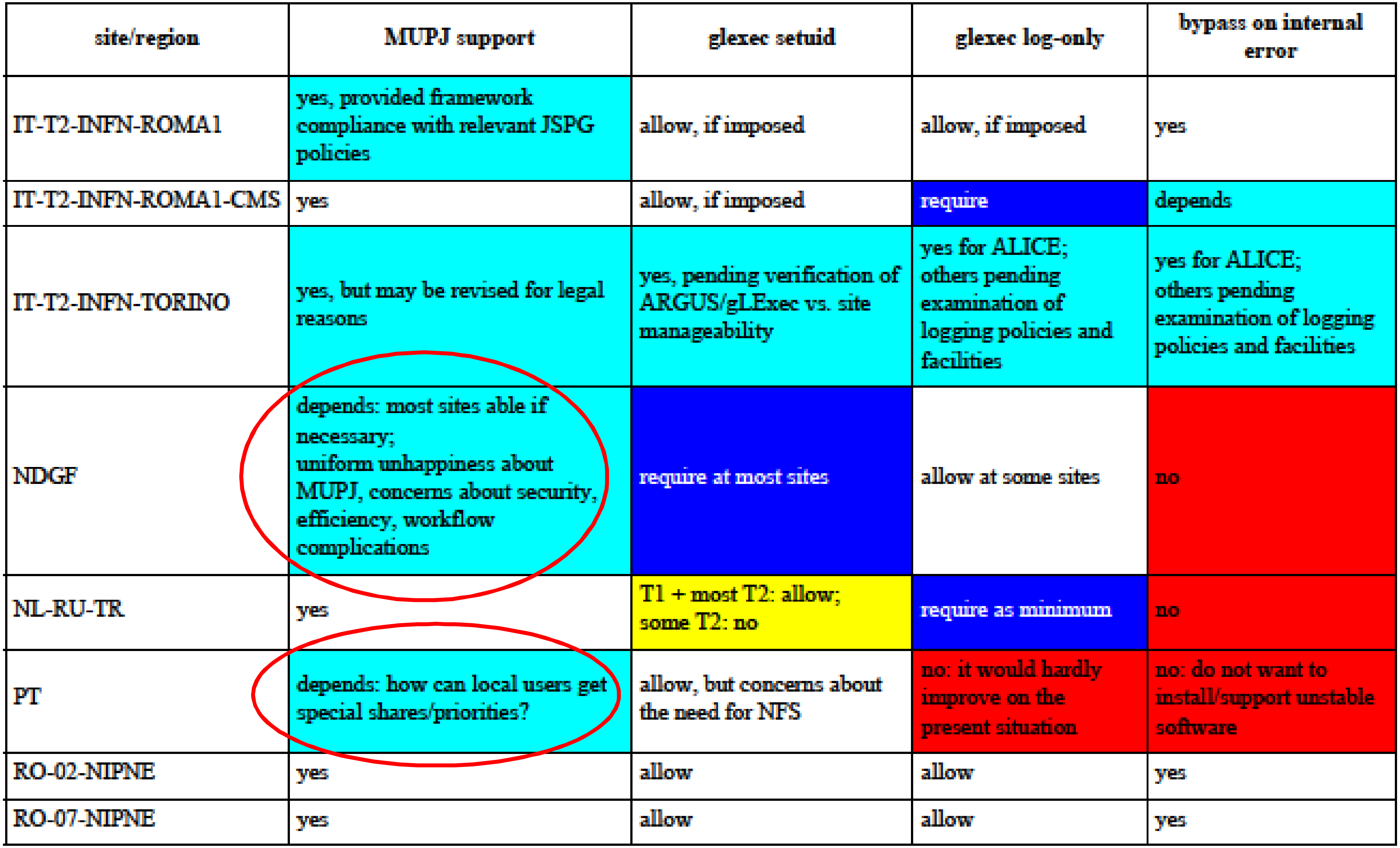
site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
BE-BelGrid-UCL	no: all the security issues	no: setuid root executable incompatible with batch system	no: pilot proxy must be protected	no: it could open access for banned users
CERN	depends: compliant experiments are supported	allow → require	allow for now, VO might be blocked on incidents	no
CH-CSCS	depends: without identity change VO must assume responsibility	allow	allow	depends: without identity change VO must assume responsibility
DE-EGEE-ITWM	yes, with restrictions	no: setuid mechanisms not allowed	require	not sure
DE-T1-KIT	yes	allow	allow	yes
DE-T2-DESY	yes	no: solution needed for job cleanup	allow	not sure
DE-T2-GSI	yes	allow	allow	yes
DE-T2-TUDresden	yes	no: policy and security issues	allow	yes
DE-T2-Wuppertal	yes	allow	allow	yes
ES-T1-PIC	yes	allow	allow	yes, for now
ES-T2-CIEMAT	yes	allow	no: essentially does not improve the situation	yes, for now
ES-T2-UB	yes	allow	allow	yes
ES-T2-USC	yes	allow	allow	yes

Questionnaire responses (2/6)



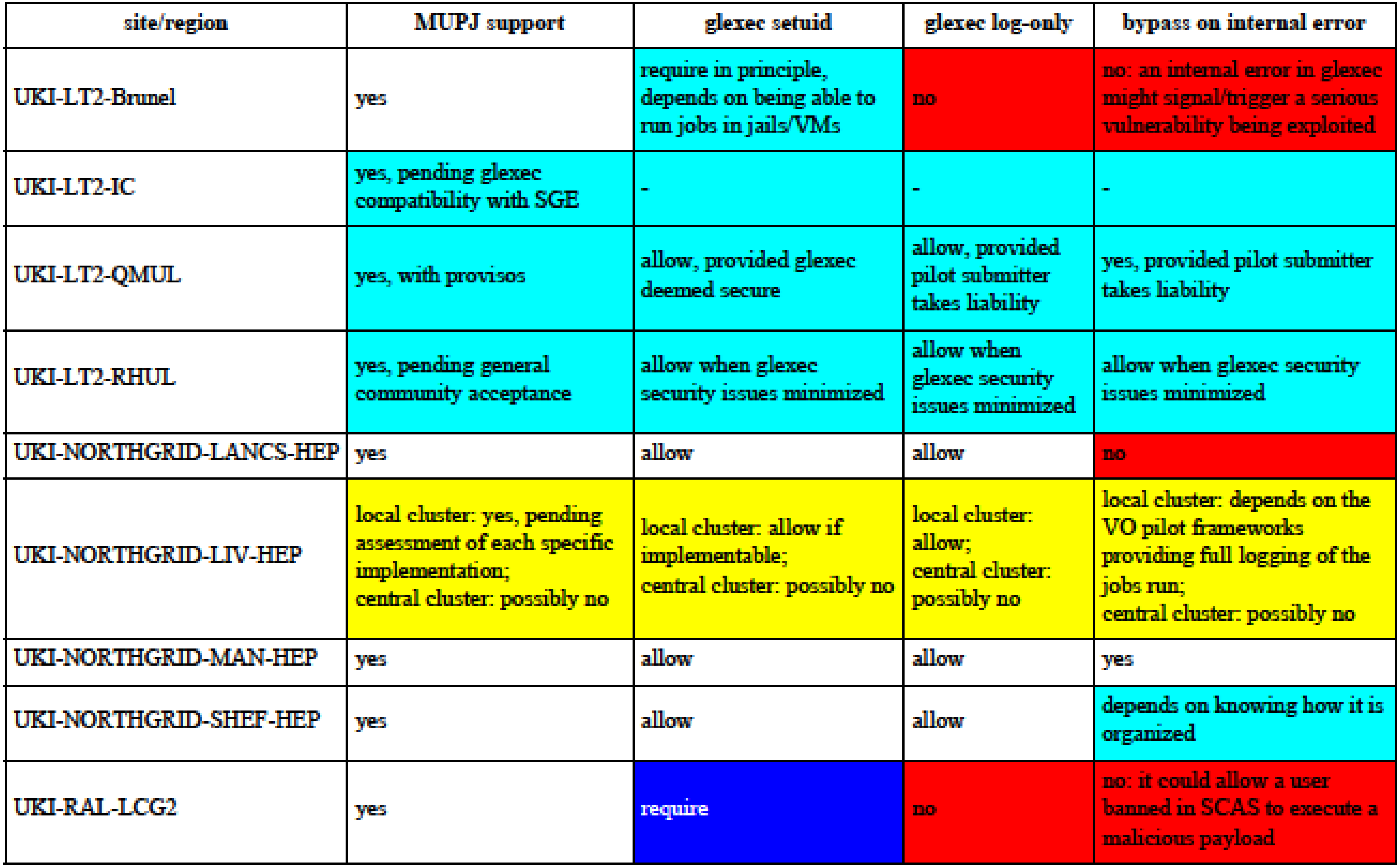
site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
FR	yes, provided the experiments enforce the relevant JSPG policies	allow	require as minimum	no
HU-Budapest	yes, pending glexec compatibility with SGE	-	-	-
IT-T1-CNAF	depends: only experiment frameworks supporting glexec	require	no	no
IT-T2-INFN-BARI	yes	allow, if imposed	require	yes
IT-T2-INFN-CATANIA	yes	allow, if imposed	allow, if imposed	yes
IT-T2-INFN-CNAF-LHCB	depends: experiment framework must support glexec	require	no	no
IT-T2-INFN-FRASCATI	yes, provided framework compliance with relevant JSPG policies	allow, if imposed	allow, if imposed	yes
IT-T2-INFN-LNL	yes	allow	allow	no
IT-T2-INFN-MILANO-ATLASC	yes, provided framework compliance with relevant JSPG policies	allow, if imposed	allow, if imposed	yes
IT-T2-INFN-NAPOLI-ATLAS	yes, provided framework compliance with relevant JSPG policies	allow, if imposed	allow, if imposed	yes
IT-T2-INFN-PISA	yes	allow, if imposed	require	yes

Questionnaire responses (3/6)



site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
IT-T2-INFN-ROMA1	yes, provided framework compliance with relevant JSPG policies	allow, if imposed	allow, if imposed	yes
IT-T2-INFN-ROMA1-CMS	yes	allow, if imposed	require	depends
IT-T2-INFN-TORINO	yes, but may be revised for legal reasons	yes, pending verification of ARGUS/gLExec vs. site manageability	yes for ALICE; others pending examination of logging policies and facilities	yes for ALICE; others pending examination of logging policies and facilities
NDGF	depends: most sites able if necessary; uniform unhappiness about MUPJ, concerns about security, efficiency, workflow complications	require at most sites	allow at some sites	no
NL-RU-TR	yes	T1 + most T2: allow; some T2: no	require as minimum	no
PT	depends: how can local users get special shares/priorities?	allow, but concerns about the need for NFS	no: it would hardly improve on the present situation	no: do not want to install/support unstable software
RO-02-NIPNE	yes	allow	allow	yes
RO-07-NIPNE	yes	allow	allow	yes

Questionnaire responses (4/6)

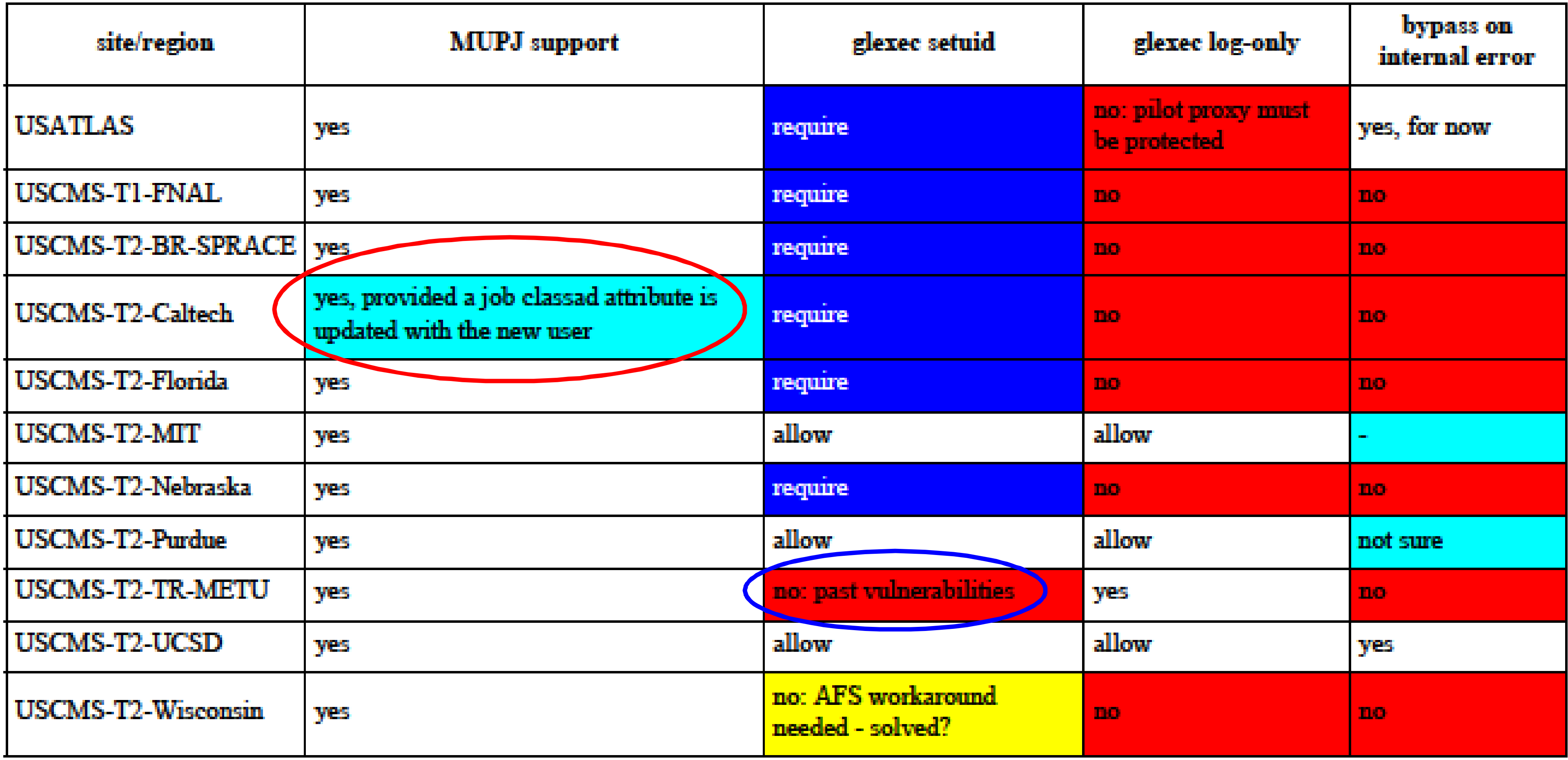


site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
UKI-LT2-Brunel	yes	require in principle, depends on being able to run jobs in jails/VMs	no	no: an internal error in glexec might signal/trigger a serious vulnerability being exploited
UKI-LT2-IC	yes, pending glexec compatibility with SGE	-	-	-
UKI-LT2-QMUL	yes, with provisos	allow, provided glexec deemed secure	allow, provided pilot submitter takes liability	yes, provided pilot submitter takes liability
UKI-LT2-RHUL	yes, pending general community acceptance	allow when glexec security issues minimized	allow when glexec security issues minimized	allow when glexec security issues minimized
UKI-NORTHGRID-LANCS-HEP	yes	allow	allow	no
UKI-NORTHGRID-LIV-HEP	local cluster: yes, pending assessment of each specific implementation; central cluster: possibly no	local cluster: allow if implementable; central cluster: possibly no	local cluster: allow; central cluster: possibly no	local cluster: depends on the VO pilot frameworks providing full logging of the jobs run; central cluster: possibly no
UKI-NORTHGRID-MAN-HEP	yes	allow	allow	yes
UKI-NORTHGRID-SHEF-HEP	yes	allow	allow	depends on knowing how it is organized
UKI-RAL-LCG2	yes	require	no	no: it could allow a user banned in SCAS to execute a malicious payload

Questionnaire responses (5/6)

site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
UKI-SCOTGRID-Durham	yes	allow	allow	depends on further investigation
UKI-SCOTGRID-ECDF	yes	allow	allow	yes
UKI-SCOTGRID-GLASGOW	yes	allow	allow	yes, provided the VO takes responsibility for jobs
UKI-SOUTHGRID-BHAM-HEP	yes	local cluster: allow; shared cluster: possibly no	local cluster: allow; shared cluster: possibly no	possibly, follow GridPP consensus
UKI-SOUTHGRID-BRIS-HEP	local cluster: yes; shared cluster: possibly no	local cluster: allow; shared cluster: possibly no	local cluster: allow; shared cluster: possibly no	local cluster: probably, for now; shared cluster: possibly no
UKI-SOUTHGRID-CAM-HEP	yes	allow	allow	depends on better understanding
UKI-SOUTHGRID-OX-HEP	yes	allow	allow	probably, for now
UKI-SOUTHGRID-RALPP	yes	require	no	probably no

Questionnaire responses (6/6)



site/region	MUPJ support	glexec setuid	glexec log-only	bypass on internal error
USATLAS	yes	require	no: pilot proxy must be protected	yes, for now
USCMS-T1-FNAL	yes	require	no	no
USCMS-T2-BR-SPRACE	yes	require	no	no
USCMS-T2-Caltech	yes, provided a job classad attribute is updated with the new user	require	no	no
USCMS-T2-Florida	yes	require	no	no
USCMS-T2-MIT	yes	allow	allow	-
USCMS-T2-Nebraska	yes	require	no	no
USCMS-T2-Purdue	yes	allow	allow	not sure
USCMS-T2-TR-METU	yes	no: past vulnerabilities	yes	no
USCMS-T2-UCSD	yes	allow	allow	yes
USCMS-T2-Wisconsin	yes	no: AFS workaround needed - solved?	no	no

Observations (1/3)

- MUPJ in principle supported at almost all sites
  - With various provisos

- Most of the sites/resources compatible with glexec setuid
  - Many sites require it in the medium term
  - Strongly preferred by CMS and ATLAS representatives

- Various sites do not like glexec setuid
  - Some cannot or will not install it
    - Batch system managed by a different group
    - Security concerns, even when:
      - Glexec code has been reviewed by EGEE and OSG
      - Glexec need only be executable for privileged roles

Observations (2/3)

- Many sites allow glexec log-only
  - Some require it as the minimum
- Many other sites do not allow glexec log-only
  - It does not separate users and does not protect the pilot proxy
- Many sites prohibit bypassing glexec on internal errors
  - Others would tolerate it at least for the time being

Observations (3/3)

- DESY not (yet?) compatible with glexec setuid:
  - Complications for cleanup of batch jobs
    - Auxiliary scripts have been provided in the WG
- NDGF have strong concerns about MUPJ concept
  - Security, efficiency, workload complications
- Portugal: how can local users get special shares/priorities?
  - ATLAS: that case is foreseen in PanDA
  - CMS?
- Caltech want Condor job class-ad to be updated with new user
  - Unknown if that is feasible
- Wisconsin deemed glexec setuid incompatible with AFS
  - Discussion in WG suggested this was solved

Proposed course of action (1/2)

- Sites should configure glexec with SCAS/Argus/GUMS backend and advertise it with a CE capability:
  - GlueCECapability: glexec
    - Capability was favored over a run-time environment tag
    - No details about the mode of operation or supported VOs

- Experiments should try out their glexec workflows at such sites as they become available
  - First as a background activity, like SAM tests

- When an experiment finds a particular site reliable with glexec, ideally all its MUPJ for that site would start using it there

Proposed course of action (2/2)

- Some sites might start with log-only and switch to setuid later
  - CMS and ATLAS representatives have argued that the log-only mode should be disallowed right away
- Some of the experiments would like to require the setuid mode in the medium term
  - Sites that do not support it would no longer receive MUPJ
- Other technologies may affect medium-term strategies
  - When each job is launched in a fresh VM, the risks are much reduced and more setups may be acceptable for MUPJ
- The WLCG policies on MUPJ are expected to become stricter as more experience has been gained by sites and experiments