18th August 2009

# Next Generation Events (NGE) - BLACK HOLE ConOp

**Author:** [redacted] (TDB), [redacted] (TSE), [redacted] (TDB)

| Issue | Date | Author | Amendments |
|---|---|---|---|
| 0.1 | 10/06/2009 | [redacted] | First draft |
| 0.2 | 30/06/2009 | [redacted] | Incorporating doc comments |
| 0.3 | 10/07/2009 | [redacted] | Further updates |
| 0.4 | 28/07/2009 | [redacted] | Updates to reflect Stakeholder feedback. |
| 1.0 | 18/08/2009 | [redacted] | Document updated and issued. |

## Distribution

[redacted] (TSE)
[redacted] (TDB)
[redacted] (TDB)
[redacted] (OPD-GTAC)
[redacted] (OPD-DRFC)
[redacted] (ICTR-FSP)
[redacted] (ICTR-MTR)
[redacted] (OPD-GTE)
[redacted] (OPC-TDSD)
[redacted] (GTE-BUDE)
[redacted] (GTE-BUDE)
[redacted] (OPP-LEG

[redacted] (TSE)
[redacted] (TSE)
[redacted] (OPD-GTAC)
[redacted] (OPD-GTAC)
[redacted] (CTOR)
[redacted] (ICTR-HSP)
[redacted] (ICTR-DMR)
[redacted] (OPD-GTE)
[redacted] (OPC-TDSD)
[redacted] (GTE-BUDE)
[redacted] (ITSERVC)
BLAZING SADDLES team

**Location:**
VOB:/[redacted]/NGE_BLACK_HOLE_ConOp.doc

## References

[a] TR BLACK HOLE
[redacted]/BLACK_HOLE

[b] BLAZING SADDLES -Legal and Policy requirements
[redacted]\BLAZING SADDLES -Legal and Policy requirements - 20090625.doc

## Introduction

1. This document aims to encapsulate the initial concept of BLACK HOLE for BLAZING SADDLES under the Next Generation Events (NGE) project.

## Brief overview

2. BLACK HOLE is essentially a flat file store, housing data from a wide range of feeds. A small amount of automated processing is applied to some of the data when it arrives. It was originally developed by Technology Research (TR) and is currently being “tech transferred” by TDB-Events as part of the BLAZING SADDLES work package under the NGE project.

3. It should be understood that BLACK HOLE is not a queryable repository of data e.g. It is not possible to select all data for one of your targets. The data is stored in compressed text files in a series of directories. The data is not databased. Using the HTTP interface, it is possible to return all data of a certain type with basic parameters e.g. by date, but that will return all data of the type.
4. One of the main uses for the TR BLACK HOLE data is as the source of data for populating key operational Internet profiling tools such as MUTANT BROTH, KARMA POLICE and SOCIAL ANIMAL. These applications will also be "Tech Transferred" as Query Focused Datasets (QFD) by TDB-Events as part of the BLAZING SADDLES project.
5. The tech-transferred QFDs will be populated from a Buffer repository, which will hold all QFD data sources for 5 days. In this way the new BLACK HOLE and QFDs will be independent of each other, and the new BLACK HOLE is removed from the processing chain.
6. The following business areas require direct data access to BLACK HOLE data:
    - TR
    - OPD-GTAC
    - OPD-GTE
    - GTE-BUDE
    - OPD-CRFC
    - OPC-TDSD
    - CTOR (part of OPIX-IT)
7. The original TR BLACK HOLE will remain in the TR laboratory for research purposes only i.e. it will no longer provide direct operational use. See Reference [a] for more information about the current TR BLACK HOLE.
8. A general overview of the system being delivered by BLAZING SADDLES can be found in Appendix A.

18th August 2009

## Next Generation Events (NGE) BLACK HOLE

The delivery of the NGE BLACK HOLE will include two instances based at Benhall and Bude. See Figure 1.

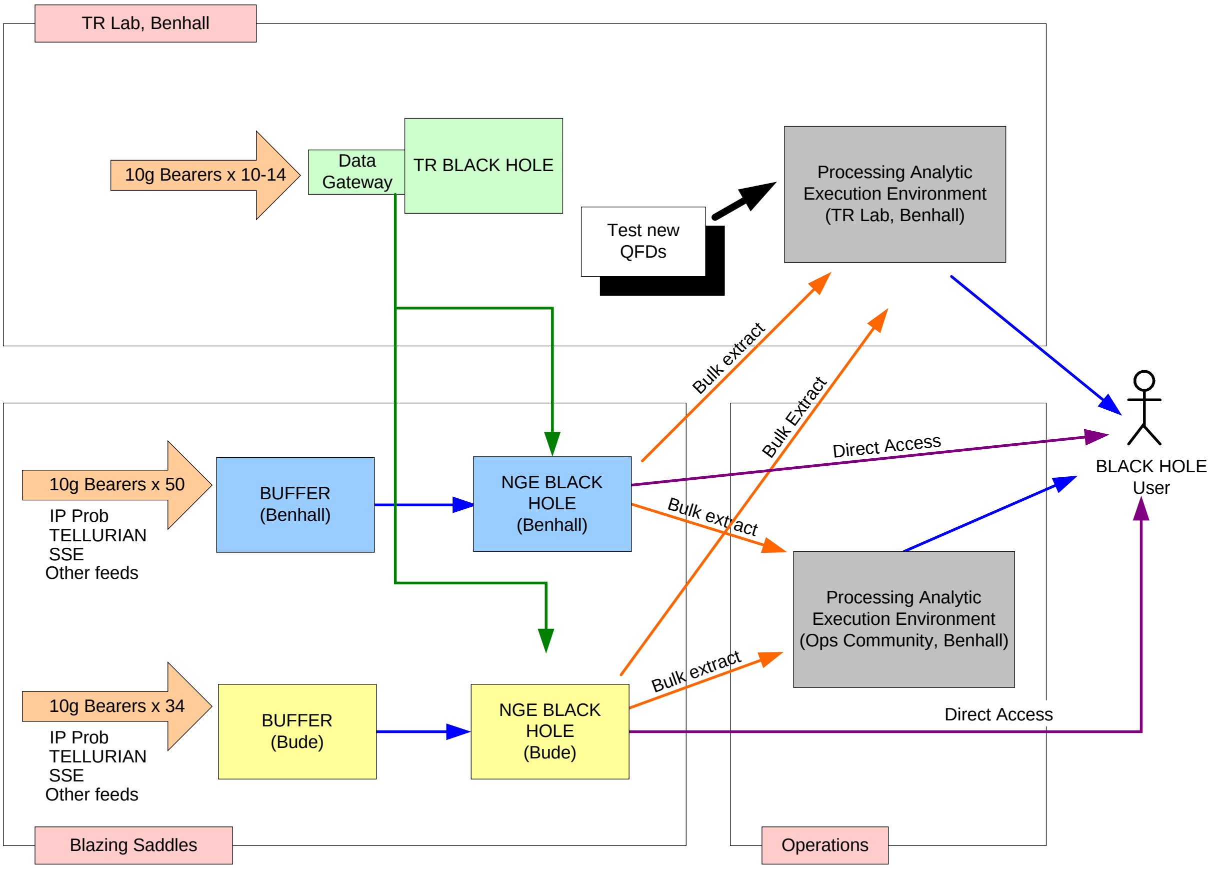


**Figure 1**

**NGE BLACK HOLE interaction**

## Business benefits

9. The Blazing Saddles BLACK HOLE will offer the following business benefit as shown in Table 1:

| Business Benefit | How it will achieve this |
|---|---|
| Enable new QFDs to be prototyped using BLACK HOLE data, then to be added to the existing Blazing Saddles QFD suite for use in Operations. | • Provide TR, and other areas, access to BLACK HOLE data, to help towards the development of new QFDs. (This includes the ability to extract data onto an external system.)<br>• Enable TR to trial new QFDs (initially developed on the TR BLACK HOLE) against BLACK HOLE.<br>• Introduce a process to manage the QFDs throughout their lifecycle on the Blazing Saddles environment. (From Tech transfer to |

| Business Benefit | How it will achieve this |
|---|---|
| | decommissioning.)<br>• Need to work out support arrangements for all this kit and the QFDs. |
| Enable the trial of new bulk analysis ideas and apply existing statistical methods in order to better understand data and develop analytical tools. | • Permit users to access data directly on BLACK HOLE.<br>• Provide processing resource for analytic trialling (part of the Capability Developer Workspace) |
| Enable new sources of data to be introduced quickly into existing QFDs. | • Introduce a standard agile process of adding new data sources to existing QFDs. |
| Enable users to understand the BLACK HOLE data to look for particular patterns and behaviours for target discovery | • Permit users to access data directly on BLACK HOLE.<br>• Allow users to perform minor processing on BLACK HOLE Management Server, or using processing resource. |
| Enable TR, GTAC and GTE access to more data for research purposes, which are not QFD related. | • Permit users to extract BLACK HOLE data onto an external system.<br>• Permit users to access data directly on BLACK HOLE.<br>• Allow users to perform minor processing on BLACK HOLE Management Server. |
| Enable the easy pull-through of new versions of BLACK HOLE from TR. | • Work with TR to keep the BLAZING SADDLES BLACK HOLE in step with the TR BLACK HOLE.<br>• Carry out development drops to enhance BLACK HOLE capability when necessary. |
| Enable increased supportability of tools and infrastructure. | • Use of corporately supported processes and infrastructure, e.g. Analytics Agility Service (AAS). |
| Removal of support burden from TR staff, allowing them to concentrate on research, rather than support. | • Support is transferred from TR staff to First Line support, Second Line support and also Deep support (TDB). |
| Engineered future expandability to cope with new protocols or volume increases. | • Scope to expand is built into the NGE BLACK HOLE system design architecture. |

**Table 1 Business Benefits**

## Data Feeds in BLACK HOLE

10. The NGE and TR BLACK HOLEs will store files containing a variety of meta data, including Target Detection Identifiers (TDIs) and encrypted content.

11. The feeds into NGE BLACK HOLE Benhall will initially be:

    - GTE TELLURIANs
    - MTI TELLURIANs
    - TR DEBIT CARD
    - MTI IP Probes
    - MTI Debit Card
    - Adhoc MAILORDER feeds from TR, OPD-GTE and OPC-TDSD

12. The feeds into NGE BLACK HOLE Bude are:

    - MTI TELLURIANs (Via MAILORDER from Bude light distribution.)
    - MTI IP Probes
    - MTI Debit Card

13. It should be noted that the NGE instances of BLACK HOLE will be fed a massively increased volume of data compared with the existing TR BLACK HOLE.

14. The aim is to feed 50 * 10Gs to Cheltenham and 34 * 10Gs to Bude, with Bude feeds possibly increasing to 50 * 10Gs.

15. Data stored on the NGE BLACK HOLE is protectively marked up to TOP SECRET STRAP2 CHORDAL

16. COI data feeds currently include:

    - CHORDAL COI

17. The retention period for all BLACK HOLE data (including Meta data considered as Content) is 6 months, although this is subject to change if the nature of the feeds going into BLACK HOLE changes significantly. The reason for this is that although most of the content QFDs have a retention period limited to a month or less, one - MARBLED GECKO has been assigned a retention period of 6 months. As most of the access to the content data is via the QFDs which have the individual controls applied, we are happy to grant the longer retention period for the content data in BLACK HOLE itself for the purposes of developing the new QFDs. Note: OPP-LEG will be kept abreast of new data types going into BLACK HOLE in case they impact on the current balance of necessity and proportionality that has led to the 6 month retention period being agreed.

18. There is a NGE (SMOKING SADDLES) future requirement for CIRCUIT and SALAMANCA data feeds into BLACK HOLE.

## Access Controls and Security

19. Users will access BLACK HOLE primarily through the BLACKFIND and BLACKCAT HTTP interface. This interface enables users to request a set of data, limited by type, name, date etc, then have that data streamed back to them.

20. Initially, data access will be coarse-grained. Only users who are members of the CHORDAL COI will be given access to the new interface. These users will have access to all data stored on the system.

21. Access to the Bude BLACK HOLE instance will be independent of access to the Benhall BLACK HOLE instance and vice versa.

22. Users will not be able to write onto BLACK HOLE disk, i.e. it is read only.

23. The Blazing Saddles BLACK HOLE is expected to have an approximate initial user base of 40 people.

24. Approval from the BLACK HOLE Manager (see BLACK HOLE Manager Role, page 12) is required before a new account is issued. Where required this will be done in consultation with OPP-LEG.

25. In accordance with Legal Policy guidance (see Reference [b]):

    - Where a user's business requirement for BLACK HOLE lapses, e.g. job/role change, their account shall be withdrawn from the system.
    - Logs of extractions must be kept and the associated HRA justification.

26. The users are expected to come from the following areas (not exclusive):

    - Technology Research:
      - ICTR-MCT (Formally B13A)
      - ICTR-FSP (Formally B13B)
      - ICTR-HSP (Formally B13C)
      - ICTR-MCA (Formally B14)
      - ICTR-NE (Formally B16)
      - ICTR-DMR (formerly B17)
      - ICTR-CISA (Formally B18)
    - Operations:
      - OPD-GTAC
      - OPD-GTE
      - GTE-BUDE
      - OPD-CRFC
      - OPC-TDSD
      - CTOR (part of OPIX-IT)

### <u>Technology Research (TR)</u>

27. The primary uses of BLACK HOLE for TR are:

    - To test and add new QFDs into NGE environment. Note: All new QFDs will be evaluated against each other, in terms of Operational benefit, to determine its pull through priority onto the NGE environment.

- To test existing QFDs with new data feeds deployed to BLACK HOLE Management Server or equivalent (See Figure 2 - NGE BLACK HOLE Data Access).
- To help upgrade/enhance the BLACK HOLE instance itself.
- To use the BLACK HOLE data for research purposes. In particular, to provide large volumes of data for ICTR-DMR to research, develop and trial new data mining analytics.
- To put data back into BLACK HOLE via MAILORDER. – (Implementation Detail - there is currently no MAILORDER capability directly onto either of the BH servers (mgmt or storage ExDS). It is envisaged that, at least initially, all data will be provided through the Buffer.)

28. Any new QFD and associated application being considered for tech-transfer will have previously been deployed within the TR area. Once it has been decided that the QFD is providing, or has the potential to provide enough business benefit, then TR and T will get together to discuss whether it is a good candidate for tech-transfer.

29. The discussion must, amongst other things, consider whether there is an existing QFD that could satisfy the data requirements of the new application. Although a new QFD may have been created in support of the new application, it may be possible, through re-engineering of the application, for it to use one of the existing QFDs.

30. It is envisaged that there will be a number of servers within the BLAZING SADDLES environment that may be used for trialling new QFDs. Initially these would be individual servers, but the aim is to be able to use some of the processing provided with the new BLACK HOLE storage.

31. QFDs and applications coming out of the TR labs will have been developed using the TR BLACK HOLE as their data source. To enable the rapid transition from TR to the NGE environment, these new QFDs will initially be tech-transferred with the NGE BLACK HOLE as their data source. This simply reinforces the importance of keeping the TR and NGE BLACK HOLES in step, as far as is possible.

32. Once the QFD/Application have proved that they provide enough business benefit, then work will start to migrate them onto more suitable hardware. This migration should include re-engineering of the QFD and its associated loading software to use the NGE Buffer as its data source, rather than the NGE BLACK HOLE.

33. For a new QFD to be trialled on BLACK HOLE, TDB-Events would expect the following to occur:

    - TR and T need to work closely together to ensure that both have a good understanding of what capacity is available, and what new QFDs are on the horizon. There is not an unending supply of processing and storage in the NGE environment. At some stage, a priority call will need to be made concerning new QFDs coming into the new environment. Decisions will need to be made by a group including T, TR, Ops, etc.
    - The new QFD will follow a template agreed between TR and TDB-Events. This will help ensure the QFD is at a consistent and suitable level for trialling

and will also help speed up the transfer of QFDs into the NGE environment. (This template will be delivered with the help of TR, and agreed by all stakeholders in QFD development. It will be lightweight.)

- The time to transfer a new QFD into the NGE environment should be approximately of the order of 1-2 weeks.
- Resource effort to transfer new QFDs will come from TDB-Events.
- The co-ordination effort between Operations trialling the new QFD will come from the BLACK HOLE Manager (see BLACK HOLE Manager Role, page 12). This also includes the establishment of a QFD lifecycle process to manage the QFDs throughout its lifecycle on the NGE environment. (From Tech transfer to decommissioning.)

34. To enable future enhancements to the BLACK HOLE system it is expected that TR and TDB-Events shall be committed to have a common development and deployment platform.

35. To feed an existing QFD or making a new QFD based around a new data source of metadata that is available via MAILORDER the following is expected to occur:

   i. The Researcher[1] contacts the data owner to obtain information about the classification of the data and the likely volumes.
   ii. The Researcher contacts the BLACK HOLE Manager to negotiate an ’incoming’ directory.
   iii. The Researcher submits an RFC to DATAFLOW CAB to get a MAILORDER feed of the data source into BLACK HOLE, referencing the BLACK HOLE Manager and the agreed incoming directory.
   iv. The Researcher develops an ingest script that can be cron-jobbed by the BLACK HOLE Manager to process the incoming data and place it in an appropriate part of the main BLACK HOLE file system. A suitable case notation is agreed.
   v. The Dataflow team liaises with the BLACK HOLE Manager to set up and test the MAILORDER feed.
   vi. The Researcher is informed when the feed is established, and they can view the statistics and status of their feed via the BLACK HOLE stats webpage.
   vii. The data either:
      a. Finds its way automatically into an existing QFD (by virtue of filenames and locations on BLACK HOLE. Note: This may be QFD dependant and may actually need a config change on the QFD’s loader.

      or

      b. The researcher uses BLACKFIND and BLACKCAT to extract the data to perfect their new QFD in experiment space.

36. For sources that are not MAILORDER, the Researcher is encouraged to use MAILORDER as this will allow control of the incoming data via a single managed interface.

[1] The Researcher role includes individuals from TR, GTE, CNE, etc.

### Operations

37. The primary uses of BLACK HOLE for Operations are:

- Run bulk queries on selectors, or geographical regions, or some other criteria, to help assist Ops when they have a surge or crisis.
- Try new bulk analysis ideas and apply existing statistical methods in order to better understand data and develop analytical tools.
- Understand the BLACK HOLE data and to use it to look for particular patterns and behaviours for target discovery.
- Bulk access/extraction to specific logs.
- Monitor hits against various TDIs, both to inform the Protocol Prioritisation List (PPL) and to ensure that TDIs remain up to date and relevant.

38. For Operations BLACK HOLE use as described above, TDB-Events would expect the following to occur:

- A BLACK HOLE Manager role will govern the management and day to day running of the BLACK HOLE. (See BLACK HOLE Manager Role for more information, page 12.)

### Crypt Target Discovery and SIGINT Development (OPC-TDSD)

39. The primary uses of BLACK HOLE for Crypt Target Discovery and SIGINT Development are:

- Off-line processing of packets and/or events on BLACK HOLE, to generate more (or enriched) events, which are put back into QFDs.
- Test and add new “Crypt” QFDs into NGE environment. Note: All new QFDs are be evaluated against each other, in terms of Operational benefit, to determine its pull through priority onto the NGE environment.
- Access Content (encrypted data samples) to aid research and capability development plus target development, e.g. attempting to decrypt traffic samples.
- Access to crypt events to aid target discovery and target development
- Monitor hits against target to inform PPL and future Crypt research and capability development.

40. For a new “crypt” QFD to be trialled on BLACK HOLE TDB-Events would expect the same principals for the trialling of TR QFD to occur.

41. In the long term future the following Crypt areas would be interested in BLACK HOLE/QFDs:

- OPC-CDP – Analytical interest.
- OPC-CDP – Development of Crypt QFDs.

## Getting Access to the BLACK HOLE Data

42. As well as being a source of data for the various QFDs, the NGE BLACK HOLE will be a source of data for TR, Ops and SD people wanting to perform investigatory and data-mining tasks on the data.

43. The task is complicated by the fact that there will be 2 instance of the BLACK HOLE – one at Cheltenham and one at Bude.

44. The option for getting access to the data and running analytics against it is:

- Extract the data from the BH storage onto the storage attached to the analytics platform.

45. Figure 2 shows the methods of getting access to the BLACK HOLE Data.

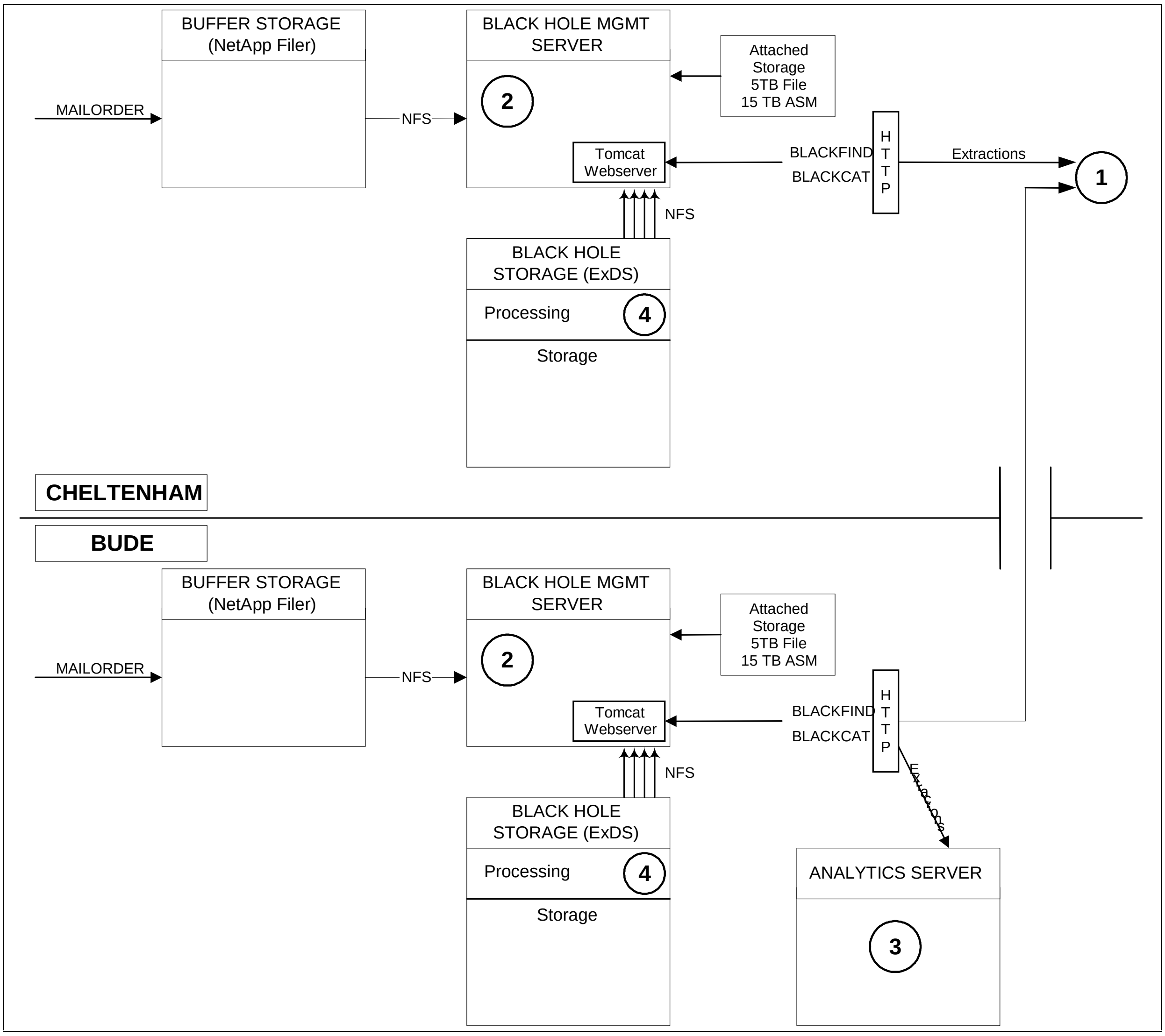


**Figure 2 - NGE BLACK HOLE Data Access**

46. Access methods are described below:

    1. The main method for getting access to BLACK HOLE data, from both Cheltenham and Bude will be via the new HTTP interface. This comprises 2 calls: BLACKFIND takes a list of criteria e.g. data type, date etc. and returns a list of the files that meet the criteria. BLACKCAT takes the list of files, and streams the data back to the user.
    2. As part of trialling new QFDs, data could be extracted from the BH storage via NFS to the BLACK HOLE Management Server. This method of access would need to be restricted, and not open for general access, as there is no way to restrict access to the data. This method would be the same at both Cheltenham and at Bude.
    3. To negate the need to transport data from Bude to Cheltenham, it may be possible to perform analysis of the data at Bude. This would entail using the HTTP interface detailed in 1.
    4. In the future, it may be possible to house some of the processing of BH data on the ExDS. However, this not possible at present.

47. Obviously the data extracted using method 1. has to go somewhere. It is anticipated that extractions would either go onto the user's own platform, or onto one of the servers in the new AAS Experiment Environment.

**Data usage scenarios:**

48. Bulk extraction for further analysis: It is envisaged that the vast majority of analysis of the BLACK HOLE data will take place off of the BLACK HOLE platform e.g. on one of the servers in the new AAS Experiment environment. To extract data from the NGE BLACK HOLE, the user will utilise the BLACKFIND and BLACKCAT procedures described above in method 1. This will enable the user to stream a set of data back to their platform for further analysis. The procedure will be valid for extracting data from both the Cheltenham and Bude BLACK HOLE instances. The analysis platform would normally be housed at Cheltenham. However, in the future it may be possible to provide analytic platforms at Bude to reduce the amount of cross-site network traffic.

49. Direct access to BLACK HOLE data: For the trialling of new QFDs, the BLACK HOLE disk would be directly NFS mounted to the QFD platform (shown above as the BLACK HOLE MGMT Server). Unfortunately, it will not be possible to NFS mount BLACK HOLE disk directly to analytic platforms due to the large number of security concerns that this approach raises. The majority of these concerns are removed by the use of the new HTTP interface, as described above.

## BLACK HOLE Manager Role

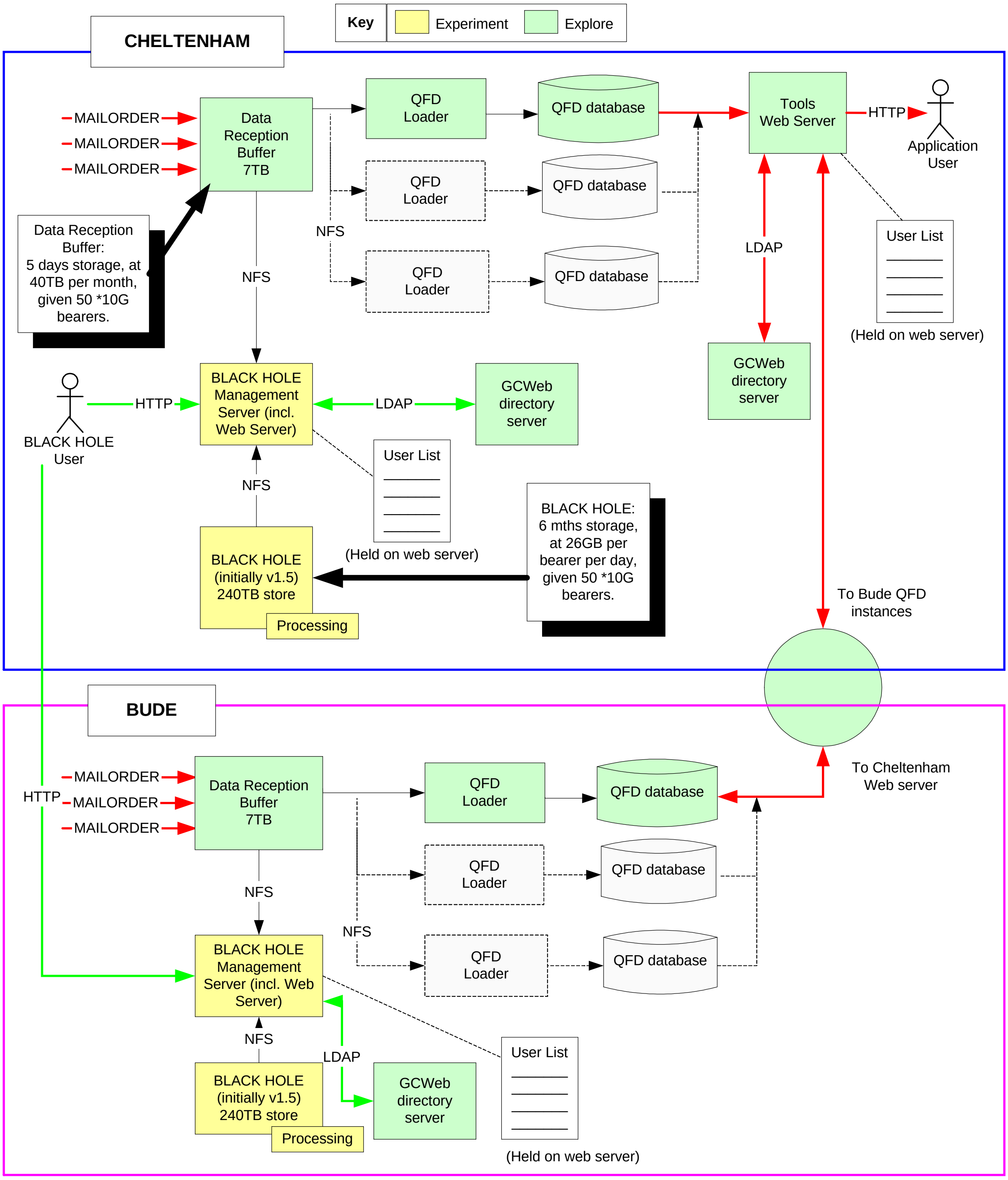
50. To enable the BLACK HOLE environment to operate in an effective and efficient manner a BLACK HOLE Manager Role shall be introduced.

    Note: No one has yet been identified for this role.

51. The BLACK HOLE Manager role and responsibility includes:

    - General maintenance of user accounts and file stores. This includes the recording of each user's business cases for access to BLACK HOLE.
    - Providing support to Users wishing to extract bulk data from BLACK HOLE onto an external Capability Developer Workspace for further analysis.
    - Ensuring ongoing legalities and security compliance.
    - Working with the BLACK HOLE Design Authority in TR to pull-through functionality from new versions of BH.
    - Provide support for new QFDs coming through for trialling.

## Appendix A – BLAZING SADDLES System Overview


Key
🟨 Experiment
🟩 Explore
CHELTENHAM
MAILORDER
MAILORDER
MAILORDER
Data Reception Buffer 7TB
Data Reception Buffer: 5 days storage, at 40TB per month, given 50 *10G bearers.
QFD Loader
QFD database
QFD Loader
QFD database
QFD Loader
QFD database
NFS
NFS
Tools Web Server
HTTP
Application User
User List
(Held on web server)
LDAP
GCWeb directory server
BLACK HOLE User
HTTP
BLACK HOLE Management Server (incl. Web Server)
LDAP
GCWeb directory server
User List
(Held on web server)
NFS
BLACK HOLE (initially v1.5) 240TB store
Processing
BLACK HOLE: 6 mths storage, at 26GB per bearer per day, given 50 *10G bearers.
To Bude QFD instances
BUDE
To Cheltenham Web server
MAILORDER
HTTP
MAILORDER
MAILORDER
Data Reception Buffer 7TB
QFD Loader
QFD database
QFD Loader
QFD database
QFD Loader
QFD database
NFS
NFS
BLACK HOLE Management Server (incl. Web Server)
NFS
LDAP
BLACK HOLE (initially v1.5) 240TB store
Processing
GCWeb directory server
User List
(Held on web server)