PRESTON Architecture

Version 3.0

Synopsis

This document presents an architecture for PRESTON.

		Signature	Date
Creator:	[redacted] (PRESTON SDA)		
Approver(s):	[redacted] (MoMo TD)		

PRESTON Architecture Sigmod/00007CPO/4502/SIG011000/05
5 July 2007

Distribution

By email/softcopy:

[redacted] (H/OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (OPA-SSOS)
[redacted] (TFE)
[redacted] (TFE)
[redacted] (TDB)
[redacted] (TDB)
[redacted] (NTAC)
[redacted] (ITIP)
[redacted] (TDB)
[redacted] (TDB)
[redacted] (TPS)
[redacted] (OPD-GTE)
[redacted] (NTAC)
[redacted] (NTAC)
[redacted] (NTAC)
[redacted] (TPM)
[redacted] (TFE)
[redacted] TPM)
[redacted] (TDB)
[redacted] (TPS)
[redacted] (TDB)
[redacted] (TPS)
[redacted] (TIS)
[redacted] (TDB)
[redacted] (TPS)

Document Amendment History

Version	Date	R&A	Amendments
1.0	10 May 2006	yes	First formal release
2.0	4 October 2006	yes	Second formal release (reference RFC136 Sigmod/00080CPO/4502/SIG006400/12)
3.0	5 July 2007	yes	Third formal release

Contents

Reference Documents

[a] PRESTON Vision
vob: preston/tech/reqts/preston-vision.doc
[b] PRESTON Business Processes
vob: preston/tech/business/preston-business-processes.doc
[c] PRESTON System Requirements Specification
vob: preston/tech/business/preston-system-requirements-specification.doc
[d] PRESTON Volumetric Model Overview
vob: preston/tech/analysis/preston-volumetric-model.doc
[e] FAST GROK
[redacted]

1 INTRODUCTION

1.1 Purpose

The purpose of this document is to present the PRESTON architecture.

Version 3 incorporates developments for NHIS 2 processing.

1.2 Scope

From DCSD web site:

> "PRESTON collection is the warranted intercept of UK line access. It covers fixed and mobile communications; and voice and data. Each target must be covered by a RIPA 8(1) warrant. GCHQ is one of eight intelligence and law enforcement agencies involved in this type of collection."

RIPA 8(1) provides for warrants to be placed on a person or organisation.

This document provides an end-to-end architecture for deriving intelligence from packet-based intercept acquired from RIPA 8(1) warrants. Intercept acquired in this manner is often referred to as Lawful Intercept (LI).

This PRESTON architecture is future facing, so the following things are in scope:

- C2C intercept.
- LI streams and processed data, which are delivered using the LI, network. (a.k.a. NTAC strategic network)

and so the document excludes:

- Circuit-switched intercept. Much of this is likely to be superseded when the UK infrastructure migrates to VoIP. VoIP intercept is in scope. Circuit-switched intercept which is translated to an IP intercept handover is in scope.
- BOXSTER, MARMION, GENTIAN. These systems are legacy circuit switched and line access solutions.

1.3 Glossary

GCHQ	Government Communications Headquarters
NTAC	National Technical Assistance Centre, and is responsible for maintenance of the LI capability on the UK on behalf of the intelligence and law enforcement agencies.
LI	Lawful Intercept refers to intercept gained as a result of the use of a legal instrument to obtain interception services from a CSP.
Active LI	LI intercept which doesn't follow the standard passive LI intercept model. Interaction with the CSP may be required e.g. cloned email accounts.

NHIS	National Handover Interface Specification – the UK-mandated handover mechanism for warranted intercept.
TERRAIN	C2C Processing system used at GCHQ.
CSP	Communication Service Provider

2 ARCHITECTURAL REPRESENTATION

The features which deliver the PRESTON architecture are largely realised in the PRESTON Processing system. Thus, the representation of the PRESTON Processing system forms the largest part of this document.

The PRESTON architecture is represented using the following views:

- **Section 4** presents a **Use-Case view**. This is a brief overview of the use cases which are architecturally significant.
- **Section 5** presents a **Logical view** overview.
- **Section 6** presents a **Logical view** of the **Lawful Intercept capability**. This describes the functionality which implements the warrants derived from the *Enable collection* business use case and delivers the intercept to GCHQ. This part of the system is provided by NTAC.
- **Section 7** presents a **Logical view** of the **Processing capability**. This describes the part of the system which integrates the NTAC-managed delivery network, applies processing and derives intercept items for storage.
- **Section 8** presents a **Logical View** of the **Offline processing and storage** aspects which store and render intercept for analysts.
- **Section 9** presents a **Deployment view** which describes how the facilities described in the Logical views are deployed.

3 ARCHITECTURAL GOALS AND CONSTRAINTS

3.1 Features

This table is a copy of the Features from the PRESTON Vision (reference [a]):

Id	Description	Priority	Derived
FEAT2	The system will be deployed diversely across two sites to allow operations to continue on failure of one site.	Highly Desirable	NEED7
FEAT3	The system will support processing of broadband[1] intercept.	Essential	NEED7
FEAT4	A comprehensive LI interception capability will be provided by NTAC.	Essential	NEED7
FEAT5	The system will present intercept to analysts in a manner consistent with other accesses.	Highly Desirable	NEED7
FEAT6	The system will support processing of mobile intercept.	Essential	NEED7
FEAT7	The system will support the processing of data which is received from NTAC's PDO (Processed Data Output) service.	Essential	NEED7
FEAT105	The MTRR of the operational LI processing service shall be less than 24 hours.	Highly Desirable	NEED7
FEAT9	The legacy PRESTON delivery infrastructure shall cease to incur maintenance costs.	Highly Desirable	NEED7
FEAT10	A secure managed interface will provide an accredited interface to the NTAC LI network.	Essential	NEED7
FEAT106	The system shall ensure the integrity of data so that data loss is no worse than 0.05% by volume.	Desirable	NEED7
FEAT12	The system shall support processing throughput up to 34 Mb/s.	Essential	NEED7
FEAT103	The system shall support processing throughput up to 100 Mb/s.	Highly Desirable	NEED7
FEAT14	The system will support delivery of LI streams to SD analysis labs for analysis.	Highly Desirable	NEED8

[1] broadband: IP stream accesses with bandwidth above 128k e.g. ADSL or Cable modem.

FEAT19	The system will support delivery of LI streams to CESG analysis labs for the application of intrusion detection techniques.	Desirable	NEED9
FEAT109	The system will support delivery of LI streams to A&R labs for the purpose of applied research.	Desirable	NEED9
FEAT15	The system will be capable of processing VoIP presented as NHIS 2.0.	Essential	NEED7
FEAT16	The system will support application of VPN decryption techniques.	Highly Desirable	NEED7
FEAT17	The system will have processing for IP multimedia services.	Desirable	NEED7
FEAT18	A Memorandum of Understanding will be agreed with other agencies to support sharing of LI processing products which are sharable.	Desirable	NEED9
FEAT19	LI processing solutions will be made available to NTAC so that NTAC can provide LI services to all the UK intelligence and law enforcement agencies.	Highly Desirable	NEED9
FEAT107	The system will support delivery of events to event repositories in accordance with defined events interfaces.	Desirable	NEED7

3.2 Volumetric requirements

Ref [d] presents a volumetric model for PRESTON, which describes the volumetric and performance requirements on the system. It is useful to present the findings from that document. The document describes a requirement for egress to deliver at up to 72 Mb/s:

Totals	**Long-term average bandwidth:**	**6.3 Mb/s**
	Max bandwidth:	**71.7 Mb/s**
Fractions	% of bandwidth which is broadband:	50.1 %
	% of bandwidth which is voice:	48.8 %
	% of bandwidth which is mobile:	1.1 %

Clearly this quantity of data will be a considerable problem for processing and storage, and so the volumetric model makes an assumption of 95% filtering of collected C2C intercept.

Collection	Voice filtering rate:	0 %	
	Voice collect:	1.40 Mb/s	35.00 Mb/s
	C2C filtering rate:	95 %	
	C2C collect:	0.24 Mb/s	1.83 Mb/s

95% de-selection is a reasonable estimate of what is possible based on current understanding of how BLACKNIGHT selectors can be used to reduce data rates. This allows us to reduce the collection rate of C2C rate to around 1.9Mb/s. For further discussion of selection and filtering see section 7.4 on page 35.

The following storage volumes were also presented in the volumetric model, derived from the other statistics.

C2C collection	Storage duration:	6 months
	Storage rate:	**2.6 GB/day**
	Storage required:	**461.6 GB**
Voice collection	Storage duration:	6 months
	Storage rate:	**14.8 GB/day**
	Storage required:	**2657.8 GB**
Survey store	Storage duration:	6 months
	Storage rate:	**4.2 GB/day**
	Storage required:	**748.8 GB**

4 USE-CASE VIEW

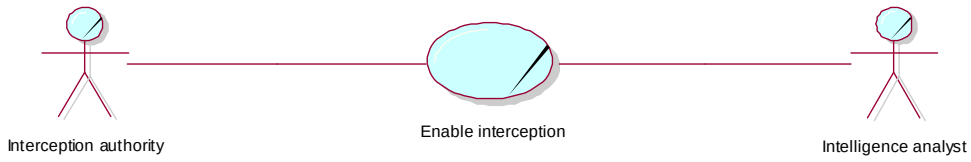
This section discusses the use case requirements gathered, and describes how they contribute to shape the architecture.

The use cases are contained in two documents:

- Business use cases and analysis were presented in the *PRESTON Business Processes* document. Use cases for the *Collection system* and *Stream routing* functions were thus derived from the business use cases.
- Use cases for the *Collection system* and *Stream routing* were presented in the *PRESTON System Requirement Specification* document.

4.1 Enable interception

The *Enable interception* use case describes how an intelligence analyst can achieve interception against a target.

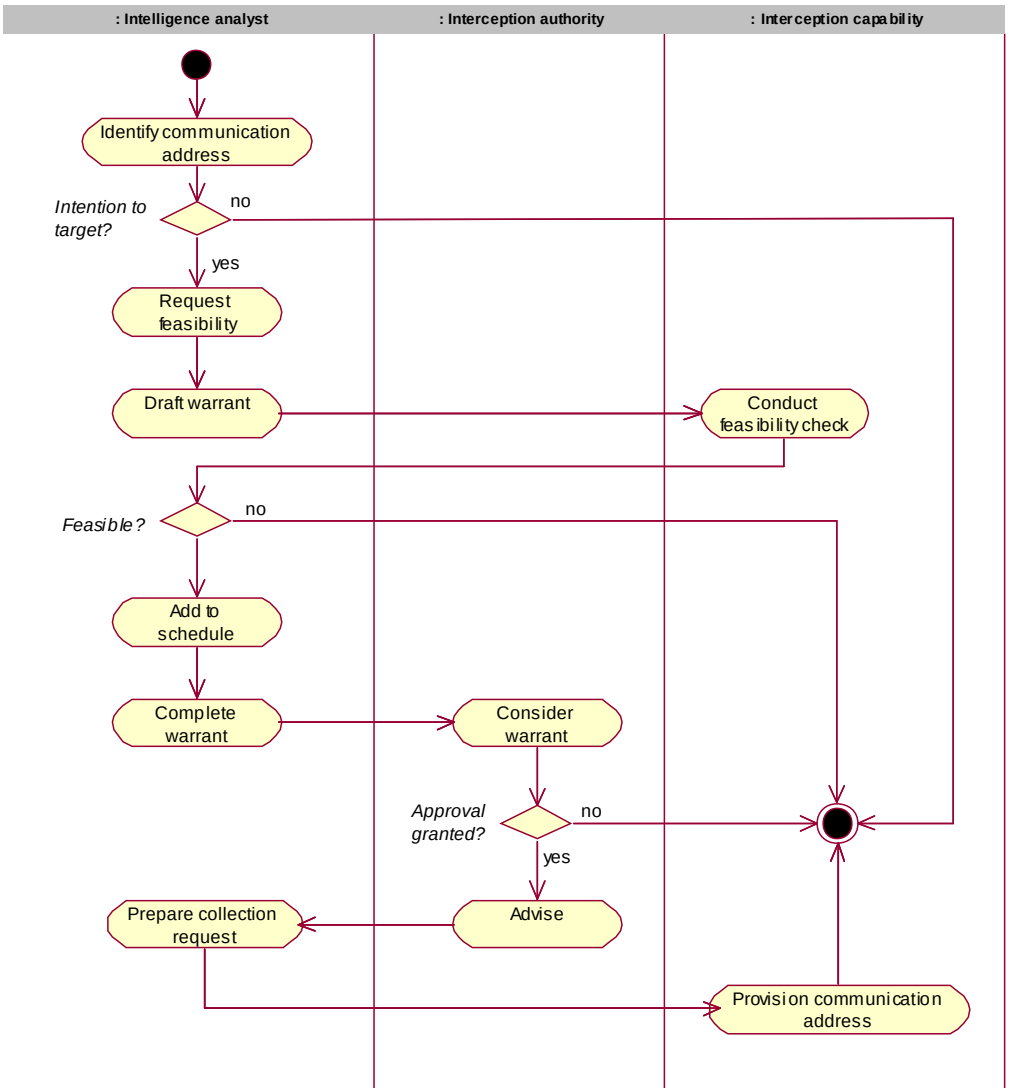


The use case describes how the actors apply a formal process to achieve the interception. In short, the process covers preparing a case for interception, application for a warrant, warrant review, and (if successful) the provisioning of the communication intercept.

It is immediately clear that this process completely covers the legality requirements:

The process is compliant with RIPA, as no communication interception can occur without a warrant in place. The RIPA requirements ensure compliance with HRA, in particular, the case for interception must be strong for the warrant to be obtained.

For reference, a collaboration diagram representing the *Enable interception* use cases is given below.



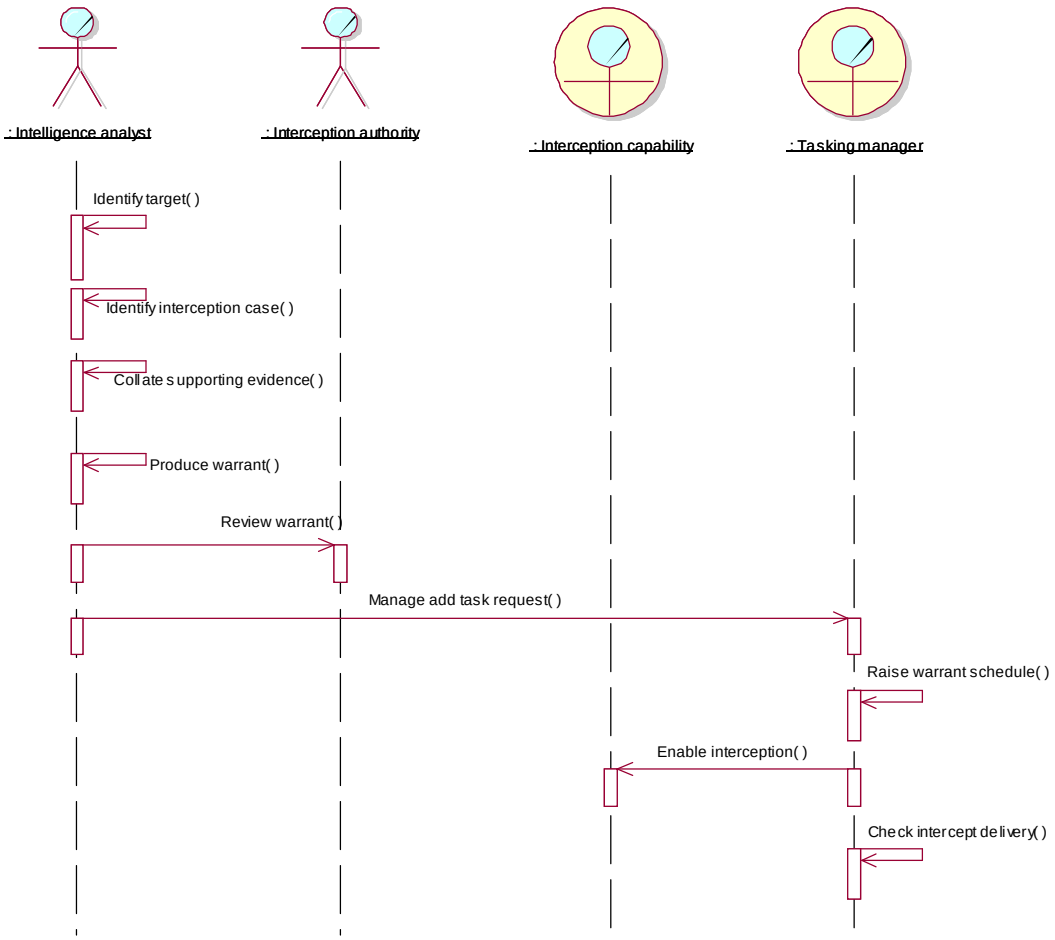
4.2 Business process detail

It is worth looking at a few aspects of the business process in detail (see ref [b]). The business process used in GCHQ differs considerably from the process employed in the other UK agencies. Some LI streams can be transitioned to collection processing immediately if the structure of the data is well understood, and no survey is necessary. For instance, a voice call warrant which delivers only VoIP is likely to need no survey, as the structure of the delivered stream (64k timeslot) needs no analysis.

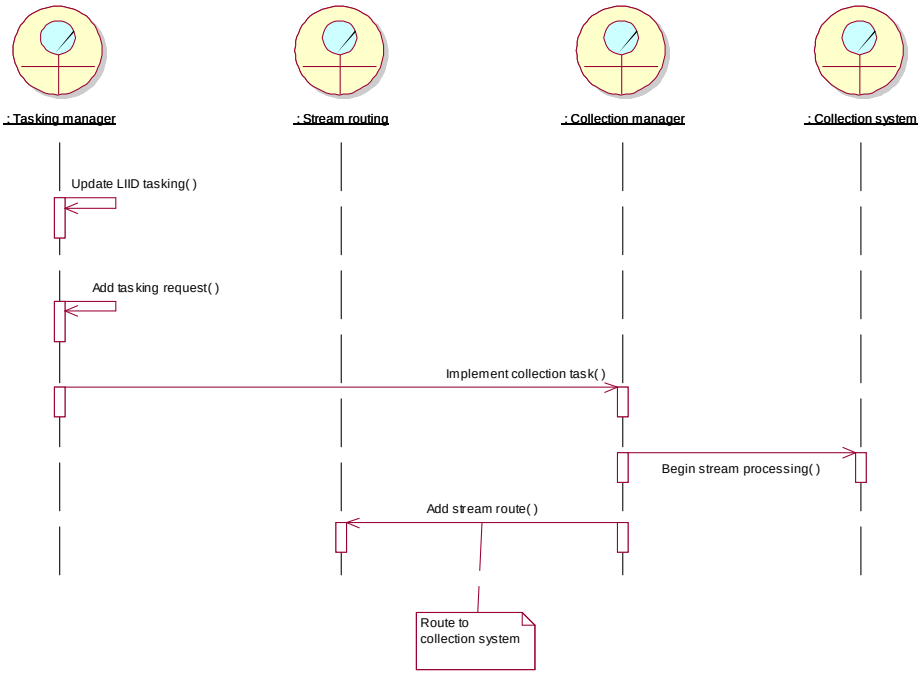
Other streams, such as broadband lines need analysis, as there is no obvious default processing configuration which can be applied to these lines. All users use of a broadband line is different. The internet is flexible, and supports a myriad of

protocols, and so a more complex process is used.

The sequence diagram for the grant of a warrant is shown below:

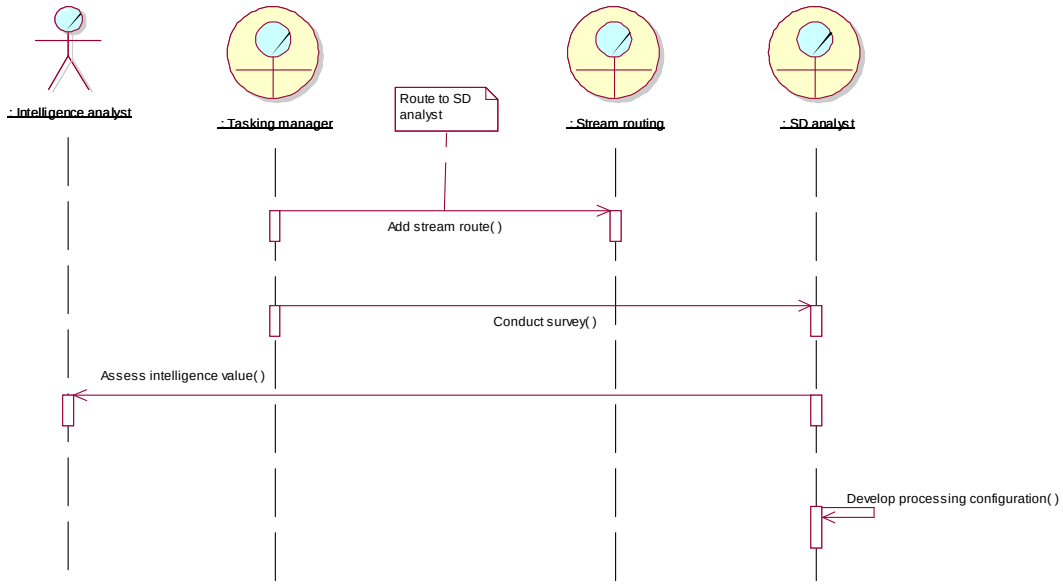


This part of the business process takes the warrant management as far as ensuring that the new intercept stream is delivered to GCHQ correctly. Once stream delivery has been configured, the stream can be routed to collection or survey. Configuration of collection processing is shown in the sequence diagram below:



Putting a stream onto collection is a simple process. The Collection manager manages the *Stream routing* and *Collection system* at the request of the *Collection manager*. The fulfilment of the *Tasking manager* and *Collection manager* roles at GCHQ is likely to be PRESTON ops, and GSOC, respectively.

Putting a stream through a survey is a key part of the process. This is depicted in the sequence diagram below:



The stream is routed to the *SD analyst* who performs a survey (technical characterisation). The *Intelligence analyst* is consulted to assess the intelligence value which is used to assess whether the warrant is maintained.

If the warrant is to be maintained, the stream is migrated onto the collection processing. A key product from the survey process is the processing configuration to be used for collection processing. This is developed by the SD analyst – it is derived from the technical characterisation of the new stream.

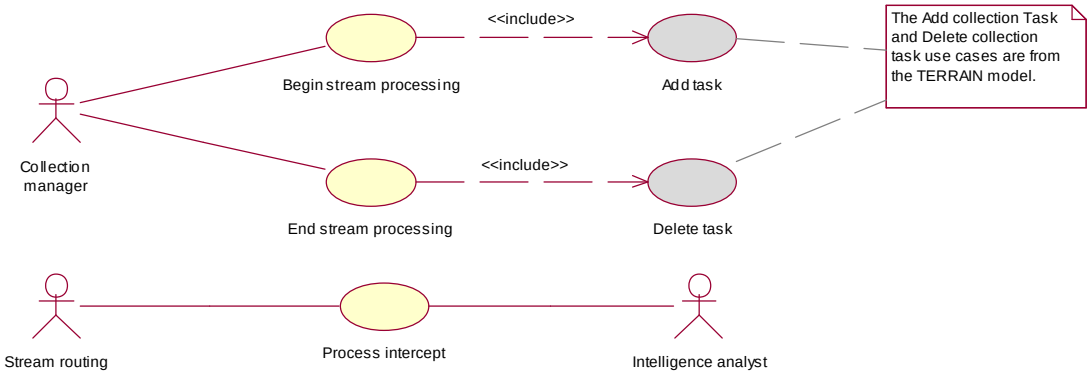
The survey process and intelligence assessment is what GCHQ provides a high quality approach to ensuring that warrants are maintained only for appropriate sources of intercept.

The development of the processing configuration is also a key part of managing the data volumes. GCHQ maintains a considerably high level of volume reduction by applying selectors and filters developed at this stage. To compare with a similar operation, volume reduction is not a feature offered by the NTAC DPC. The DPC applies some basic filters to eliminate protocols (such as file sharing) that none of the agencies could possibly want. No doubt offering a filtering service for all agencies for all targets would require a significant amount of management time.

When intercept is initially enabled, the selector or filter terms are not well understood. They are thus developed during the survey phase, by Intelligence analysts (as part of intelligence assessment) or by SD analysts (as part of the development of the processing configuration). Filters developed by SD analysts are more likely to be technology filters, whereas those developed by Intelligence analysts are likely to be target selectors. Selectors and filters are currently managed using CORINTH.

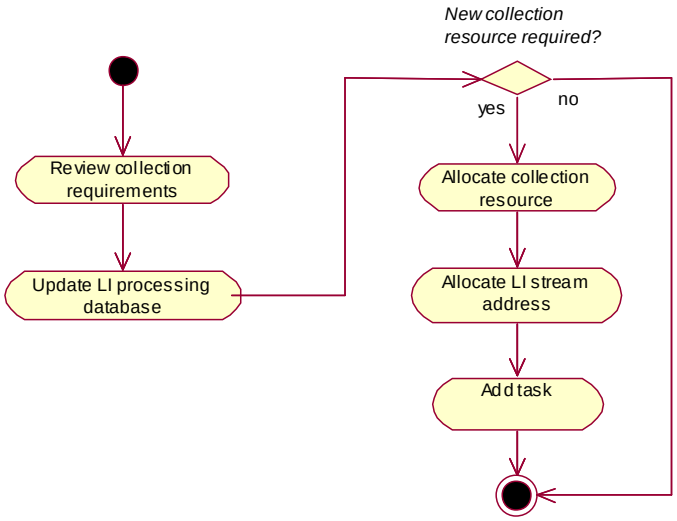
4.3 Collection system

The use case model for the *Collection system* is pictured below:



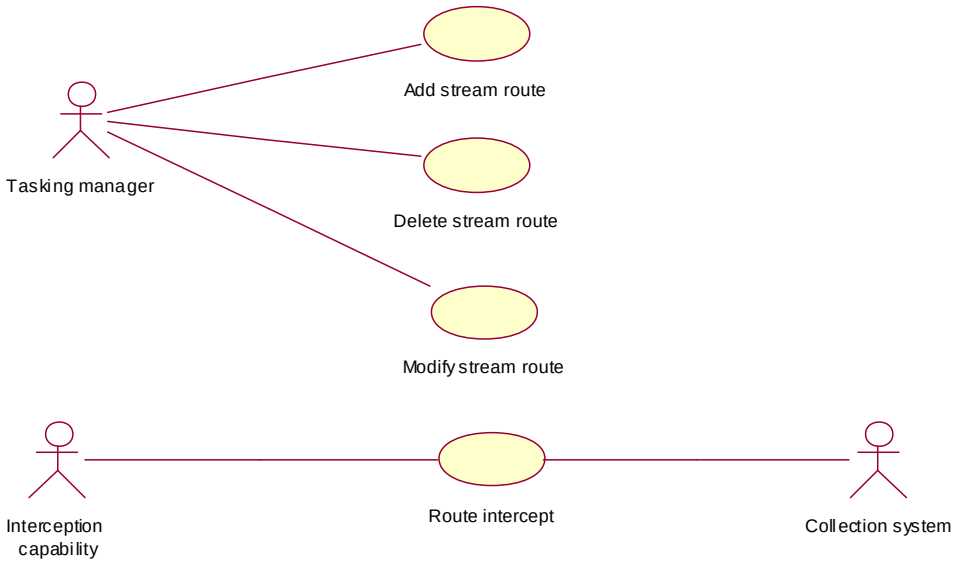
The *Begin stream processing* use case is illustrated below. The Collection manager is required to manage a number of things in order to successfully implement a stream processing task:

- The stream addresses assigned to each task must be managed to ensure each collection task has a unique address.
- The collection resources must be managed.
- The LI database maps LIIDs in the input streams to metadata for GCHQ's processing.



4.4 Stream routing

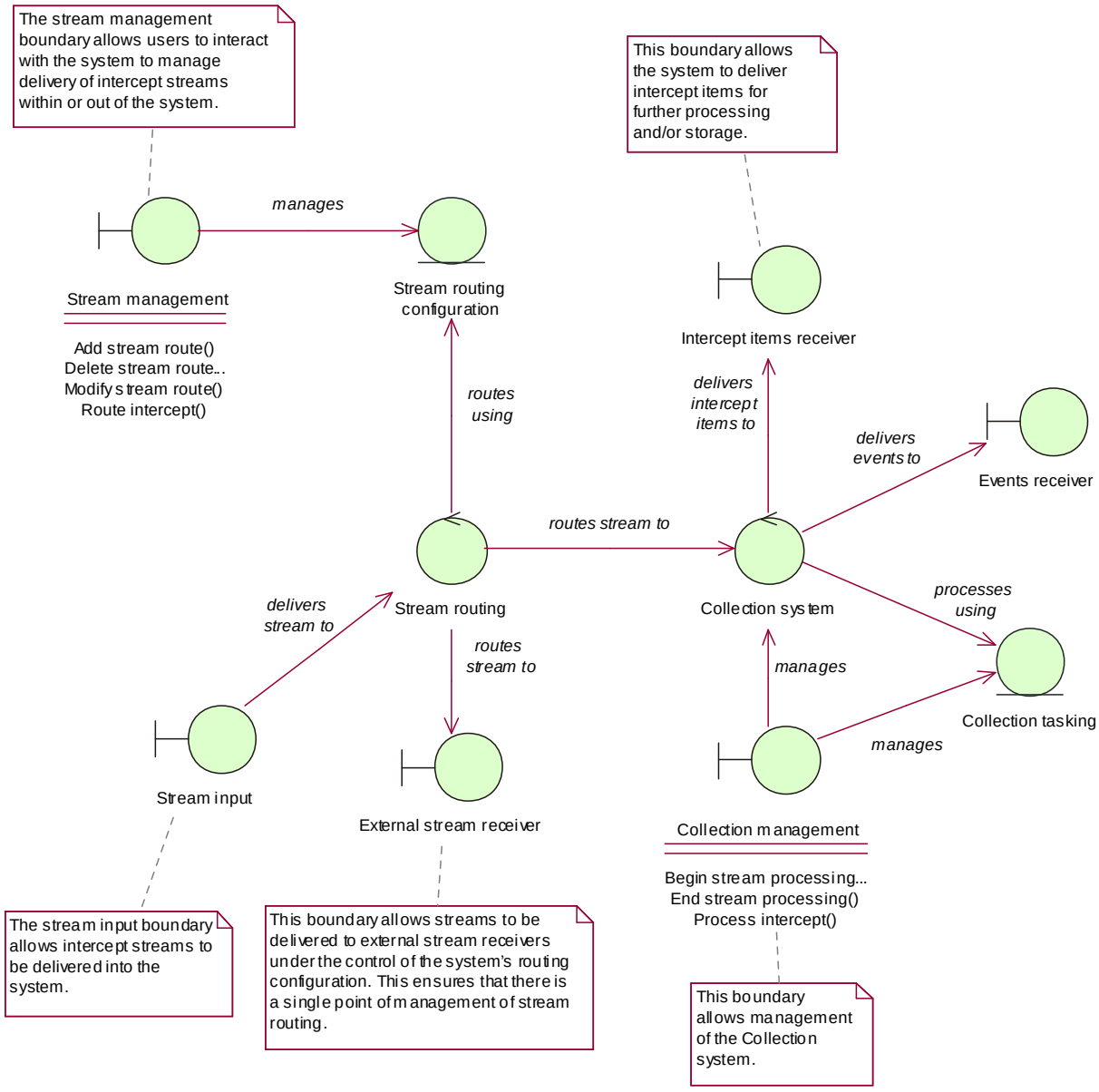
The use case model for the *Stream routing* is pictured below:



The detail of the use cases is not significant at an architectural level, and won't be discussed further in this document, although the use cases will be used to drive the development and testing.

4.5 Analysis model

The architecture is derived from an analysis model which captures the concepts of behaviour and responsibility in the system. The first diagram presents the use cases mapped to logical components which deliver a stream processing functionality.

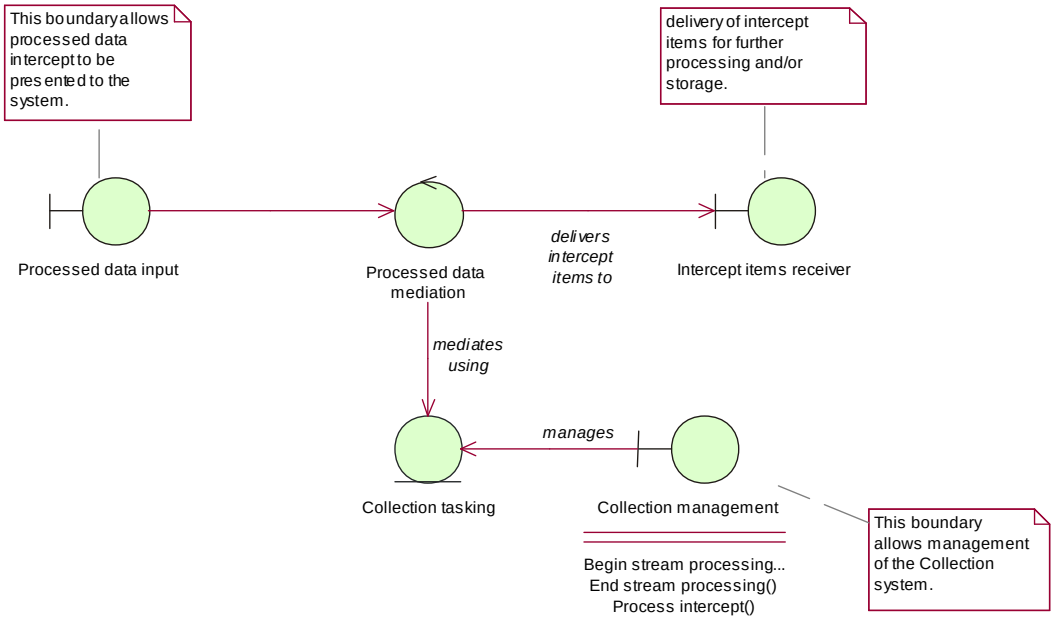


Stream management The stream management boundary allows users to interact with the system to manage delivery of intercept streams within or out of the system.

Stream routing configuration This entity represents the current routing table - the instructions for the routing which should be applied to each stream.

Stream routing	This control accepts intercept streams and implements the route which has been requested for the stream.
Stream input	The stream input boundary allows intercept streams to be delivered into the system.
External stream receiver	This boundary allows streams to be delivered to external stream receivers under the control of the system's routing configuration. This ensures that there is a single point of management of stream routing.
Collection system	This control processes intercept streams and delivers intercept items.
Intercept items receiver	This boundary allows the system to deliver intercept items for further processing and/or storage.
Events receiver	This boundary allows the system to deliver events for query and/or storage.
Collection tasking	The collection tasking information specifies the processing which should be applied to each stream by the collection system.
Collection management	This boundary allows management of the Collection system.

The system requirements (ref [c]) articulate the need for the handling of processed data. The diagram below shows the components which deliver this functionality. The two diagrams have *Collection management* and *Collection tasking* in common. The functionality embodied in these concepts will be common in both processing systems.

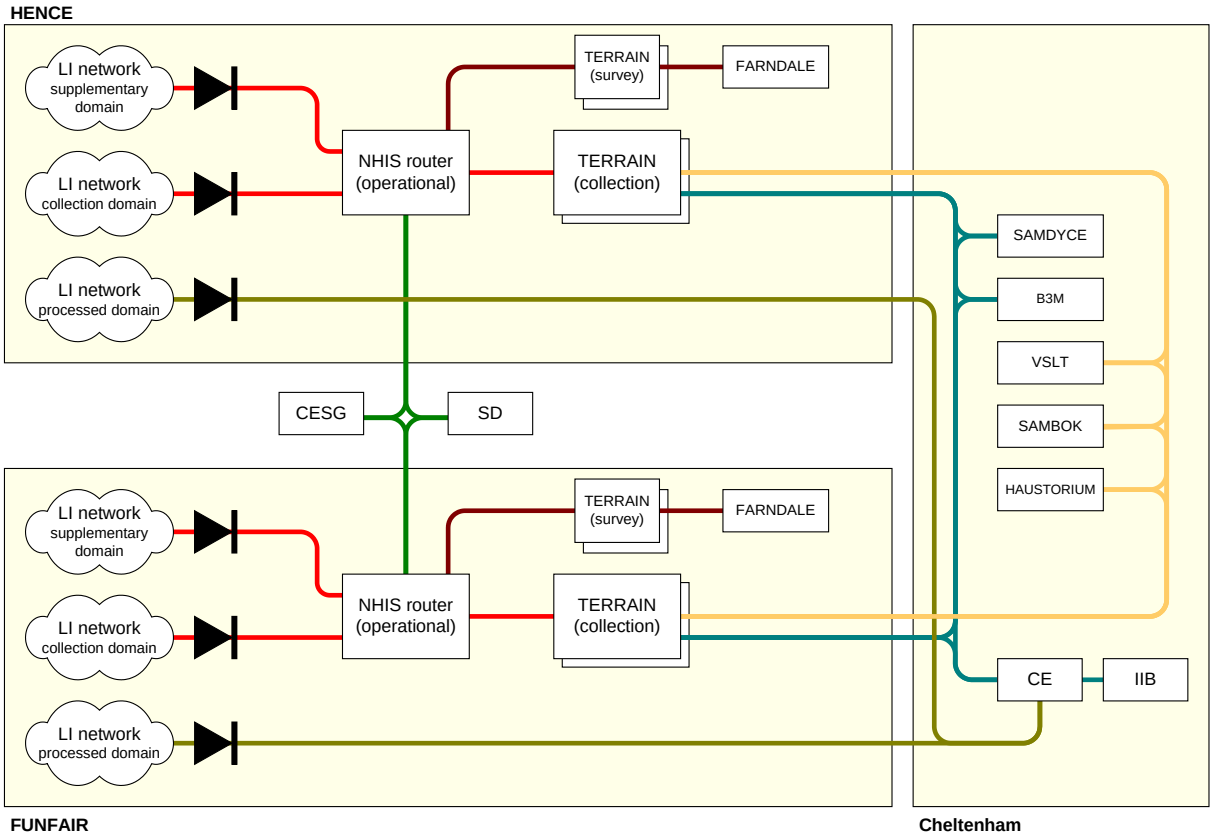


Processed data input This boundary allows processed data intercept to be presented to the system.

Processed data mediation The processed data mediation applies any conversion required to be able to deliver processed data to the Intercept item receiver.

5 LOGICAL VIEW: OVERVIEW

The diagram below presents an overview of the complete system.



The system consists of two processing centres deployed across two geographically separate sites.

5.1 Delivery

Delivery to GCHQ's processing occurs on 3 domains. Streamed data is presented on the Supplementary and Collection domains. Processed data is delivered on the Processed domain. All domains are connected to GCHQ's networks via data diodes for security reasons.

5.2 Streamed data

Data from the Supplementary and Collection domains is presented to the NHIS router which can route individual intercept streams. Streams can be delivered to TERRAIN for collection processing, or to non-operational areas such as SD or CESG.

A separate NHIS router and TERRAIN cluster are provided for survey processing at each site.

The TERRAIN system can deliver to all repositories.

5.3 Processed data

Processed data is delivered to Content Enhancement for processing prior to delivery to the IIB.

5.4 Offline processing and storage

The TERRAIN system can deliver events:

- VoIP events in telephony form to SALAMANCA.
- Geo events to SAMBOK.
- C2C events to HAUSTORIUM.

The TERRAIN system can deliver content:

- Operational collect data to Content Enhancement for processing prior to sending to the IIB.
- VoIP call data to B3M.
- Survey or target development data to FARNDALE for analysis by SD analysts.
- SMS content to SAMDYCE.

6 LOGICAL VIEW: INTERCEPTION

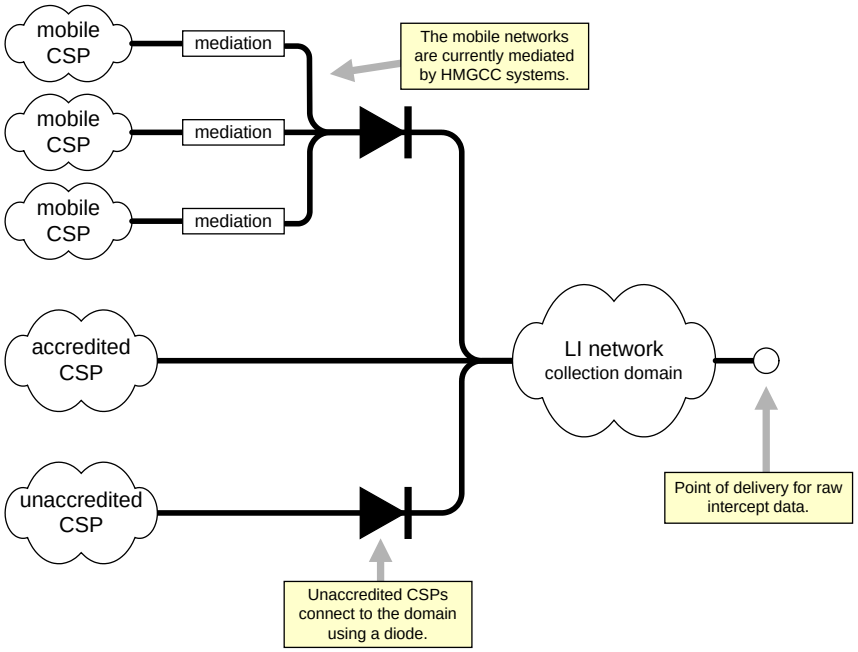
6.1 Overview

The lawful intercept capability is currently managed and provided by NTAC. GCHQ fully supports NTAC's mandate, and there is no intention of changing this arrangement.

As well as the set of obvious set of benefits it offers HMG (e.g. pool the costs to get a better system) there are benefits in having NTAC as the organisation which approaches and manages the relationship with CSPs, which reduces GCHQ's association with CSPs.

For completeness, the following sections describe the parts of the system which comprise the LI capability.

6.2 Collection domain

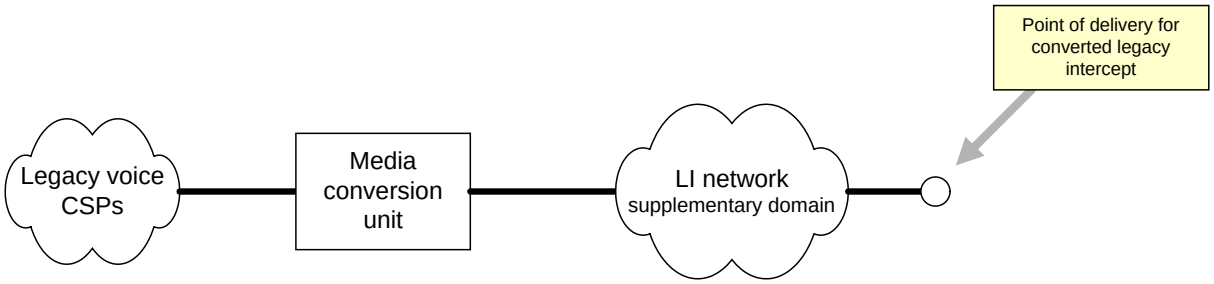


The LI collection domain provides a network domain for CSPs to deliver intercept data. CSPs which gain SECRET accreditation for their interception services may connect directly to the LI network, while those that cannot be accredited are connected using a diode. A number of mobile operators are connected the collection domain. No common handover agreement has been made with mobile providers, so each provider connects to the collection domain via a HMGCC managed mediation unit.

The NHIS 1.1, NHIS 1.5 and NHIS 2 handovers are approved for delivery of intercept on this network.

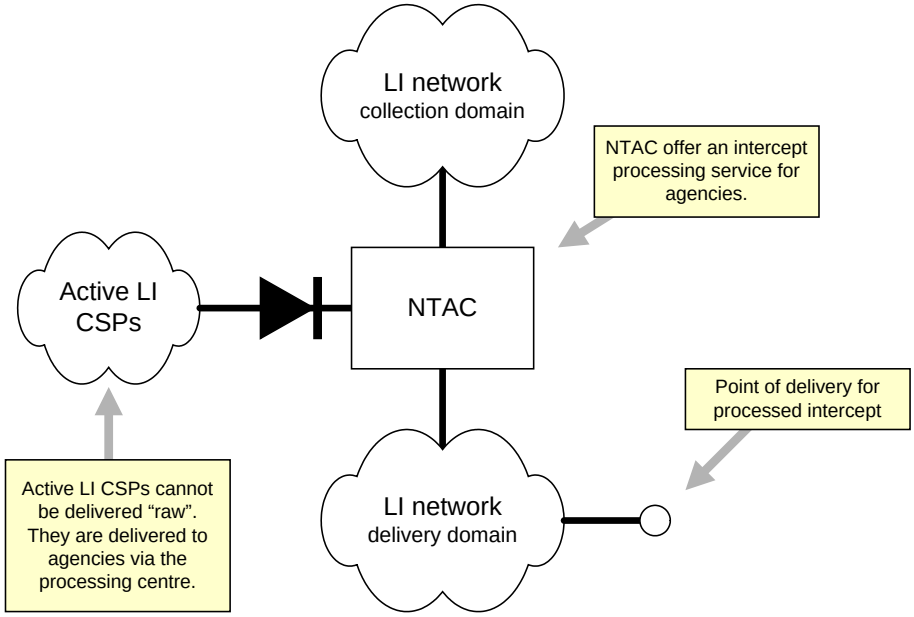
6.3 Supplementary domain

The LOCHNVAR project is intending to migrate circuit-switched intercept from existing (circuit-switched) handover to an NHIS 2 handover.



The supplementary domain is intended for this purpose. A separate domain is used for security reasons.

6.4 Delivery domain



The processed domain allows delivery of processed data to agencies. NTAC offer a processing service for agencies, capable of processing raw data from the collection domain, and delivering it on the delivery domain. The delivery domain is also used for Active LI sources which are delivered to NTAC in processed form. Such sources have no raw form, and so cannot be presented in accordance with an NHIS handover.

6.5 Other sources

The interception of a leased line under RIPA 8(1) is a more complex warrantry scenario. There is no standard LI handover for a leased line[2], and so delivery would not take place via NTAC-managed interception processes. Typically, a leased line is delivered on a bespoke network, or copied through the MONACO delivery networks to the narrowband processing systems. There are no specific PRESTON features to this processing scenario, and no new capabilities are required as a result.

GCHQ has, as an option, the ability to use RIPA section 11 to deploy an intercept probe outside of NTAC's management. A bespoke delivery network would be needed to get the data to us.

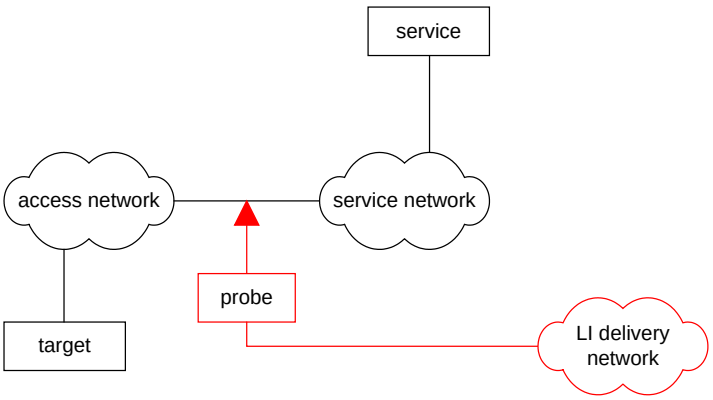
6.6 Application layer handover

There is presently no support for a number of application layer handovers for which requirements exist. For example, in certain CSPs, email or HTTP proxy interception is performed, however there is no defined handover for this form of intercept: NHIS 1.1 is used currently.

The definition of a handover standard achieves two key things: Firstly, there is an agreed mechanism to present the intercept data to the LEMF from the CSP. Secondly, there is a clear set of requirements on the CSP which specify what must be intercepted.

As there is no defined handover for e.g. an HTTP proxy interception, there is no way to be certain that the intercept will be correctly presented, or that the intercept will be derived in a useful way.

Consider an application layer probe deployed on a network which is removed from the target.



The probe is likely not integrated with the ISP network, and so must deduce the presence of a target's communications by studying the protocols which make up the

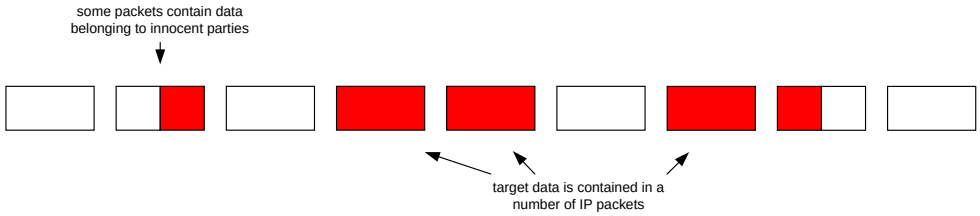
[2] ETSI TS 102 815 defines a layer 2 handover mechanism which may be appropriate for a leased line, but this is not a standard handover in the UK.

target communications e.g. by studying the application layer protocols.

Consider a typical protocol stack.

The application layer protocols may be at the top of stack of complex protocols, and so a number of packet decoding steps may be required to get from network packets to application layer PDUs.

In practice, this means that a target's communications may be dispersed across a number of network packets, and that the target's communications may be interspersed with communications which are not from the target.



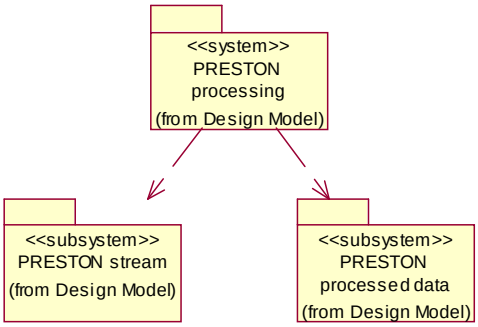
The typical LI handover is at the network packet layer (i.e. IP packets), and yet this handover may not offer the correct precision required to exactly capture a target's communication.

I must conclude that, with the increasing use of application layer probes, that we need to work with NTAC and the other agencies to ensure handover mechanisms are defined which allow capture of target communications with application layer probes.

7 LOGICAL VIEW: PROCESSING

7.1 Processing

The *PRESTON processing* system consists of two subsystems: *PRESTON stream* and *PRESTON processed data*.



PRESTON stream is the primary system: it realises the bulk of the use cases, and meets the majority of the system requirements. *PRESTON processed data* is used to handle the cases where intercept cannot be delivered in a stream, perhaps because a file or message-based interception is used, in which case it is not possible to intercept the target's raw IP streams. The *processed data* subsystem exists to receive data from NTAC's PDO (Processed Data Output) service. The PDO service supplied by NTAC is positioned by NTAC to be the primary mechanism for delivery of intercept to NTAC's agency customers.

For this purpose, NTAC makes available to its customers a specification for the service interface. The output format is referred to as MIME/CCDF, and is derived from the GDR ICD used internally within GCHQ. MIME/CCDF carries CCDF 4.1 payloads to describe meta-data.

While NTAC positions PDO as the primary service available to agencies, GCHQ's needs are such that it primarily intends to use NTAC's raw data service, where raw data is delivered to the agency directly from the point of intercept.

Our justification for using the raw service is that:

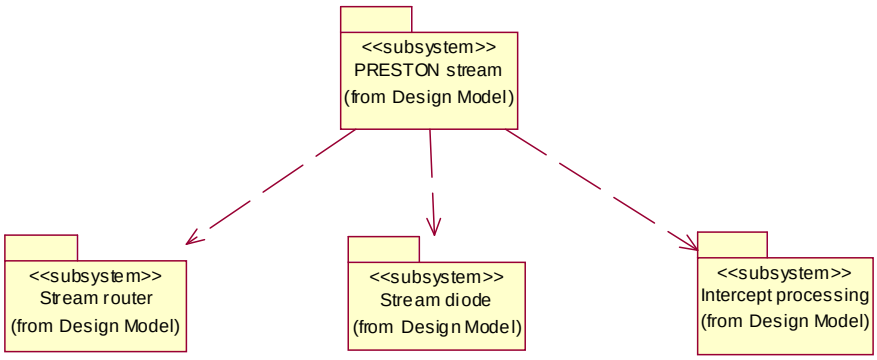
- We have more advanced processing facilities, which are necessary for us to use, since our PRESTON targets typically require more complex processing than that required for other UK targets.
- We use (broadly) the same GCHQ-developed processing equipment to that in use by NTAC, so our use of the raw service doesn't undermine NTAC's investment in our products, as the majority of the developments we produce in support of the raw processing will be made available for NTAC to use.
- Our need to analyse the data using SD techniques requires that raw data be made available to the SD community. GCHQ regularly performs collection or

survey development activities on behalf of other agencies, and this is only possible by having a dedicated processing system.

7.2 Stream Subsystem

7.2.1 Overview

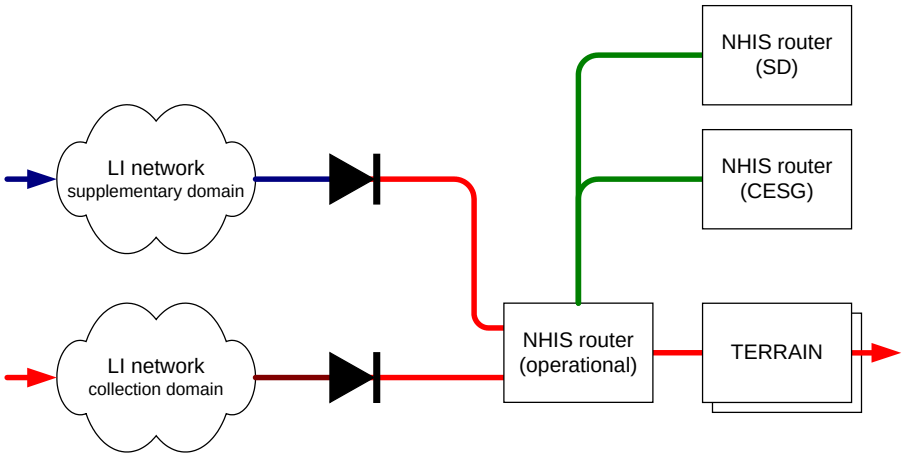
The purpose of the *stream* subsystem is to accept raw unprocessed data from the LI network Collection Domain (CD), apply processing, and deliver the processed data to *Offline Processing*.



The stream subsystem consists of three components:

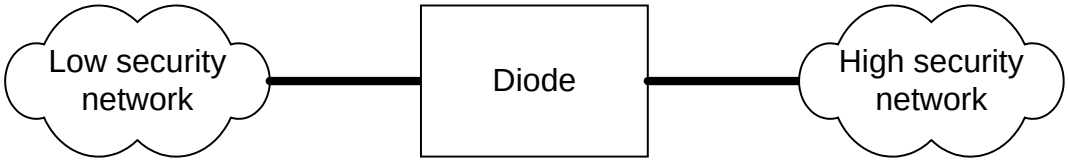
- The stream diode, which provides the security protection between the LI network, and GCHQ's internal networks.
- The stream router, which routes the LI, streams to the various processing elements.
- The intercept processing system (TERRAIN) which processes the LI stream to resulting intercept items.

The stream input diode interfaces directly with the LI network. The stream output diode interfaces directly with the NHIS router via GCNet. The two nodes are connected via an optical one-way interface allowing communication in one direction only. The one-way network is managed using the NHIS diode software developed by TPS.

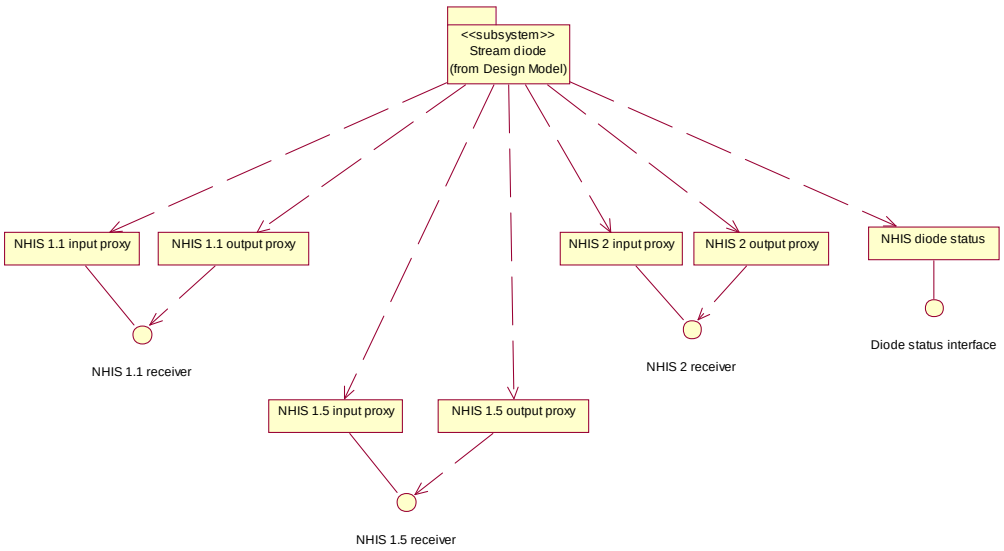


7.2.2 The stream diode

The stream diode, is a managed interface which permits NHIS intercept streams to be delivered from a lower-security to a higher-security network. Intercept is allowed to pass to the higher security network, but data may not pass in the other direction. Thus, the security of the higher security network is not compromised.

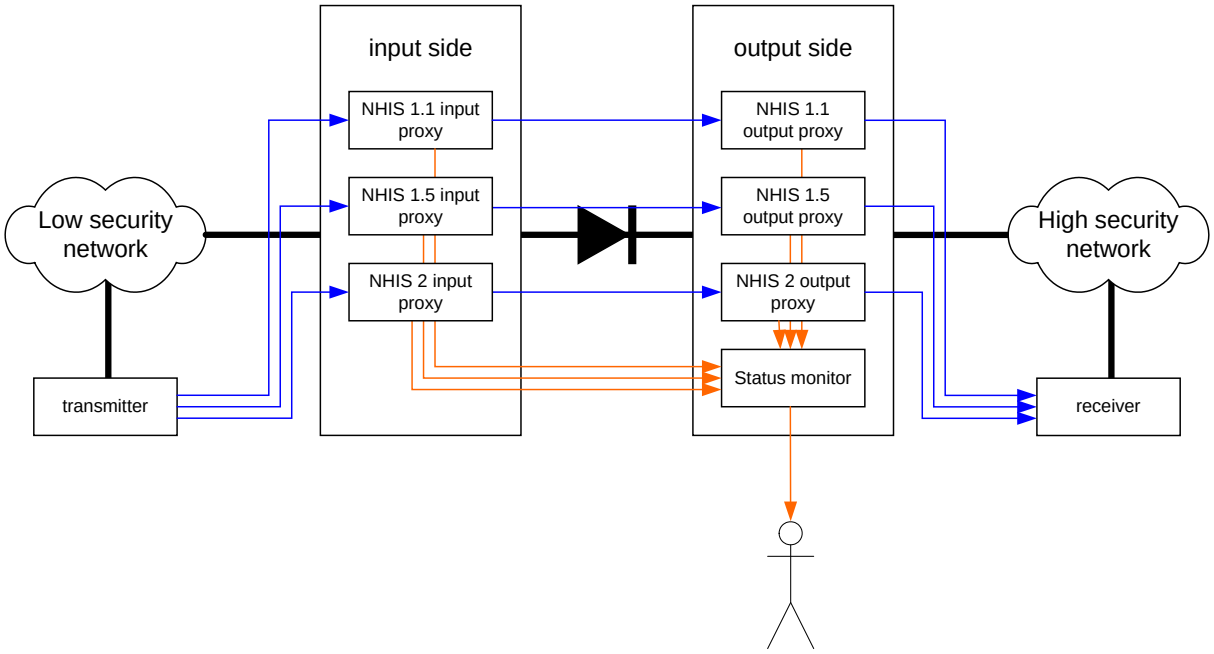


The components of the stream diode are shown below:



The stream diode software consists of a set of proxies, which allow NHIS streams to transit over a one-way optical network. There is a set of proxies (input and output) for each NHIS protocol variant. The input proxies convert the NHIS streams to a private protocol form which can pass over the one-way network. The output proxies convert the private protocol back to the correct NHIS stream format.

The input and output proxies communicate with a monitoring process which tracks the state of all processes. The status monitor can thus report on the failure of a process on the input and output sides of the diode, or the failure of the one-way optical link.

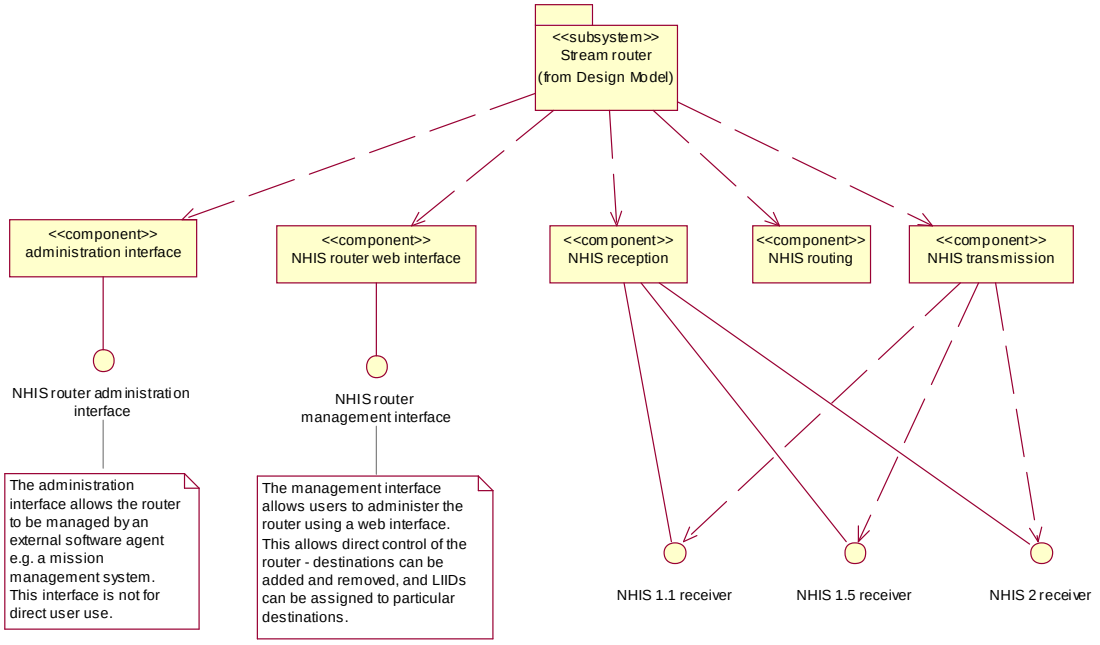


The diode terminates NHIS streams on the input side, acting as a standards-compliant NHIS receiver. The NHIS streams transit the one-way network in an internal representation of the NHIS PDUs and are reconstructed on its output. The diode must act as an NHIS compliant sender on the output side.

The status interface allows the diode to be monitored manually by operators. Alternatively, a monitoring facility such as HP Openview can be used.

The diode software is a GCHQ-developed solution called the *NHIS diode*. It is currently managed by TPS, but is being transitioned to an external contract.

7.2.3 The stream router



The stream router is implemented by the *NHIS router* product. The purpose of the router is to accept (terminate) NHIS streams on its input, and then copy the streams to one or more destinations as specified by an operator.

The NHIS router presents an NHIS interface for NHIS 1.1, NHIS 1.5 and NHIS 2.0 data. The NHIS connections are terminated on the NHIS router. New NHIS transports are used to deliver the NHIS input to destinations according to the routing rules, which are edited using a local web interface. The NHIS router is thus required to implement fully-compliant NHIS sender and receiver stacks to perform this role.

The NHIS router can be configured to offer an alternative delivery mechanism which is raw Ethernet packets to an Ethernet network. This mode of operation is not intended for use in the operational system, but can be used in SD areas to deliver NHIS-packaged data to systems which can only accept raw Ethernet input.

The stream router will be configured to deliver data to the following destinations:

- Operational TERRAIN processing at HOTLINE or Benhall.
- SD analysts in GTE survey. Two destinations will be configured – one for NHIS 1.1, one for NHIS 1.5. These destinations will be on an NHIS router deployed at GTE.
- CESG intrusion detection. Two destinations will be configured – one for NHIS 1.1, one for NHIS 1.5. These destinations will be on a network endpoint to be specified by CESG.

- Applied Research. Two destinations will be configured – one for NHIS 1.1, one for NHIS 1.5. These destinations will be on a network endpoint to be specified by CESG.

The stream router implements a buffering resource, which serves as part of the NHIS, features to replay data for failed connections. As this buffering resource is limited, it will be assigned only to the operational processing streams, so that data will not be buffered when streams to AR, GTE or CESG fail.

The stream router is implemented by the NHIS router product, which is managed and developed for GCHQ by LogicaCMG.

7.2.4 Intercept processing

The intercept processing is performed by the TERRAIN system. The TERRAIN system is delivered with an LI configuration component, which contains configuration specific to the LI environment. TERRAIN supports processing of NHIS 1.1, NHIS 1.5 and NHIS 2 intercept streams.

The TERRAIN system can integrate with a number of systems at GCHQ: FARNDALE for survey data, BLACKNIGHT for selection, GDR/CE/IIB for delivery of collection data, BRIGHTON for legacy delivery. SALAMANCA, HAUSTORIUM, SAMDYCE, SAMBOK for events.

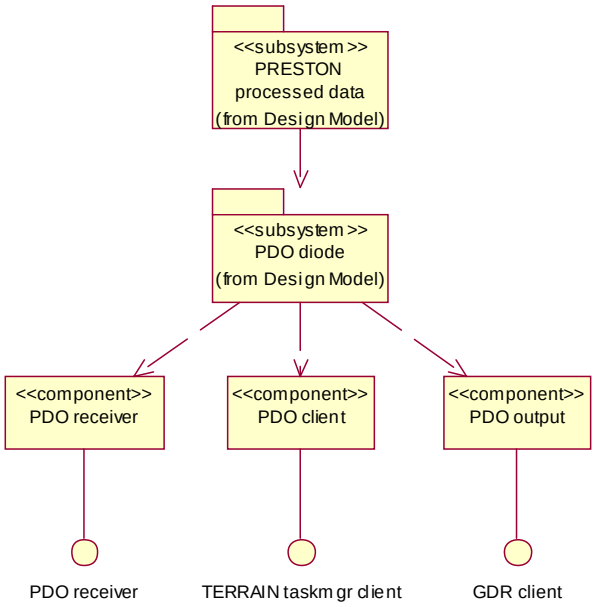
7.3 Processed data subsystem

7.3.1 Overview

The purpose of the *Processed data* subsystem is to facilitate the delivery of processed intercept from the LI network Delivery Domain (DD) to GCHQ's processing facilities.

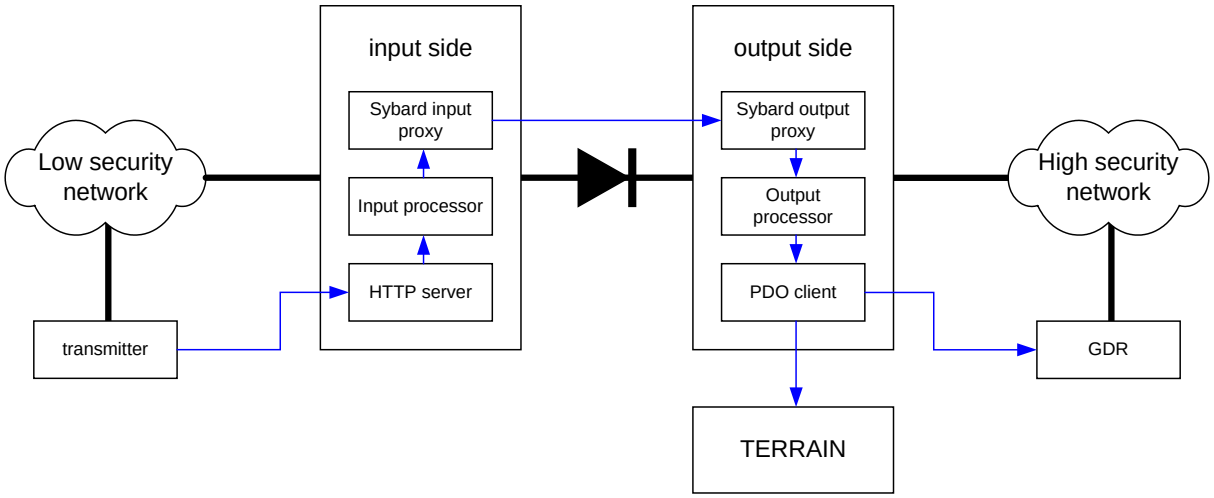
The data conforms to the GDR ICD, but the data model inside SPQR currently varies from that used in NTAC, so it is necessary for the data to be converted to conform to SPQR meta-data management.

The connection between GCHQ and NTAC must be managed, therefore it is necessary for a security barrier to protect GCHQ's networks.



The processed data subsystem logically offers three components:

- A PDO receiver accepts the data from NTAC it must offer an interface which conforms to the PDO ICD.
- A PDO client performs any alterations to the data so that it can be presented to the GDR system.
- A PDO output delivers the data to the GDR system.



The system is pictured in detail above. The system consists of two nodes connected by a one-way optical network. The one-way network is managed by the Sybard diode product which manages the delivery of files over the diode network.

The PDO reception interface is implemented using a standard HTTP server (Apache) and a simple input processor which transfers the files from the HTTP server to the Sybard diode to deliver across the network.

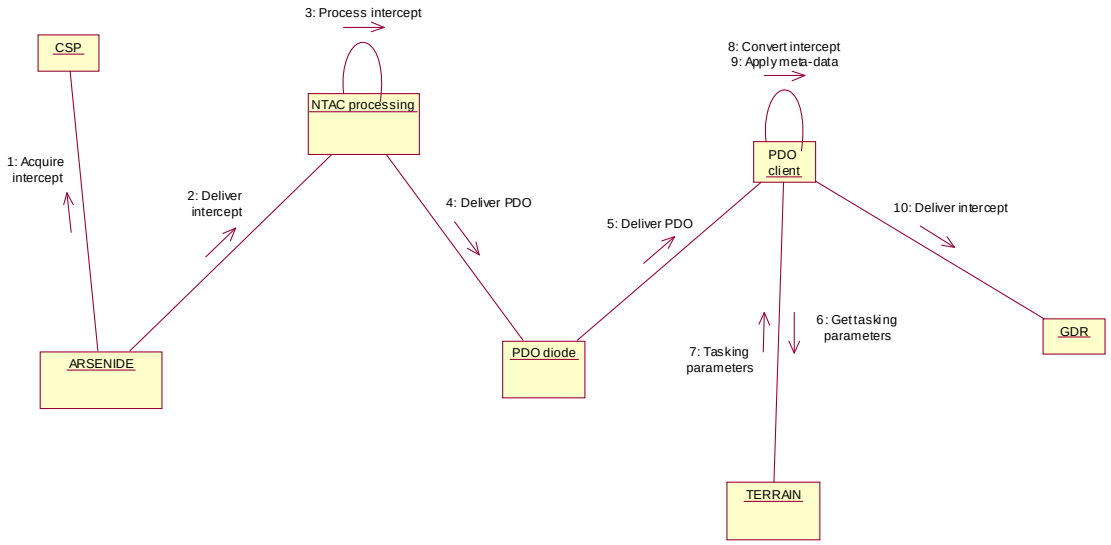
The delivery interface consists of the Output processor, which bridges the delivery from the Sybard diode to the PDO client. The PDO client performs the data conversion, and uses the tasking information from an operational TERRAIN to apply meta-data (case notation, PDDG etc.) that is needed for the data to conform to the SPQR data model.

The reformatting of meta-data, and the use of the TERRAIN tasking information is a short-term measure which can be improved by aligning the GCHQ and NTAC intercept data models.

Monitoring can be performed using HP Openview software on the output diode to monitor the Sybard software which allows the status of both sides of the diode to be ascertained.

7.3.2 PDO client

The collaboration diagram shows how the components interact to deliver the intercept items to the GDR.



7.4 Filtering and selection

7.4.1 Overview

PRESTON presents a considerable data management problem to GCHQ. On the face of it, the intercept data is well selected: a number of targets are selected, and only intercept from those targets is delivered to GCHQ. By definition, Strong Selectors are used to manage the volumes – we are only able to intercept targets for whom a strong business case (warrant application) can be made for cover, as mandated by RIPA.

The PRESTON Volumetric Model (see ref [d]) describes the impact of accepting data from our warranted intercept targets. The intercept from that number of targets is considerably more than our databases can store. We are currently not able to put on cover the full set of desired broadband targets, for a number of reasons, a big one of which is bandwidth to the databases. FARNDALE is currently overloaded, but without selection, these feeds cannot be transferred directly to the IIB, as the network path to the IIB will not support the appropriate load.

Dataflow have demonstrated that considerable bandwidth reduction can be gained by applying selection and filtering to the data streams. Two types of volume reduction are used:

- *Selection*: Further strong selectors can be applied to the LI stream to select particular targets. This can be used where the warranted entity is an organisation, and the selection of particular targets within that organisation can be particularly effective.

- *Filtering*: Filtering can be applied to remove items with low intelligence value. Such data can include SPAM, adverts, viruses, pornography and open source material which will not be of interest to us.

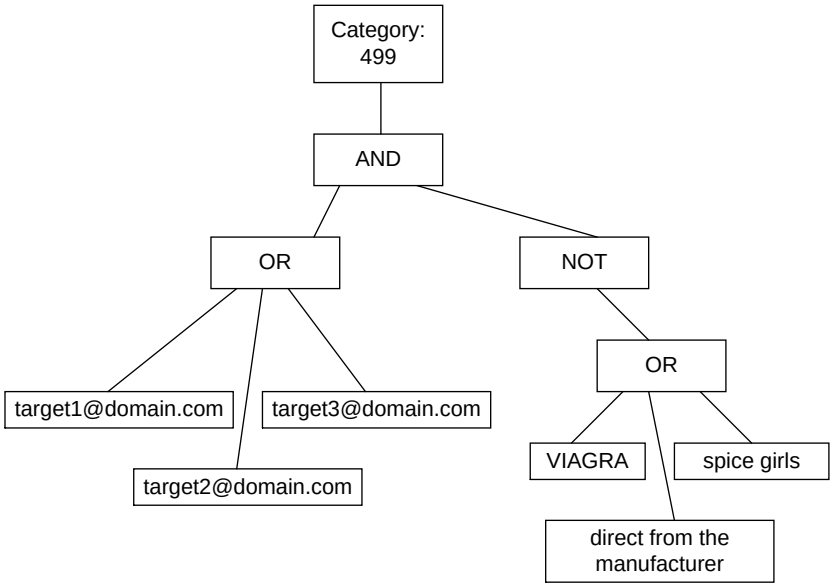
It is thus necessary to apply both *Selection* and *Filtering* to the intercept to reduce the volumes to a manageable level. The volumetric model (ref [d]) used an estimate of 95% de-selection rate to derive the data volumes to the collection repositories. Further study may be required to establish if this level is achievable.

This approach is at odds with the generic MoMo approach, where positive strong selection is the only selection mechanism, and deselection is not a supported process.

With warranted intercept, many of the filter/selector terms are derived from SD analysis of the target data during surveys – it is essential these selectors are derived before the intercept streams are transitioned to collection otherwise the data volumes cannot be managed.

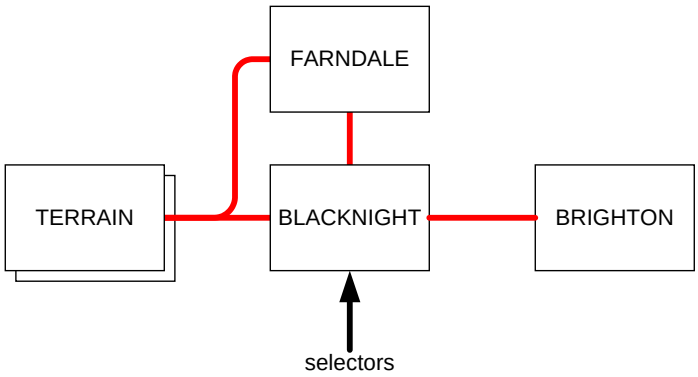
The diagram below gives a representation of a typical selector in the current PRESTON dictionary from CADENCE. Selectors are grouped and combined with defeats which are likely to detect traffic which is not of interest.

Filtering terms are typically crafted as part of development of the processing configuration, and are thus specific to one target line. In the current dictionary, there is little category re-use, as categories are crafted for each target line. It isn't clear if things would be managed when we have many more broadband targets intercepted under PRESTON.



I look at the *Filtering* and *Selection* options available.

7.4.2 Option 1: Current system



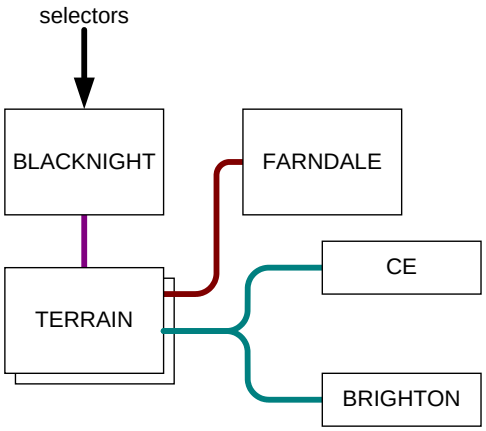
The current system uses BLACKNIGHT for selection. Survey data can be routed to either BRIGHTON or FARNDALE by the BLACKNIGHT system.

Selectors are delivered to BLACKNIGHT using the CADENCE dictionary management system using an established dictionary management process.

Pros: Established delivery path for selectors. Established business process for managing filtering.

Cons: BLACKNIGHT is no longer supported. BLACKNIGHT will not support the data rates from high-bandwidth sources (high speed DSL, BT21c). BLACKNIGHT will not integrate with the SPQR systems (GDR, CE, IIB) and a convoluted delivery path (via BRIGHTON) is needed to deliver data to IIB.

7.4.3 Option 2: Use TERRAIN to route



We have the option to use the BLACKNIGHT system in an alternative scanning mode, so that it performs selection, but does not perform routing. This has the advantage that it could be used to provide selection on data which is to be delivered

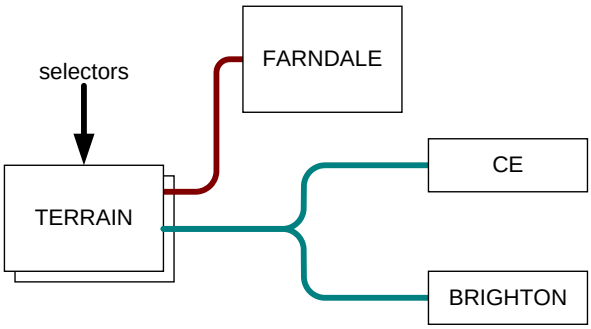
as CCDF to Content Enhancement. Data could thus be delivered to SPQR after BLACKNIGHT scanning.

This does not solve the bandwidth limitation problems with BLACKNIGHT.

Pros: Established delivery path for selectors. Established business process for managing filtering. Sane delivery path to SPQR systems.

Cons: BLACKNIGHT is no longer supported. BLACKNIGHT will not support the data rates from high-bandwidth sources (high speed DSL, BT21c).

7.4.4 Option 3: Use TERRAIN for filtering and selection



TERRAIN is able to apply boolean selectors to intercept for selection and filtering purposes. The selection engine, FAST GROK is able to apply strong selection and filtering in a manner which is similar to the BLACKNIGHT functionality (see ref [e]). FAST GROK is a low cost scanning algorithm. The FAST GROK engine was developed as part of TERRAIN 9 to replace use of the 6 selection engines in use, and the plan is to ensure that this engine deprecates all other use.

The FAST GROK engine will happily work with any of the selector sets (TACHO, CORINTH, TRAFFIC MASTER) which TERRAIN receives, and thus it replaces the 6 of selection engines in TERRAIN with a single high-speed dictionary which meets all requirements. The algorithm can optionally use a dictionary format (FAST GROK dictionary) which has never been exposed outside of TERRAIN.

The FAST GROK dictionary format is semantically, nearly equivalent to the BLACKNIGHT Netlayer dictionary format, and thus a conversion from Netlayer to FAST GROK may be achievable.

It is possible to have TRAFFIC MASTER distribute FAST GROK dictionaries, although work would be needed on TRAFFIC MASTER to achieve this.

Pros: Selection and filtering will work at required rates. Sane delivery path to SPQR systems.

Cons: No established delivery mechanism for management of the dictionary.

7.4.5 Selected approach

There is no clear solution to the Filtering and Selection problem. The use of BLACKNIGHT cannot continue in PRESTON because of the increase in data volumes which BT21c promises, and the increase in target coverage which CSIP require. The lack of support for BLACKNIGHT is a major problem with us continuing to use the system.

None of the potential replacement systems (KEYCARD, X-KEYSCORE, COURIERSKILL) can be used to integrate with SPQR as CCDF is not on NSA's roadmap for these systems.

My recommendation is that a selector distribution mechanism be used for strong selectors e.g. MONKEY PUZZLE/TACHO.

My recommendation is that FAST GROK dictionaries are maintained on TERRAIN for target-specific filtering terms, if they are required. This should be considered an initial response to the increasing rates.

Further study work is needed to ascertain whether an appropriate selector management system is needed for filter terms, and what support (e.g. TRAFFIC MASTER management) needs to be created.

7.5 Events

The stream subsystem will be able to use TERRAIN to integrate with SAMBOK, SALAMANCA and HAUSTORIUM to deliver events.

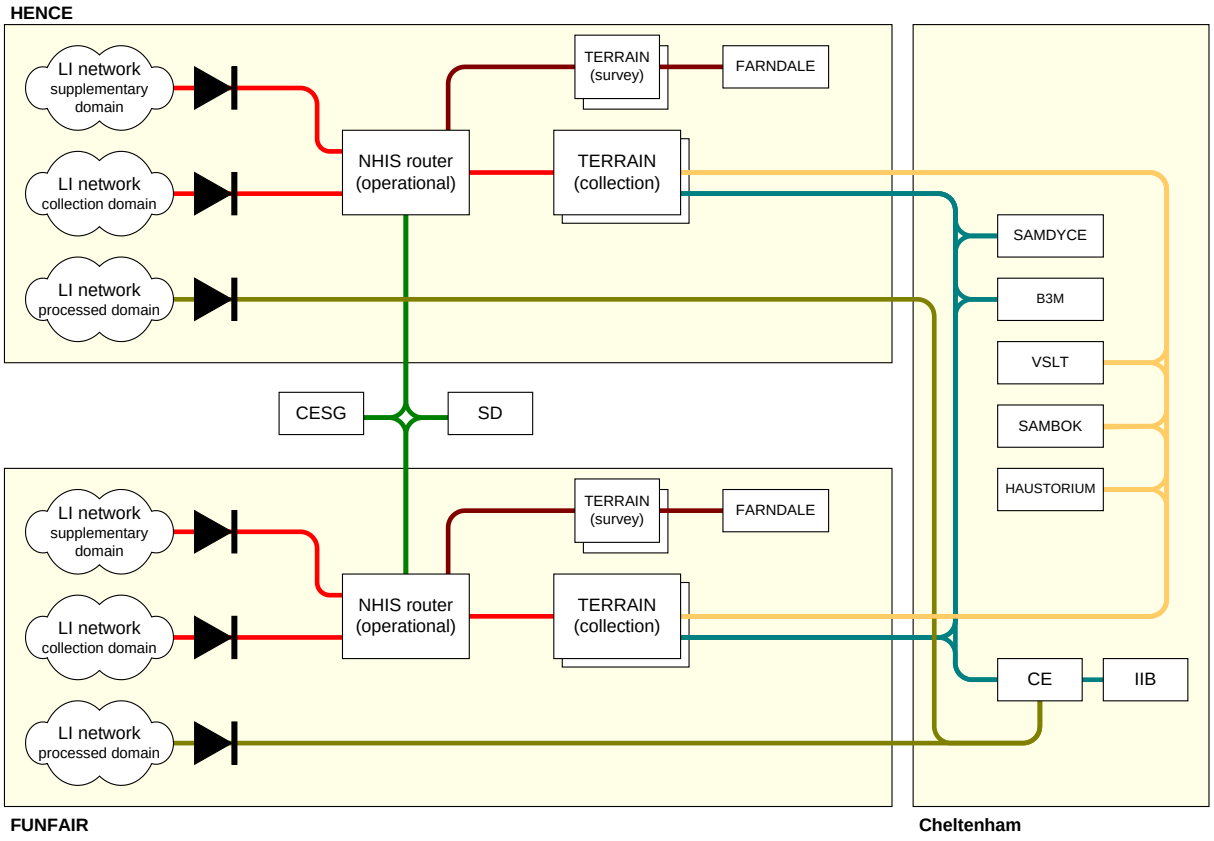
There is no offline event extraction facility, and there is no intention to build one. We will thus not be able to extract events from the data delivered on the Processed domain unless NTAC undertake to provide an events capability in the processing centre.

8 LOGICAL VIEW: OFFLINE PROCESSING AND STORAGE

8.1 Offline processing

I return to the diagram presented in section 5 to illustrate dataflows. This architecture intends to see all collection C2C delivered to Content Enhancement for delivery to the IIB. Survey data is intended to be delivered to FARNDALE, as at present.

The filtering approach is required in order to be able to manage the volumes which are derived from the considerable input to the processing.



8.2 Dataflows

The following dataflows exist for processed data:

Source	Destination	Interface	Description	Estimated volumes per day
Processed domain	PDO client	NTAC PDO	Delivery of processed data from NTAC.	200MB, 1500 items
PDO client	CE	CCDF/MIME (GDR ICD)	For offline processing.	200MB, 1500 items
TERRAIN	FARNDALE	CCDF/MIME (GDR ICD)	Survey data.	4.2GB, 40 000 items
TERRAIN	CE	CCDF/MIME (GDR ICD)	Processed intercept for offline processing.	2.6GB, 27 000 items
CE	IIB	GDR baton	Intercept for storage.	2.8GB, 28 500 items
TERRAIN	SAMBOK	SAMBOK ICD	Target location events.	
TERRAIN	SAMDYCE	SAMDYCE ICD	SMS content.	
TERRAIN	B3M	B3M ICD	Voice content.	15GB, 2500 items
TERRAIN	VSLT	SALAMANCA ICD	Telephony events.	
TERRAIN	HAUSTORIUM	PILBEAM ICD	C2C events.	

The volume estimates are derived from the volumetric model.

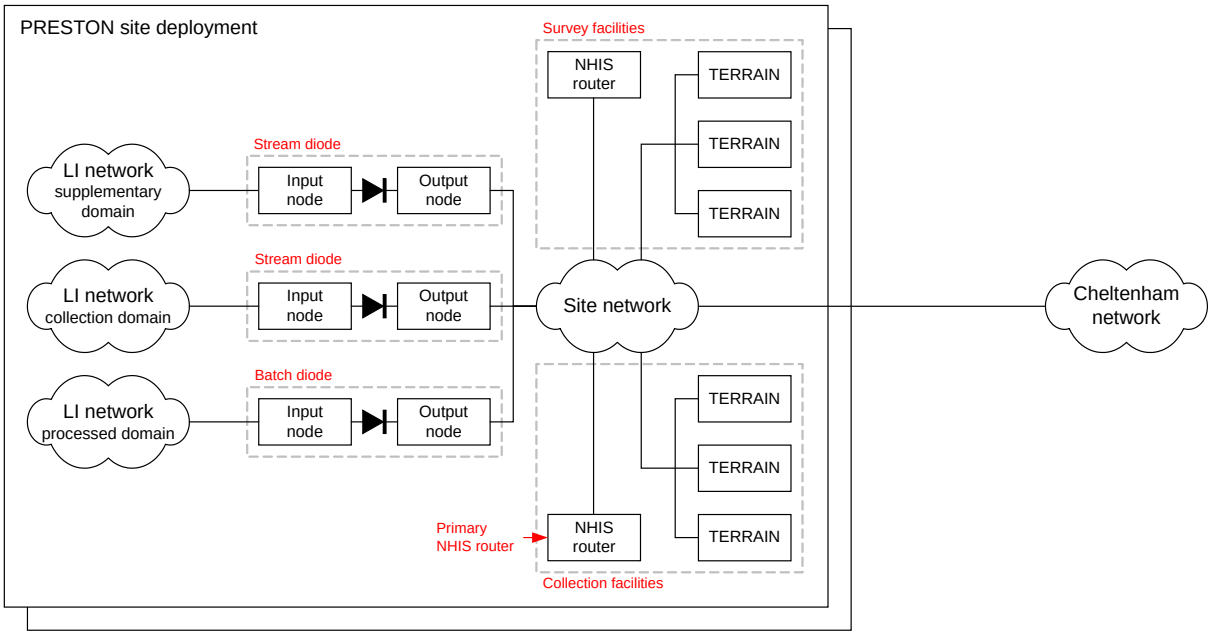
The volumetric volume has no information on events or SMS. Volumes for SMS content and Telephony events are not expected to change significantly from what they are now, although analysing this information remains important to understand the impact from the business on deploying this architecture.

9 DEPLOYMENT VIEW: PROCESSING SYSTEM

9.1 Deployment overview

The deployment consists of a number of site deployments, each of which is capable of interfacing with the NTAC networks, and delivering data to the Cheltenham network repositories.

The deployment is pictured below:



The current proposal is to deploy two sites.

The site deployment consists of:

- a stream diode for each LI network stream delivery. We currently plan for two stream network connections for the Supplementary domain, and the Collection domain.
- a batch diode for each LI network processed data delivery. We currently plan for one for the Processed domain.
- a collection facility consisting of a single NHIS router, and a number of TERRAINs.
- a survey facility consisting of a single NHIS router, and a number of TERRAINs.

The PRESTON systems could use TERRAIN capability from existing deployments. There is an advantage of deploying specific PRESTON TERRAIN systems for processing in collection and survey, as this allows for more agile deployment of PRESTON-specific features in the future.

Also, there may be a requirement in the future to provide for stricter auditing controls over PRESTON intercept as a result of changes to UK legislation. Separate PRESTON systems allow for the easier introduction of PRESTON-specific auditing requirements.

9.2 Deployment details

At each site, it is proposed to deploy 10 TERRAIN systems for collection purposes, and 5 TERRAIN systems for survey purposes.

Purpose	System	Hardware	Software
Supplementary domain	Network connection	100 Mb/s network connection	
	Stream diode	NHIS diode hardware (2 servers)	NHIS diode
Collection domain	Network connection	100 Mb/s network connection	
		NHIS diode hardware (2 servers)	NHIS diode
Processed domain	Network connection	100 Mb/s network connection	
	Batch diode	2 servers	Sybard software, PDO client
Operational collection	NHIS router	1 server	NHIS router
	TERRAIN	10 servers	TERRAIN
Survey	NHIS router	1 server	NHIS router
	TERRAIN	5 servers	TERRAIN

10 SIZE AND PERFORMANCE

10.1 LI network

Returning to the values from the volumetric model:

Totals	**Long-term average bandwidth:**	**6.3 Mb/s**
	Max bandwidth:	**71.7 Mb/s**
Fractions	% of bandwidth which is broadband:	50.1 %
	% of bandwidth which is voice:	48.8 %
	% of bandwidth which is mobile:	1.1 %

It is clear that the delivery network must be able to sustain burst loads of around 75Mb/s. This load is expected to be delivered across the LI network's Collection Domain and Supplementary Domain. It is expected that the load will be roughly equally spread across Collection Domain (for broadband) and the Supplementary Domain (for voice).

To deliver this architecture, a number of features must apply to the end-to-end system, including the LI network. The following features cannot be delivered unless there is support in the LI network:

FEAT105	The MTRR of the operational LI processing service shall be less than 24 hours.	Highly Desirable	NEED7
FEAT106	The system shall ensure the integrity of data so that data loss is no worse than 0.05% by volume.	Desirable	NEED7
FEAT12	The system shall support processing throughput up to 34 Mb/s.	Essential	NEED7
FEAT103	The system shall support processing throughput up to 100 Mb/s.	Highly Desirable	NEED7

10.2 Volume management

The components of the system manage volumes thus:

Feature	Benefit
NHIS router	LI streams can be delivered to multiple processing systems to share the load.
NHIS router	Manages delivery of LI streams so that only wanted streams are delivered to e.g. CESG.

TERRAIN	Multiple TERRAIN systems can be deployed to ensure that the TERRAIN system is not a bottleneck.
FARNDALE	Provides a local repository for survey data is not a burden on the delivery networks or the collection systems.
TERRAIN	Applies selection so that volumes are considerably reduced presentation to Content Enhancement. This is particularly important since that delivery route may traverse a WAN link.

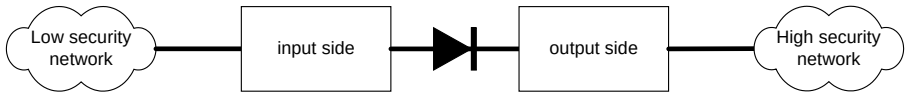
10.3 Diode availability

The availability requirements present us with a problem:

FEAT106	The system shall ensure the integrity of data so that data loss is no worse than 0.05% by volume.	Desirable	NEED7

This requirement can easily be met by most of the components specified in the architecture, and all delivery mechanisms (including NHIS) support reliable delivery of data.

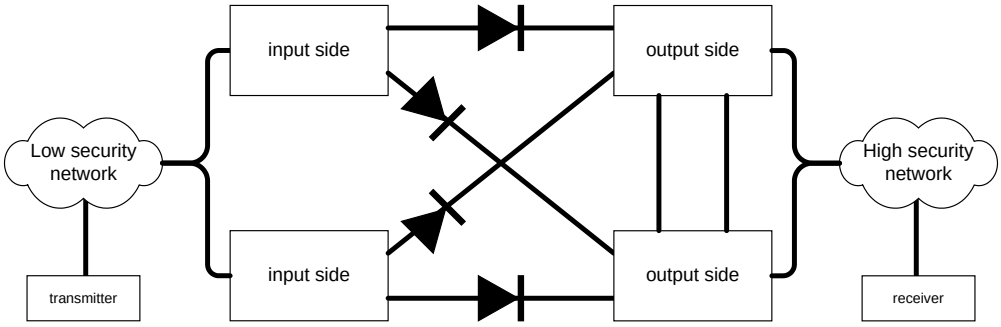
However, the use of data diodes presents a problem here. As the input side can only deliver to the output side, if there is a failure of the one-way link, or the output side, there is no way for the input side to determine that there is a failure and buffer data.



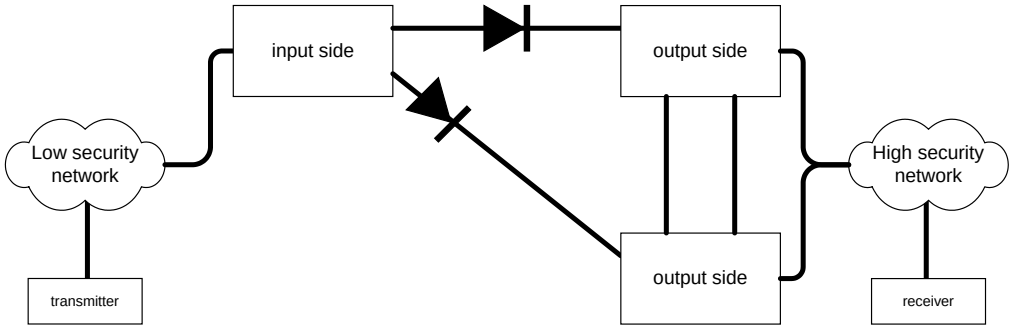
This failure mode is different from others, as while the diode is in this mode, data is lost forever. Contrast this with failure of e.g. the LI network, where the CSP will buffer data and re-deliver when the LI network returns.

Although the failure of the one-way link or the output side will likely be detectable by management systems, there is still a high likelihood that the standard support response (in the region of a few hours) be too long. Schemes have been proposed to deal with this data loss problem, such as, allowing something on the high security side to switch off the input side.

Other possibilities include constructing a highly available diode with redundant components:



The increased resilience of this configuration would considerably reduce the risk of failure, and thus reduce data loss. A fully redundant diode may be expensive: a cheaper configuration might be introduced by removing the redundancy of the input side:



Failure of the diode input side does not result in data loss for NHIS delivery, as the NHIS protocol includes features which provide resilience against data loss.

The design of a highly available data diode is beyond the scope of this document, and is a matter which requires further study. This topic is something which GREENHEART phase 3 may accept as a study task.

11 QUALITY

Maintaining the quality of NHIS processing products is a challenge. NHIS protocols are intercept protocols. Although NHIS 1.5 and 2.0 are based on ETSI standard protocols, LI protocols are still a “niche” market for protocols.

It is thus important for the engineering community to invest in test and analysis products which can be used to test and diagnose problems with NHIS delivery.

Two key products are:

- NHIS emulators which are able to present an NHIS receiver or sender interface.
- NHIS analysers which can passively report on the behaviour of observed NHIS communications.

GCHQ maintains an emulator product under the GREENHEART contract. We should consider how to engage with the LI community on acquiring an analyser product.

We may consider whether the open source product Wireshark may well make a good base for such a product. Addition of NHIS plug-ins to Wireshark would be able to harness Wireshark’s analysis capabilities, so that a complete protocol stack e.g. VoIP in NHIS could be analysed.