Tab 27

LME Infrastructure Solution Design
1095 Select7

Creation Date	*17th May 2010*
Last Updated	*2nd September 2011*
Version	*2.0*
SAP Reference	*IFMMETA001GW*
Requirements	n/a

Review	**Architecture:**	Ian Hersey	√
	Unix:	Douglas Furlong	√
	Application:	Tony Sims	√
	Network:	Adrian Jenkinson	√
	Implementation:	Shane Baker	√
Sign Off	**Lead TA:**	David Spencer-Nixon	*dd/mm/yy*
	Production (app):	Philip Tagg	*dd/mm/yy*
	Production (infra):	Nick Thomas	*dd/mm/yy*

1 Table of Contents

1.1 Diagrams

1.2 Figures

Document Location

The released copy of this document can be found in the document repository at

\\lmefp1\it\Document Repository\Projects\SELECT 7\3. Design

This design also references inter-related projects that are out of scope of this design.

- DC Link

\\lmefp1\it\Document Repository\Projects\DC Real Time Replication Network\3. Design

- Magnifix

\\lmefp1\it\Document Repository\Projects\MagniFIX\3. Design

- Enterprise Monitoring

\\lmefp1\it\Document Repository\Projects\ Enterprise Monitoring\3. Design\3.2-High Level Design

2 Revision History

Version	Update Date	Summary Of Changes	Updated By
0.1	29/06/2011	Initial Version	David Spencer-Nixon
0.2	05/08/2010	Added UNIX requirements	James Barnard
0.3	11/08/2010	Added Network Detail	David Spencer-Nixon
0.4	17/08/2010	Added Monitoring Information	David Spencer-Nixon
0.5	09/09/2010	Updated Unix detail & clarified some text in response to Cinnober feedback	David Spencer-Nixon
1.0	12/10/2010	Document Sign Off	Terry Fenn
1.1	09/11/2010	Remove change bars & correct some graphics text	David Spencer-Nixon
1.2	22/11/2010	General typo corrections. Removal of references to FE servers for clarity, now all “application”	David Spencer-Nixon

1.3	09/02/2011	Diagram update & clarified detail on low level flows	David Spencer-Nixon
1.4	14/06/20111	Update for hardware refresh, and aligned IP addresses	David Spencer-Nixon
1.5	29/06/2011	Updated to incorporate addition of SelectMD	David Spencer-Nixon
1.6	11/08/2011	Feedback	David Spencer-Nixon
2.0	02/09/2011	No further feedback so released	David Spencer-Nixon

3 Introduction

3.1 Business Requirement

The LME are looking to upgrade their TRADExpress Select platform. This is the first upgrade the LME are performing with Xchanging with a view to driving improvements and taking advantage of newer code, rather than out of pure necessity.

The Select platform performs the Trading function for the electronic side of the Metal Exchange (as opposed to the Voice and Ring trades). This is therefore a critical component of the infrastructure for the LME.

The Select platform is based on the TRADExpress solution from Cinnober, with specific customisations made by the vendor to meet LME needs.

3.2 Project Objectives

- Upgrade of Select to leverage TradeExpress8. (otherwise known as Select 7)
- Replacement of communication paths to/from Select from using the XI+ protocol to using FIX
- Improvement in resilience and recovery over that available in Select 6

3.3 Scope

High level project scope includes:

- Implementation of the Select 7 application
- Re-Configuration of the network to support Select 7
    - Allowance for users running FIX directly to the FS servers rather than via TAX
    - Provision of Unicast as well as Multicast communications LANs

- Re-Configuration of the servers to support Select 7
  - Upgrade to current standard OS provision
  - Hardening of boxes to meeting current security recommendation
  - Upgrade of JDK, MySQL server, MySQL client packages
- Security testing & verification

3.4 Purpose of This Document

This document provides the high level design required to implement the deliverables as listed in the main body of requirements and proposed design, aligned with the 'in scope' section of □

3.5 Intended Audience

This document is intended to be used by Implementation and Data Centre staff as a reference for when they deliver the project. It should also form an agreement between the Project Manager and the technical resources on what is to be provisioned.

4 XTS Infrastructure Principles

For each design produced common or shared infrastructure and/or software will be used in line with current XTS Standards Framework unless otherwise stated.

Table 2 below contains specific project information at a glance. This should enable the delivery resource to proceed with the Implementation knowing environment locations, storage type, and connectivity requirements etcetera.

Specific build information for each technology team can be found in Section 0 of this document. When using Hyperlinks please note careful attention to build versions should be adhered to.

Within the Standards Framework a COMMON folder exists. COMMON contains Naming Services such as NTP, DNS and hostname standards deployed within Xchanging which are used by all Technology teams hence the Common folder name.

5 General Overview

Select 6 Production is provisioned in Basildon and Slough with the ability of one site to continue should the other site fail once certain actions are performed by the Administrators.

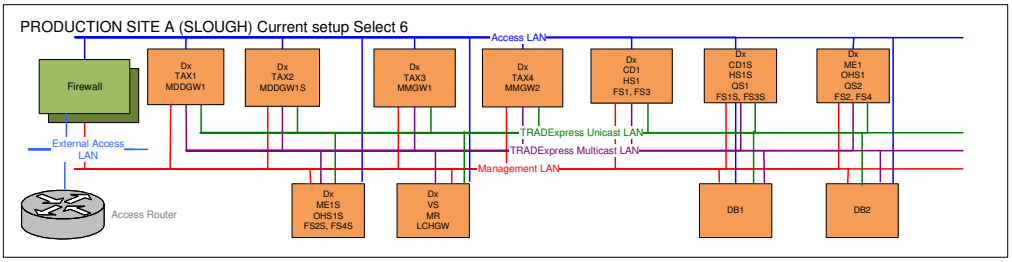


Diagram 1: Select 6 (existing layout)

Failover between sites is achieved through a manual process of shutting down the current active site, transferring the data and bringing up the current inactive site. This is an extended process, during which the market will be unable to trade.

In the event of an unplanned failover event (e.g. loss of the current active data-centre) the trades performed during that working day will be “lost” from a Select perspective, with the resultant impact on LME workloads to recover market to a known point.Select 7 will be built to better handle component failures and improve on the current manual intervention between Data Centres

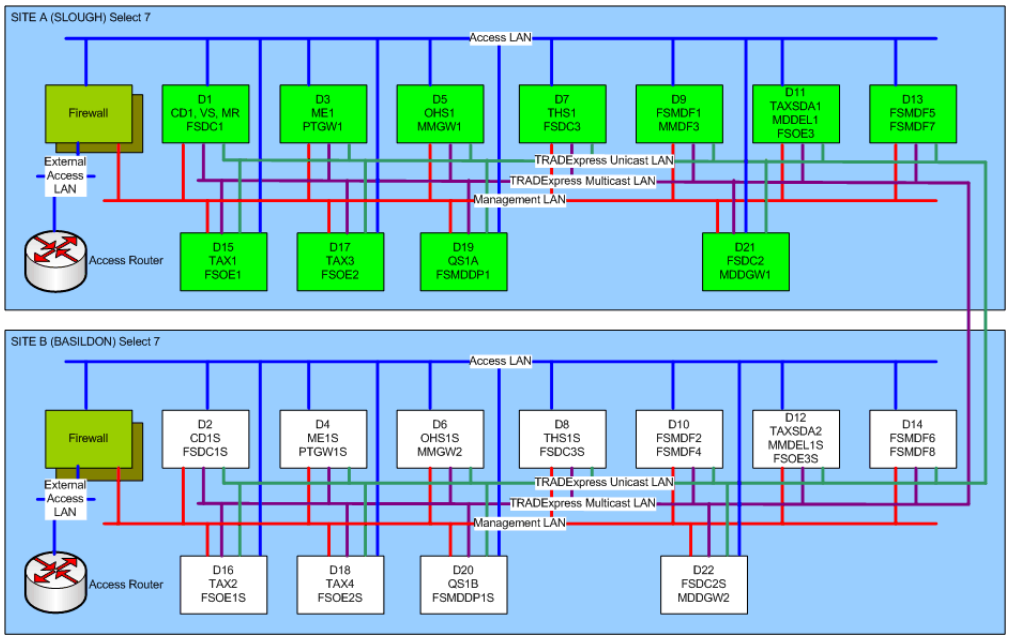


Diagram 2: Select 7 (proposed layout)

Under normal operation both the data centres will be active, however with a nominated primary data-centre carrying the primary machines providing service. A secondary data-centre containing identical Secondary (or redundant) machines and services that are able to run in the event of a component (or greater) failure at the primary site.

In the event of a single machine failure, the appropriate secondary systems will be able to run at the remote location. However, in the case of certain key server failure e.g. to FS, the complete range of FS servers will need to failed in order to ensure consistency of performance across all the Members (a fundamental obligation of the LME).

This set-up however has the benefit that in the event of an unplanned failover due to loss of the current active site, the second site is “in-sync” and up to date, and therefore able to carry on trading with minimal impact.

This does however carry a requirement for:

1. minimal latency link between the matching engine (ME) which is constrained by the current distance between the data-centres (noted risk)
2. the client to support and follow the failover as appropriate (noted risk)

Alternatives discussed, however removed after further discussion with the LME were

a) Similar provision to the current Select 6 installation – e.g. two complete installations, however this removes the benefits of improved failover in this proposal.

b) Migration of one of the DCs to remove the constraints of distance on latency – somewhat cost prohibitive!

c) Provision of an Emergency Matching Engine (EME) on the second site, with the normal secondary servers remaining on the active site. However whilst this is a middle ground, this still results in an period of unknown market status after a failure due to the EME lagging behind the ME(s).

Select7 will now incorporate the MDD data feed into the solution, due to the expected increase in MDD volumes potentially exceeding the capacity of the current MDD solution.

MDD data will be collected from

- Select
- SMART via IGW
- MOIC via MDDEL
- MIQ via MDDEL
- LCH via IGW
- LMEprice via IGW
- CSS via IGW
- SWORD via IGW

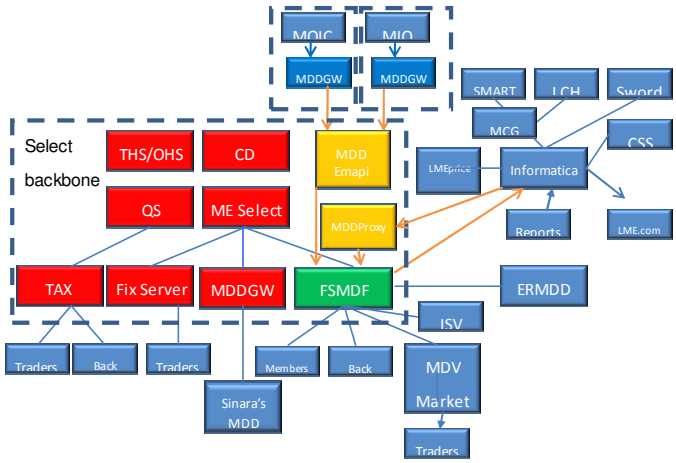


Diagram 3: Proposed MDD System Overview

6 Unix Design

6.1 Unix Requirements

- Provision of a new standards based build as far as feasible.
- Provision of build to effectively meet Cinnober requirements.
- Provision of additional network interfaces to support Unicast as well as Multicast LAN requirement.
- Provision of upgraded packages to support the Select 7 application as outlined by Cinnober.

6.2 Proposed Unix Solution

6.2.1 Hardware

The new Select 7 machines will be a new hardware provision, in order to handle the additional requirements upon select, and to better support future growth.

These new machines will be of the following specification

Part No.	Description	Qty.
583917-B21	HP DL380G7 LFF CTO Chassis	1
587498-L21	HP X5680 DL380G7 FIO Kit	1
587498-B21	HP X5680 DL380G7 Kit	1

500662-B21	HP 8GB 2Rx4 PC3-10600R-9 Kit	6
516814-B21	HP 300GB 6G SAS 15K 3.5in Dp ENT HDD	4
534562-B21	HP 1G Flash Backed Cache	1
512485-B21	ProLiant Essentials Integrated Lights-Out Advanced Pack	1
512327-B21	HP 750W CS HE Power Supply Kit	2

Figure 1: Server Model Specification

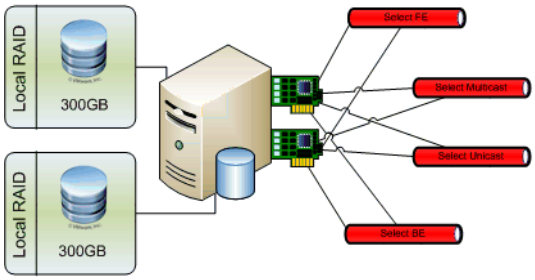
The four hardrives will be paired of in a RAID configuration. This leaves two 300GB logical drives for use. (600GB usable from 1.2TB actual)

There will be 11 machines provide into each environment. All machines will be of the same specification.. This will permit complete flexibility in the allocation of services across the machines should either of the following arise:

- Re-provision of services to support changing load-profiles for best performance
- Re-provision of services in event of machine failure, should work beyond expected recovery scenarios arise.

The same hardware (11 machines per environment, 22 in total) will also provisioned into the Pre-Production environment (logically pre-production 1 and pre-production 2) supporting a like-live pre-production requirement.

The following machines will be required in total



Name	QTY	Description
CPU	2	Six-core Xeon X5680 3.33GHz
RAM	48GB	RAM
NIC	6	Gigabit NIC, 2xO/B + 4xPCI
DVD	1	DVD ROM
Storage	Sys	2 x 300GB RAID
	Data	2 x 300GB RAID
Software	OS	RHEL 5.x x64
	Other	JRE 1.6.0-2
		MySQL client 5.1.5047.1
		HTTPD 2.2.3-22

Diagram 4: App Server (e.g.)

Site	*Environment*	*No. Physical*	*CPUs per Phy*	*RAM per Phy*	*O/S*	*Disks*
BAS	Prod1	11	2 x 6 core	48	RHEL 5.x x64	2 x 300GB 2 x 300GB
SLO	Prod2	11	2 x	48	RHEL 5.x x64	2 x 300GB

			6 core			2 x 300GB
BAS	Pre-Prod1	11	2 x 6 core	48	RHEL 5.x x64	2 x 300GB 2 x 300GB
BAS	Pre-Prod2	11	2 x 6 core	48	RHEL 5.x x64	2 x 300GB 2 x 300GB

Figure 2: Select Servers To Be Provisioned

6.2.2 Operating System

Cinnober recommendations are for a 64bit version of the operating system to be used across the board. RHEL 5 x64 will be provisioned on the machines. In the first instance this will follow the base (minimal) XTS standard installation. Above this installation, the additional packages as specified in each section below have been requested by Cinnober. Further detail on Cinnober build recommendations are available in a separate document in the same location as this design, this document therefore just summarises these as appropriate below.

The Kickstart servers will be used to install each server. A separate Kickstart configuration file, based on the XTS standards but tailored for each server, will be created to allow quick re-installation if the need arises.

With the exception of anything that impacts specific Cinnober requirements (as listed below), all available patches and security updates from RedHat should be applied after the OS installation.

Logical Volume Manager (LVM) will be used for file system volume management, allowing dynamic disc space allocations to the OS file systems as and when required. The standard XTS OS layout within Volume Group (VG) 0, which will exist on the first RAID pair of 300GB drives. The minor exception being that `/tmp` will be expanded to 10GB

The remaining space on this first pair of drives will be left for future flexibility and production maintenance support.

The second 300GB pair, will be provisioned as VG1 and Logical Volumes (LVs) assigned according to the server function listed in the next section.

The following OS modifications are recommended for all machines. The kernel parameters must tally with the Select configuration for optimal performance.

- In `/etc/sysctl.conf`
    - Change the ephemeral port range
    - Increase the UDP core buffers

- Increate the TCP core buffers

```
#allowed local port range
net.ipv4.ip_local_port_range = 49152 65535

#increase UDP buffers
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216

#increase TCP buffers
net.ipc4.tcp_wmem = 16777216
net.ipc4.tcp_rmem = 16777216
```

- Increase the MAX open files
  - In `/etc/security/limits.conf`

```
<select user> hard nofile 32768
```

  - In `/home/<select_user>/.profile`

```
Ulimit -n 32768 > /dev/null 2>&1
```

- Remove limit for core file sizes in `/home/<select_user>/.profile`

```
ulimit -c unlimited > /dev/null 2>&1
```

In a move from the current Select6 installation, all servers are “application servers” and equally specified. This allows full future flexibility of migration tasks between machines if Cinnober recommend this for performance improvement.

Additional packages (beyond standard build) required on the application servers will be:

- Inclusion of updated Java JRE version 1.6.0 Update 21
- MySQL connectors jar (e.g `mysql-connector-java-5.1.10.jar`)
- Installation of updated MySQL client to version 5.1.5047.1
- Installation of X-server in order to run SOPS and TxLog viewer from the servers
- Installation of a Web-Browser in order to access SOPS (and associated help-files) locally as an

alternative to client based access.

Comment [d1]: Queried by Douglas, passed to Cinnober.

- Installation of Apache to version 2.2.3-22
- Installation of updated MySQL server to version 5.1.50.1
- Installation of updated MySQL client to version 5.1.5047.1

The first 300GB drive will configured with the following, in addition to the standard layout

- A 1GB LV mounted into /home/selectadm (this is to assure this system user has working space)
- A 50GB LV mounted into /var/lib/mysql

The second 300GB drive will be configured to provide:

- A 200GB LV mounted into `/opt/select/`

6.2.3 User Accounts

Individual user accounts will be created with the expectation that they will be maintained through an LDAP directory services solution. Due to provisioning timescales, so as not to impact this project negatively, the pre-production machines will initially have individual user accounts, however configured such that the LDAP server can be plugged in relatively simply.

Individual users will be expected to login with their own credentials, and then use SUDO as necessary to access privileged commands that need to be run as another user (e.g. “selectadm”). SU should only be used where absolutely necessary by the system administrators, and this should not be necessary in the vast majority of cases

6.2.4 NIC Configuration

For the new Select provision, an additional network will be required when compared to the existing solution.

The new Select solution will have to support 4 VLANS, from 3 bonded pairs of NICs. (each member of the bonded pair being provisioned to alternate switches as per standard design). Access to the fourth VLAN can be achieved by extending the standard configuration to use VLAN tagging on the servers themselves.

- Bond0 has been specified to be the Application LAN for all systems
- Bond1 has been specified to be the Multicast LAN for the Application servers
- Bond2 has been specified to be the Management LAN for all systems
- Bond3 has been specified to be the Unicast LAN and is created using Virtual NICs sharing the Multicast physical connections (bond1).

6.2.5 DNS

It is expected that the server hostnames will be manually populated into the DNS system through the AD; this will need to be provision with both the A & PTR records.

Customer access to `select.lme.com` will be handled through the public DNS – this is expected to move onto the F5 GTM (dynamic DNS) solution, however this is a dependency upon the DCLink project provision and out of scope of this document.

7 Database Design

The database design is part of the Select set-up, and therefore the layout of this is handled as part of the Cinnober works.

It should be noted that the database provision will be installed into its own partition in order to ensure the integrity of the data and that the system is not impacted by (or causes impact to) any other part of the server. This partition will be mounted into `/var/lib/mysql` in order to ensure a standard location from the MySQL perspective.

For reference, the following layout is expected across the two database servers

Database Function	Database Schemas
Primary	LMESELECT_CD / LMESELECT_THS / LMESELECT_MR / LMESELECT_OHS
Secondary	LMESELECT_CDS / LMESELECT_THSS / LMESELECT_OHSS

8 Storage Design

n/a all storage will be local to the servers.

9 Backup Design

9.1 Backup Requirements

- Effective backup of all data partitions
- Effective backup of database data.

9.2 Backup Solution

NetBackup should be installed on all the machines. However as the machines can be quickly rebuilt using

effective KickStart configuration files, only the data partitions `/opt/select` need be backed up directly. Not including the system partitions results in an overall saving of >600GB per environment per backup cycle.

For the database servers, a cron job will need to ensure a dump of the data in the database on a regular basis *ahead* of the backup cycle occurring. The backup function will include this file, therefore ensuring a state-consistent DB backup exists.

There is the expectation that these areas will have nightly incremental backups, with a full weekly backup

Server	SAN	Internal	OS	Volumes	Service	Notes
Prod App	n/a	200GB	RHEL 5.x	/opt/select	Bronze	Start after 7pm
Prod DB	n/a	50GB	RHEL 5.x	/var/lib/mysql	Bronze	Start after 7pm and start after mysqldump

Please note that there is no backup provision in the pre-production or functional test environments at this time, therefore the above configuration will be applied to production only.

10 Network Design

10.1 Network Requirements

- Provision of new Select 7 environments
- Provision of additional Unicast LAN
- Provision of site-site communications
- Support of direct communication to the FIX servers.

10.2 Proposed Network Solution

10.2.1 General Overview

[Note the following discusses FIX and TAX servers, however in practice they are all APP servers for future flexibility. The TAX and FIX references certain functions that an APP server may take on, subject to configuration of the solution]

There has been some discussion around a desire to minimise the impact of any component failure to the client/end-user. However, there has (as yet) been no formal view on the abilities of the application, more specifically internal processes & the clients interacting with the application, to support this; E.g. the statement is made that if one FIX server fails, then all four must be failed over – in conjunction with a

single IP solution, the network could fail-over the communications for all servers automatically. However manual process invocation would result in the remaining three good servers becoming unavailable until such time as the manual involvement has been completed. Until examples such as this have been cleared up and resolved (these are application/user/process level queries) then it is dangerous to make the infrastructure too dynamic. Therefore in this edition of the design, the infrastructure (specifically the network) will stick with the "keep it simple" philosophy, and omit to add any extra intelligence until such time as the suitability of it can be assured.

That said, the network provision for Select 7 is largely similar (from a high level) to that for Select 6, however with a couple of notable additions/differences that are discussed in further detail later.

1. The traffic profiles for the users will change, in that they will now require a direct connection to the FIX (FS) servers, rather than today's provision where they connect via the TAX server.
2. In addition to the current provision of Front End (Access), Back End (Management) and Multicast LANs, there will need to be a fourth Unicast subnet.
3. The Select 7 implementation will be spread across both Production locations; therefore there will be a need for site-site communications.

10.2.2 RAG Overview & Select provision

Select 7 will be provisioned following the RAG model, as is effectively the case for Select 6 today.

RAG, is a Red, Amber Green classification of network "zones" within the estate. Each zone having a logically equivalent (consistent) provision, however the use of (and access levels to) each zone will vary depending upon the classification.

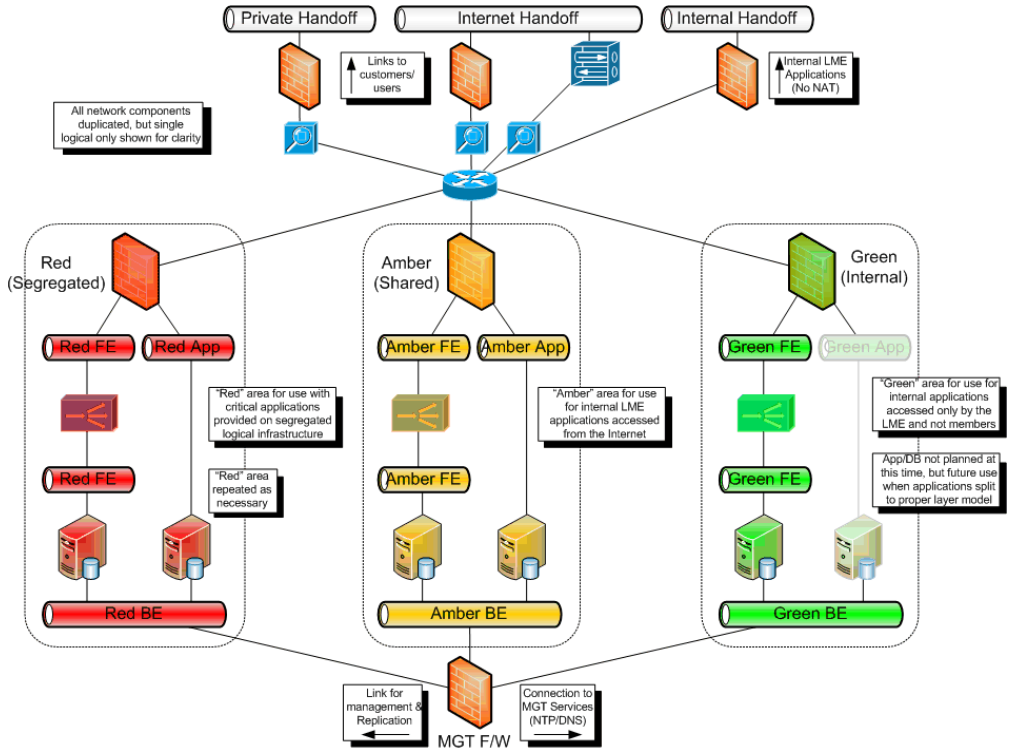


Diagram 5: RAG High Level Layout

Red, which is effectively achieved by Select 6, and what will be used for Select 7, means that the zone is dedicated to the given application only. This means it will have its own VLANs, firewall and load-balancer provision, which is not shared with any other application. There can be more than one Red zone in existence (there is also a Red-Smart zone for example), however these are only provision where necessary, typically in the case of high-criticality of the application to the operation of the LME business.

For reference only, Amber is a single zone, which will host multiple applications. These applications are those that are used by the LME Members; however they do not warrant the same levels of criticality that would require separation of the firewall and load-balancer configurations. Green in the meantime is also a single zone, similar to Amber, except it is for the provisioning of non-member facing services (also, though not quite accurately, referred to as the back-office applications).

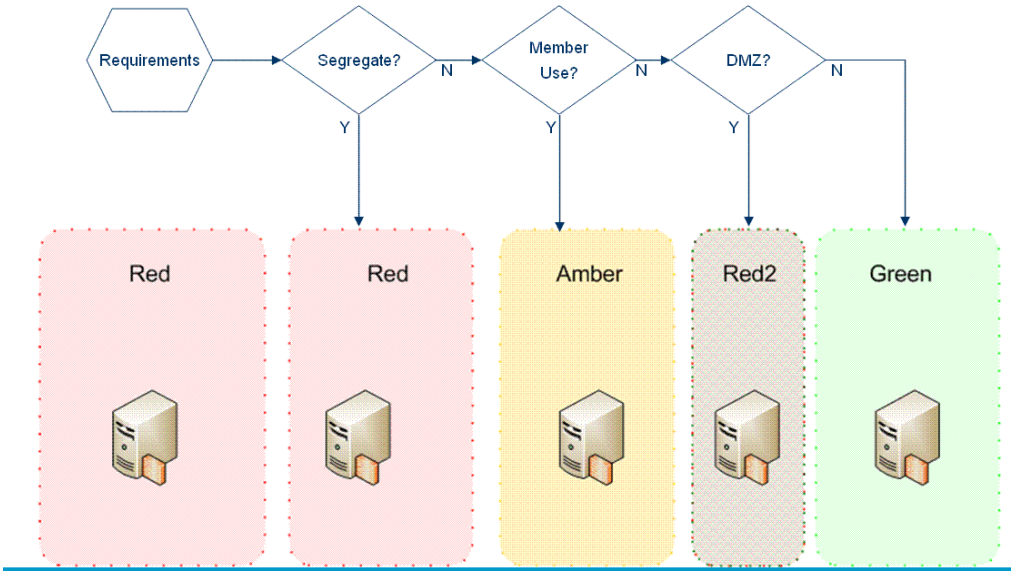


Diagram 6: RAG Decision Matrix

Each zone, is provision with a routed firewall context as the gateway, and to control access to and from the other zones. By virtue of the firewall being routed, it is possible to split of multiple LANs from the firewall, permitting provision of separate DMZ, App and DB layer LANs as appropriate.

If necessary, and it is not always the case, each zone can be provisioned with a load-balancer context, for use with the servers inside that zone.

Due to Select 6 having been one of the original application on the estate, it was built pre-RAG, therefore Select 6 however isn't quite a pure Red in terms of current design guidelines, however in essence it effectively achieves the requirements of a Red zone provision. As a result there will be a subtle difference with Select 7 to simplify the provision and provide consistency with the RAG style, this would make no fundamental difference to the function or provision of the Select application function, as the impact of this will be contained within the Network Team. This change does still warrants noting for reference however.

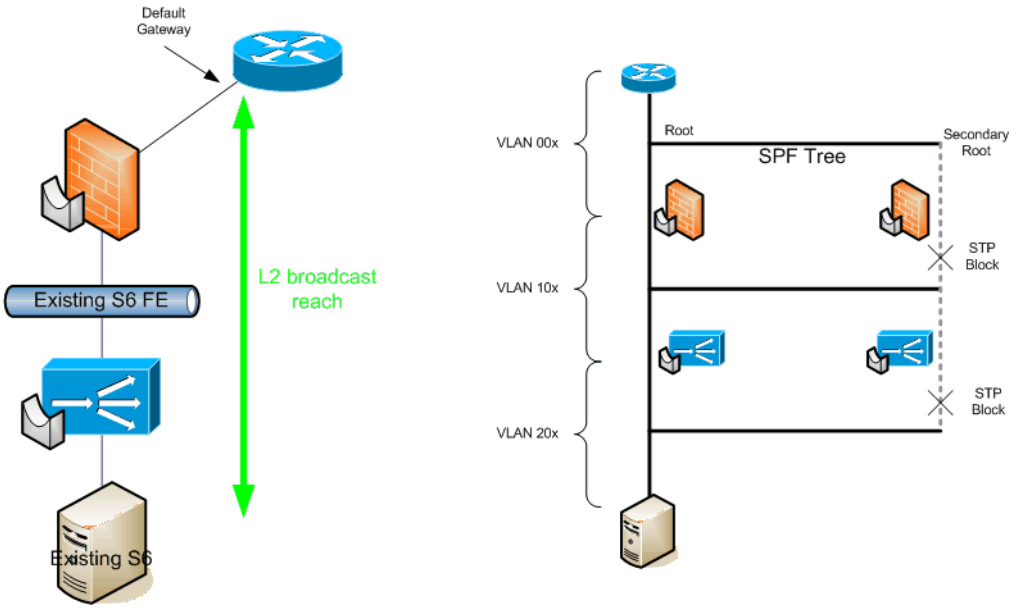


Diagram 7: Legacy L2 STP

At the time of the original Select 6 deployment, the model was for the provision of transparent firewalls, however as the estate has grown beyond original expectations, along with the associated increase in complexity, the RAG model now in place will use routed firewalls amongst other improvements that reduce the complexity and improve the scalability of the estate considerably.

This results in the ability to provision a single firewall for the Red deployment, rather than multiple as in the case of Select 6. This also permits constraining the reach of the layer 2 broadcast domain, simplifying low-level troubleshooting and reducing related spanning tree considerations.

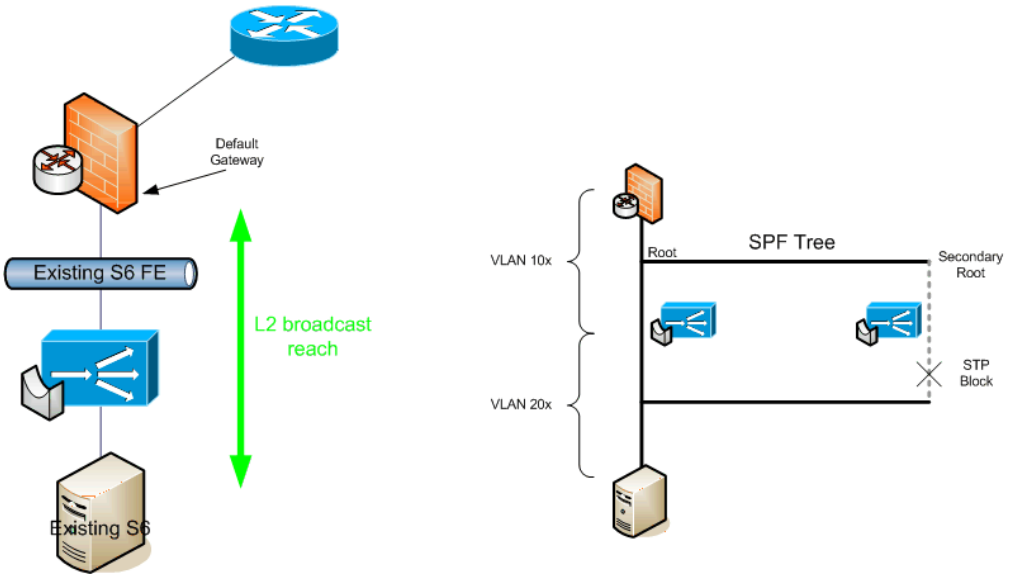


Diagram 8: Reduced STP

10.2.3 Changes in FIX protocol access

In the Select 6 environment, Members using the FIX protocol to communicate, would do so by sending their FIX traffic to the TAX servers, which also double as the web servers, and provided a form of firewalling. The main benefit of the TAX servers though, was the ability to then send the FIX traffic to the appropriate FIX server (FS).

Each FS is configured to operate for a particular group of users, therefore any given user can only access a particular FS – any attempt to access a different FS will result in denied login attempt.

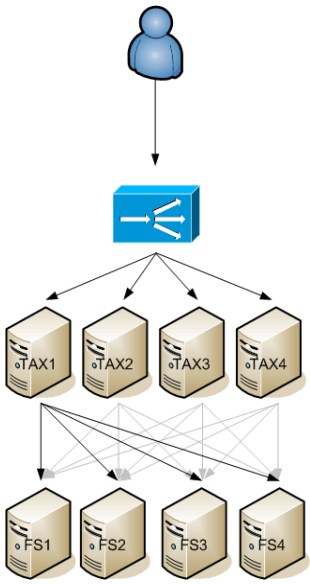


Diagram 9: Connection to FIX in Select6

As part of the drive to improve the performance of the Select application, the latency of the FIX traffic is an area that has been impacted. With the new edition of the software, it is expected that the FIX clients will be able to access the FIX servers directly, rather than going via a TAX server.

This has improved the latency of the application by removing a processed hop from the transaction flow.

However, the need for users to access only a particular FIX server has not been removed. Therefore load-balancing of these servers by the infrastructure is not immediately an option, instead this (as was the case previously) will continue to be achieved through use management dictating how many accounts are assigned to each server. It has been investigated, with the benefits of the additional intelligence that can be applied with the Cisco ACE, as to whether this is an option to identify something unique in the TCP stream set-up, that would permit the ACE to direct the traffic to the appropriate server accordingly. The application Vendor (Cinnober) after consideration, have advised this not to be feasible.

As a result of this, each FIX server will need to be given a unique tuple (IP & TCP Port combination) to the outside world. In conjunction, each member will need to know the details of their correct tuple in order to access the service. Should a member attempt to connect to a different tuple, they will be unable to login.

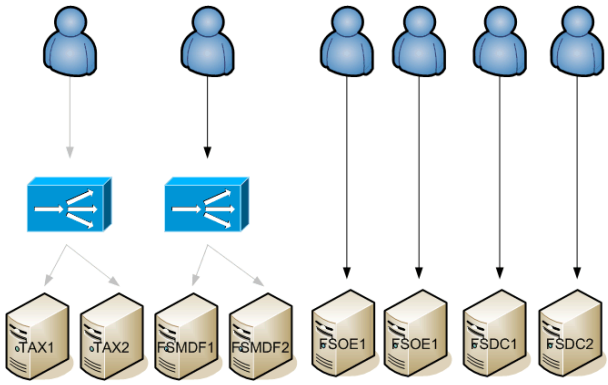


Diagram 10: Connection to FIX in Select7

Further as part of the Select 7 design, the FSxS (backup server for FSx) will now be located on the alternate production site (discussed later). Therefore failover between the FS servers will need to be handled by the clients. This removes any requirement/need for a load-balancing to present a single IP/port to the customer.

This means that the ACE will provide no benefit in this location, and therefore will not be used for the FIX servers – on the other hand, this does result in an increase management workload for the LME who will need to manage the Members independently, with a noted risk that in the event of requiring to adjust the Member distribution on the servers, will require additional planning and notice periods to achieve. This is all outside of the infrastructure scope however.

To support direct communication of the FIX traffic directly to the FS servers, provision of additional PAT's on the gateway firewalls will be provisioned. (Note that in practice in some case multiple FIX servers, e.g. FSDC1 & FSMDF3 in failover, will share a server, with a single IP and unique port each – the logical diagram below shows them independently for clarity.)

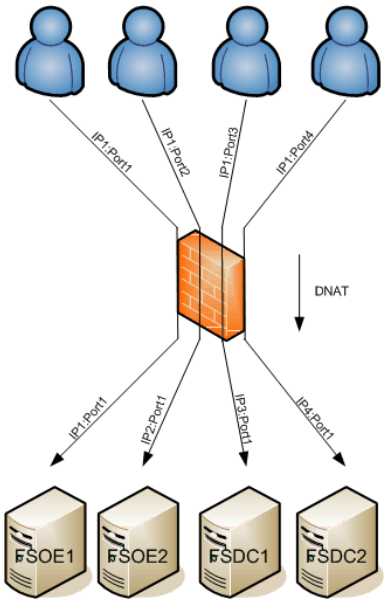


Diagram 11: FIX PATs

10.2.4 Additional Unicast LAN

Select 7 will now have a requirement for an additional VLAN above that provided to the Select 6 servers. This will be for the purposes of Unicast communication between the servers, which will operate in conjunction with the existing Multicast network.

The Unix portion of this design, has made the call to provision this additional network via use of a trunk link, as the servers are only currently provisioned with three pairs of NICs – these is seen as lower risk/cost in comparison with the installation of additional NICs; it has been noted that this is subject to testing of course.

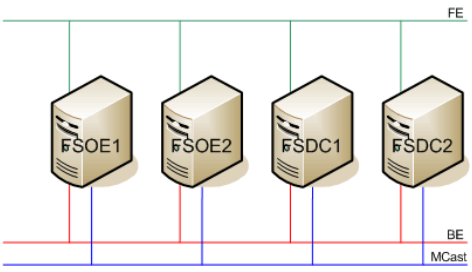


Diagram 12: VLAN Provision in Select6

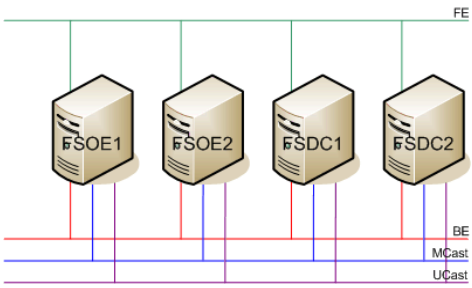


Diagram 13: VLAN Provision in Select7

Therefore, the three network ports will be provisioned on each switch to each server. Two as access ports and one 802.1Q trunk port. The Unix design has specified (again subject to testing) that the Unicast and Multicast VLANs will be shared on the trunk port, with the front end and back end VLANs provided on access ports.

Should the testing result in this proving inadequate, and additional NICs become a requirement, then the network configuration will be modified to provision four access ports on each switch for each server; one access port for each VLAN.

10.2.5 Stretching across two data-centres

The legacy Select 6 installation achieved resilience through having two complete installations – one at each data centre. This has the drawback however, that in the event of an intra-day failure preventing the active Select 6 installation from operating, the Select knowledge of the day's trading has been lost – the time therefore to recover this and initiate operations from the secondary site is therefore significant.

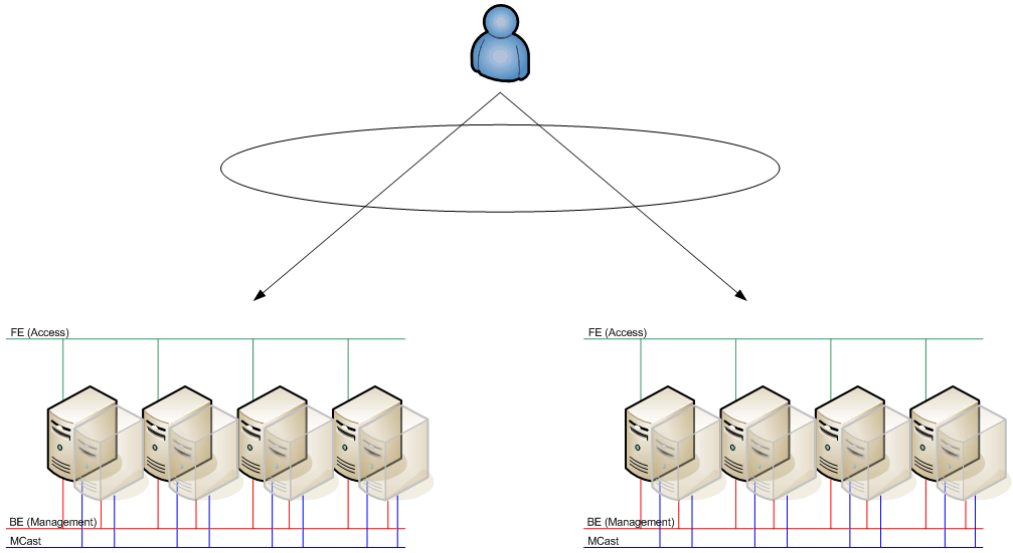


Diagram 14: Select6 Seperate Installations

To mitigate against this in Select 7, the vendor has proposed stretching a single installation across both data-centres. This will be achieved by keeping the primary servers at the active data-centre, whilst moving the secondary servers (which existing in Select 6 to mitigate against node failure) to the secondary site. The secondary servers will maintain communication with the primary servers in the same fashion as Select 6 – i.e. over a layer 2 link.

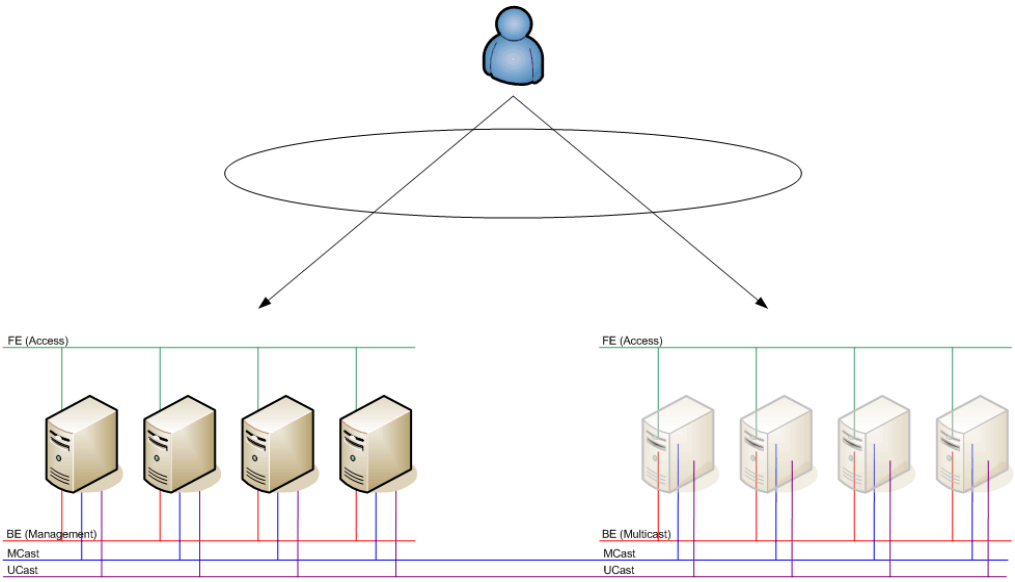


Diagram 15: Select7 Single Installation

For the provision of site-site communications, this design will have a significant dependency upon the provision of a high speed (low latency) site-site circuit. Provision of this is out of scope of this project, but timely delivery under separate project cover will be expected.

For this project's consideration, will be the provision of two "stretch VLANs" – the natural risks of stretching VLANs will be mitigated through an expected resilient site-site communication infrastructure, and through the fact that these two VLANs will have no gateway. This VLANs will be the Multicast and Unicast Select communication LANs.

The FE access VLAN will not be stretched – this is to avoid gateway issues. This will instead maintain "independence", with IP addresses for their local respective subnets. Conversation with Cinnober has indicated that this will be acceptable. The primary servers on the primary site, will have a default gateway of their local firewall. The secondary servers on the secondary site will have the default gateway of their local firewall.

It should be noted that the Vendor preferred recommendation for latency between the two sites (or specifically between the servers on each site) would be to achieve a RTT < 400µs – however due to the distances involved, this will not be physically possible (one way propagation delay alone as the crow flies would be ⩾400 µ s). This response time has been requested, in order to prevent the network aspect being a cause of delay in the matching engine – the match time being the slowest component of

i) processing - typically at ~400µs

ii) disk write time

iii) network response – talking between the primary and secondary servers.

From initial potential supplier conversations, a discussion of achieving a 1.5-1.85ms response has been mooted as achievable. Actual response time cannot be assured until such time as a circuit has been installed and tested, so this figure is taken as a rough guide only.

As a result of this, the actual performance of the matching engine will be subject to test. A risk is therefore noted against the acceptability of the performance of the service in this configuration. Should this prove unacceptable, then provision of the entire service on one site, with a emergency engine on the second site will become necessary – this latter option has not been taken at this time, due to the fact that the emergency server may be a few seconds behind the operational matching servers; enough to cause an unknown market position in the event of unexpected site failure.

10.2.6 Resilience & Redundancy

Resiliency in the Select 7 provision will be achieved by having half of the installation at the secondary site. The primary site will host all of the active servers, with each server having a secondary or alternative provision on the secondary site.

Due to the nature of the application build, if the primary server fails (where applicable) the secondary server will automatically take over.

In the position where the secondary server is on a second site, and is a server directly access by the customer (e.g. the FIX server), this will mean that the secondary server will have a separate unique external IP address & TCP port tuple. It will be dependent upon the client to detect the failure and to reconnect to the secondary server when necessary.

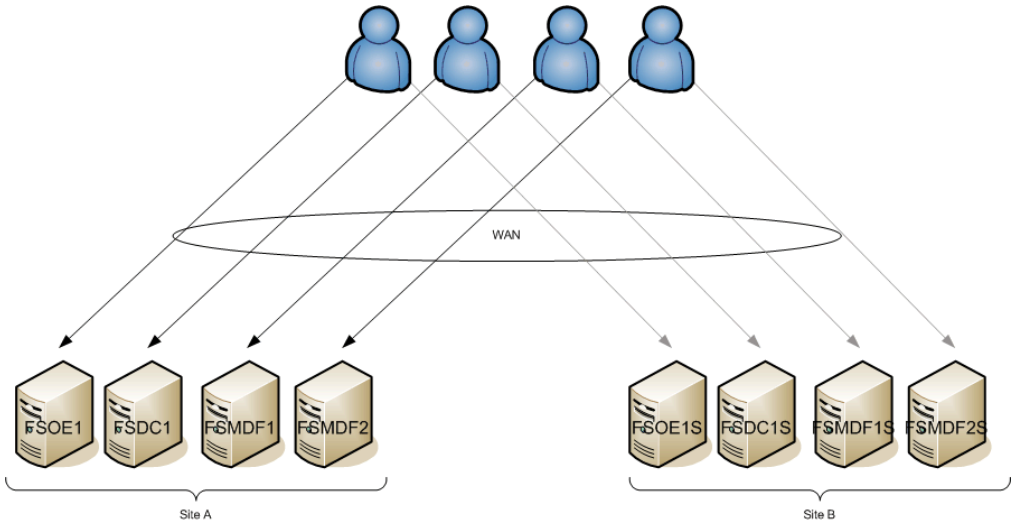


Diagram 16: FIX Redundancy

The same process will be carried out for all other server functions, where a secondary exists. The exception being the TAX servers which do not have a secondary, instead this will be split equally with two on each site.

This carries a risk, in the event of failure, that the users connected to that server may not have access to the secondary server suitably configured – this is requirement of the FIX specification, and is subject to the client, which is out of scope of the network design.

This also carries the risk, in the event of a single FIX server failing, that the clients using this server will connect to the secondary site with impacted performance, in the mean-while the clients on the remaining servers will still be getting normal performance, which will result in an unbalanced market. The statement has been made, that this will be mitigated by manually failing over the remaining FIX servers, so that all clients are provided with an equal service. The implementation of this mitigation, is subject to production processes, and also out of scope of the network design.

Finally this carries the issue that a single component failure can result in cross-site dependencies for a period of time. This goes against the LME desire, that any single component failure may be handled locally. Unfortunately, due to the way the application is built, it is not possible to achieve this desire at the same time as having the option for site-site failover without loss of data. The closest alternative would be through provision of the secondary servers on the primary site, and a emergency server solution on the secondary site, however this emergency server will have a few second lag behind the production service, which maintains a risk of loss of trade data in the event of a major site failure.

There are potential options for the network to add further intelligence at this point to meet each of the above risks, for example a single-IP can be fudged over the existing infrastructure, to avoid the client having to look for a second IP – however depending on the implementation, this will either

i) Have a detrimental impact on the service performance (!) due to the client still connecting to the primary site, getting redirected to the secondary site for the server, then coming back to the primary site for matching (and return). This would also not resolve the secondary issue.

ii) Alternatively, this will be in the position, where the IP switches to the second site, but then if only one server has failed, this would be recovered, whilst any access to remaining working servers would get "broken" until such time as production process has failed over the remaining FIX servers. This just changes the impact of the second issue, but doesn't really resolve.

In conclusion, the benefits of the extra complexity are not clear cut, and indeed (at this time) carry their own risks and confusion, therefore these have been omitted at this time; with the preference to keeping a simple (and faster) solution that is easier to maintain and support.

One effect of the above solution, is the load-balancers will no longer provide any value add to the FIX servers, therefore these will (logically) be removed from this part solution. This will remove a layer of processing delay, which all helps to improve the performance of the solution.

The load-balancers will remain however to support the TAX & FSMDF servers, in the same manner as they do today.

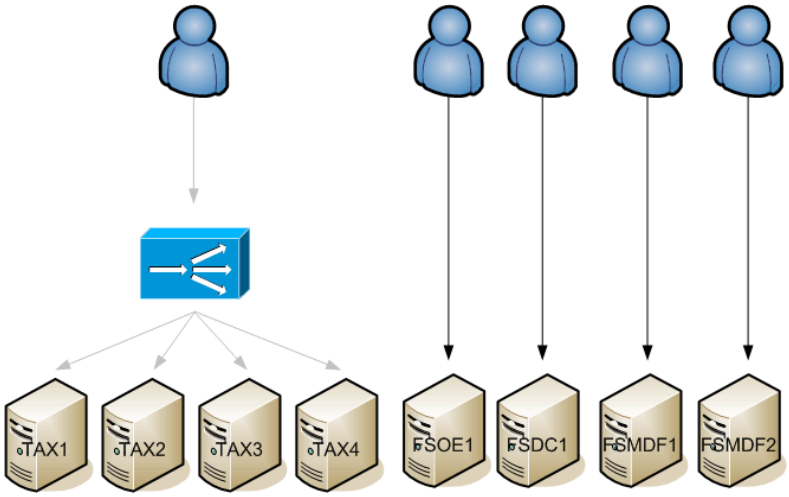


Diagram 17: LB Remains for TAX flow, but not seen in the FIX flow

This is because the TAX servers do have the ability to be balanced, and avoids the need to provision

multiple IPs on each site for the web-interface part of this solution.

10.2.7 Bandwidth requirements & scalability

The client access bandwidth is not mooted to be changed with the upgrade from Select 6 to Select 7. Therefore the configuration for client access will not be modified. This currently sits as a 100Mbps assurance on the Packeteer, which will in due course be backed up by the QoS configuration on the Savvis network (under separate project cover, and out of scope of this document).

For the site-site communications, the application vendor have stated a minimum of 300Mbps availability per NIC (VLAN) – although they would ideally recommend allocation of 1Gbps per interface. In practice, the FE and BE will only be site-local traffic, and therefore will have the full capability of the switching infrastructure in that site (the switch at 720Gbps, the ACE at 4Gbps (upgradeable to 10Gpbs), the FWSM at 5Gbps. The slowest part of this being the client access provisioned at 100Mbps as noted above.

This leaves the Multicast and Unicast networks requiring site-site bandwidth in addition to today's current provision; minimum requirements here would then dictate 600Mbps availability between sites. Discussion with the site-site circuit provisioning project, has indicated that this will be looking at 10Gbps, therefore it is envisage that 1Gbps will be available to reserve for Select 7, exceeding the minimum value stated by the vendor.

10.2.8 Environments & Lower Levels

Select 7 will be provisioned with a stretched Production environment across both Basildon and Slough production solutions.

Pre-production however, exists only in one location (Basildon). For effective testing of the Select 7 solution, this ideally should also be like-live and exist across both sites. Unfortunately the costs of provisioning this are such that the LME do not wish to provide this ability at this time.

Instead the recommended solution, will be for the Basildon Pre-Production environment, to be configured to provide two logical installations of the RAG provision. This therefore can then support two logical installations of Select (representing either site) along with two logical installations of any dependant applications that may also be required to support testing.

This is quite possible, due to the pseudo-virtualisation of the network stack permitting this isolation.

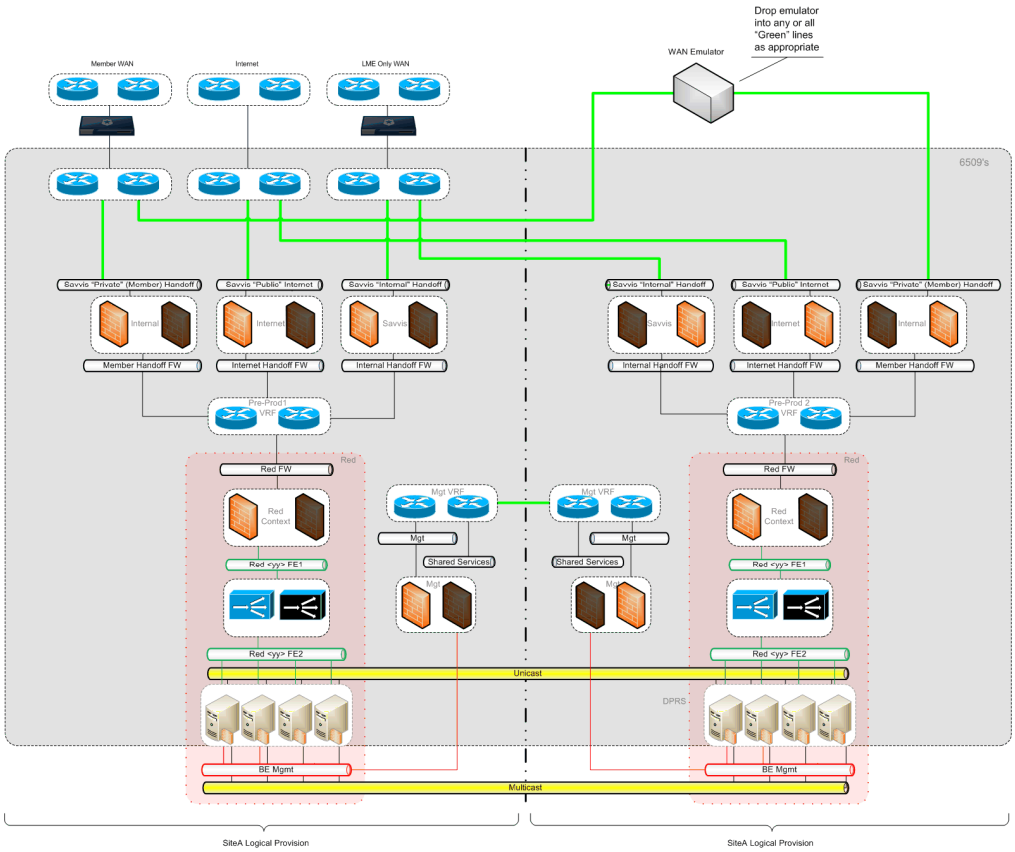


Diagram 18: Pre-Production Dual Split Logical Layout

The only potential down-side with this solution, will be the ability for true performance testing – i.e. when the logical "siteA" provision talks to the logical "siteB" provision, it can do so at switch speed. The resolution for this will be for the provision of a WAN simulator between the two logical installations. This may either be through the occasional use of the Xchanging simulator or through the LME's purchase of a dedicated unit.

The actual provision of the pre-production environment (as a whole) is happening under separate cover, and will not be discussed further here beyond the essentials mentioned above. However it remains a dependency of this project, that a suitable pre-production environment, with access to a WAN simulator, is available to effectively replicate the dual-site provision nature of Select 7.

10.2.9 IP Addressing

New VLANs & IP addresses will be provided for the new installation.

The legacy installation IPs will be recovered after completion of migration and settling in period.

The new Internal IP ranges will be sized as follows (actual IPs to be defined during low level works)

Slough User LAN (FE) - /26

Slough Management LAN (BE) – /26

Basildon User LAN (FE) - /26

Basildon Management LAN (BE) - /26

Shared Multicast LAN – 224.23.23.0/24 (Cinnober defined)

Shared Unicast LAN - /24

10.2.10 Firewall configuration

A new Select 7 RED firewall will be built as part of these works.

The firewall will need to contain the following configuration as a starting guide. Actual rules will be defined during low level works.

Inbound

USERS -> Load-balancer -> TAX http, https, 5025/tcp

USERS -> (direct)FSOE1 5555/tcp

USERS -> (direct)FSOE2 5556/tcp

USERS -> (direct)FSDC1 7980/tcp

USERS -> Load-balancer -> FSMDF (7990/tcp)

NAGIOS -> all systems/ports

Management Access -> 22/tcp & all of above + SOPS 22780/tcp (to be verified)

Outbound

-

10.2.11 Load Balancer configuration

A new Select 7 RED load-balancer will be built as part of these works.

The load-balancer will effectively follow a similar configuration to that existing for Select 6, however without the addition of the FIX servers.

Under normal operation

USERS -> VIP -> TAX http/https & 5025/tcp, sticky

USERS -> VIP -> FIX MDF (7990/tcp), sticky

Probes are to run syn checks on the tcp port being balancer (e.g. check for the presence of 5025/tcp for the TAX servers etc).

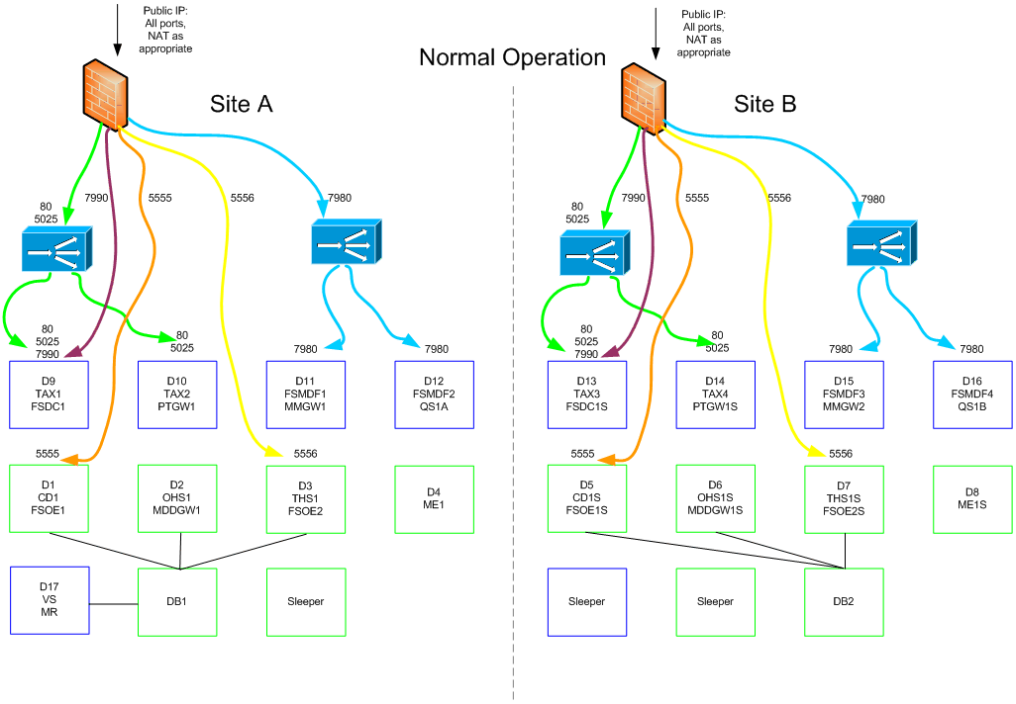


Diagram 19: S7 Normal Operation Incoming Flows

To support disaster recovery (see section 12) the following are also to be pre-configured on the balancer, however not used under normal operation

VIP -> FSOE1 & FSOE1s(sleeper) 5555/tcp

VIP -> FSOE2 & FSOE2s(sleeper) 5556/tcp

VIP -> FSDC1 & FSDC1s(sleeper) 7980/tcp

11 Migration

Provision of new hardware will simplify migration requirements.

11.1 Application Migration

The machines to be provision in Production, will be installed into a new Red-Select zone – independently from the existing installation (and therefore without impact to the existing service).

At time of go-live, the gateway firewall rules (Savvis) will be updated to point to the new servers, and permit the additional ports required for the new configuration.

The new servers will take over the load of the Select traffic, with the legacy Production servers moving to a “standby” capacity in the case of the need for rollback. Finally, after a suitable bedding in & confirmation period, the legacy Red-Select environment will be decommissioned.

11.2 Network Migration

To support the migration from Select 6 to Select 7, and the requirements of the Unix/Application build (noted above), the Select 7 environment will be built as a new “Red” installation, this will then support the integration of the new Select servers (from a network perspective) without risk of impact to the existing Select 6 service provision. When the Select 6 servers have been decommissioned, the Select 6 VLANs, Subnets, Firewalls and Load-Balancers will be decommissioned.

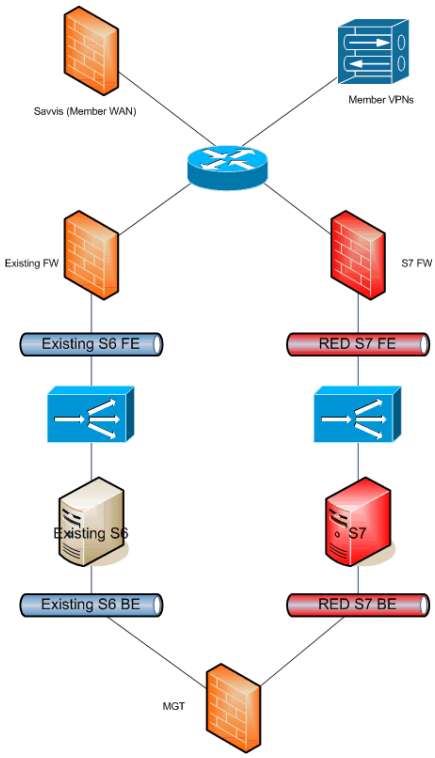


Diagram 20: Logical Network Addition (Abstract)

This method of provisioning the new service, will have an implied dependency upon the availability of an additional firewall context and load-balancer context within each Production environment to support the new build until such time as the legacy build can be decommissioned.

Prior to implementing the Select 7 provision in this manner, there will also be a require to test destructively and at load, the Select 7 implementation. This will have an obvious impact to the Production network, resulting the requirement for a major restriction on the ability to test this service.

Therefore, it has been agreed that the Select 7 will be provision within the Pre-Production environment in the first instance. The configuration can be tuned, and the application thoroughly tested without impact to the production environment.

When Select 7 has completed testing, the machines will be cleaned of test data and then migrated (re-racked/re-patched) into the proposed Production environment. Testing at this point merely needs to assure connectivity and the performance of the inter-site link.

At the completion of the testing period, and subject to acceptable results and customer sign-off, the

Select 7 service will go live – an external NAT migration for the TAX servers being relatively simple at this point. Initially the Select 6 server will remain in situ to support emergency regression should the need arise. After a suitable bedding in period, the Select 6 servers will be decommissioned and then re-connected into the Pre-Production environment, to provide a Pre-Production solution going forward.

12 Replication & Failover Design

Replication is handled within the Select application itself. No additional infrastructure tasks are required to support this, aside from the availability of the stretched Unicast & Multicast VLANS between the two production sites.

Failover from SiteA to SiteB (and vice-versa) in the event of a component or site failure will be handled through dynamic DNS. The configuration of this will be detailed in the DC Link design as provision of the F5 GTMs for this are under that project scope. However, in summary, the GTM devices will probe the Select servers in the same fashion as historically completed by the ACE. In the event of a loss of probe, the GTM will update the DNS to advertise the IP address of the remote site. Actually auctioning this update can be either automatic or manual, subject to Production procedures & requirements.

In the event of an extended total site failure, the firewall NAT's for the FIX servers (FSOE1, FSOE2 & FSDC1), as part of disaster procedures, will be updated to point to the balancer VIP rather than the real servers. This is to recover machine resilience. Under normal circumstances FSOE1 will have the resilient FSOE1s on the remote site, however upon loss of the remote site this will not be possible, therefore FSOE1s will need to be brought up on the remaining active site. This applies vice-versa of course in the event of the loss of the active site.

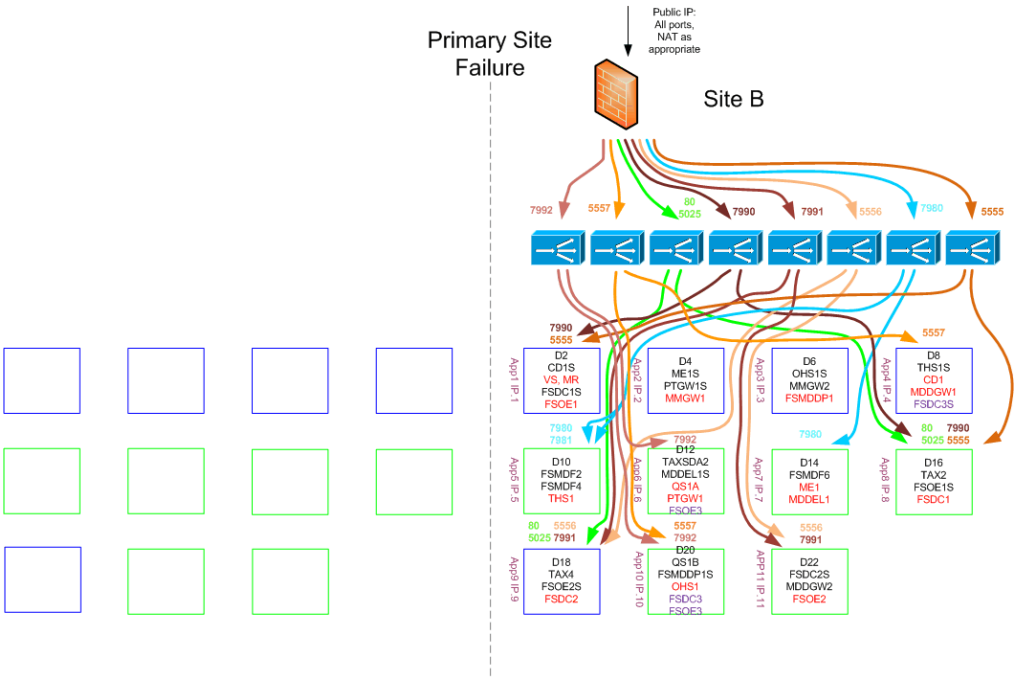


Diagram 21: Select7 Inbound Flows with Extended Site A Failure

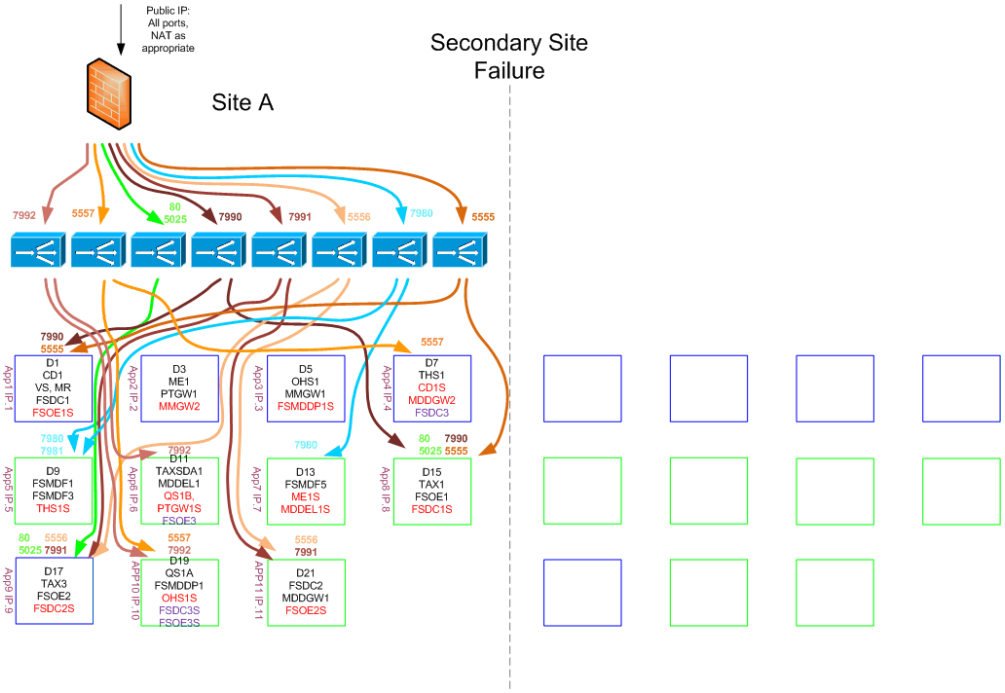


Diagram 22: Select7 Inbound Flows with Extended SiteB Failure

13 Monitoring

Monitoring should be implemented in the standard fashion for the servers hardware and O/S, specific application rules will need to be design to support the run characteristics of the Select7 product, although a starting position can be taken from the existing Select6 monitoring configuration. Information for the application will be provided by the Application Engineering team to Production.

It will be expected that the equivalent monitoring will be provided in the pre-production environment. This will enable verification of the monitoring templates as well as a documented baseline of application performance during normal testing and load testing.

For this specific build, the following monitoring templates would be expected for the infrastructure.

- RHEL5 x64 (11 per environment)
- Apache (2 per environment)
- MySQL (1 per environment)

- ACE Load-Balancer Instance (1 cluster per environment)
- ACE VIPs (5 per environment)
- FWSM Instance (1 cluster per environment)

13.1 Extended DR

Please note that in the event of an extended DR resulting the application having to use the alternative DR configurations, then the monitoring templates will need to adjusted accordingly.

13.2 Corvil

Further to the standard monitoring provision, the Corvil BQM solution is being updated as part of the Enterprise Monitoring project. Therefore the new Corvil provision will work with the Select7 installation, removing the need for the migration of the legacy Corvil provision. Upon removal of the Select6 installation, removal of the legacy Corvil installation will also occur. Detail on the provision of the new Corvil solution will be covered in the Enterprise Monitoring design.

13.3 Magnifix

Along with Select 7, the LME are also implementing the Magnifix product from Greenline – only general detail is documented here however, as this is being provisioned under separate project cover.

The Magnifix project will take a SPAN monitor session from the Select 7 FE VLAN – the data captured will be passed into an RSPAN to which Magnifix will be listening. This will grant Magnifix visibility of all the user FIX traffic to and from the Select estate.

Magnifix will then provide application level reporting on FIX price variation etc to the LME.

14 Risks, Issues & Assumptions

14.1 Risks

- Dependency upon available of FWSM & ACE (firewall & load-balancer) contexts for use during migration.
- Dependency upon separate project provision (DCLink) of suitable site-site communication link.
- Dependency upon suitable Pre-Production provision of downstream/dependent applications
- Minimising latency between sites is constrained by distance between sites;, therefore this may have an impact to the trade-matching confirmation time. Actual impact, and acceptability therefore, is subject to testing.

- The clients must implement the APIs correctly in order to follow the LME in the event of failover, and they must also ensure that there network security settings are configure to permit access to both data centre locations.
- Choice of VLAN tagging on the Bond interfaces, is subject to testing, and actual allocation may be modified if testing indicates a better performance with a different configuration. This is a relatively simple change, as long as it is done in conjunction with both Networks and Unix implementation teams.
- The application services may be moved between different physical devices according to the results of the performance testing and Cinnober recommendations. This may result in a requirement for additional testing to re-check and ensure best performance solution is provided.

14.2 Issues

- There is a risk that by using VLAN tagging we are effectively halving the network speed for each network configured as a VLAN tag – this configuration will be proven through testing. In the worst case, additional NICs may be required, however the Vendor has agreed with the use of VLAN tagging. Due to the lead times & impact of provisioning additional NICs, it is recommended that this aspect of performance testing occurs early on in the process.
- Assumption that the current LME patching solution is suitable for RedHat Enterprise Linux updates and fixes – this should be the BigFix solution which carries Production concerns on lack of clarity, an intermediate solution (manual) will be provisioned, however this needs to be resolved properly separately (out of scope of this project)
- Due to lack of clarity/formality around the interaction of the Clients and internal processes with the Application & infrastructure, certain discussed options for improving the intelligence of the infrastructure have been omitted. This is following the “keep it simple” methodology, and to mitigate against the not inconsequential risks of any extra intelligence back-firing and causing as many, if not more, problems than it resolves.

14.3 Assumptions

- It is assumed in this design, that the latency actually achieved between the two sites (which won’t be known until circuit delivery) is suitable to operate the application in a performant manner.
- It is assumed that clients implement the FIX API correctly and can therefore follow the failover proposed.

15 Appendix A – Standards Framework

Table 1 below contains Hyperlinks to the Technology Standards Framework. These documents provide the reader with all project specific information from build standards, hostname standards through to naming services such as DNS, NTP etcetera.

Sharepoint Folder	Description	Sharepoint Link
High Level Estimate Template	Proposal costs	XTS High Level Estimate Template
Standards Framework	Methodologies, Templates and documentation control	XTS Methodologies, Templates and documentation control
COMMON	Naming services NTP, DNS.Hostnames	XTS Naming services NTP, DNS
Windows Standards	Microsoft and VMware builds	XTS Microsoft and VMware builds
Solaris Standards	Sun Computer builds	XTS Solaris builds
Linux Standards	Redhat Linux builds	XTS Redhat Linux builds
MS SQL Standards	Microsoft SQL Server builds	XTS Microsoft SQL Server builds
Oracle Standards	Oracle Corporation builds	XTS Oracle Corporation builds
Storage	IBM SAN, Netapp	XTS Storage Area Network and NAS builds

Standards	Filers	
Backup Standards	Symantec Netbackup	XTS Netbackup builds
Network Standards	Cisco, Firewall and other network documents	XTS Cisco, Firewall and network documents
XTS Naming Abbreviations	Locations, Customers, Environment details,	XTS Naming Abbreviations

16 Appendix B – Useful Links

Description	Sharepoint Link
Xchanging Wiki	XTS Wiki
Xchanging Design Process	XTS Design Process
Xportal link to Technology Services	XTS Technology Services
Senior Management Team Org Chart	XTS SMT Org Chart
Xchanging Quality Management System	XTS Quality Management Systems
PMO Repository	XTS PMO Repository

17 Appendix C – Glossary

ACE – Application Control Engine *(load balancer)*

API – Application Programming Interface

APP – APPlication server

BAS – Basildon

BE – Back End *(Management Access)*

BQM – Bandwidth Quality Manager

CPU – Central Processing Unit

DB – DataBase

DC – Data-Centre

DMZ – De-Militarized Zone

DNS – Domain Name System

DR – Disaster Recovery

EME – Emergency Matching Engine

FE – Front End *(User Access)*

FIX – Financial Information eXchange

FQDN – Fully Qualified Domain Name

FS – FIX Server

FT – Functional Test

FWSM – FireWall Services Module

GB – GigaByte

Gbps – GigaBits Per Second

GTM – Global Traffic Manager

HP – Hewlett Packard

HS – trade History Server
(now THS to avoid confusion with OHS)

HTTP – Hyper Text Transfer Protocol

HTTPD – HTTP Deamon *(HTTP server)*

HTTP – Hyper Text Transfer Protocol Secure

IP – Internet Protocol

ISD – Infrastructure Solution Design

JAR – Java ARchive

JDK – Java Development Kit

JRE – Java Runtime Engine

LAN – Local Area Network

LB – Load Balancer

LDAP – Lightweight Directory Access Protocol

LME – London Metal Exchange

LV – Logical Volume

LVM – Logical Volume Manager

Mbps – MegaBits Per Second

MDD – Market Data Distributor

MDF – Market Data Feed

ME – Matching Engine

MS – MilliSecond *(1/1,000 second)*

µS – MicroSecond *(1/1,000,000 second)*

NAGIOS – Nagios Ain't Gonna Insist On Saint
(monitoring tool)

NAT – Network Address Translation

NIC – Network Interface Card

NTP – Network Time Protocol

OE – Order Entry

OHS – Order History Server

OS – Operating System

PAT – Port Address Translation

RAG – Red / Amber / Green
(Infrastructure zoning framework)

RAID – Redundant Array of Inexpensive Disks

RHEL – RedHat Enterprise Linux

RTT – Round Trip Time

S6 – Select 6

S7 – Select 7

SLO – Slough

SNMP – Simple Network Management Protocol

SOPS – Select OPerationS
(Select Management & Monitoring Tool)

SPF – Shortest Path Forwarding

SQL – Structured Query Language

STP – Spanning Tree Protocol

SU – SuperUser
(command to elevate privileges)

SUDO – SuperUser DO

TA – Technical Architect

TB – TeraByte *(1024 GB)*

TAX – web front end service

TCP – Transmission Control Protocol

THS – Trade History Server

VG – Volume Group

VIP – Virtual IP

VLAN – Virtual LAN

VS – Vote Server

WAN – Wide Area Network

XTS – Xchanging Technology Services