WITN05970142

RESTRICTED - COMMERCIAL

PWY/JFO/006

POCL Infrastructure/End-to-End Demo - Meeting Report

Supplier: Pathway/Escher **Date**: 11/12 December 1995

Location: Escher Group Ltd, Cambridge, MA USA

Attendees:

BA/POCL	**Supplier**
Jeremy Folkes	Mike Murphy (Escher)
Naresh Mohindra	Drew Sutherland (Escher)
	Liam Church (An Post)
	Liam Foley (Pathway)
	John Dicks (Pathway)

Purpose:

Visit to Escher as part of Demonstrator Reference Site visit programme.

Escher are the developers of Riposte, the package which in Pathway's proposal provides the middleware functionality for both the Office Platform and TMS layers of the POCL Infrastructure and also provides a major component of the End-to-End benefit payment system. Riposte was, at the time of the visit, the subject of the severity A/5 risk PWY009, which was the most serious outstanding technical risk against Pathway. This risk was downgraded to B1/5 in January 1996 and subsequently to C/5 in February.

Caveat:

The meeting was over-shadowed by the issue of a Non-Disclosure Agreement which Pathway/Escher wished the programme to sign. As a result, some time was lost during the meeting for protracted telephone conversations with BA/POCL in London to try to resolve the issue. By the end of the first day in Boston it became known that BA/POCL were unwilling to sign the NDA as it stood, which caused the atmosphere in the meeting to be somewhat strained. As a result, the meeting was fairly disjointed - as portrayed by these notes - although a considerable amount of information was still gathered.

Items of Note:

1. Background on Escher

- Mike Murphy - previous experience:
    - Background in Data Management facilities, wrote first system (sold to NCR) when still at high school.
    - Anti-submarine warfare database
    - Rug servicing company (order scheduling, cleaning etc) - late 70s
    - Housing partnership database (matching properties to investors)
    - 1-800 order direct database
    - Saddlebrook - company writing software for the Thrift Industry. MM was Vice President for Technology. Brought database products "up to date". Product called DBX - an object store for very large databases, had "Strong Versioning", the ability to view the database at any point in time as consistent.
- Drew Sutherland recruited to Saddlebrook in 1987.
- Saddlebrook bankrupt in 1989 (due to the crash in the US thrift industry).
- MM formed Escher Group from the high tech group at Saddlebrook (including Drew Sutherland). Apparently Escher started on funding from Digital, via the An Post

WITN05970142

connection (An Post are traditionally a Digital shop) and Mike Murphy's consultancy role.

- Took forward Saddlebrook's DBX as a product called "Matisse", now sold to various French companies (nuclear, gas/chemical plus military for naval system), plus 40-50 major clients in US (including military), plus now Tokyo Gas.
- Matisse now developed/marketed by French company Intellitic International; specifically marketed in US by company called ODB. However Escher still collects royalties on all sales.
- Matisse and Riposte "*both come from the same intellectual exercise*" but are independent products. Riposte is built on the experience from Matisse.
- Matisse was a kernel rather than an end product, designed to be used as the core of systems. Aimed for use in a specialised, controlled environment using skilled developers. With Riposte the aim is that a standard VB programmer ("*a COBOL programmer of the 90s*") should be able to write applications.

2. Riposte Philosophy and Background

- Mike Murphy involved in evaluation of proposals for An Post system (as "chief technical consultant" to An Post, a role he apparently still holds). Found none were suitable/affordable, so Escher developed Riposte in very short timescale to fill this need.
- Riposte was designed to "*never deny a payment*" - had to give a very low probability of failure and a high probability of recovery.
- Escher have a close relationship with Microsoft and Bill Gates; positioning Riposte to add value to the Microsoft product line. However, Mike Murphy also stated that *architecturally* they are not forced to Microsoft - the link is commercial.
- Riposte:
    - interfaces at a high level (applications can write to it)
    - truly peer to peer distributed messaging architecture
    - robust, assumes little about the environment and a low skill level
    - secure in a vulnerable environment
    - cheap support costs (very little support needed)
- Riposte 32:
    - improved communications/resilience
    - improved datastore at Correspondence Server; data management facilities *"Industrial Strength"*
    - using Microsoft NT certified access control (with Windows 95 would have to "do it yourself").
- Riposte 2 and Riposte 32 are designed to interwork (as will indeed happen in An Post during cutover)
- In later meeting Mike Murphy introduced the concept of two release of Riposte 32:
    - Riposte 32 Rel 1 - Jan96 - the version described here
    - Riposte 32 Rel 2 - Jan97 - NT/CAIRO integration, better backend facilities (distributed security and transaction processing, extending further into the central systems)
- Policy is not to use third party products - only to use using NT facilities - to avoid additional licence fees and additional dependencies. Aiming to lever best value out of NT and to "converge" with NT; (eg with NT 3.51 Microsoft have added

WITN05970142

underlying facilities, designed for use by SQL 6, which are of use for any database developer). [But they do use PKWARE compression/CRC libraries, as POCL do in AP]

- Escher do not use "unpublished interfaces" with NT

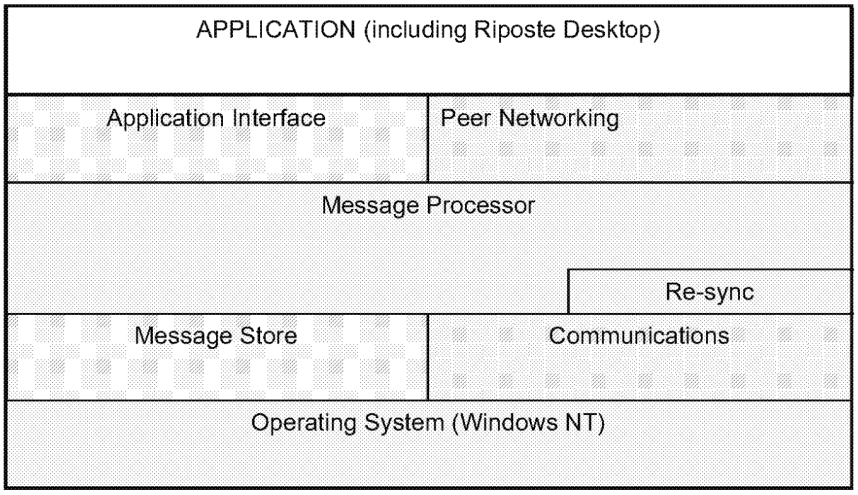
3. Differences between Riposte 2 and Riposte 32

Feature	Riposte 2	Riposte 32	Reason
Peer to Peer LAN	NetDDE	UDP/IP	Performance; IP-Multicast support; Standard TCP/IP
Application Interface	DDE	RPC/OCX (OCX custom controls for Visual Basic, calls DLLs)	Performance (especially at Correspondence Server)
Event Logging	JOURNAL.LOG file	Integrated with NT event management	Efficiency; integrated audit log (may also feed up events through Riposte if available)
Execution Environment	Windows application/task (icon on screen etc)	NT Service (started by NT)	Security
Performance Monitoring	Custom VB Utilities	NT Performance Monitor (hooks in Riposte)	Integration; scales to large network
Configuration	CONFIG.TXT file	NT Registry	Standard/tools and security
Security	Limited access controls (digital signatures, IFF, DES)	NT features, digital signatures, IFF. Only Riposte can get at the files.	Security
Time Synchronisation	Partial	Full (using NT facilities)	Note: Riposte itself is not based on times, as it uses message sequence. NT has local election of a "time authority" is office off line. Clock changes probably handled via GPS clock on one of the CSs.
CHANGES TO HERE HAD BEEN COMPLETED ALREADY; FOLLOWING ONES HAD YET TO BE COMPLETED.			
Call Management	Connectionless Message Passing paradigm.	Connection Management (UDP over PPP)	ISDN-B for BA/POCL. Note: still needed, but thinking slightly changed as Pathway moved to standard IP/Routers

WITN05970142

Feature	Riposte 2	Riposte 32	Reason
Message Store	Text based. External indices	Binary, Random Access (region of disk, bucket of storage, possibly across multiple disks). Parallel access. Internal indices. Ability to selectively compress or encrypt in file store.	Performance/Scalability. Facilities enabled by NT and the NT File Store NTFS. Concept of never overwriting still maintained even in new file store.
Synchronisation Protocol	Single level marker exchange	High level marker messages (Escher were unwilling to talk more in detail without NDA)	Tuning/Scalability Allows sync between overlapping groups
Archiving	Off-line, overnight - need to take down and restart	Background task	Possible due to the parallel access to the new message store
Persistent Journal Objects	Limited. Tended to use .INI files for application config. Had to reboot to load.	PJOs in the Riposte journal. No need to reboot. Application can ask for notification of updates	More efficient. Ability to use "on the fly"

4. Structure of Riposte



WITN05970142

5. Relationship Issues - Escher/Pathway

Second day started with a debate between John Dicks and Mike Murphy regarding the programme risks - Mike Murphy: "*are we aware of these?*". Mike felt that Pathway had failed to "sell" the advantages and strengths of Riposte and Escher to BA/POCL, asking "*Who is remiss at getting the message across?*".

Mike outlined the strengths of Riposte:

- he views Riposte as *reducing risk* - use of NT, IP messaging, convergence with Microsoft plans for Cairo and Talon.
- close links with Microsoft, moving towards Message Orientated Middleware, Bill Gates aware and supportive
- considers Riposte to reduce dependency on other products, messaging architectures will become a "*commodity item*".

Mike Murphy to John Dicks: "*Why did you leave it so late in the game before bringing us in?*".

- re manageability risk, Mike Murphy's view is that using Riposte *"its replacing 20,000 servers with 16 servers and so much easier to manage"*. They had considered NT Server in each office, but how do you manage 20,000?
- Pathway had apparently originally considered a centralised system as less risk for the bid

6. Relationship with An Post

- Escher said they were "avoiding" transfer of the Riposte IPR to Pathway
- An Post is considered a "marketing partner"
- Apparently Escher, An Post and POCL are working together on Hungarian bid (appears some work by Systems Consultancy in support?)
- Escher had worked with both IBM and with Anderson Consulting as well as Pathway - and say *they* chose Pathway. Regret the amount of documentation they gave IBM.
- Re NT - Commented on experience of Nat West as an early adopter of NT Server - long painful process, due to lack of support tools etc. Now running well after significant Microsoft involvement/money.
- *Conversation with Liam Church on the plane back: Riposte is apparently jointly and severally owned by Escher and An Post - with An Post funding the development up to June/July 1995. Pathway are now part funding the development.*

7. Synchronisation/Replication

Escher gave a primer of Sync/Resync but based on the technology of Riposte 1. This is said to be much optimised for Riposte 2 and Riposte 32 but lack of NDA etc made them unchain to go into too much detail on 32.

- messages - written to local store, broadcast to neighbours
- when message processor receives a message, it checks last message number from that node (which may have been *forwarded* via a third party - irrelevant). If not ready (ie hasn't got all preceding messages) sends a request back to the *forwarding* node. If no response it takes no action. But the original (potentially out of sequence) message is still forwarded, even if this node is sick (including via ISDN in BA/POCL).

WITN05970142

- Riposte maintains two counters - the last message stored (ie last in true sequence), plus a high-water mark. Apart from these markers, Riposte is stateless.
- Every node broadcasts the state of all the nodes (latest in store plus high water mark). Markers are send on a time interval plus whenever a machine re-establishes a network connection.
- Messages are protected by a CRC (implemented using the standard PKWARE algorithm). All CRCs are checked on receipt *and* retrieval from the message store. If one is discovered bad on retrieval, assumes the entire store to be corrupt and marks it as such, recovering from elsewhere.
- Marker messages - Down to how quickly you want to detect a missing message, based on how often a message will go missing (latency). Worst case shown as:

  Assume 40,000 terminals * 32 bit message sequence number = 160kB
  Assume 60 second exchange interval
  => 24kbit/second (<<8Mbit/second)

  As must journals will not have changed, can send only the changed data
  => 6kbit/second which is <1% of the network capacity.

  Once a missing message is found, it will generate traffic to request the missing data.

8. Mortality

- mortality is either based on an explicit mortality date, on a default expiry, or based on PJO being superseded (later version of the same message)
- to maintain sequencing, a message which is behind one to be killed is re-written (*re-incarnated*) with a new number and a pointer back to the old one, so that there is always a dense set of messages in the store. This is only done for "own messages". [Appears that there is also a low water mark concept].
- When rebuilding from another node (eg a CS) Riposte checks the mortality rules before writing.... this would appear to force all (including those ready to be killed) messages to be transferred over the network. Apparently this is one of the "optimisations" covered by NDA.

9. Recovering Own Node Messages (different)

- looks for markers, to get own high water mark, then goes into recovery mode (cannot store transactions at this point)
- how does it decide when fully recovered (the risk that there could be transactions hidden on a disconnected terminal, for instance)? Number of options are possible:
  - warn if not all terminals have been seen (manually check physical receipts)
  - refuse transactions until all machines back
  - use a dongle to record the high water mark and id of the terminal.
- The dongle was Drew Sutherland's preferred option, and anyway they may want a dongle for security. Mike Murphy*: "There are numerous other reasons why you want a dongle*". John Dicks retorted that there are other ways, must remember cost etc.

- Escher seemed to fully understand the scenario we were proposing, unlike Pathway in the UK who were keener to rubbish the idea. It seems that the terminal will always know which terminals it hasnt spoken to, so it would be possible to detect the danger cases, but will this be strong enough?

WITN05970142

- Liam Church commented that this scenario had happened twice in 2 years in An Post. They now have a procedure in place (since last Easter) to force all terminals to be "seen" before a replacement can serve.

10. Security

- Escher have access to MIT, and to Rivest of RSA fame (had worked with Drew Sutherland). Comments made that academics are interested in Escher as means of establishing credibility in commercial world.
- Riposte does not support Clipper; does support Kerberos (will be in NT 4.0) [ICL need to get Kerberos project for Sesame EU work?] and are planning to use Kerberos for BA/POCL (or ICL will propose an alternative?)
- An Post are using 1000 bit RSA keys for digital signatures, with no encryption. For UK, Pathway are proposing 64 bit DES key.
- An Post do not use MAC-ing from office to CS - seen as an overhead.
- Escher preference for cryptography is to supply source to client for them to compile, to avoid trap doors.
- Escher mentioned they have cryptographically strong random number generator available if strong cryptography required.
- Digital Signature planned for Benefit Payments; keen to use secure sequencing (again pushing for dongles) but Pathway negative to the idea.

11. Indexing

- can either commit indexing at the time, or can blast data in and build the indices later.
- hierarchical CRCs for indexes
- partitioning of indices in Rip32 - so can do parallel updates using RAID etc.
- change to caching algorithm and how the indices are stored
- background threads in Rip32 to maintain indices. Indices are never force committed at CS level. Application can force; in offices transactions are committed in an atomic fashion.
- two types of index - attribute index (permanent), retrieval index (temporary, read only, never updated, equivalent to an SQL query/view.
- same software but different code paths between CS and office - no compile time switches (and some code never executed, eg office doesn't expect data from another office)
- if need compression, would use PKWARE libraries... achieve >8:1 on attribute grammar.

12. Agents

- Agent uses the Riposte 32 API to communicate with data store via the Topology Manager/Supervisor. The agent is the means by which external systems (including PMS/CMS, TIP) are linked into Riposte.

WITN05970142

- Topology Supervisor knows which CSs are able to respond to a call (could be a number). Allows dynamic load balancing. The TS is strictly a dynamic entity, no storage of its own.
- Three types of agent:
    - On-line - listens to messages and acts. Specifies the type of messages it want to be notified of. Aims to be a simple state machine, does not need to keep records of outstanding messages.

      Concept of "tokens" to handle contention and locking - bronze for R/O, silver for possibly unable to write, gold for exclusive.

      Topology Supervisor will map across the CSs to get a full view of messages. May end up getting multiple copies of message, but this is ok.
    - Harvester. Concept of office sending an "end of day" to show cut-off, to define reversal constraints etc. Harvester does not release a batch of transactions until it gets end of day, using "markers" in the CS store, replicated back to the office.

      This would mean although a batch of transactions had been trickled to the CS, you would still be dependent on an end of day "poll". Purely done to police reversal policy? Apparently could do either (harvester code change).
    - System Management - to look at state of individual CS etc.

13. Miscellaneous

- Topology Supervisor - one per cluster, probably a separate machine . Handles requests from agents to appropriate CS via RPCs, to avoid external agents from having to know the topology. [300 msgs/sec experienced on "old" technology].
- Actuarial Broker. Interesting concept, of assigning value to a message or transaction computing value vs cost for where to authorise etc. John Dicks not keen to discuss, cut debate short!
- Wholesale Broker. Takes transactions in bulk, managing the initial disbursement of transactions over the network (+ other info and emergency payments etc). [Pathway think 900tps is the maximum push and pull load].
- Single Position Offices. Mike Murphy still proposing PCMCIA disk; John Dicks suggestion dongle (to hold recent txns) or flash memory. Would need a special driver for message store to truncate to last (n-k) messages.

14. Call Management (Mark Jarosh, Pathway)

- Pathway now looking at using IP Routers with implicit connection to PPP router - passing of a IP frame causes the call to be made.
- Considered that using NT Server for comms would be high risk - now prefer Cisco IP routers say 3 PRIs.

15. Escher

- ~14 staff working at Escher offices Cambridge, including 2-3 imported from Pathway for BA/POCL.

WITN05970142

- Escher have several different rooms located in communal office building - Suite 2200 One Kendal Square. The building is used by large number of hi-tech companies formed out of MIT/Harvard.
- Mike Murphy tends to work from home or globetrots (did not seem familiar with the latest offices layout)
- Drew Sutherland splits time between home and Cambridge office

16. Demonstrations Given

- Demo of Message Spy/Object Browser tools.
- Saw ISDN (PRI and BRI) line simulators etc. Test lab is set up consisting of a number of Compaq servers, LAN etc.
- Demo of Riposte Cashier - screen based transaction definition tool, allows simple transactions to be added, including effect on ledgers, ensuring session ends balanced, etc.

Papers Received:

- Matisse brochures
- "*Distributed Algorithms*" by Nancy Lynch (Draft of Ch1-4 and Ch5-6)
- "*Small-Depth Counting Networks and Related Topics*" by Michael R Klugerman (MIT Research Paper)

Jeremy Folkes, 29 Jan 1996