NORTHROP GRUMMAN

Engineering Report: Lantern Performance and Scalability

17 July 2007

Programme Name: IDENT1
Contract No.: 3333
Document Set: Project Management
Document ID: CCN014R2-020.2-1.0

Prepared for:
National Policing Improvement Agency (NPIA)

Prepared by:
Northrop Grumman Information Technology

Approval Signature

Engineering Report: Lantern Performance and Scalability

17 July 2007

Programme Name: IDENT1
Contract No.: 3333
Document Set: Project Management
Document ID: CCN014R2-020.2-1.0

Prepared for:
National Policing Improvement Agency (NPIA)

Prepared by:
Northrop Grumman Information Technology

Prepared by:
Lee Whitney

Approved by:
Jerry Munizza
Development Engineering Manager

Approved by:
Daryl Salmons
Systems Engineering Manager

Approved by:
Rod Forry
Programme Manager

Approved by:
Mark Kieffer
System Assurance Manager

1

Revision History

Date	Document ID	Revision/Change Description
17 July 2007	CCN014R2-020.2-1.0	Initial release for NPIA comments and/or acceptance

Table of Contents

List of Figures

List of Tables

1. Introduction

Requirement Pilot176 states: “The pilot architecture will be used to create a performance model of the behaviour of the system under increasing loads and associated cost implications as the usage of the capability rises.”

This report contains a description of the required performance model and the results of analysis to predict the behaviour of Lantern under increasing loads. The time-variant nature of usage during the pilot is an important consideration in this prediction. Empirical data from the pilot operation is analysed to gain insight into the time-variance of search rate. Actual performance results from this usage are evaluated as a basis for scaling up capacity to handle more users.

Section 3 of this report describes the architecture as it influences the capacity of the Lantern system and its performance. It identifies attributes of the architecture that define its scalability.

Section 4 presents empirical data collected during the pilot that indicates actual patterns of Lantern usage. Long-term variations are described in terms of a ratio of peak hour to average hour usage. This is applicable to establishment of an hourly capacity specification. Short-term variations are also described through analysis of peak usage over shorter 5 minute periods. This guides the sizing of the search rate needed to meet the 5 minute response time requirement given a specified hourly volume.

Section 5 reports the results of an analysis of the estimated capacity of the pilot system to process a burst of search request transactions without exceeding the response time requirement for any of them. This is in response to a question from NPIA.

Section 6 extends the foregoing analysis and actual performance results to model the expected behaviour of Lantern augmented with more capacity and servicing larger workloads. This section develops guidance for procurement of additional Mobile Fingerprint Readers (MFR) and recommends the volume that will likely produce search volume that requires added capacity at Central. It also describes the implementation of such capacity increases in stages as needed.

Section 7 summarises the conclusions and recommendations for consideration in defining the requirements for a post-pilot Lantern capability. Qualitative cost implications are described, although quantitative cost figures are outside the scope of this document.

1

2 # 2. Reference Documents

Document	Document ID	Author	Date Issued
Lantern Accuracy Analysis Engineering Report	CCN014R2-020.1-2.0	Northrop Grumman	23 February 2007

3

3. Lantern Performance Architecture

This section describes the architecture as it influences the search capacity of the Lantern system and its performance. It identifies attributes of the architecture that define scalability, noting in particular where parallel paths provide redundancy and the ability to share load.

3.1 Response Time Requirement

Requirement Pilot157.1 targets a 5 minute response time at 2 searches per hour per MFR, i.e., 200 per hour for the 100 MFRs deployed in the pilot. This combined response time/capacity requirement is the only quantitative performance requirement imposed on the Lantern pilot system. It formed the basis of design in determining sizing of capacity for the pilot. Higher capacities of 400 and 700 searches per hour for all 100 MFRs were also analysed and optional cost data was provided to PITO in accordance with Pilot158, but these options have not been exercised for the pilot as of this writing.

3.2 Identification of Potential Bottlenecks

The architecture of Lantern is scalable and contains various nodes where the impact of usage will be felt as it reaches higher levels. The nodes in the Lantern Central architecture are shown in Figure 3-1 as they appear in the pilot configuration. Upgrades to increase capacity can be implemented at any of these nodes if it becomes a bottleneck limiting performance. The more MFRs are added, the more usage is likely to increase on the average. The MFRs drive the usage, but each one only processes its own searches; so the aggregate MFR capacity increases with usage as the MFR quantity increases. Therefore, they are not considered to be bottlenecks.

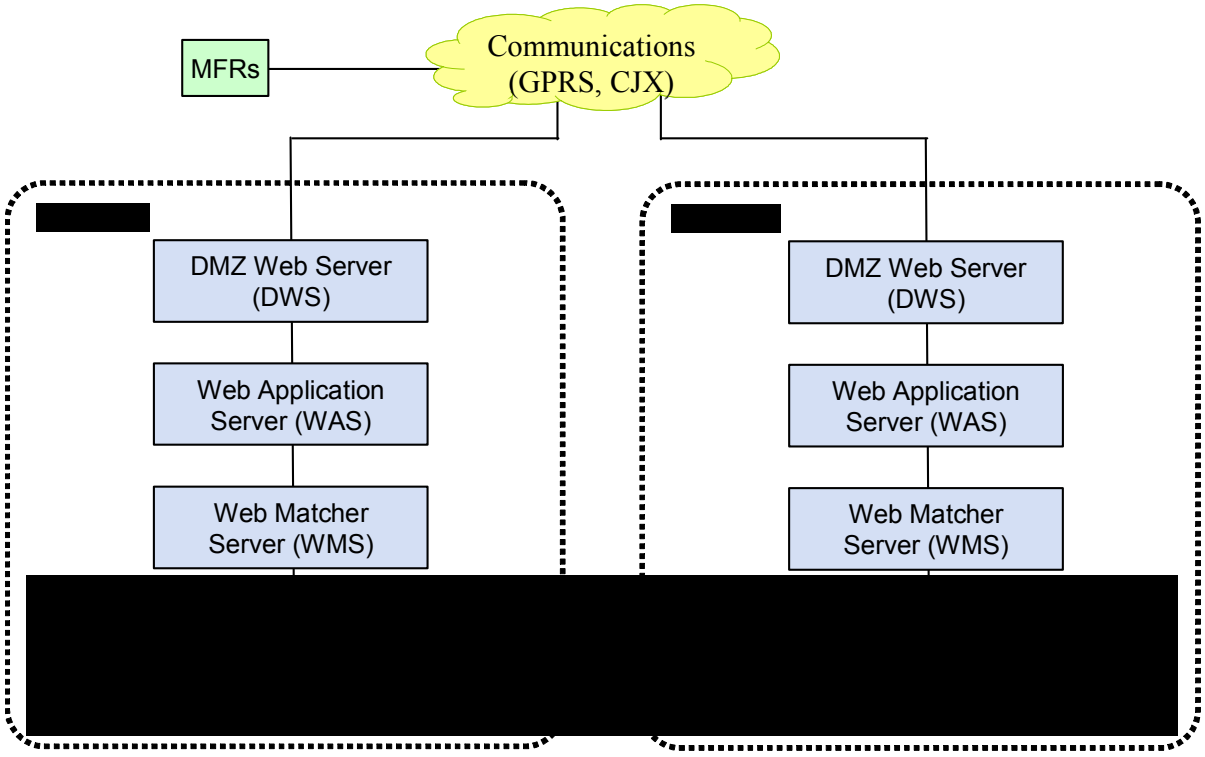


Figure 3-1 Lantern Pilot Central Architecture

The Lantern system may be thought of as a chain only as strong as its weakest link. Thus, if a bottleneck is modified to accommodate increased usage, another node may emerge as a new

bottleneck. The Lantern pilot system was purposely sized such that the factor limiting the rate of search completions was the number of matchers, because they are a cost-driving element of the architecture. The objective was to avoid over-designing the matching capacity while limiting the search rate by under-designing some less cost-sensitive node.

Projecting the capacity required for processing Lantern searches expected from various quantities of deployed MFRs depends on several factors:

- Establishing the desired response time (defined by maximum, median, mean, and/or variance)
- Estimating the time distribution of search submittals (how "bursty" it is)
- Identifying which sub-processes incur delay or limit throughput rate at each step in the process, from search submittal by the MFR to response arrival at the same MFR
- Understanding the parallelism of the Lantern architecture and its ability to process more than one search simultaneously
- Considering the operational missions that Lantern supports and the rate of searches resulting from them
- Determining the percentage of deployed MFRs actually in use.

The impact of bottlenecks is mitigated by the parallelism of the Lantern architecture. Concurrent processing of searches occurs in parallel paths using identical banks of matchers at [redacted] and at the duplicate site in [redacted]. They are designed to share load as well as provide continuity of operation in the event of failure of an entire site.

4. Empirical Data from Pilot Usage

This section presents empirical data collected during the pilot that indicates actual usage of Lantern and resulting performance statistics.

The system was sized to handle searches from 100 MFRs simultaneously, each submitting 2 per hour for a total of 200 per hour. Actual usage did not reach this rate for any hour, so the ultimate capacity of the Lantern model-based sizing has not been conclusively verified. However, the corollary estimate of response time that was done prior to Lantern implementation proved to be reasonably close to that observed.

The empirical data provides two important types of information for estimating the needed capacity of a Lantern configuration with more MFRs:

- It gives insight into actual average search volumes per MFR at least for the pilot forces
- It examines the distribution of searches over time to improve the ability to predict future peak search rates that may occur during the busiest minutes of the hour. Long-term and short-term variations are each examined separately, and their effect on specification of capacity requirements is described.

4.1 Volume of Police Usage

Actual usage during the pilot operation has been regularly reported to NPIA (formerly to PITO) since the Lantern rollout to the first pilot force. Table 4-1 documents the go-live date for start of operational usage at each of the ten pilot forces. The initial deployment of 93 MFRs did not include all 100 available. The remainder have been used on an ad hoc basis for demonstrations, special events, and subsequent force allocations in response to usage. Spare MFRs were also supplied to allow quick replacement of units that need repair, but spares are not counted in this analysis because they are not used simultaneously with operationally deployed units. For the purpose of the analysis, a quantity of 100 active MFRs is assumed.

Table 4-1 Actual Go-live Date for Rollout to Each Force

Force	Go-live Date	Number of MFRs
Bedfordshire	31 October 2006	7
West Midlands	16 November 2006	10
London Metropolitan	21 November 2006	12
British Transport Police	28 November 2006	8
West Yorkshire	5 December 2006	10
Lancashire	12 December 2006	12
North Wales	14 December 2006	10
Essex	16 January 2007	10
Hertfordshire	23 January 2007	8
Northampton	25 January 2007	6
TOTAL INITIAL DEPLOYMENT	**AS OF 24 APRIL 2007**	**93**

In order to visualise the characteristics of Lantern usage, data was gathered for a period of operation of almost 12 weeks, starting just after the last force went live, February 1, 2007 and concluding on the date of this writing, April 24, 2007. This period is believed to be representative of Lantern user behaviour during the pilot because it is a prolonged span (83 days), and it omits atypical ramp-up activities such as training and test searches.

Figure 4-1 documents the actual quantity of search requests received each day during the period to show short-term and long-term variations.

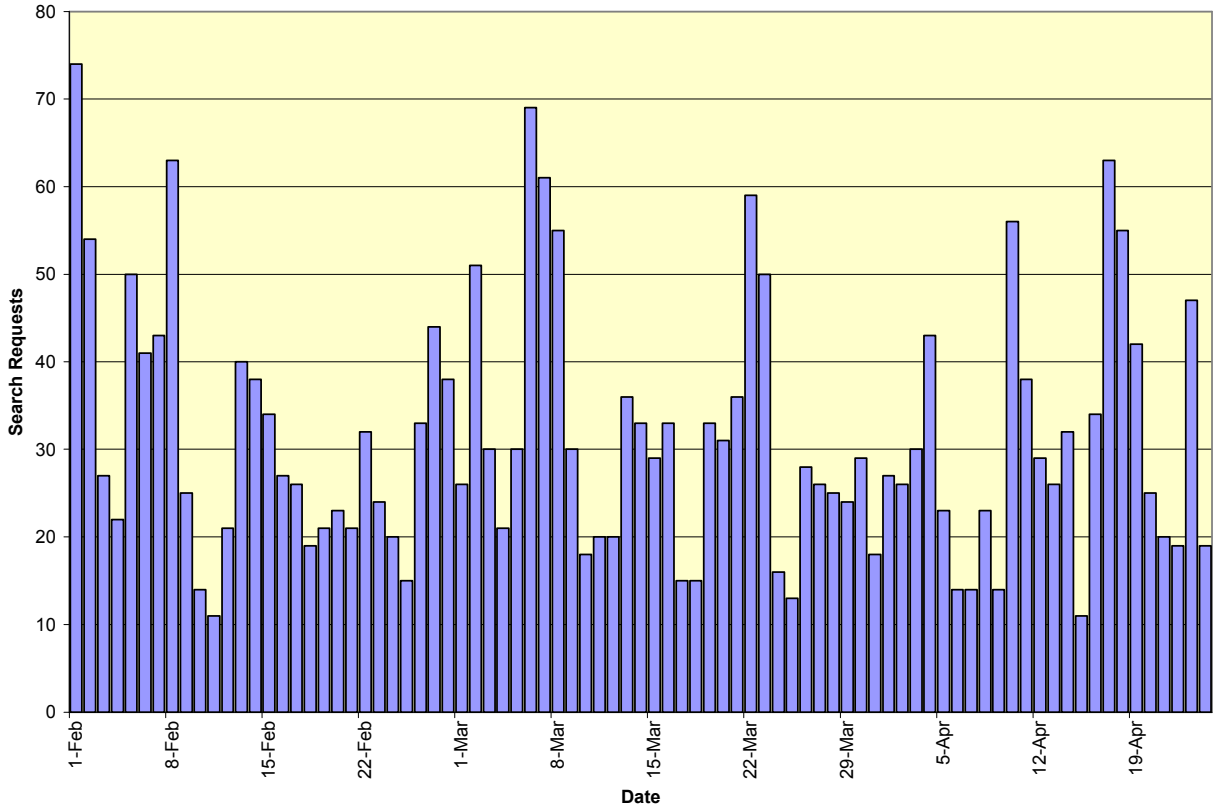


Figure 4-1 Daily Search Requests in Lantern Pilot

4.2 Long-term Variability of Search Rate

It is notable that the daily variation shown in Figure 4-1 is substantial, probably reflecting the use of Lantern pilot capability for specific exercises planned by a force and involving a number of officers and MFRs for a limited period of time rather than prolonged use in the course of all police work. While the mean daily search rate during this period was around 31 per day, on 4 days it exceeded double that number. In addition to apparently random variations, cyclic variations are obvious on a weekly cycle and a daily cycle. Variations that occur on these time scales cause peak hours that need to be considered when specifying hourly capacity for future Lantern configurations with more MFRs and higher expected workloads. Hourly capacity should be specified to be substantially higher than the expected average. This is explained below.

4.2.1 Weekly Cycle

A weekly cycle can be seen in Figure 4-1, particularly by noting the daily volume on weekends tends to be less than during the work week. This weekly cycle is even more clearly visible in Figure 4-2, which is a polar or “radar” plot of the same data. The cumulative count

of searches on each day of the week throughout the period is indicated by the plot's distance from the origin. Tuesday, Wednesday, and Thursday extend the farthest from the centre, showing that these mid-week days are typically the busiest. Usage decreases markedly on the weekends. This might change to some degree in the post-pilot era as the use of Lantern becomes more integrated into a wider range of policing, but this cannot be assumed.

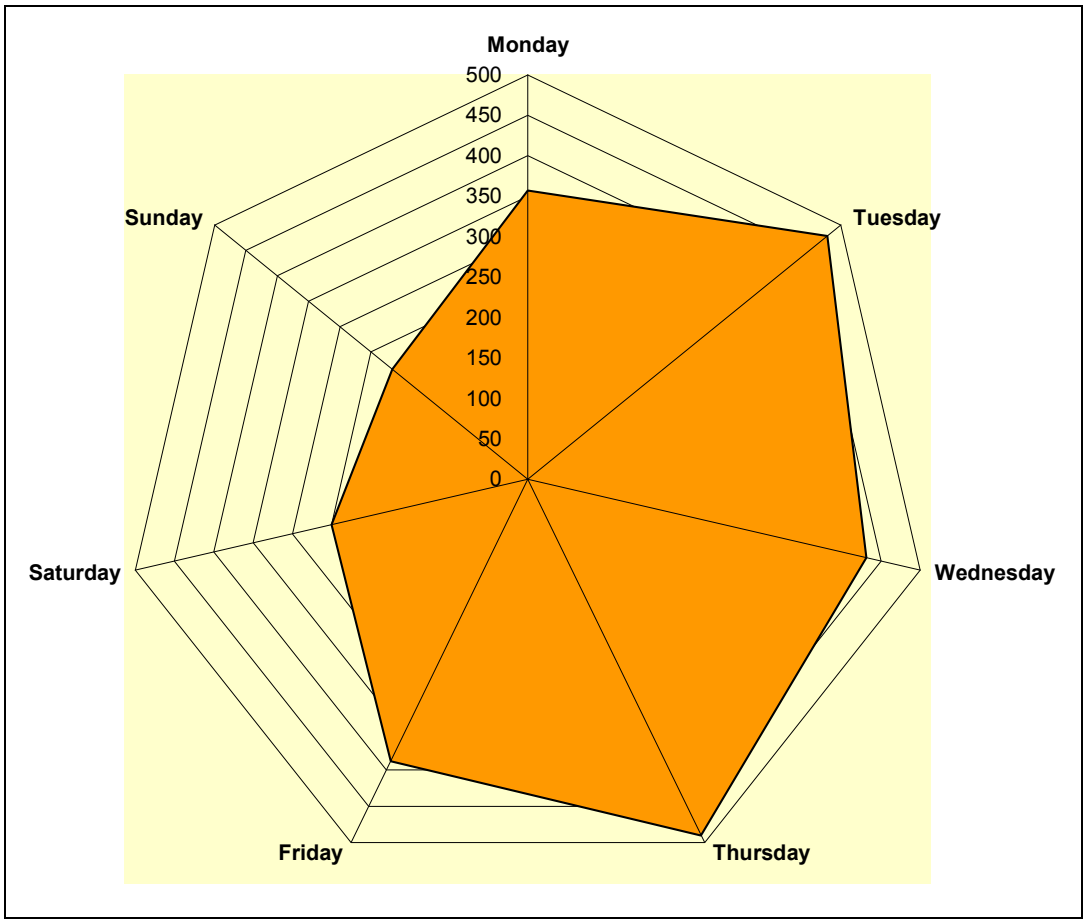


Figure 4-2 Weekly Search Volume Variations by Day

4.2.2 Daily Cycle

In order to peer deeper into the variability of search submittal rate during the pilot, Figure 4-3 was prepared to show the number of searches for each hour of the day around the clock cumulatively during the period. On this time scale, a cyclic pattern is also quite clear. Most of the searches occur during the day shift between 10:00 a.m. and 5:00 p.m. Few occur between 2:00 a.m. and 6:00 a.m.

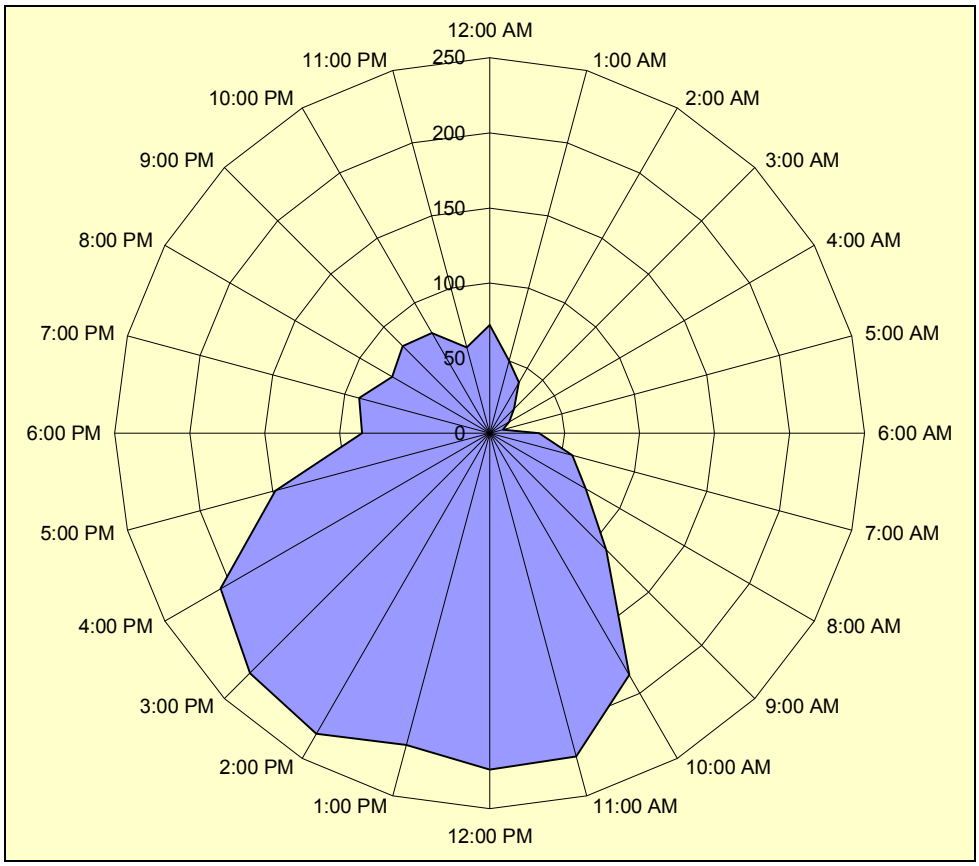


Figure 4-3 Daily Search Volume Variations by Hour (on 24-hour clock)

4.2.3 Long-term Peak Factor

The weekly and daily search rate variations described above are likely to continue in somewhat similar ways as Lantern operation matures and grows. If search rate capacity is specified as a number of searches per hour, variations at these longer time scales should be considered in setting the specification for hourly capacity. The capacity should be high enough to handle most peak hours, not just for the average search rate. If a peak factor that accounts for these long-term variations in submittal rate is built into the requirement, it becomes less likely that exceptionally busy hours of the day or the week will cause queues to build up and response times to grow. Although the peak factor needed for future operation cannot be predicted with certainty, applicable estimates based on the observed pilot operation are summarised below:

- Peak day of the period: total searches = 2,610 in 83 days of data (31.4 searches/day average) with 74 on the peak day. Peak factor to allow for busiest day of the period = 74/31.4 = 2.4
- Peak hour of the day: total searches = 2,610 in all 24 hours of the day (an average of 108.8 searches in each hour) with 231 searches in the busiest hour (2:00-3:00 p.m.). Peak factor to allow for busiest hour = 231/108.8 = 2.1.

Combining these two peak factors by simply multiplying them is a simple yet safe method of establishing an aggregate peak factor. It intrinsically allows for the busiest hour of the busiest day. It relates the average hourly search rate to the peak hourly rates expected in actual use.

That factor would be 2.4 × 2.1 = 5.0. In other words, peak hours are typically five times busier than the average hour.

If the hourly search rate capacity requirement is specified to be sufficient for a peak hour 5 times higher than an average hour, then an average hour is likely to only use one-fifth or 20 percent of the capacity and some hours even less. This may seem wasteful on the surface, but if capacity is specified just for the average hour, then any hour when usage is above average will overtax the capacity and many searches may incur delays.

This methodology is inherently geared toward a worst-case scenario. A more pragmatic treatment of the peak factor might conclude that a multiplier somewhat lower than five is acceptable based on willingness to allow response times to grow during high usage hours. However, this incurs risk of user dissatisfaction if delays are not tolerable during important police operations. Therefore, the empirically determined peak factor of five will be used to relate the average hourly rates to the capacity required to handle peak hours.

4.2.4 Hourly Capacity Actually Used

The observed variability in search rate suggests a question about how much of the provided capacity was actually used in peak hours. Answering this requires examining the hours with the largest numbers of searches, how busy they were, and how often they occurred.

Figure 4-4 is a graph of the quantity of occurrence of busy hours during the course of the 83 days of data, and the volume of searches in those hours. Hours are defined as 60 minute periods beginning at any time, not just those that start at the top of the clock. Busy hours are defined arbitrarily as those with 12 or more searches requested. Hours with 12 searches occurred 113 times, many of which overlapped. The busiest hour had 22 searches. That occurred only once. With a capacity of 200 searches per hour required, the actual hourly load never exceeded 11 percent of the hourly requirement. However, this is true only when viewed on an hourly basis. Search request rate is not uniform during the hour. So when viewed with finer granularity, as in the next section, it becomes apparent that short-term peaks in search rate occur within the hour and additional peak capabilities may be needed to handle them.

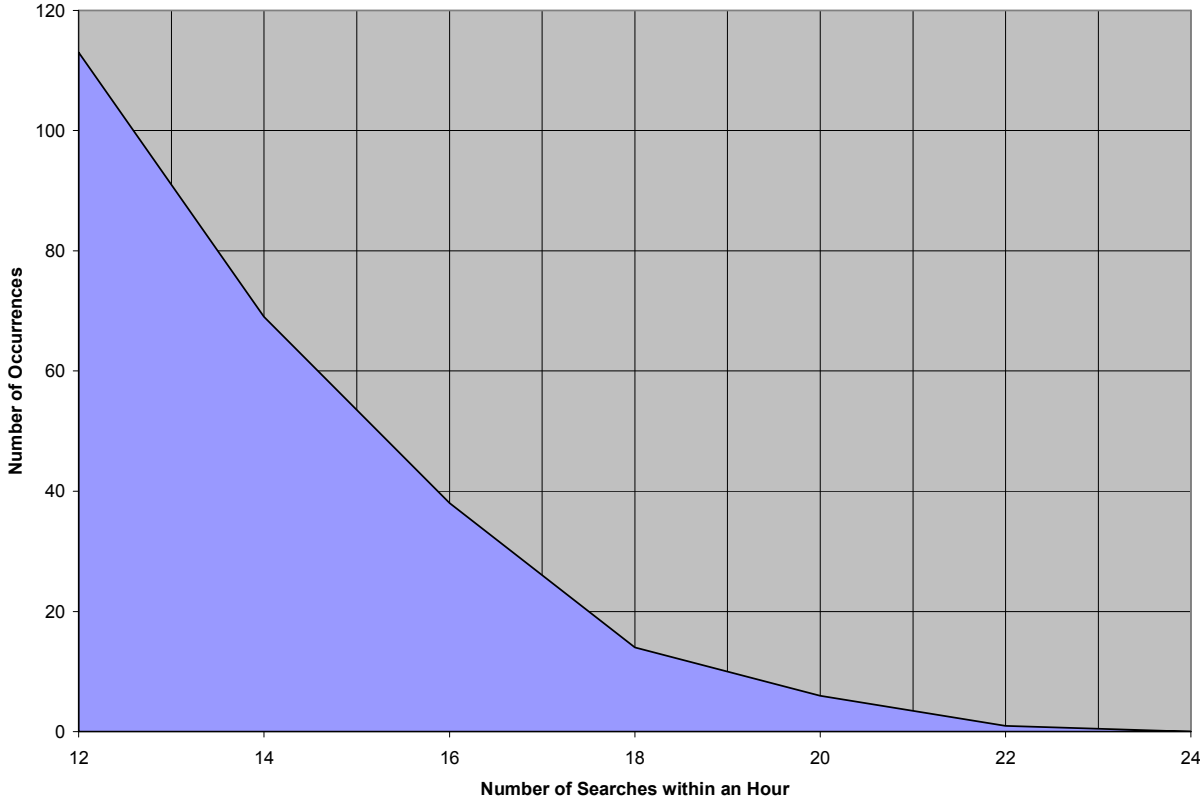


Figure 4-4 Busiest Hours

4.3 Short-term Variability in Search Rate

The search submittal rate averaged over an hour is a convenient way to specify workload, but it does not fully define the capacity needed in order to keep up with typical search request input and avoid queuing delays. Since the specified target search response time is 5 minutes, and typical response times during normal operation have been observed to be 2 minutes or less, it is important to know what percentage of an hour's search requests may be expected not only during the busiest hour, but also during the busiest minutes of that hour or any hour.

Hypothetically, if the specified 200 searches/hour were uniformly distributed, there would be one every 18 seconds, but it is possible for a burst of searches to be submitted within a shorter period. Analysis of actual search request submission time data from pilot operation allows a measure of the non-uniformity of search rate demands within an hour. This produced the following results.

There were 2,610 searches in 83 days, so the mean rate was 2610 / (83 × 24) = 1.31 per hour. This would be just over 0.1 searches in each 5 minute period if they had occurred uniformly. However, they did not occur uniformly. In reality, many 5-minute periods had none; and the busiest 5-minute periods had up to 9 searches as recorded below:

- 191 instances with three searches submitted within 5 minutes, including:
  - 25 instances of four searches within 5 minutes
  - 9 instances of five searches within 5 minutes
  - 3 instances of six searches within 5 minutes
  - 1 instance of seven searches within 5 minutes
  - 1 instance of nine searches within a 5-minute period.

This means that during the entire 83 days of data there were only 14 instances when 5 or more searches were sent within a 5 minute period. Some of these instances may have been ad hoc tests or training even though most of that activity was finished before the data period began. Other submissions may have been repeated due to perceived response problems. Regardless, it illustrates how short-term variations can occasionally place extraordinary demands on the search rate capacity.

5. Burst Analysis

This section reports on an analysis that estimated the Lantern pilot system can theoretically handle up to 19 simultaneous search requests completing them all within the 5 minute response time requirement. This analysis was undertaken to examine the theoretical response time capability of Lantern under peak loads.

The method employed was to calculate the system response to a burst of searches tracked through the modelled search process for 5 minutes. The largest burst that can be entirely finished within 5 minutes is the search rate burst capacity, and represents performance under the worst case of searches arriving all at once rather than distributed more uniformly over time. Designing for a uniform arrival rate would be overly optimistic because, as shown in the previous section, real world operations do produce "bursty" arrivals.

5.1 Background

A question was raised by PITO during the Operational Readiness Review about the peak number of searches that could be handled by Lantern with all of them still meeting the 5 minute response time requirement. This is particularly of interest because actual Lantern search rates peak during short bursts of intense activity in police operations.

It would be difficult to ascertain the peak search capability empirically using pilot resources because it would require submitting a coordinated burst of search requests simultaneously from a number of MFRs with controllable GPRS connectivity, and measuring the results. Approximately 100 MFRs are deployed, but they are allocated to geographically dispersed forces that use four different GPRS providers whose instantaneous performance cannot be controlled. Hence, this theoretical analysis was devised to predict the maximum capacity for simultaneous searches that could be completed within 5 minutes under ideal conditions. It should be emphasised that ideal conditions did not exist during generation of the empirical data analysed in the preceding section, and may never exist; but their assumption is nonetheless useful for establishing an upper bound.

In this analysis, a hypothetical scenario was developed to represent the conditions of simultaneous searches all completing within the same 5-minute period. Simply counting search completions for a 5-minute period of steady state operation was considered to be too optimistic because it necessarily would include searches that started before the period began. Instead, this analysis determined the maximum number of search requests in an instantaneous impulse that would meet the condition that all of them are finished 5 minutes later.

The Lantern Central architecture has the capability to queue searches and direct them to search engines as these critical resources become available, at a rate limited only by how fast they can complete them. The search engines in the pilot comprise four parallel banks of matchers, each bank able to process a different series of searches at the same time.

The Lantern architecture is also characterized by a minimum time for a lone search transaction to propagate through the system from request to response. This is defined by a critical path through the processing of a single search in isolation without queuing delays.

This analysis postulates that, with an instantaneous impulse of input to a dormant system, the output will be nothing until the propagation time has elapsed and the first search comes out. Thereafter, it will continue for the rest of the period at the throughput rate of the slowest bottleneck. By design, that bottleneck is the search engines.

5.2 Requirements

The Lantern pilot requirements that relate to search capacity and response time (cited in Section 3.1) are reviewed and restated as follows:

- Peak input required is 200 searches/hour = 18 seconds/search (optionally 400 or 700 searches per hour)
- End-to-End response time required = 5 minutes or less
  - With only one search at a time active on any MFR
  - Assumed to include communications links (GPRS and CJX both ways).

The number sought by this burst analysis is not a Lantern requirement.

5.3 Search Engine Sizing

The timing and rate data used in this analysis was built on a foundation of model results presented at the Lantern CDR, validated and refined using actual measurements where available from the initial Lantern pilot operations.

The Lantern capacity is dominated by the search engines. The deployed search engine sizing is based on the 200 searches/hour option and was described at the CDR as follows:

- PerfSim model parameters were adapted to reflect Lantern search load
- Analysis was based on performance runs using Alpha software version (complete implementation of parsing, storage, feature extraction, search preparation, and results formatting software)
- Regardless of workload, each bank of matchers needs 12 partitions (i.e., three 4-CPU quad matchers) to contain the full national database
- A 2-finger Lantern search by one bank of matchers takes a predicted [redacted] seconds
- 200 searches/hr. (18 seconds/search) throughput requires [redacted] seconds/18 seconds = [redacted] banks
- At least [redacted] (integer) banks needed to meet [redacted] bank throughput
- [redacted] banks were actually provided - configured as [redacted] banks per site (2 symmetrical sites)
- Therefore, the installed pilot configuration of [redacted] banks should finish a search every [redacted] = [redacted] seconds on average.

Matcher complements required for this throughput and for optional higher capacities are summarised in Table 5-1

Table 5-1 Matcher Sizing Summary

Searches /hour Requirement	Banks Needed	Matchers Needed	Spares Included	Total Matchers	Total Banks
200	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
400	[redacted]	[redacted]	n/a	n/a	n/a
700	[redacted]	[redacted]	n/a	n/a	n/a

In addition to the [redacted] second matcher time, other components of the search process were predicted and reported at CDR, although the matching was the dominant contributor. Others included feature extraction, polling interval, and message transmission time over GPRS and CJX networks, etc.

The GPRS estimate, while significant, was based on the condition that the link was solid and no significant number of packets were dropped nor were retransmissions needed. This was the biggest unknown at the outset because the geography of each of the forces and their choice of GPRS carrier are different. Other relevant factors such as GPRS delays caused by a busy network were also uncontrollable and assumed to be not a factor in system response time predictions.

5.4 Analysis Scenario: Impulse of Search Requests

The hypothetical scenario used for this analysis is summarized below. Details of the estimation process for each scenario item are in Section 5.5.

- Postulates an impulse containing an unknown number of search requests sent at the same instant
- Prior to the impulse, no queues exist and all Lantern resources are idle
- Assumes for the calculation that no errors or dropped links occur, so no retransmissions add to the time
- Result is the number of searches that can be in the impulse for the last one to finish within 5 minutes
- If more searches come in after the impulse, by definition they will not finish within 5 minutes after the impulse. Lantern does not have a prioritisation scheme that could cause some searches to be processed before other searches that arrived earlier. Later arrivals therefore are irrelevant to the timing of searches in the impulse
- Matching is the defining throughput rate "bottleneck" because the matchers are predicted to take [redacted] seconds per bank, longer than any other concurrent process; and the number of banks, being a cost driver, was sized at the minimum needed for the 200 search/hour throughput plus a safety margin
- [redacted] banks of matchers each able to do a search in [redacted] seconds => [redacted] every [redacted] seconds. => [redacted] searches finish every 5 minutes (300 seconds) in a steady state
- Each search has to run the gauntlet of one-time delays totalling an estimated [redacted] seconds. So, the first search is completed [redacted] seconds after its request is sent. After that, search results pour out of the matchers one every [redacted] seconds.
- Therefore, after 5 minutes (300-[redacted]=[redacted] searches have been completed.

5.5 Estimation of One-time Delays and Rate at Bottleneck

Table 5-2 shows each significant step in the search thread and its predicted impact on timing. The first column shows steps executed on the MFR, while the second column shows steps executed at Central sites. Two Central sites together must process the searches submitted by all MFRs; in this case 100. Each serial step in the thread adds to the search time no matter how many resources can process searches concurrently.

This is shown in the Step Impact on Delay column. Some steps are off the critical path and do not add to the total search time because they occur concurrently with other steps. From a

different perspective, each step has the potential to become the bottleneck that limits throughput rate causing queues to build up as the volume increases. This is shown as Step Impact on 5-Minute Output Capacity.

Table 5-2 Impact of Each Processing Step in Sequence

PROCESSING STEP		STEP IMPACT	
At MFR (100 Units Operating)	**At Central (2 Sites Sharing the Load)**	**On Delay (Propagation)**	**On 5-Minute Output Capacity**
Sends connection request. Receives connection response		Precedes search thread. No impact	Precedes search thread. No impact
Captures two index finger prints		Precedes search thread. No impact	Precedes search thread. No impact
Sends Search Request containing (NIST Criminal Print to Print Search [CPS] message) via GPRS link		25 seconds. assuming good connection-no retries needed	100 deployed MFRs support 100/25=4 search requests per second (other MFR comms add negligible load). Not the bottleneck
	DMZ Web Server (DWS) receives and logs incoming transaction from CJX	2096 kbps into each site. Search request = 30kB. Full utilisation = 0.05725 seconds/search	17.5 search requests/seconds or 5,240 in 5 minutes both sites balanced—not the bottleneck
	Scheduling/parsing by the WAS	2 seconds	2 parallel sites—not the bottleneck
	WMS performs decompression and feature extraction	[redacted] seconds ([redacted]finger × 2 fingers)	2 parallel sites—not the bottleneck
	Matchers perform template matches against national database	[redacted] seconds through one bank (modelled)	[redacted] parallel banks complete a search every [redacted] seconds average or [redacted] searches/seconds. **This is the bottleneck**
	WMS formulates report based on matcher scores (CRO, confidence, may be up to 3 respondents)	2 seconds	2 parallel sites—not the bottleneck
	WAS processing. If search result is a hit, WAS accesses AFR server to get demographics of respondent(s).	5 seconds (estimated) but only affects hits (approximately 41% of searches based on data from November 2006 through February 2007). Average 2 seconds	2 parallel sites—not the bottleneck
	DWS logs and sends outgoing transaction via CJX	2096 kbps into each site. Search response ≈ 0.5kB (est.). Full utilisation ≈ 0.002 seconds/search	2 parallel sites—not the bottleneck

PROCESSING STEP		STEP IMPACT	
At MFR (100 Units Operating)	**At Central (2 Sites Sharing the Load)**	**On Delay (Propagation)**	**On 5-Minute Output Capacity**
Receives Search Request Accepted message		Concurrent branch not additive to search thread time	Utilisation split among 100 MFRs-negligible impact on one
Waits for status poll interval , then requests results of Search Request		15 seconds concurrent with Central processing. Not additive to search thread time	Not limited in number of simultaneous searches waiting
Receives status indication of whether or not a processing error has been detected		Concurrent branch not additive to search thread time	Utilization split among 100 MFRs-negligible impact on one
Waits for response poll interval, then requests results of the search		Suggested wait depends on workload and is concurrent with Central processing	Not limited in number of simultaneous searches waiting
Receives Search Results Complete Response (containing NIST Search Response Electronic [SRE] message)		Concurrent branch not additive to search thread time	Utilisation split among 100 MFRs-negligible impact on one
Sends Response Acknowledgement		Concurrent branch not additive to search thread time	Utilisation split among 100 MFRs-negligible impact on one
	Receives Response Acknowledgement	Concurrent branch not additive to search thread time	Utilisation split among 100 MFRs-negligible impact on one
Aggregate impact		[redacted]seconds	1 every [redacted] seconds, [redacted]in 5 minutes; [redacted]excluding [redacted]seconds propagation time

5.6 Conditions

The foregoing analysis assumes that the following conditions apply:

- Estimated timing values are accurate estimates. They are consistent with the model estimates presented at CDR with the exception that the polling interval is concurrent with the Central processing (matching, etc.) not additive. Also, common empirical time observations from early Lantern operations have not conflicted with any of them.
- Polling intervals are set optimally; that is, polling occurs as soon as results are ready but needless premature polling does not sap processing or communications resources. Tuning of the algorithms that determine polling intervals can be done at any time (with PCCB concurrence) to balance the goals of fast response time and low polling overhead.
- The GPRS connectivity is solid so that no retransmissions are needed due to dropped packets, and no dropped calls. The empirical data shows that this is generally a good assumption.

- AFR access time to obtain demographics (5 seconds) is short compared with the search, and is optional for search completion in any event. A search can be completed without it, so it cannot be the bottleneck.
- All elements of the system as actually configured are operating at both sites.

Deviations from these ideal conditions may reduce the number of searches that can be processed in 5 minutes.

6. Behaviour with Larger Workloads

This section extends the foregoing analyses to model the behaviour of Lantern when augmented with more capacity and servicing larger workloads. The key determination is that, based on usage patterns in the pilot, up to three times the current number of MFRs can be accommodated with confidence before it becomes necessary to add capacity to Central. The derivation of this is presented in Section 6.2.

The data set available to analyze from the pilot is not very large and contains very few response times that exceeded the spec. Furthermore, many of them were not even at busy times but occurred for other reasons. Very few of the searches whose responses were delayed beyond the 5-minute response time spec even coincided with load peaks. Therefore, with the limited available data it is not feasible to derive a precise response time probability as a function of MFR quantity. So the analysis derives the estimated probability and draws conclusions based on a combination of factual analysis and judgements about risk tolerance in Lantern operation.

The pilot requirement for search rate was based on an estimated two searches/hour/MFR. The Lantern Central capacity was sized to ensure that. Actual pilot usage on average has been much less. It certainly appears that more usage from more MFRs could be handled without adding capacity, but how much more?

The busiest hour of the whole period had only 22 search requests. One could simply say that if 100 MFRs produced 22, then on average each produced only 0.22 searches in that hour.

However, the maximum response time spec is 5 minutes—much less than an hour. So the hourly average is not an applicable criterion. The search engines have to keep up with a reasonable expectation of search requests in a peak 5-minute period. They cannot take the rest of the hour to catch up. If they did, some searches would have to wait in a queue, and the response time would more often be exceeded.

There are [redacted] banks of matchers, each of which can do searches independently of the others and simultaneously share the load. Therefore, the analysis focused on how often more than [redacted] search requests came in at about the same time, for these purposes within 5 minutes.

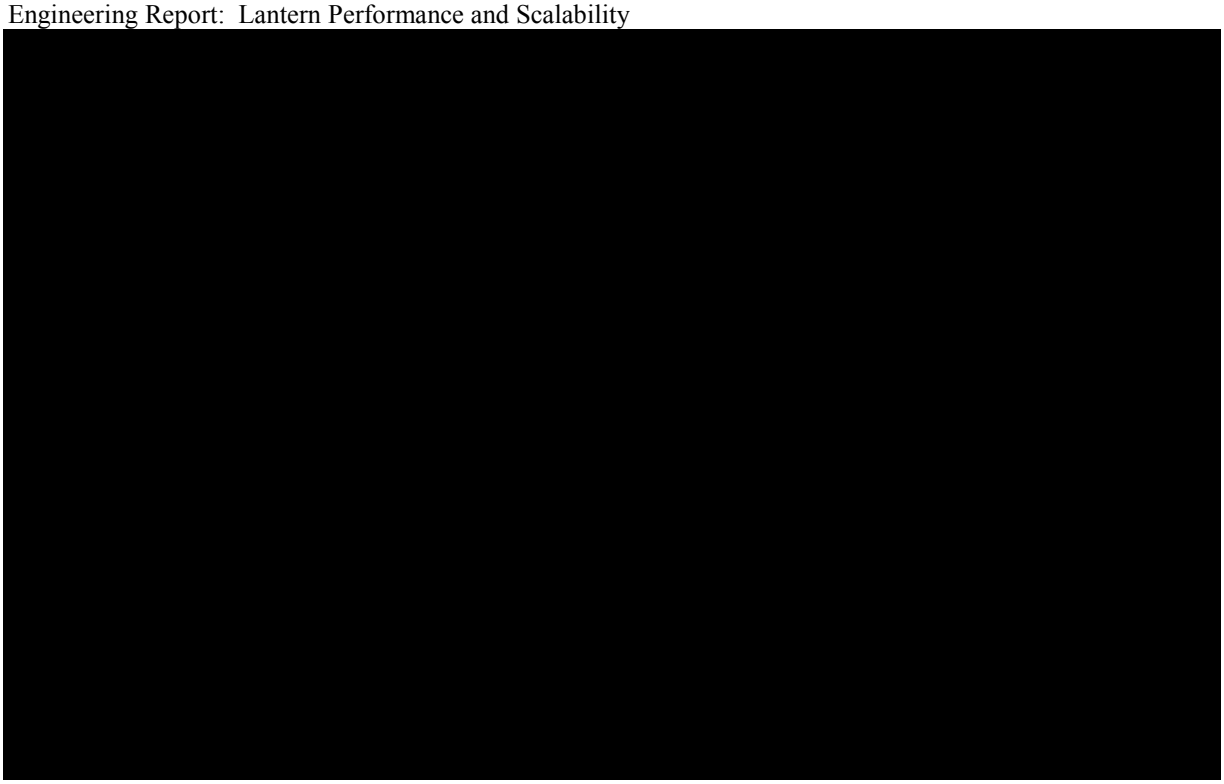
Based on actual pilot data, this occurred only 14 times during the 83 days of data. The average was [redacted] searches in these 14 busiest 5-minute periods. So if the capacity can handle that, it should be adequate most of the time.

Instead of [redacted] searches per 5 minute period, the specified capacity is 16.7 searches in a 5 minute period (200 per hour × 5/60 of an hour). So it can handle about [redacted] times what it has been getting from 100 MFRs in all but the 14 busiest times [redacted].

The conclusion of the analysis is that up to [redacted] times the current MFR count) can be deployed without adding to the Central capacity. The details and methodology are described in the following sections.

6.1 Response Time Actuals

Figure 6-1 is a cumulative distribution graph of response time data from actual Lantern pilot operation that has been shown at Program Status Reviews (PSR). It shows that each month, 94 percent to 96 percent of searches finished in less than the target 5 minutes, and 70 to 85 percent finished in less than 2 minutes.



The way that response times are distributed in Figure 6-1 provides a check on previous predictions. The response time distribution for each month consistently shows the median end-to-end search time (50 percent above and 50 percent below) to be on a plateau occupied by a large percentage of searches, so it is typical. It is [redacted] minutes or [redacted] seconds for a single search. This empirical data is in reasonably close agreement with the prediction of [redacted] seconds given at the CDR and the refined [redacted] seconds derived in the burst analysis in Section 5.5.

When search requests are received at a rate less than or equal to the maximum capacity, as in the pilot data, the response time is stable, though not invariant. (Even if the arrival rate was perfectly uniform, random variations in the actual search duration naturally occur due to the matching process, memory conditions, etc.) If higher average arrival rates are sustained, the response time becomes unstable and rises without bound as queues build.

Of course in reality, arrival time is not perfectly uniform at an average rate but has a strong variable component. It is even more bursty than perfectly random (Poisson) arrivals would be due to the nature of the exercises in which Lantern is used. So, avoidance of queuing delays requires a processing rate margin well above the steady state average rate.

The actual capacity of the Central configuration that was fielded for the Lantern pilot is somewhat greater than the required 200 searches/hour because of rounding the number of matcher banks up to an equal integer at each site. As explained in Section 5.3, the actual search engine sizing provides a sustained throughput capability of:

- One search every [redacted] seconds (steady state)
- 3600[redacted] searches/hour (compared to the requirement of 200)
- [redacted] searches/hour × 24 hours/day × 30 days/month = [redacted] searches/month.

In contrast, the monthly volumes on which the response times in Figure 6-1 were attained never exceeded 1,973, which is less than 1 percent of the [redacted] searches/month capacity. Therefore, the pilot response times are not indicative of what they would be if loads were greater. In the pilot, searches seldom had to queue waiting for service because the designed rate was exceeded only on rare occasions and then only briefly. So the system capacity has not been seriously stressed during pilot operation.

This is borne out by examining actual response times[1] for the searches that arrived in bursts (the bursts introduced in Section 4.3). The Lantern pilot uses four parallel banks of matchers that can process four searches simultaneously. In the infrequent instances when more than four searches were requested in rapid succession, some may have had to wait for available matching resources. The outcomes below show the variability of response times for search requests that arrive in bursts:

- On 18 April 2007, five searches were submitted within 5 minutes, 4 of them from the same MFR. All responses came back in less than 5 minutes. Queuing obviously was brief and the returns met the 5-minute requirement despite the burst.
- In a 9-search burst on 17 April 2007, searches were submitted from 3 different MFRs, all 9 within about 2 minutes, and the response times were less than 5 minutes for 6 of the searches. The other 3 took between 6.8 and 8.1 minutes.
- Also on 17 April 2007, 3 MFRs submitted 5 searches in 5 minutes, all of which completed in less than 2 minutes.
- On 16 April 2007, seven searches were submitted from 2 MFRs within 2 minutes. Each completed with a response time of less than 3 minutes.
- On 21 March 2007, 5 searches were submitted in a 5-minute period from 4 different MFRs, resulting in 4 responses in less than 2 minutes and 1 timeout without a response.
- On 13 March 2007, 5 searches from 5 MFRs were submitted within 5 minutes. Three of them finished in less than 3 minutes but the other 2 timed out due to unknown problems.
- On 7 March 2007, 5 searches were submitted from 5 different MFRs within 5 minutes. Four finished in less than 2 minutes but the 5th took almost 3 hours.
- On 2 March 2007, after 2 consecutive bursts of 4 searches, 5 more were submitted from 5 different MFRs within 5 minutes. All but 1 finished within 5 minutes.
- On 15 February 2007, 5 searches submitted from 5 different MFRs within 5 minutes all returned a response in less than 5 minutes.
- On 8 February 2007, 6 searches from 1 MFR all submitted within 4 minutes each returned a response in less than 3 minutes.
- On 7 February 2007, 5 searches from 1 MFR were submitted within just over 2 minutes and all completed within 3 minutes.

[1] Response times are end-to-end including GPRS paths and Central processing as measured by the difference between search request and response acknowledgement times, both on the same MFR clock.

- On 5 February 2007, 5 out of 6 searches from 2 MFRs submitted within 4 minutes of each other took longer than 5 minutes to return, the longest two taking over 20 minutes.
- Also on 5 February 2007, 5 searches were submitted within 3 minutes from one MFR and only the first 3 finished within 5 minutes; the other 2 took over an hour.
- On 1 February 2007, 6 searches from 1 MFR were submitted within 3 minutes but did not receive responses. Therefore, there is no response time value. Searches from other MFRs submitted just before and after were getting responses. So, it appears that this MFR was shut down or had a unique problem, possibly a lost GPRS or VPN connection, after it sent a burst of search requests.

These few examples comprise all bursts of 5 or more searches within 5 minutes, and are the most extraordinary instances of bursts during the 83-day period. Five bursts had a total of 9 searches that exceeded the 5 minute response time requirement, neither frequent nor serious enough to warrant an increase in capacity. However, when workload is scaled up, such bursts will become more frequent and less likely to be able to be processed without some searches being delayed.

Interestingly, these bursts were not the only instances when searches took longer than 5 minutes to complete. Of the 2,610 searches requested (2,413 of which were completed) during the period, 109 had response times over 5 minutes. This is 5 percent, which agrees well with the monthly response time statistics in Figure 6-1. Only 27 of these 109 search requests arrived in bursts of 3 or more within 5 minutes. Therefore, most searches with excessive response times were not caused by heavy load conditions.

Occasional peak load demands such as these should be duly considered in the design of the Central infrastructure to ensure that its capacity has sufficient margin to meet not only average demands but some level of peaks as well. However, that is not to say that the design must be capable of peak factors to cover every possible occasion. It should be a pragmatic trade-off between peak capacity and cost. The next section provides more information on that trade-off.

6.2 Increasing Volume with the Same Matchers

As noted, there is a serious paucity of data on pilot searches that were delayed by other searches arriving in rapid succession and getting first access to the search resources. This makes it much more challenging to extrapolate the data to larger search volumes. It would be easier and less heuristic if peak loads had more often taxed the capacity and caused response times to increase.

To use the data available for estimating the capacity needed to scale up to larger workloads, the following imperfect but reasonable assumptions are necessary:

- Average Central workload as measured by mean arrival rate scales up linearly with the addition of MFRs. That implies that additional users to whom Lantern capability is deployed will settle into usage patterns similar to those of the existing ten pilot forces.

- The shape of the arrival rate distribution is not appreciably affected by the addition of large numbers of new users even though the mean arrival rate increases.[2]
- The variance of arrival rates about the mean in the pilot usage data is representative of the variance expected in post-pilot usage.
- The sharing of search workload among search engines is optimal. That is, none is idle while queues exist with search transactions are waiting to be processed.

As the number of MFRs increases, the stress on the system to keep response times below 5 minutes rises for two reasons: more searches occur in bursts that cause queuing, and queues become longer so more of the searches in the queue have to wait for shared resources before they can complete. The impact is not linear and the distribution cannot be precisely defined from the sparse pilot data.

As the data analysis in Section 4.3 concluded, with the 100 MFRs deployed, there were only 14 instances when [redacted] or more searches arrived within a 5-minute period during the entire 83 days of data. The average number was [redacted] searches within each of those 14 periods.

If considered a tolerable frequency for bursts that might tax the search capacity, then it can be related to the capacity actually specified—200 searches/hour or 16.7 searches every 5 minutes. The specified capacity is more than [redacted] times (16.7 [redacted] what the incoming search requests required in all but the 14 instances. The implication is that about [redacted] times the current number of MFRs could be used without frequently taxing the capacity or causing excessive response times. In other words, [redacted] MFRs (an additional [redacted]) could be deployed before a capacity increase at Central needs to be considered.

One way to evaluate the tolerability of that trade-off is by considering the probability that implemented capacity will keep up with search requests during any 5-minute period and meet the response time requirement. Figure 6-2 shows this graphically.

[2] The Central Limit Theorem predicts that as more and more random arrival inputs are added, the result approaches a standard normal distribution. However, this effect is insignificant at expected MFR quantities. It makes the assumption slightly pessimistic.

All 5-minute periods in the dataset were grouped by the number (N) of search request transactions that they contained, and their total instances shown as bars on the graph. Then the cumulative percentage was plotted at each value of N to show the probability that a 5-minute period contained N or fewer searches.

Two operating points are highlighted. One is for N= which represents use of the current pilot configuration for MFRs[3], but only 92.7 percent of searches meet the N= criterion. This would heighten the risk that delay of the rest of the searches could prove operationally significant. The other operating point is at N= searches, which results from using the pilot configuration for MFRs as recommended. This usage criterion describes 99.8 percent of the searches.

6.3 Increasing Volume with Added Matchers

With the addition of matcher capacity at Central, the search engines can process more searches from more MFRs. For the previously defined upgrade steps to 400 and 700 searches/hour[4], the number of MFRs can be calculated using the same logic as in the previous section. It is directly proportional and builds on the recommendation of MFRs as the limit for the current pilot matchers. With an increase from 200 to 400 searches/hour, twice the number of MFRs can be handled, that is × (400/200) = MFRs. At 700 searches/hour, × (700/200) = MFRs.

Figure 6-3 graphically shows the hourly search rate needed for Central capacity as a function of the number of MFRs deployed. It shows the capacity steps at the operating points where

[3] The probability observed for 100 MFRs submitting 2 searches in a 5-minute period is the same as 800 MFRs submitting 8 times as many or 16 searches, just within the specified 200 searches/hour or 16.7 searches per 5-minute period.

[4] These rates are arbitrary to correspond with the previous analyses. They are not recommendations.

each of the two specified upgrade options could be exercised in order to process the estimated search workload from added MFRs.

The graph also shows the estimated MFR quantity [redacted] at which the 700 search/hour Central capacity is no longer adequate. Above that, the straight line sloping upward indicates a constant proportionality. Additional operating points can be established on this line by adding matcher capacity at Central.

The line ends at 1,200 searches/hour, the point where the number of matchers is no longer the limiting factor for capacity. Operating above the line decreases the likelihood that searches arriving in rapid succession will have to queue for so long that they fail to finish within the 5 minute target. Operating below the line increases this likelihood.

6.4 Increasing Volume Beyond Added Matchers

At still higher workload levels beyond the straight line in Figure 6-3, even if more banks of matchers are added to raise capacity above 1,200 searches/hour, other nodes begin to show up as potential bottlenecks and require capacity increases. The burst analysis in Section 5 calculated the output capacity as time per search for each potential bottleneck. The significant ones are summarised below for a single Lantern Central equipment stack:

- GPRS link – Each MFR is a separate user and link bandwidth up to saturation of the cell is assumed not an issue at any envisioned MFR concentration
- CJX/DWS – 0.057 seconds/incoming search request; outgoing search responses are shorter and negligible; not a bottleneck until over 60,000 searches/hour if Lantern is the only user. Although Lantern has its own dedicated DWS, IDENT1 also shares the same CJX connection, and the combination should be assessed for network loading

- WAS – Scheduling and parsing estimated at 2 seconds/search
- WMS fingerprint image decompression and feature extraction – [redacted]second/finger = [redacted] seconds/search
- AFR query for demographics – 5 seconds/search on 41 percent of searches = 2 seconds/search on the average.

With a processing time of [redacted] seconds per search, the WAS, WMS, and the AFR connection become saturated at [redacted]searches/second or a sustained [redacted]searches per hour.

The WMS performs decompression and feature extraction on search prints, and assign searches to matcher banks. Each matcher bank has sufficient memory to store the entire Unified File of fingerprint templates. All banks can process searches simultaneously and independently, and each bank passes its search results back to the WMS for reporting.

In order to assure that the WMS does not become a bottleneck due to its feature extraction load combined with its management of the interfaces to an increased number of matchers, it is recommended that not more than [redacted] banks of matchers be operated from a single WMS. This number can process [redacted] searches/hour, leaving a reasonable margin before WMS or WAS processing becomes the bottleneck at [redacted]searches/hour.

Figure 6-3 indicated that a search capacity of [redacted]searches/hour is sufficient for up to [redacted] MFRs, each being used similarly to usage in the pilot. The equipment configuration to provide this capacity at Central is shown in Figure 6-4 as a Lantern Stack.

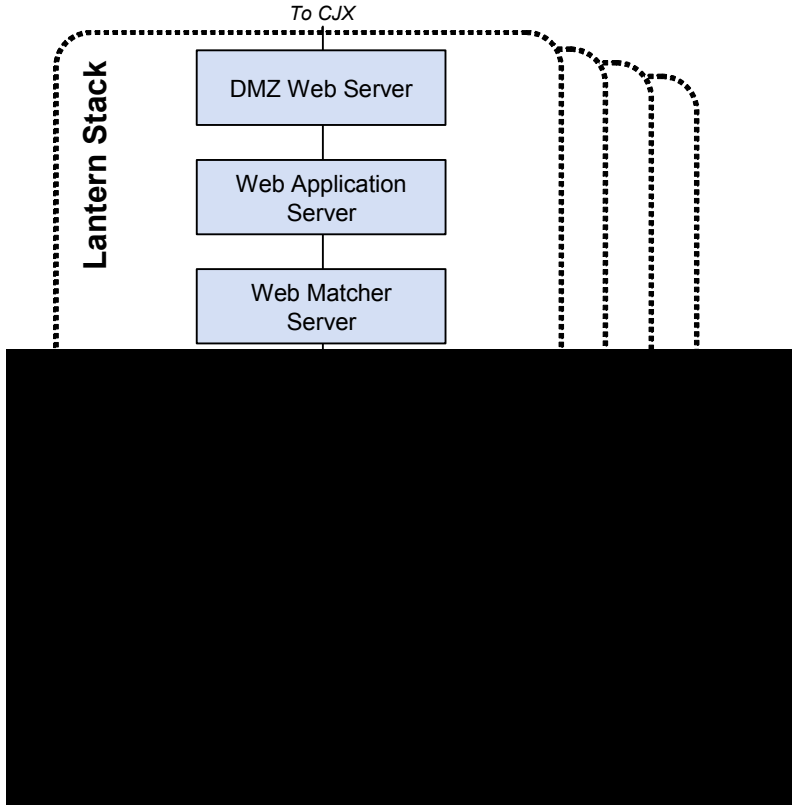


Figure 6-4 Lantern Stack—Building Block for Higher Capacities

Partially or fully populated stacks can be configured at both Central sites by increasing the number of matchers with the existing servers. Fully populated stacks at both sites would provide [redacted] searches/hour with the load split optimally between sites. If even larger search

capacities become necessary, the Lantern Stack becomes a building block whose quantity can be increased, each stack adding capacity for another [redacted] searches/hour. Similar measures may be needed in the IDENT1 AFR Server to handle the demographics requests resulting from an increased number of search hits.

7. Conclusions and Recommendations

This section presents conclusions and recommendations for NPIA consideration in defining capacity requirements for the post-pilot Lantern capability. They are based on estimates of the impact of deploying additional MFRs to police forces, as shown in the previous section.

While the MFR quantity actually deployed must be driven by user requirements, the Lantern architecture can be adapted to the addition of MFRs and the search rate that results, because the architecture is scalable. Central search rate capacity can be augmented as needed in steps.

The foregoing analysis initially focused on the current pilot configuration, and two options for 400 and 700 searches/hour. Matcher quantity and the resulting search capacity in this range are well understood through modelling and observation in operation. These steps are estimated to provide sufficient capacity for [redacted] and [redacted] MFRs respectively without negative impact on Lantern performance.

Higher Central matcher capacities needed for workloads up to [redacted] MFRs were also estimated. The matchers, Web Matcher Server (WMS), Web Application Server (WAS), DMZ Web Server (DWS), and associated firewall and router equipment form a modular stack, which can be equipped with a maximum of [redacted] banks of matchers as shown in Figure 6-4. This fully populated stack provides the capacity to handle up to [redacted] MFRs and the stacks at both sites can be fully populated and share the load from up to [redacted] MFRs. Capacity growth beyond that can be achieved by replicating the Lantern Stack at one or both sites.

As the MFR quantity grows, cost factors other than the MFRs themselves come into play. Equipment procurement costs should be planned for added Lantern stacks or matcher banks as needed for Central sites. MFR and Central costs include not only the operational equipment, but also an appropriate percentage of spares to assure the ability to promptly remedy any malfunctioning units.

There will also be costs associated with Lantern growth that are outside the tangible items of MFRs, matchers, and servers. Added Central kit will draw more primary power and need more cooling and ventilation, possibly requiring a facility upgrade. Network usage will also rise, and consideration must be given to available capacity and incremental costs for CJX and GPRS links. The capacity and scalability of alternate communications infrastructures such as Airwave should also be evaluated.

Inevitably, other service-related costs will grow with increasing MFR deployment including:

- MFR build process
- MFR deployment
- Central kit installation
- Testing
- Trainer training
- User Training
- Force IT workload
- Service Desk workload and staffing
- Preventive maintenance

- Remedial maintenance
- Subcontractor services and warranties.

All sources of cost sensitive to MFR quantity or workload should be considered with any planned expansion of the Lantern scale from the initial pilot upward. Adding MFRs incurs cost beyond the unit cost of the MFR. Nevertheless, because the pilot system is scalable and robust enough to absorb substantial increases in load with only small financial outlays and low risk, scaling up to meet growing needs represents good value for money.