LSNARP Technical Working Group Meeting
December 6, 1999
Las Vegas, NV

Attending:

Dan Graser	NRC/ASLBP	(301)415-7401	djg2@nrc.gov
Glen Foster	NRC/Labat	(703)598-3759	gfoster@gfoster.com
Harry Leake	M&O/YMP	(702)295-5531	harry_leake@ymp.gov
Kazem Taghva	UNLV-ISRI	(702)895-0873	taghva@cs.unlv.edu
Chris Berlien	Nye Co/TSG	(702)795-8254	chris.Berlien@terraspectra.com
David Hunt	MTS/YMP	(702)794-5571	david_hunt@ymp.gov
Tom Nartker	UNLV-ISRI	(702)895-0848	tom@isri.unlv.edu
Dennis Bechtel	Clark Co	(702)455-5178	dax@co.clark.nv.us
Jill M. Schrecongost	DOE/YMP	(702)794-5436	Jill_Schrecongost@notes.ymp.gov

In opening remarks, Dan Graser discussed the two documents provided to the TWG prior to the meeting. The first document discussed was the fleshed-out description of the third and fourth scenarios developed at the previous TWG meeting. The second document was a "strawman" discussion version of revised functional requirements. Additionally, reprints of an InfoWorld™ article (also found at http://archive.infoworld.com/cgi-bin/displayTC.pl?/991122comp.htm) comparing portal software products was made available. He noted that the purpose of this meeting was to use these documents to help formulate the materials to be used to present technical alternatives to the full Advisory Review Panel. He noted that an action item from the previous TWG meeting was as yet not addressed: discussions with DOE's ES&H organization about their experiences with developing a portal site and promised to do that quickly. It was noted that the strawman functional requirements still needed some work in the area of developing performance metrics, and it was suggested that perhaps the ES&H portal at DOE could be characterized as a best practices case against which to develop baseline performance expectations.

General discussion of the two scenarios developed at the last TWG meeting focused on the fact that in any system that will link existing repositories, bandwidth will be the over-arching issue and that any implementation strategy will need to address this. In the two viable scenarios, it was noted that servers can be scaled to size and enhanced to address performance issues. Additionally, it was noted that bandwidth sizing is optimized at a point just above the usage spikes. In the discussion, it was also noted that the bandwidth issue was most likely to be impacted in trying to deal with image handling.

These observations led into a general discussion of the system's architecture attributes that would most directly bear on its performance in handling requests for large files. Having large text and image files reside on participant maintained storage devices provides a "multi-pathway" capability, thereby spreading out bandwidth impacts to some degree although with 85% of the documents being at the DOE site the impacts may still be felt, thus placing a larger burden on DOE's bandwidth capacity. Multi-pathway is a predominant feature of the original scenarios #1 & #2 which were, for the most part, discounted at the previous TWG meetings.

Conversely, developing the system in a central campus means that only one feed will need the higher bandwidth, minimizing the set of connections needed, localizing the area, and requiring dedicated lines. The campus approach is simpler to design in a way that will ensure

performance, has bounded costs, and is more manageable for backup, recovery, load-balancing, etc. It was noted that caching is what creates a localized effect, and enhanced performance is not based on where documents are located. This led to the proposition that another architecture could be considered, a distributed portal approach which retains a complete cache of each participant's holdings. In that approach, the cache needs to be at a high-speed location, right at the entrance to "big bandwidth". Approaches that do not heavily utilize a cache require proxies over to a participant operated storage device and then use the multi-path approach to directly delivering files to the requestor.

The general distinctions were then categorized as being:

- **A Comprehensive Distributed Portal with Participant Remote Storage** typified by a remote portal with software that only maintains indexes, and by participant sites in which the participants each maintain their collection. Their several collections represent the single source of document, header, and image files (except for backup).

- **A Comprehensive Campus Portal** typified by a centralized portal with software that only maintains indexes, and by participant maintained file storage and backup devices that are proximate to the portal device. [1] Participants provide for and maintain their collection, and their several collections represent the single source of document, header, and image files (except for backup).

- **A Comprehensive Distributed Portal with Enhanced Central Storage** typified by a remote portal with software that maintains indexes and a cached copy of all document, header, and image files. Participant collections are downloaded and the portal caches a copy of the participants' files and thereafter uses the cached items exclusively for general search and retrieval.

DOE representatives then introduced a discussion in which the essential technical attributes of the LSN system were identified. These included the ability to 1) provide a high degree of control that can be exercised by the LSNA; 2) ensure timely availability of the system to support the licensing process; and 3) deliver the highest performance at the least cost. NRC noted that these technical attributes reflect the basic mission of the LSNA: 1) to deliver a web-based system that makes all documents equally available in a uniform way, 2) do so in an environment

---

[1] "Close" does not mean geography. Close is defined by the nature of the communications between the machines, specifically that it is quick (high bandwidth and low latency), predictable, and private. For our purposes high bandwidth is somewhere over 25Mbps and low latency is less than 5 ms average with less than .5 ms std. deviation.

It is certainly possible to achieve these performance figures with geographically dispersed systems through the use of appropriate technologies (e.g. 100baseF, a FDDI ring, or DS3 telco circuit) but the latency requirement limits the total distance that can be spanned (to about 100 miles) and the type of circuits that can be used (e.g. no satellite circuits need apply).

In the specific instance of Summerlin to UNLV, this could pretty easily be accomplished with a DS3 or ATM circuit leased from SW Bell (or whoever). This is not cheap, a SWAG is $10-20K per month plus $150K-$300K equipment at each end.

If line-of-sight can be established, it would be possible to use microwave or laser equipment at each end with no recurring costs (uwave = $50K-100K, laser = $15K-$50K). Of course, LOS technologies are subject to weather disruptions but that is probably not too much of a problem in LV. Air rights may have to be secured to avoid disruptions from construction.

It is not realistic to expect to be able to pull a single fiber cable between the sites. Any other hard-wire approach depends on the nature of the rights of way that can be secured and the specific physical topology of interconnections that would result. You would need repeater equipment at each interconnection.

Traditionally, use of the term "campus" has indicated that a single entity controls the physical plant that the gear and interconnections occupy. This includes the ability to trench and install cable. What you are describing is a "multi-campus" situation.

that can be independently audited for compliance, and 3) which provides the tools for ensuring that the system overall performs with acceptable responsiveness.

The group discussed backup/redundancy and noted that the presence of an enhanced central storage facility would lessen the participants' requirement to implement rigorous backup and disaster-recovery procedures (since the central storage facility would be an implicit backup). However, this does not alleviate them of their responsibility to provide and preserve the "true copy" of a document.

The group discussed performance enhancement and noted that this is easier to accomplish via a campus approach, especially if the portal server is modular and multi-processor based.

The group noted that centralized cache storage in a campus location provides the best control, the cheapest overall storage-per-document, and was more predictable.

The group noted that the scenario where the portal is remote and the participants maintain their own collection storage servers will cost more to fix if there are performance problems which should be anticipated especially in large text file and image file transfers to users.

The group then discussed the issue of caching: the distinction between what it will take for the LSNA to ensure system performance and responsiveness viz NRC providing a capability which the participants are required to deliver. There is still also an open issue of certification of records for use in hearing and other legal proceedings which must be done by the submitter - and the fact that the chain of custody goes through the portal site (and the LSNA) in any option where the portal caches everything and that is the file being relied upon. It was decided this is an issue for the full LSNARP to consider.

The group finally discussed overall cost elements in the life cycle and noted that while it may be cheaper to ease in the door with the scenarios that do not rely on centralized cache/memory, that in the long term the solutions where participants maintain decentralized data stores may prove much more labor intensive on an ongoing basis to ensure system control and performance. DOE representatives noted that the cost of memory in the terabyte range has gotten down to the $300-400K range. NRC noted that a recent RAID implementation in that class cost in the $700-800K range; but all agreed that memory/storage costs were declining and could be expected to continue doing so when equipment purchases occur next year. Dr. Nartker observed that delivering and sustaining performance will be the biggest technical problem confronting the operational phase of the LSN. DOE noted that it was easier to initially over-engineer the system rather than to try to remediate performance on a system that is architecturally constrained.

At this point the group decided to start developing a presentation chart which could be used in presenting the issues and recommendations to the full LSNARP. It was decided that the two scenarios discarded at the initial TWG meetings should be included in this chart so that the TWG's evaluations could be documented with the same detail as those options still in consideration. See the charts on the pages following.

Discussion closed on the issue of functional requirements and the difficulties that were being encountered. It was noted that something would have to be done because they will be needed for procurement and also for acceptance testing. DOE representatives made an observation that detailed capability requirements such as print, deliver paper, storing canned queries, etc.,

were, of course, causing problems because the nature of the system is now connecting diverse collections and we are looking at the technologies to do that which are commercial-off-the-shelf (COTS). [E.g., we're purchasing a method to connect existing collections so the FR's need to reflect that as opposed to reflecting the attributes of a licensing methodology management system. The one could be COTS but the other is definitely custom. If we have FR's for a licensing software environment, when we try to do test and acceptance against the COTS portal, we will have disconnects and failed requirements, or, we will walk into a commitment for high degrees of customization that may preclude any COTS portals. So, the revelation was that NRC will have to spec to meeting a different mission and will rework the FR's.] NRC will rework the functional requirements to reflect the mission of providing connectivity and performance in a web environment, rather than focusing on the attributes of a legal support environment.

GENERAL ATTRIBUTES OF ALTERNATIVES

	I Simplified Strategy	II Moderate Strategy	III Comprehensive Distributed Portal with Participant Maintained Remote Storage	IV Comprehensive Campus Portal with Participant Maintained Proximate Storage	V Comprehensive Distributed Portal with Enhanced Central Storage
Description **Criteria**	**Homepage with Pointers to Other Homepages**. **Each Participant Maintains Fully Capable Storage, Search, Retrieval Capability.**	**Centralized Search Interface.** **Each Participant Maintains Fully Capable Storage, Search, Retrieval Capability.**	**Remote Portal Software Indexes.** **Participants Maintain Single Set of Files.**	**Proximate Portal Software Indexes and Data Stores.** **Participants Maintain Single Set of Files.**	**Remote Portal Software Indexes.** **Portal Downloads and Caches a Copy of Participants' Files and Uses Cached Items Exclusively.**
Ability for LSNA to Exercise High Level of Control	No Systematic Controls Each site Varies	Rudimentary Controls on Interface and Search "Passing"	Search, Interface, Security & Access, and Monitoring &Tuning Tools Provided	Search, Interface, Security & Access Enhanced Monitoring & Tuning Capability	Search and Interface Enhanced Security & Access Enhanced Monitoring &Tuning Capability
Ability for LSNA to Ensure Overall Configuration Performance	Performance is Highly Variable LSNA Unable to Respond Quickly to Performance Problems	Performance is Highly Variable Normalized Search "Passing" Still Does Not Guarantee Performance	Performance of Interface Dialogs are Less Variable Fetching Text & Image Files are Constrained	Assured Interface Performance Assured File Delivery Performance	Assured Interface Performance Assured File Delivery Performance
Schedule Risk to LSNA Having Operational to Support Licensing	Low Risk	Moderately Low Risk	Moderate Risk	Moderate Risk	Moderate Risk
Implementation Complexity Risk to LSNA	Low Risk	Moderately Low Risk	High Risk	Moderately High Risk	Moderate Risk
Overall Cost for NRC to Develop	Very Inexpensive	Inexpensive	Expensive	Highest Expense	Very Expensive
Participant Burden to Exercise Controls	Participant Implements within Highly Structured Guidelines and Procedures and is Heavily Audited	Participant Implements within Highly Structured Guidelines (esp. Tech Guidelines for Query Processing) and Procedures (esp. For Change Notification) and is Heavily Audited	More Coordination and Integration Required (ex. When Site Gets Crawled) but More Flexibility is Allowable	Campus Administration Imposes More Restrictions in Format Standards, Population of Collections, Security Access than a Distributed Portal	No Responsibility for Controls Except Change Notification within 5 Day Window

	I Simplified Strategy	II Moderate Strategy	III Comprehensive Distributed Portal with Participant Maintained Remote Storage	IV Comprehensive Campus Portal with Participant Maintained Proximate Storage	V Comprehensive Distributed Portal with Enhanced Central Storage
Participant Burden to Ensure Performance	Totally Responsible for Availability, Performance and Bandwidth	Totally Responsible for Availability and Performance Relieved of Search Interface	Highly Responsible. Portal Provides Some Availability Features. Participant Ensures File Delivery and Bandwidth	Highly Responsible. Portal Provides Some Availability Features. Participant Ensures File Delivery Relieved of Bandwidth	No Responsibility Except for During Initial "Crawling" or Loading
Schedule Risk of Participants' Having Operational to Support Licensing	Moderate	High	Higher	Highest	Moderate to High (Affected by Transmission Security)
Implementation Complexity Risk to Participants	Low	Moderately Low	Moderate	Moderate to High	Low
Cost Burden to Participants	Minimal A portion of a body or outsourced	Variable Requires Comprehensive System Administration, Depending on Participant System. A portion of a body or outsourced	Variable Requires Comprehensive System Administration, Depending on Participant System; More Difficult to Outsource, More Data Management. A portion of a body or outsourced	Variable Requires Comprehensive System Administration, Depending on Participant System; More Difficult to Outsource, More Data Management. Requires Some Personnel Resource at the Campus Location. A portion of a body or outsourced	Minimal A portion of a body or outsourced
User Flexibility to Tailor Desktop/Interface	Relatively Inflexible	Relatively Inflexible	Very Flexible	Very Flexible	Very Flexible
Ease of Use	Hard Variable Interfaces, per Each Collection/Server	Relatively Easy Query Screen is Consistent	Easy	Easy	Easy
Availability to Users	Acceptable One or Two Participants Down Leaves the Rest Still Available	Acceptable One or Two Participants Down Leaves the Rest Still Available	Most Available	High Availability	High Availability

	I Simplified Strategy	II Moderate Strategy	III Comprehensive Distributed Portal with Participant Maintained Remote Storage	IV Comprehensive Campus Portal with Participant Maintained Proximate Storage	V Comprehensive Distributed Portal with Enhanced Central Storage
Response Time Performance	Variable Depends on Participant Resources	Variable Depends on Participant Resources	Somewhat Variable Image & Text Delivery Depends on Participant Resources	Very Timely	Very Timely