National Institutes of Health/Office of Extramural Research

Final Report on Feasibility Study to Assess Moving eRA Non-production Environments to the Cloud

Reference Number: 12-6016 OD-OER-ORIS

Version 0.4
Jan 30, 2014

Document History

Document Location

The source of this document is located in the ClearCase repository at: **<path>**.

Revision History

Version Number	Revision Date	Author	Summary of Changes
0.1	01/10/2014	Dmitriy Kokiyelov	First draft
0.2	01/24/2014	Dmitriy Kokiyelov	Updated based on first round of reviews
0.3	01/29/2014	Dmitriy Kokiyelov	Re-worked section 5
0.4	01/30/2014	Dmitriy Kokiyelov	Incorporated Al's comments

Reference Documents

Number	Document Name	Version, Date
1	\FunctionalGroups\architecture\docs\Initiatives\Top 10 Strategic Technologies for Smart Government.doc	1
2	\FunctionalGroups\architecture\docs\Initiatives\Cloud\Request for Approval Memo for ESA.docx	1
3	\FunctionalGroups\architecture\docs\Initiatives\Cloud\Cloud Selection - Scope.doc	6
4	\FunctionalGroups\architecture\docs\Initiatives\Cloud\eRA Cloud Provider Evaluation.pptx	1
5	\FunctionalGroups\architecture\docs\Initiatives\Cloud\AWS SOW 2013.docx	1
6	\FunctionalGroups\architecture\docs\Initiatives\Cloud\Project Plan.xlsx	4
7	\ProjectsAndInitiatives\Initiatives\Cloud Provider Evaluation\Cloud Provider Evaluation Initiative_120613 v1.0.docx	1
8	\FunctionalGroups\architecture\docs\Initiatives\Cloud\Cloud Provider - Intermediate Report.doc	1
9	\FunctionalGroups\architecture\docs\Initiatives\Cloud\ASSIST in AWS Cloud.docx	4

Key Terms

The following table provides definitions and explanations for terms and acronyms relevant to the content presented within this document.

Term/Acronym	*Definition*
Cloud, a.k.a. Cloud computing	Cloud computing is a term used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network. The term also commonly refers to network-based services, which appear to be provided by real server hardware, and are in fact served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up or down on the fly without affecting the end user, somewhat like a cloud.
AWS	Amazon Web Services (abbreviated AWS) is a collection of remote computing services that together make up a cloud computing platform, offered over the Internet by Amazon.com. The most central and well-known of these services are Amazon EC2 and Amazon S3. The service is advertised as providing a large computing capacity much faster and cheaper than building a physical server farm.
EC2	Amazon Elastic Compute Cloud. Amazon Elastic Compute Cloud (EC2) is a central part of Amazon's cloud computing platform, Amazon Web Services (AWS). EC2 allows users to rent virtual computers on which to run their own computer applications.
S3	The Amazon Simple Storage Service (Amazon S3) is a scalable, high-speed, low-cost Web-based service designed for online backup and archiving of data object. The S3 is intentionally designed as an object store with a minimal feature set.
SES	Amazon Simple Email Service (Amazon SES) is a cost-effective outbound-only email-sending service built on the reliable and scalable infrastructure that Amazon.com has developed to serve its own customer base. Along with high deliverability, Amazon SES provides easy, real-time access to your sending statistics and built-in notifications for bounces and complaints to help fine-tune email-sending strategy.
SNS	Amazon Simple Notification Service (Amazon SNS) is a fast, flexible, fully managed push messaging service. Amazon SNS makes it simple and cost-effective to push to mobile devices and internet connected smart devices, as to other distributed services. To prevent messages from being lost, all messages published to Amazon SNS are stored redundantly across multiple availability zones.
VPC	Amazon Virtual Private Cloud (Amazon VPC) lets you provision a logically isolated section of the Amazon Web Services (AWS) Cloud where an AWS user can launch AWS resources in a virtual network. Customers complete control over virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways. Amazon Virtual Private Channel (sometimes also referred as VPC) is an integral part of Virtual Private Cloud providing secure connectivity between on-site services and resources in the cloud. Such connections over VPN are used to create private or hybrid clouds.

Term/Acronym	*Definition*
eCBS	eRA Central Build System (eCBS) supports common, consistent, repeatable and traceable Build, Unit-Test and Publishing (Staging) processes for all Software artifacts across the eRA program. It also offers limited Deployment capabilities to 'DEV' environment as part of automated in-container testing.
VAP	Value-Add Provider is an independent commercial organization providing sales and limited support of AWS services to public and private customers. Amazon sales practices require use of VAP for sales to Federal Government.
ATO	An Authorization to Operate (ATO) is a formal declaration by a Designated Approving Authority (DAA) that authorizes operation of a Business Product and explicitly accepts the risk to agency operations.

Table of Contents

1. OVERVIEW 7

1.1 BACKGROUND 7

2. ACCOMPLISHMENTS 8

2.1 INFRASTRUCTURE 8

2.2 PROCEDURES 8

2.3 IMPLEMENTATION 8

3. FINDINGS 10

3.1 CURRENT CAPABILITIES 10

3.1.1 Infrastructure 10

3.1.2 Applications 11

3.1.3 Database 12

3.1.4 Storage 12

3.1.5 Performance 12

3.1.6 Costs 13

3.1.7 Other 14

3.1.7.1 NIH SSO bypass 14

3.1.7.2 Session replication 14

3.1.7.3 Application changes 15

3.2 PROCUREMENT AND SUPPORT 15

3.3 PLATFORMS AND SERVICES 16

3.3.1 VPC 16

3.3.2 EC2 16

3.3.3 RDS 17

3.3.4 S3 17

3.3.5 SNS 18

3.3.6 SES 18

3.3.7 DNS 18

4. STRATEGY 18

4.1 APPROACH 18

4.2 POSSIBLE USE 19

4.2.1 Platform 19

4.2.2 Services 19

4.3 INFRASTRUCTURE 20

4.4 PROCESSES 20

5. NEXT STEPS 20

1. OVERVIEW

1.1 BACKGROUND

eRA is the program within NIH's Office of Extramural Research (OER) that is responsible for developing, managing, and supporting the NIH Enterprise systems used to manage the NIH Extramural Grant Program. eRA offers vital solutions to manage the receipt, processing, review, award and monitoring of over $40 billion in research and non-research grants awarded annually by NIH and other grantor agencies in support of the collective mission of improving human health. It is used by Federal staff and applicants at over 9,500 institutions worldwide. eRA continuously strives to lower cost, greater flexibility, and improved continuity of operations and disaster recovery for the NIH IMPAC II system.

The increasing use of cloud-based services and infrastructure has been a sustained trend in the industry for several years. IT organizations have achieved greater flexibility, shorter delivery times, and lower operational costs by shifting from in-house infrastructure to virtual environments and, ultimately, cloud-based environment. eRA has been relying on a limited use of virtual infrastructure and services for certain parts of the business. The Cloud Evaluation project ultimate goal is to evaluate the use of virtualization and cloud computing to support the organizations mission and business goals.

The evaluation of cloud-based infrastructure and services for eRA business needs targets the following primary goals:

- Migrate eRA "home" pages eRA Home Page and http://inside.era.nih.gov using cloud infrastructure and make these resources available to the appropriate user groups
- Investigate the possibility to have backup eRA Central Build System using cloud infrastructure fully integrated with eRA configuration management systems maintained in-house
- Develop capability to deploy mid-tier portion of one or two eRA applications using in-house virtual infrastructure for non-production use
- Develop capability to deploy an "enhanced" eRA application using cloud infrastructure for non-production use
- Identify technical requirements for eRA applications to be ready for migration to cloud-based infrastructure
- Prototype a notification solution for the use of cloud-based services for the needs not addressed by currently available eRA or NIH offerings
- Prototype a solution for the use of cloud-based storage infrastructure for document management and develop approach for integrating it with eRA Document Service

In support of the primary goals, the following additional goals have been defined:

- Develop the infrastructure architecture for integration of the existing eRA non-production environment and cloud-based systems and services for the applications
- Work with various NIH departments and groups to implement necessary infrastructure and validate its correct and reliable functionality
- Identify additional offerings by the cloud provider(s) that may supplement or enhance the existing infrastructure support and management capabilities

2. ACCOMPLISHMENTS

2.1 INFRASTRUCTURE

1. A dedicated virtual private cloud has been setup for eRA home pages. The VPC is "public" so the deployed resources are accessible from the public Internet. The connectivity to the VPC is limited to administrative interfaces required to manage AWS resources, operating systems, and deployed applications. No permanent connectivity to NIH network is required due to the nature of this applications. Additional details can be found in [8].
2. A dedicated virtual private cloud has been setup for eRA applications. The VPC is "private" so the deployed resources are accessible from NIH network only. The connectivity between NIH and AWS is established via dedicated highly available VPN tunnel setup in cooperation with DNST. Additional details can be found in [8].
3. The private cloud has been successfully converted to hybrid cloud with multiple controlled subnets. The resources in the "web" subnet can be accessed from the public Internet to allow deployment of eRA external applications in the cloud. The "application" subnet and "data" subnet are isolated from the Internet and can be accessed from NIH network only. Additional details can be found in [9].
4. DNST has presented a new approach for NIH-to-AWS integration. The proposed network architecture has been partially implemented by NHLBI. However, the implementation by NHLBI does not cover some eRA needs. A number of business and technical questions have been raised by eRA to DNST during the presentation. See the "Next Steps" section below for additional information.

2.2 PROCEDURES

1. We have met with AWS representatives at the beginning of the project to identify possible procurement options. Amazon does not sell AWS services to US Federal Government directly. After evaluation of available AWS partners and re-sellers, eRA BPO has established the 1-year contract with A&T Systems. The contract covers service procurement and business support.
2. Over the course of the project we have contacted both AWS Technical Support and A&T Systems for issue resolution and escalation. AWS support center provides multiple options including e-mail, phone, chat, and others. We have found that use of AWS ticketing system in combination with the integrated online chat is convenient and efficient way to resolve issues, typically in very short time. The copies of tickets and chat logs are preserved for future use and references.
3. We have worked with DNST to allocate network resources and establish connectivity between NIH and AWS VPC. The two-channel VPN tunnel between NIH border router and AWS virtual private gateway was setup per DNST guidelines. The initially design of the network routing table has been simplified to work around constrains of the network equipment. The traffic control is done using NIH and eRA firewalls (on NIH side) and VPC routing tables and EC2 security groups (on AWS side).

2.3 IMPLEMENTATION

1. Migration of eRA "home" site(s)

The eRA “home page” sites have been deployed in the public VPC using two EC2 virtual servers behind load balancer (ELB). During the project, eRA staff has successfully deployed two version of the site to ensure the cloud-based deployment matches one in-house.

The sites have been fully functional in the cloud in highly available configuration. During the performance testing from the public Internet, the response time was the same or better than in-house instance. No major functional issues have been found during the testing.

2. Investigate possibility to have eRA Central Build System backup in AWS

   We have performed the analysis of the eCBS dependencies and integration requirements. It has been concluded that, due to a large number of integration points and dependencies, the implementation will require significant re-engineering of supporting scripts, as well as the network and firewall configuration. Also, eCBS reliance on existing development and configuration management tools and infrastructure, including requirement for direct connectivity to eRA development servers make eCBS in its current form not suitable for deployment in AWS EC2 environment.

   The performed analysis should be used as input for the future eCBS development plans. Some aspects of eCBS design and implementation may need to be altered to minimize reliance of local resources and platform-specific code.

3. It is our conclusion that the potential benefits of migrating eCBS in its current form to cloud environment does not justify required cost and efforts. Such a move should be considered in combination with the migration of Development environment as a whole.Deploy mid-tier portion of eRA applications using in-house virtual infrastructure

   Two eRA “internal” applications (ADI and RCDC) have been successfully deployed in AWS EC2 environment. The deployment relied on in-house network edge infrastructure (BigIP, RP) and databases (OLTP, IRDB). No code changes have been required, and only minor configuration changes within EPM and BigIP were made. We have also successfully implemented a “hybrid” solution when the application pool contained both in-house and in-cloud application servers.

   There have been no unexpected functional issues found during the testing. Some business functions were not working due to application reliance on locally mounted resources such as eRA central storage file systems. This was expected from the initial analysis and review of the systems designs.

   The application performance in the cloud have not been noticeable different from in-house user experience. The special performance tests indicated the application response times are 50-400ms slower from the cloud due to network latency. Such delays will not usually impact end-user experience as typical response time of an application is several seconds. However, applications relying on multiple round-trips to the database to serve a single user action may see more significant impact on their performance.

4. Deploy an “enhanced” eRA application using cloud infrastructure

   The project initial scope has been expanded to include deployment of an eRA business system in AWS cloud including network edge infrastructure, application, and data services. ASSIST system with supporting services has been selected for this effort due to its limited dependencies on eRA common data model.

   The application has been modified to severe some dependencies on eRA infrastructure such as NIH SSO and eRA Metadata Service. The end-user authentication was relying on enhanced NIH

SSO “bypass” implementation and AAS service. The database was implemented using AWS RDS Oracle 11gR2 instance. The document management service DocServiceLite was modified to utilize AWS S3 object storage.

Limited testing of the application functionality has been performed. The main functional defect found was due to delayed authorization data synchronization between in-house and cloud-based database instances. This was expected due to known system design constrain.

As part of this effort, the application was modified to allow for in-memory session replication across multiple application server instances. The required code and configuration changes were numerous but not complex. We have found the implementation of real-time session replication addresses certain limitations of existing “failover” support but has significant impact on the application performance. Additional details can be found in [9].

5. Prototype document management solution using of cloud-based storage

   We have evaluated functional suitability of AWS S3 object store by implementing DocServiceLite using AWS S3 for the data store. The changes were localized to one part of the application and required minimal effort. The functionality of the application was not impacted after the change.

   The performance tests of AWS S3 services accessed from NIH internal network have indicated it is comparable to existing solution relying on NFS and BFILE. Due to the nature of eRA document management requirements and the document access patterns, it may be desirable to implement disk-based cache within DocumentService to reduce the associated network transfer costs. Additional details can be found in [8].

   AWS S3 can also be used as secondary storage for eRA data objects such as database backups, log archives, and similar data that do not require immediate online access. The cost of S3 storage is significantly lower than the current price charged by CIT. The data transfer can be implemented using command-line and UI-based interfaces provided by AWS.

6. Prototype end-user notification solution using cloud-based services

   An evaluation of AWS SNS service for notification delivery has been done using specially designed test program. The service supports delivery of messages by “topic” allowing users to receive notifications on a certain subject using variety of the protocols. It can be easily interated into eRA eNotification engine without requiring any significant design changes.

   We have found the SNS can be used to deliver e-mail and text messages using single API. In order to avoid use of SNS by spammers, the service requires user to opt-in by responding to standard (non-modifiable) e-mail or text message. Thi s might limit suitability of the SNS for eRA business needs to the situations where end-user opt-in for the message delivery is acceptable. AWS offers alternative SES service which does not require receiver to opt-in but this service supports e-mail notifications only.

3. FINDINGS

3.1 CURRENT CAPABILITIES

3.1.1 Infrastructure

AWS services and resources can be integrated with eRA infrastructure using VPC with network connection over VPN tunnel. The network throughput and latency are lower than

in-house resources but are adequate for practical use by eRA applications. During the trial, a dedicated VPN tunnel has been setup for eRA VPC. The existing constrains limit VPN routing table to no more than 2 entries. Additional traffic routing and controls are managed using network equipment such as firewall, routers, and security groups/rules. AWS provides more granular controls using network ACL but this methods is difficult and costly to configure, manage, and troubleshoot. As of now, eRA does not have use for this level of granular control.

The current setup can be extended to include multiple VPC(s). The present architecture requires allocation of two logical ports at the border router per VPC. Therefore, the number of VPC instances is limited by the NIH equipment capacity. DNST has future plans for the different connectivity architecture where all NIH traffic going to AWS via the single tunnel terminating on CIT-owned VPC. The traffic routing from this VPC to different VPC(s) is managed within AWS using virtual routers. This network architecture has not been implemented as intended by CIT at this time. Critical questions about network bandwidth, latency, and operational setup and troubleshooting need to be answered. It is also unclear how the network costs will be allocated within NIH for the use of shared network channel.

Note that some NIH services are not yet available in the cloud. This includes access to LDAP(s), NIH SSO, and some others. Applications and services having dependency on such elements should not be migrated to the cloud or need to be re-designed to address these aspects. An example of such solution can be eRA “NIH SSO bypass” used to successfully deploy ASSIST in the cloud without any change in user’s experience.

AWS infrastructure charges for the resource instances can be estimated with high degree of accuracy. It is more difficult to estimate the network traffic cost as eRA currently has no reliable information of the network traffic needs by different eRA applications. This may not be critical for non-production environments where the activity is low. However, the costs may be significant for production use. The estimates should be made separately for the user-to-system traffic, system-to-system traffic, and system-to-database traffic as the network capacity and pricing will be different for different routes.

Referenced documents [8,9] contains additional details outlining infrastructure setup.

3.1.2 Applications

Existing eRA applications can be deployed in the AWS EC2 cloud with minimal changes. The critical functional constrain is reliance of local resources such as mounted shared file systems or services protected by network firewalls. As eRA applications continue evolving to rely on the service-based architecture this constrains will go away. One of the key elements is the current DocumentService as applications bypass service interface to store (and, sometimes, retrieve) documents. The current initiative for implementation of the new DocumentService abstracting applications from the access to the underlying infrastructure will help address this issue.

The applications in the cloud may not be co-located with the data sources. Therefore, they should not rely on ability to “chat” with RDBMS and other infrastructure elements. While the current round-trip between an application and database servers is frequently under 100ms it may grow to 400 ms between an application in AWS EC2 and database hosted within NIH. This increase may not have significant impact for large transactions that take time to execute on the database side. However, the performance of some eRA applications

relying on multiple very small database transactions to serve a single user action may be impacted. In such cases, re-design of the corresponding actions may be needed.

Referenced document [9] contains additional details outlining application coding and setup.

3.1.3 Database

eRA applications in the cloud may rely on in-house database instances or utilize database instances setup in the cloud. The former case is simpler to setup and is more familiar to eRA staff, so it is a good starting point. However, it requires each database request to travel over WAN using secure VPN connection. This carries associated costs and may impact application performance due to network latency or insufficient bandwidth.

Alternative setup required use of database instances in the cloud. Oracle databases can be setup using customer-provided instances or “pay as you go” Oracle RDS. It is important to note that RDS does not have ATO from DHHS at the present time so its use for production purposes should be reviewed and approved on a case-by-case basis. Due to the limits of the available Oracle RDBMS license at eRA, we have used RDS for the project.

RDS is relatively new offering by Amazon and it comes with a number of constrains. For example, DBA do not have access to the underlying physical hardware or operating system. This makes some of optional Oracle features unavailable as their setup and configuration requires OS level access. We have found that RDS provides functionality required by eRA applications and allows integrations with in-house databases for data exchange over database links. Additional review by operational DBA may be desirable to identify if any operational requirements cannot be satisfied by RDS.

3.1.4 Storage

AWS S3 service provides object store API. We have used this service to satisfy application business requirements such as storing business documents as well as offline storage for archives etc. In all cases, the service has performed well and we experienced no major issues. The service ability to manage business metadata associated with an object simplifies data management and seems very promising to address certain eRA business requirements.

The service is accessed over Internet and, as such, is not very suitable for dealing with very small objects as the request overhead becomes comparable or larger than the data transfer itself. As eRA deals with a significant number of small files, a local disk cache within DocumentService may significantly enhance its performance if S3 will be used as the primary document storage.

The built-in security, high availability and reliability, and low cost of the AWS S3 make it a good candidate to store large volume of offline data such as database backup, archived logs, system images, and other similar data. The service can be used for such purposes by Operations without need for the cloud adoption by the rest of eRA.

3.1.5 Performance

eRA applications had no or limited impact on the performance when deployed in AWS cloud. The increase in the response times, when observed, was minimal and did not impact end-user experience in noticeable way. However, it should be noted that the testing was

limited to single-user experience and more through testing, including concurrent multi-user performance testing, is recommended before proceeding forward with the production use of AWS infrastructure by eRA.

AWS cloud may be very suitable for the systems that experience varying resource requirements due to peaks of user activity. The ability to scale resources, manually or automatically, very quickly can allow eRA to satisfy changes in demand quickly and efficiently, without the need to pre-provision resources.

One of the low risk adoption strategies may be use of mixed application pools combining in-house and in-cloud servers. This strategy allows increase application capacity quickly without need to change border network and database services. This approach can also be used for resource-intensive systems that require, for example, high performance CPU or high memory utilization. Allocating dedicated virtual servers to such systems reduces risk of impact for other applications and allows better match between application needs and infrastructure capabilities.

3.1.6 Costs

The potential cost savings brought by AWS cloud infrastructure and services can be significant. According to estimates based on the latest billing information, the cost of running applications servers using AWS EC2 can be half of what eRA pays today, as can be seen in [9]. The savings from the use of AWS S3 as secondary storage can be even higher. In an effort to win customers and expand the business with the public sector, Amazon offers new services and reduced prices on regular basis. However, eRA needs to do more accurate analysis and consider possible added direct and indirect costs.

One of the significant factors is changes in the system administration effort and required resources. Today, eRA delegates system administration of the Unix servers to CIT while Windows servers are administered by eRA staff. eRA SA staff includes four people, with two being responsible for Unix systems and the other two for Windows. It needs to be estimated if the staffing levels need to be changed if eRA will be administering its Unix virtual servers. Another relative question is the required skillset and processes involved into administering virtual infrastructure comparing to the physical one we utilize today.

Another aspect is the costs associated with the use of network bandwidth within and outside AWS infrastructure. Careful estimate of the required network use based on available data from production and non-production environments will allow eRA management to estimate the associated costs more accurately. Furthermore, it may be beneficial to deploy small number of systems in “hybrid” environment as described above to validate the accuracy of the cost estimates.

It is important to remember that the one of the main benefits of cloud based infrastructure is its flexibility. AWS solutions and services allow eRA to quickly adjust to the changing business needs including changes in user’s demand, ability to run business project pilots and setup isolated resources for specific time without the need to pre-provision the resources. All this gives eRA ability to better and faster respond to user needs and to provide higher quality of service.

Additional cost savings can come from allocating resources for a given phase of SDLC lifecycle, for example, by scaling STAGE or UAT environments up when they are needed and scaling them down once the corresponding phase of SDLC is competed. In order to

realize such savings, the program needs to plan and monitor the needs and usage of resources across the enterprise on a continuous basis.

3.1.7 Other

3.1.7.1 NIH SSO bypass

NIH SSO services are not available within AWS infrastructure as the present time. In order to enable end-user authentication without impacting end-user experience, ASSIST code has been modified to use enhanced version of "NIH SSO bypass" feature. The changes were made to allow use of application-specific login, logout, and error pages and relying of AAS for end-user authentication and authorization behind the scene. In order to minimize system vulnerability, the "switch user" feature has been disabled in this version of the software.

The modified code, together with recommendation for allowing "switch user" feature with minimal security risks, have been provided to eRA Team 4. Team Lead will communicate the details to eRA Program Management to decide if the enhancement should be incorporated into eRA common code base to be available for all eRA applications.

3.1.7.2 Session replication

As part of this pilot, we have implemented in-memory session replication using Apache Tomcat built-in clustered session management feature. The implementation required changes to a number of application objects to allow data serialization and changes in the session use practices and code patterns (such as setting an entire session attribute object rather than modifying its properties) to take advantage of the clustered session manager capabilities. The code changes were relatively simple but required changes in multiple parts of the application.

The modified application provides better user experience in the event of failover such as application server crash. Currently, user typically needs to log out and re-login into eRA application to continue work after a container crash or restart. This results in long time taken by operational "rolling" restart of the systems as actively used instances cannot be restarted without impacting users. With the changes, the only impact of failover is loss of user-supplied form data at the time of crash. This benefit, however, comes with some drawbacks.

First, the session replication results in noticeable impact of the application performance from user's point of view. The delays caused by the session synchronization across two instances are noticeable, and the impact will be even more significant if more instances are involved. The impact is noticeable even if the session data are limited in size and may be more significant if sessions are used to store large amounts of data such as list views. It must be noted that there are commercial products on the market that may reduce the impact of the session replication on application performance. These products are not currently available at eRA and therefore have not been evaluated.

Second, the mechanism requires use of network "multicast" for configuration and systems discovery. Multicast traffic is not allowed within AWS EC2 subnets, so we had to pre-configure application server instances manually so they know of each other. This effectively prevents Operations from dynamically allocating and removing application instances unless

application is down. These restrictions may not apply within NIH data center but needs to be considered for future network planning.

Overall, the session replication functionality works as described in vendor's documentation. However, the required application changes and impact on the system performance may overweigh benefits of the solution.

3.1.7.3 Application changes

A number of other changes have been made to ASSIST and DocServiceLite to remove dependencies on shared infrastructure elements, add missing resources, and resolve some defects discovered during the testing. The detailed list of the changes can be found in [9]/ The modified code is available in eRA ClearCase repository in "architecture" area.

3.2 PROCUREMENT AND SUPPORT

The federal government rules for cloud-based infrastructure and services, known as FedRAMP, require certification of the services provider by the federal agency. At the time of the project initiation, AWS was the only cloud infrastructure provider certified by DHHS that offered the desirable services. Amazon Federal Sales team has met with eRA and provided list of certified VAP(s) as Amazon does not sell AWS services directly to the Federal Agencies. eRA has put together the technical requirements matrix and RFP. Several companies have responded, and the one year contract been awarded to A&T Systems.

The contract put in place allowed us to execute the majority of the planned activities without major procedural difficulties. However, few lessons have been learned during the past year that might be helpful for the future work procurement of similar services and products:

1. The current contract is based on the fixed monthly price. It is not based on the actual usage and does not include provision to carry over unused funds to the next months. As a result, eRA is receiving deep discount from the list price but still pays for the services and resources not being used. It seems the actual usage-based pricing with the controlled usage levels will be more appropriate and beneficial to eRA.
2. The VAP abilities to control the flow of information need to be documented in the contract and, preferably, limited to absolute necessity. Specifically, eRA staff needs to have as-needed access to administrative console, account administration (AWS AIM), usage reports, administrative notifications, and all other information distributed by AWS. The contract should document any restrictions put in place by VAP rather than list what information will be made available to eRA staff.
3. The VAP staff ability to provide technical support seems to be extremely limited. Based on our conversation with AWS support center and other users of AWS within NIH, this situation applies to most AWS VAP(s). Therefore, it is essential to ensure the contract includes direct and sufficient support from AWS technical staff. The responsibilities of the VAP should be clearly documented and ability to deliver tested as early as possible, possibly via references or case studies.

4. The previous projects delivered or served by VAP need to be studied to see if they are similar to eRA needs and requirements. The nature of previous work as well as the scope and complexity may be good indicator of how VAP staff technical and procedural knowledge meets the program needs.
5. The current VAP, A&T Systems, seems to have limited capabilities to provide technical support and guidance with AWS implementation. Our direct contact is senior staff (CTO) and we do not have knowledge or experience if there is other qualified engineering staff in the organization. The response times are unpredictable and, in some cases, quite long. We had also experienced difficulties in our attempts to receive written guidance and recommendations. Overall, our impression is A&T Systems can barely meet eRA requirements and does not have in-house knowledge and structure to provide desirable guidance and recommendations.
6. The support provided by AWS directly has been very satisfactory. Majority of issues raised have been resolved in matter of hours, frequently less than an hour. The staff is knowledgeable and willing to involve higher tiers of support when needed. In two cases, AWS technical support willingly worked with us to troubleshoot issue that was occurring within NIH network and was not directly related to AWS infrastructure. We realize that majority of the experienced issues were relatively straightforward to resolve and the experience may be different for more complex situations. Based on our current experience, the provided guidance and support are very efficient and were sufficient to satisfy all evaluation project needs.

3.3 PLATFORMS AND SERVICES

3.3.1 VPC

The VPC setup providing efficient and reliable network interactions are key elements of the virtual infrastructure. NIH currently does not have a standard architecture for VPC setup. DNST has a high level technical proposal but more analysis and testing required to ensure it satisfies eRA needs. The current setup can be used for continuing research, prototyping, and testing but it is not sufficient for large-scale deployment due to limited CIDR block allocated by DNST for the project.

One of the most challenging parts of the project was identifying and troubleshooting connectivity issues. Due to the large number of NIH organizations and groups involved, and limitations of the troubleshooting techniques, the resolution of these issues can be time and effort consuming.

In addition, DNST and NIH Security require administrative access to eRA-owned account to perform their activities. This resulted in incorrect provisioning of resources which might cause security and financial implications. A better definition of the needs and responsibilities of these groups will help adoption of cloud solutions by eRA and NIH in general.

3.3.2 EC2

EC2 resources can be provisioned easily using administrative tools provided by AWS. We

have used different instances known as “tiny” (for administration), “small” (for infrastructure elements such as reverse proxies and small applications) and “medium” (for larger applications). Certain system administration tasks such as base monitoring, backup, and cloning can be done easily using administrative tools by staff without prior SA experience. Adjustment and tuning of specific OS features may require SA expertise. However, once created and configured to specific eRA needs, the entire configuration can be saved as the Amazon Machine Image (AMI) and easily used to create multiple servers.

During the trial we have experienced two crashes of one virtual instance over the course of four months (no other instances have failed). The investigation has shown the instance was running on defective hardware. AWS has notified eRA about the problem but the first notification was not received due to communication error between AWS, eRA, and re-seller. After the second crash, the instance has been re-located to another hardware following AWS SOP. The process took about 15 minutes done manually. AWS Support Center has told us the procedure can be automated by the customer if desired.

There have been no other issues or incidents with EC2 services during the project.

3.3.3 RDS

The RDS instance has been setup following AWS SOP. Upon setup, the instance was used by two eRA applications without any issues. Scheduled data bridge between in-house OLTP database and in-cloud instance has been setup as well. The built-in RDS backup has been used but we have not tested database recovery using the created backups.

Several limitations of the RDS have been noted. First, AWS provides no access to the underlying operating system. This limits ability to enable some optional Oracle packages as they require such access. For example, built-in Oracle SMTP mail package used by eRA today could not be used within RDS. The analysis of eRA use of Oracle features by Operations DBA is required to identify possible gaps. Second, RDS is not covered by DHHS ATO. This means its use for production purposes needs to be approved on a case by case basis. RDS application for ATO is pending but there is no definite timeline for it.

3.3.4 S3

AWS S3 object store was used for both application needs and offline data storage. Both API and command-line interface are simple and straightforward to use. The ability to manage data object together with structured metadata is very convenient and may help address some of he outstanding eRA business requirements.

The service performance is adequate for eRA needs based on our estimate of the current production use of DocumentService. It is important to note that S3 is most efficient for the medium-side objects (several megabytes in size) and its relative performance is lower for small objects.

We have noticed the service API error rate grows quickly when the number of application threads within the process writing to the store concurrently exceeds 40. The cause of the errors have not been investigates in details, partially due to other network issues experienced at the same time. As the level of “write” concurrency within eRA DocumentService is well below this number, the detected constrain may not require further analysis in the near future. If the DocumentService implementation will include local cache the issue may not apply to eRA at all.

3.3.5 SNS

AWS SNS is convenient gateway for sending notifications to the end-users via variety of protocols including e-mails and text messages. The service API is very simple and easy to use. A convenient feature allowing subscribing users to a topic of interest and associating a message with a topic allows very flexible and convenient organization of communications

SNS carries cost “per message” and “per data volume”. The cost for e-mails is quite low but the text messages are significantly more expensive. The recipients need to “opt in” by acknowledging the message sent to AWS at the time of user’s registration. The “opt in” must be done from the recipient’s address and can not be done by the system such as eRA applications.

Overall, SNS nicely complements some of the functionality missing in eRA portfolio today such as text messaging and message grouping by topic. The requirement to “opt in” by the recipients and associated costs do not make it a suitable replacement for the current e-mail based notification solutions but it can be a useful add-on to existing capabilities.

3.3.6 SES

AWS SES is Amazon’s e-mail server gateway. It provides simple API supporting all major features for e-mail message formatting and delivery in the framework of today’s standards. The service is functionally similar to NIH e-mail server and can be used as backup for it in case of outages or overload of NIH infrastructure.

3.3.7 DNS

In order to enable transparent navigation between eRA systems in-cloud and in-house and ensure proper integration, we have created the site within ***.nih.gov*** domain deployed within AWS cloud. The solution is based on registration of the new site name as “C” class record in NIH external DNS. The record is defined as synonym of existing public name of AWS ELB so, when a user enters site name, the DNS automatically resolves it to the IP address assigned at the time of request to the appropriate ELB.

Two things must to be noted. First, the “C” class record is used because IP address of an ELB is not permanent. Therefore, one cannot create “A” class record and associated with an IP address. The difference is visible to the network administrator only and transparent to the end-users. Second, due to the setup of NIH HTTP proxy, the traffic to the sites in the ***.nih.gov*** domain is not routed outside of NIH network so the desktops within NIH cannot reach the site deployed within AWS infrastructure over public IP. Two possible solutions have been identified for the issue. One has been implemented for the staff involved into the project. The second one is being pursued by Operations management with OIT.

The use of AWS DNS service known as Route53 has not been evaluated as the service does not seems to allow allocation of sites in ***.gov*** top level domain at this time.

4. STRATEGY

4.1 APPROACH

eRA applications form an integrated solution supporting grants management process for NIH as well as a number of other federal organizations and agencies. The complexity of the

business processes requires multiple integration points between the various systems. The common data model historically has tight coupling between various business domains to ensure data quality and minimize redundancy. This environment represents some unique challenges to adoption of the virtual and cloud-based infrastructure and services. However, as it has been shown during the evaluation project, there are numerous opportunities where eRA business and technical needs can be satisfied by cloud offerings in cost efficient manner allowing the program to increase is flexibility and achieve greater customer satisfaction.

The adoption of cloud-based infrastructure and services will bring some new experience to eRA staff from all divisions including business planning, system owners, development, operations, security, and others. The procurement and use of the services needs to be planned and controlled to avoid business process interruptions and avoid cost overruns. Proper coordination with other NIH departments, primarily CIT/DNST, is another key element to succeed in adoption of the cloud solutions for eRA needs.

We believe eRA should start by using cloud-based infrastructure in simple manner and gradually increase use of specific services and platforms to meet eRA business needs. It is critical to ensure the services are chosen to satisfy the business needs and not for the sake of "being in a cloud". While it may be tempting to start moving eRA system into the cloud, additional information about system usage, dependencies, and needs should be gathered to prioritize and plan the following steps. It may be convenient to start with "hybrid" solution using both in-house and in-cloud elements at the same time to minimize the potential risks and ensure smooth migration. This will include usage of EC2 virtual infrastructure and S3 data store as well as selected services offered by AWS.

The migration of the structured data services shall be considered at the later stage of cloud adoption strategy. This is driven by internal and external factors. Internally, the current tightly integrated existing IMPACII data model makes planning migration extremely challenging. As eRA will continue evolving IMPACII data model as part of system modernization efforts, the opportunities will arise which be a good fit for use of segments of the data model along business lines with limited dependencies, like it was done with ASSIST. Externally, RDS product is relatively new to the market and has a number of potential procedural and operational constrains. As AWS RDS platform matures, it may better satisfy eRA needs and simplify migration efforts.

4.2 POSSIBLE USE

4.2.1 Platform

eRA can start taking advantage of the following AWS platform offerings:

1. EC2 virtual instances, load balancers, and network management solutions for application servers and network infrastructure.
2. S3 for offline data management and, going forward, business data objects such as documents, reports, logs, and other similar objects

4.2.2 Services

The following services can be used to satisfy existing eRA needs with relatively little effort:

1. SES as backup e-mail gateway to deal with NIH mail server availability and/or overload situations.
2. SNS as solution to allow gradual rollout of SMS notification, starting with text messaging, both for eRA business users and internal staff.
3. CloudWatch for system monitoring and management, with the goal to supplement existing solutions and better manage maintenance costs over time.
4. In the long run, PaaS offerings such as RDS may be beneficial for eRA. The adoption of these services may be planned for later stages due to existing dependencies and procedural complexity of the migration.

4.3 INFRASTRUCTURE

eRA needs to work with CIT/DNST to identify the long term architecture and strategy for network connectivity and operational management of the integrated infrastructure. Depending on DNST plans and timeline, it may be desirable to undertake two parallel efforts. The existing infrastructure can be used to setup applications and services while eRA will continue working with DNST to setup and evaluate alternative architecture.

It is important to understand CIT/DNST expectations regarding the enterprise services eRA is responsible for. For example, the proposal implies having multiple geographically distributed LDAP and, in the future, database instances. This may require additional budget and operational staff from eRA.

4.4 PROCESSES

The major challenge in cloud adoption by an established IT organization lies not in the technical but in the procedural area. The new concepts may require certain cultural changes from the organization to enable benefits and minimize associated risks. We believe the following processes need to be put in place or adjusted appropriately at the very early stage of the process:

1. Ensure eRA application infrastructure needs, dependencies, and usage is estimated at the early stages of planning and monitored and evaluated on the ongoing basis. This is critical element to ensure the stability and costs of the system management.
2. Establish flexible contract with qualified VAP to ensure eRA immediate needs as well as expected growth can be met.
3. Provide eRA staff with offsite or on-the-job training to ensure sufficient understanding to develop internal processes and operational procedures.
4. Define roles, responsibilities, and controls for access, configuration, and use of the cloud resources. eRA may stay with the current department division or adopt DevOps model recommended by earlier adopters and IT industry observers.

5. NEXT STEPS

eRA can take advantage of cloud-based products and services to address certain business needs such as increasing program ability to respond to changing business needs, reducing costs, and providing new services and features to our users. The proposed strategy calls to start with cloud by supplementing, not replacing the existing solutions. As the experience and level of confidence grow, the program can take

more active steps to realize benefits of flexibility and efficiencies from offered by cloud providers.

The program should start by establishing the contract with AWS VAP that provides wide range of services and has sufficient knowledge and experience to support our efforts. One the necessary legal and financial framework is put in place, the program can start using AWS services that do not require additional efforts from the NIH divisions outside of eRA. The following initiatives can be done in parallel, independently from each other:

1. Use of S3 object store for offline data management

   This activity will be mostly limited to Operations team. This will include defining current resources and associated costs, and identify specific items and expected savings to ensure specific goals are set for this effort. This will be followed by making appropriate changes to the identified processes and procedures assisted, if necessary, by introduction of appropriate tools and training.

2. Use of S3 object store for eRA document management

   This effort should be done as part of DocumentService Re-engineering Initiative. The technical approach is outlined in the original proposal prepared for the initiative. As the new DocumentService implementation moves through the environments, the S3 services will be used to store non-production and later production documents.

3. Use of SES as backup e-mail gateway for eRA

   eRA eNotification engine can be enhanced to interoperate with multiple e-mail gateways. By doing so, eRA will obtain ability to send e-mails to user community and internal staff in the event NIH mail gateway becomes overloaded or temporarily unavailable. The implementation will be very straightforward and do not impact any other eRA systems.

4. Use of SNS to provide additional notification channels to eRA users

   eRA eNotification engine can be enhanced to take advantage of SNS to allow delivery of text messages to the users as well as internal staff. This will allow users to receive notification while not being in the office and, in some cases, reduce delay in message delivery. This functionality can also be used expand ability of eRA staff to receive system notifications and alerts. The implementation will be very straightforward and do not impact any other eRA systems.

As a parallel effort, eRA can start using AWS VPC and EC2 infrastructure to supplement in-house resources. These activities require working with CIT/DNST to define network architecture and implement necessary infrastructure. Once this is done, eRA can use it for deploying business applications, systems, and services in the cloud in a sequence of steps:

1. Use AWS VPC and EC2 resources to deploy middle tier applications in the cloud, starting with non-production environments and moving through tiers to enable use of cloud for production systems. eRA can start by supplementing in-house physical servers with cloud-based virtual ones and, over time, reduce reliance on in-house infrastructure following the equipment retirement and changes in lease agreements.

2. One the previous step is accomplished, we can use AWS to implement network edge infrastructure such as termination points, reverse proxies, etc. This will require integration with NIH SSO services in the cloud or use of “NIH SSO bypass” solution as it was done for ASSIST during the pilot implementation.

3. Once the business applications and network infrastructure utilize cloud-based infrastructure, we can consider migrating data services to cloud as well. This effort will require additional planning due to large number of consumers outside of eRA and associated dependencies. It is recommended to include “componentization” of the IMPACII8 data model and de-coupling of components into modernization efforts planned by eRA. The existing systems with minimal data model dependencies, such as ASSIST, may be the first to rely on relational data services in the cloud.

These steps will result in ability to deploy entire business line of services in the cloud or use cloud-based offerings selectively for various parts of the solution as desirable. Combined with the use of CloudWatch monitoring services, this will provide eRA will flexible and reliable platform to meet continuously changing demands of the business community.