

United States Department of Agriculture

Geospatial Data Warehouse Implementation

Kevin Clarke

USDA

Farm Service Agency

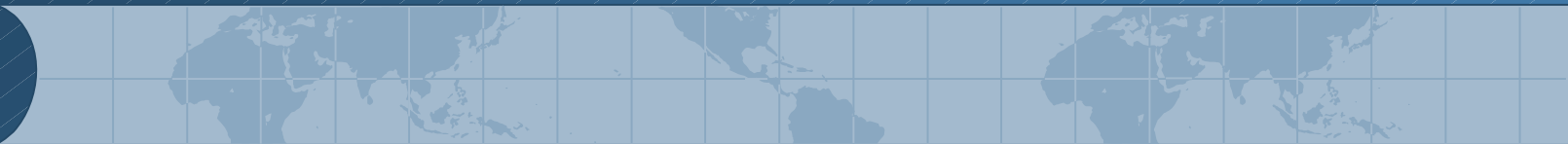
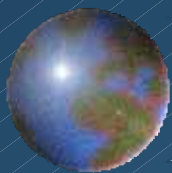
Salt Lake City, UT

Art Ullman

USDA

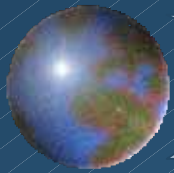
Natural Resources Conservation Service

Fort Worth, TX



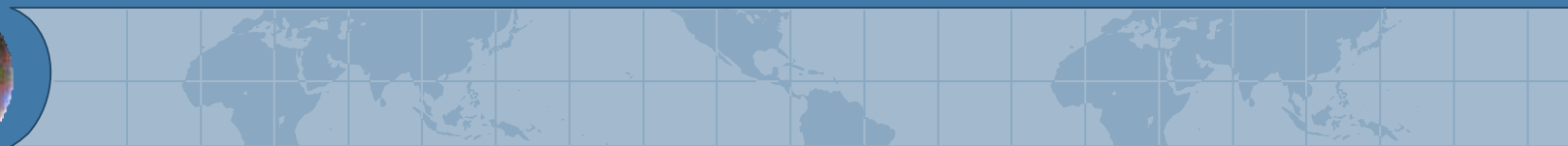
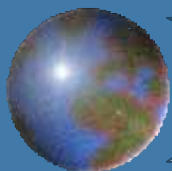
Abstract

USDA has created a geospatial data warehouse for managing and serving authoritative geospatial data to USDA agencies in support of the many agricultural and conservation based programs the USDA administers. The GDW is a multi-site, multi-server warehouse, consisting of raster and vector data sets that are managed by ArcSDE 9 in SQLServer 2000 databases hosted on Windows Server platforms. EMC SAN technology is utilized to provide data storage. All warehouse data is replicated to a fail-over site using a combination of EMC SRDF and SQLServer log shipping. The GDW currently manages in excess of 30TB of raster and vector data via ArcSDE and StorageTek near-line storage systems. This paper will provide a technical design overview of the USDA Geospatial Data Warehouse, and will discuss in greater detail several of the design techniques and implementation details utilized within GDW to maximize the benefit of the substantial cost outlay for the project.

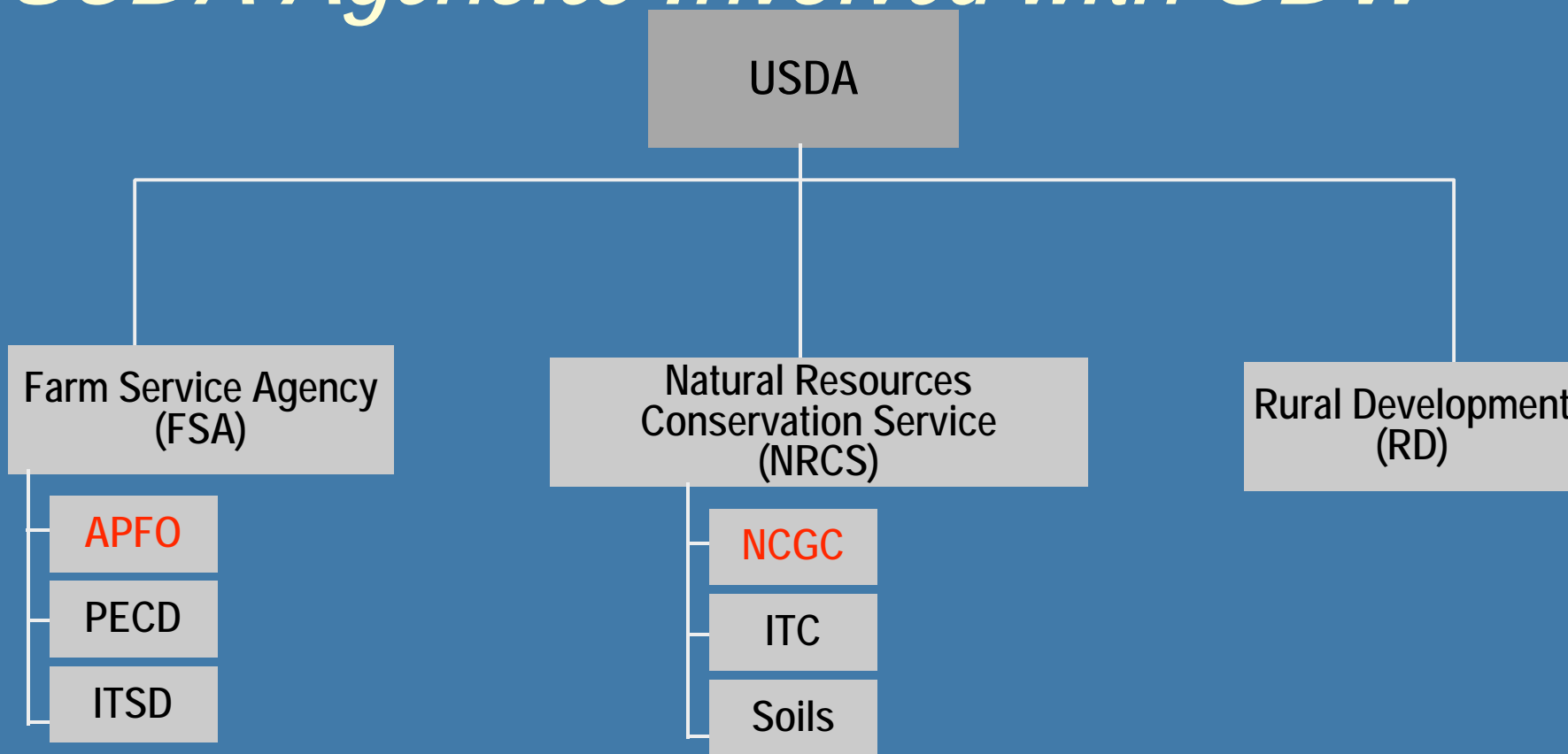


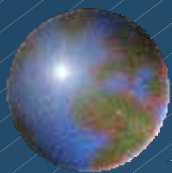
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery



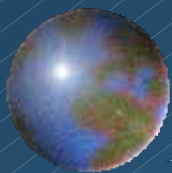
USDA Agencies Involved with GDW





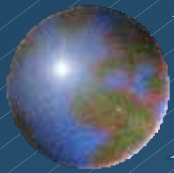
Why only FSA, NRCS, and RD?

- ❖ FSA, NRCS, and RD are considered “Service Center Agencies” and together comprise the majority of the USDA workforce
- ❖ Each of the 2700+ counties in the US has a USDA office called a Field Service Center where land owners and producers within that county conduct their business with the 3 agencies (crop reporting, loan applications and payments, and conservation enrollment and payments, etc)
- ❖ The county Service Center is where the delivery of the numerous FSA, NRCS, and RD farm programs occurs, giving producers access to the programs



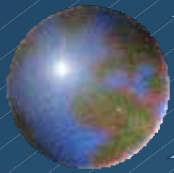
Agency GIS Requirements

- ❖ The 3 agencies acquire, enhance, store and distribute ortho imagery (3.75' and 7.5'), DRGs, and vector data sets to the Service Centers in each county to support program delivery
- ❖ Data Centers acquire and then enhance the imagery and vector layers and then distribute to the county offices for use in GIS applications
- ❖ Crop reporting, acreage adjustment, disaster assessment, conservation enrollment, soils analysis, etc



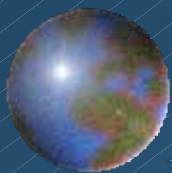
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery



What is the Geospatial Data Warehouse?

- ❖ Technical Architecture, processes, and infrastructure through which USDA tabular and spatial data is developed, managed, and distributed to the 2700+ USDA county offices and other authorized users
- ❖ Not a “Warehouse” in the traditional sense
- ❖ Plural entity comprised of 2 Data Centers and 3 Web Farms/Data Marts



GDW – A Plural Entity

❁ Two Geospatial Data Centers

- ❁ Salt Lake City, UT and Fort Worth, TX
- ❁ Geospatial Data Production and Warehousing
- ❁ Data Delivery (CD-ROM, DVD, FTP)

❁ Three Web Farms/Data Marts

- ❁ Fort Collins, CO, Kansas City, MO, & St. Louis, MO
- ❁ Data Delivery for Web Applications
- ❁ High Availability, Secure Public Access
- ❁ ETL data from Data Centers to populate the Marts
- ❁ Use web/map services to access Data Center data via an authenticated proxy request from the Web Farm

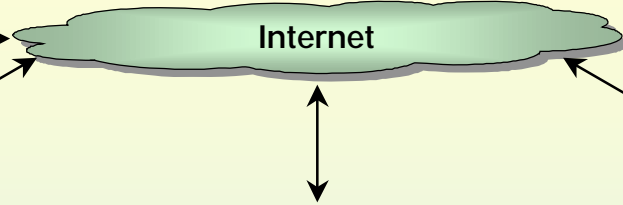
USDA Geospatial Data Warehouse



Field Service Centers



Ag Users/Partners



NRCS at Fort Collins, CO
Geospatial Data Mart in Web Farm

FSA at Kansas City, MO
Geospatial Data Mart in Web Farm

RD at St. Louis, MO
Geospatial Data Mart in Web Farm

Data Themes:

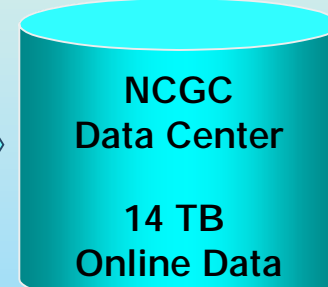
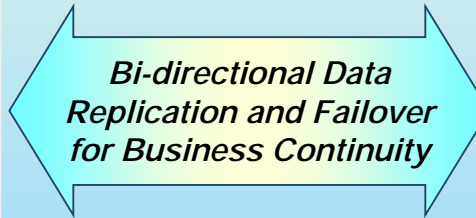
- NAIP
- MDOQ
- CLU
- DRG
- Soils (SSURGO)
- Hydrography
- Geopolitical
- Transportation
- Elevation

ESRI Tools:

- ArcGIS
- ArcIMS
- ArcSDE
- ArcGIS Server
- Python



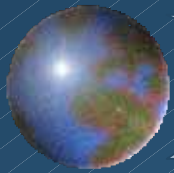
Aerial Photography Field Office
Salt Lake City, UT



National Cartography and Geospatial Center
Fort Worth, TX

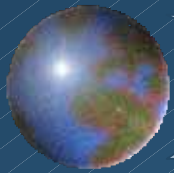
USDA Implements Geospatial Data Warehouse (GDW) as part of the Service Center Modernization Initiative

The USDA Geospatial Data Warehouse is designed to serve the business and data distribution needs of the three USDA Service Center Agencies: Farm Service Agency (FSA), Natural Resources Conservation Service (NRCS), and Rural Development (RD). Currently approximately 14 TB of geospatial data is managed in two data centers in Salt Lake City, UT and Fort Worth, TX. These two data centers populate geospatial data marts at the three web farms located in Fort Collins, CO, Kansas City, MO, and St. Louis, MO which support mission critical web based business applications for the three agencies. The GDW is designed to provide authoritative data to the 2700+ USDA county-based offices, other governmental agencies, cost share partners, and the agricultural community as a whole. ESRI's suite of GIS products provide the client toolkits, server software, web mapping services, and underlying geoprocessing engines for the data centers and data marts.



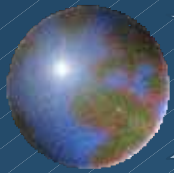
GDW Data Center Environment

- ❖ Base Servers are Dell 2650s/Windows 2003
- ❖ Raster DB Server - 8 CPU and 16GB RAM
- ❖ Most other servers - 4 CPU and 4 GB RAM
- ❖ SunFire/Solaris servers for ASM control
- ❖ SQLServer, Python, EMC DMX 3000 SAN, StorageTek L5510 Near-Line Tape Library
- ❖ Raster databases are partitioned by UTM zone (Largest Raster Catalog ~ 3.2 TB)



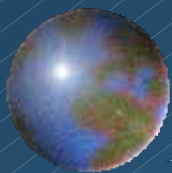
Geospatial Data Warehouse Objectives

- Provide Nationally Consistent Authoritative Data Sets
- Data is Accurate and Current
- Provide High Availability (Multi-site Failover) to support Business Continuance
- Provide a Scalable Architecture
- Provide Web Service Access
- Maximize ROI (Storage Optimization)



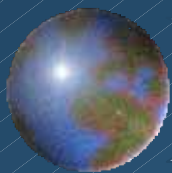
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ **ArcSDE Raster Tile Size Optimization**
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery

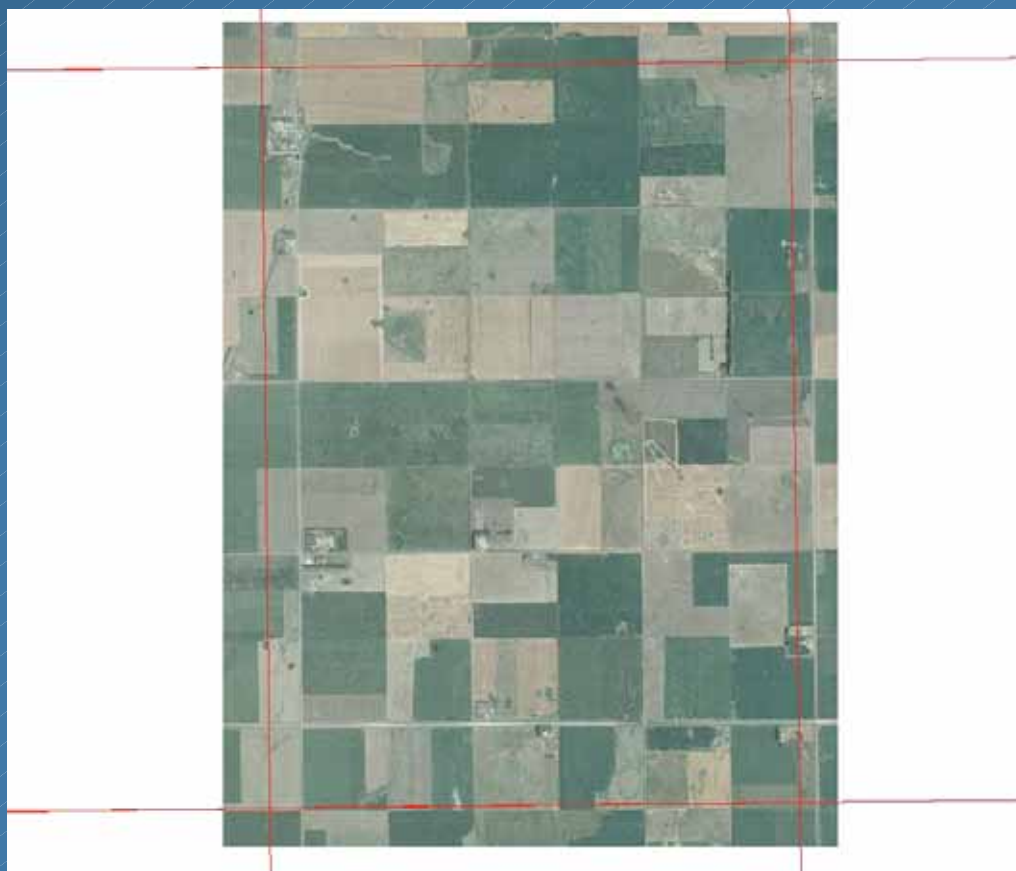


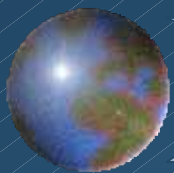
SDE Raster Tile Size Optimization

- ❖ Raw 1m USGS 3.75' Quarter Quads are acquired and mosaicked into into 7.5' Quads on a county basis
- ❖ Perform color balancing, tone matching, radiometric adjustments and edge matching as part of the mosaic process resulting in seamless MDOQ coverage
- ❖ Acquisition of NAIP "leaf on" imagery in 1m Quarter Quad format is also occurring
- ❖ Black and White, Natural Color, and Color Infrared formats
- ❖ All acquired and enhanced imagery includes a 300m buffer



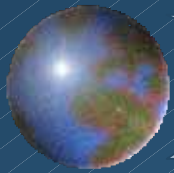
Quarter Quad Image with Buffer *USGS 3.75' (1:12,000) Topo Index*





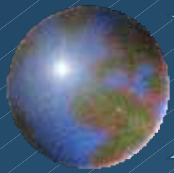
How Tile Size Affects SDE/SQLServer

- ❖ Storage – Larger tiles yield fewer database records and index entries, decreasing overhead
- ❖ Compression – Smaller tiles preserve locality of patterns and yield better compression rates with LZW, and better compression leads to efficient bandwidth utilization since SDE transfers compressed tiles to requesting clients



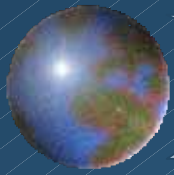
How Tile Size Affects SDE/SQLServer

- ❖ Load Time – Larger tiles can be “created” faster inside SDE since the original image needs to be cut into fewer tiles and fewer database records need to be generated
- ❖ Display/Extraction – Larger tiles can be stitched together faster to generate the image to the requestor



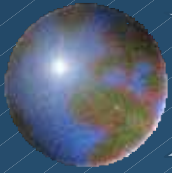
GDW Priorities were:

1. Maximize Storage Efficiency
2. Minimize Load Time



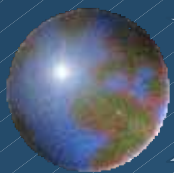
SDE Tile Size Testing Results

- ❖ ESRI recommends 128 x 128
- ❖ Our tests resulted in different tile sizes for BW, NC, and CIR
- ❖ BW - 168 x 168
- ❖ CIR - 192 x 192
- ❖ NC - 176 x 176



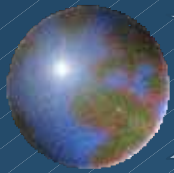
Benefits of Custom SDE Tile Size Settings

- ❖ 10% to 20% better SDE compression ratios inside SQLServer, depending on imagery type, which reduces total disk outlay
- ❖ 5% to 10% improved loading performance depending on imagery type, which decreases ingestion times



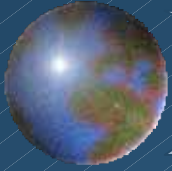
Raster Ingestion

- ❖ Developed a multi threaded Python Load Manager
- ❖ Parameterize the number of threads, etc
- ❖ Raster DB Server we can run a maximum of 8 threads (1 per CPU) without over utilizing the server
- ❖ Ability to ingest 500 GB in a 24 hour period



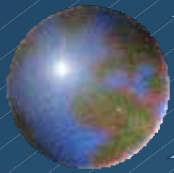
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ **Raster Data Management/Data Model**
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery



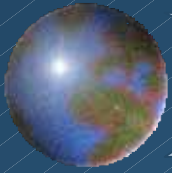
GDW Raster Data Model

- ❖ Utilize Raster Catalog
- ❖ By definition it eliminates the overlap of the 300 m buffers around each MDOQ Quad or NAIP Quarter Quad
- ❖ Employed a hybrid mosaic mechanism
- ❖ Each raster in the catalog comprises all the Quads (up to 64) of a 1 Degree Block
- ❖ Catalog contains 1 raster for each 1 Degree Block in the associated UTM zone



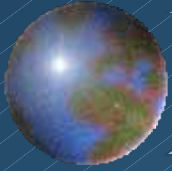
Hybrid Mosaic Mechanics

- ✿ Very first raster (ie no catalog exists):
sderaster -o import -n "35096"
- ✿ Subsequent rasters within 35096
sderaster -o mosaic -v 1
- ✿ v = select image from edoq_bw_14 where
name = '35096'
- ✿ New 1 degree block within existing catalog
sderaster -o insert -n "36100"



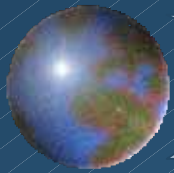
Raster Load Manager

- ❖ Each loading thread (max 8) will discover sources of imagery to load and will hold these sources as their own
- ❖ Utilizes a locking mechanism so multiple threads do not attempt to load into the same 1 degree block inside the raster catalog (SDE does not like this)



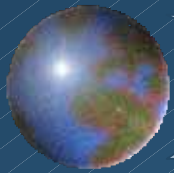
Pyramid 33% Dilemma

- ❖ Current 14TB of raster data would require an additional 5TB of disk for pyramids
- ❖ Needed to find a less expensive solution given that the warehouse was going to continue to grow to over 20TB of raster data within 2 to 3 years of coming online



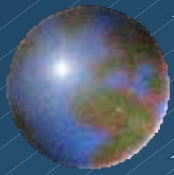
GDW Pyramid Solution

- ❖ Do not create pyramids on the full 1 meter layer
- ❖ As part of the loading procedures, create a reduced resolution 4m navigation layer
- ❖ This navigation layer resides in a separate raster catalog and all Quads/QQs for the entire UTM zone are moseiacked into a single raster
- ❖ Create pyramids on this 4m layer and present this layer at smaller scales ($< 1:10,000$) thru ArcIMS services
- ❖ Navigation layers with pyramids consume ~ 1TB (versus 5TB)

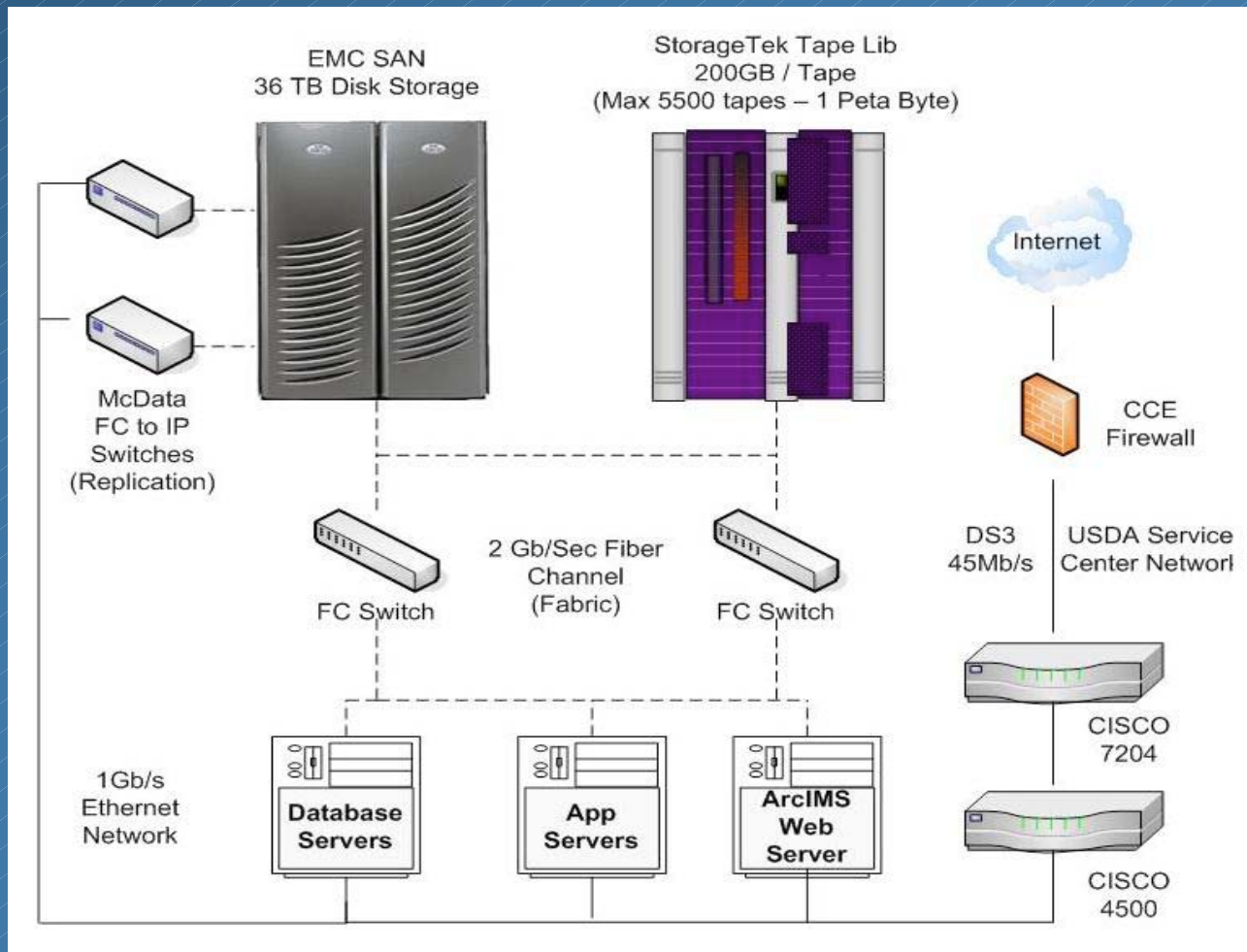


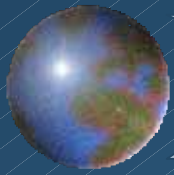
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ **SAN Architecture**
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery

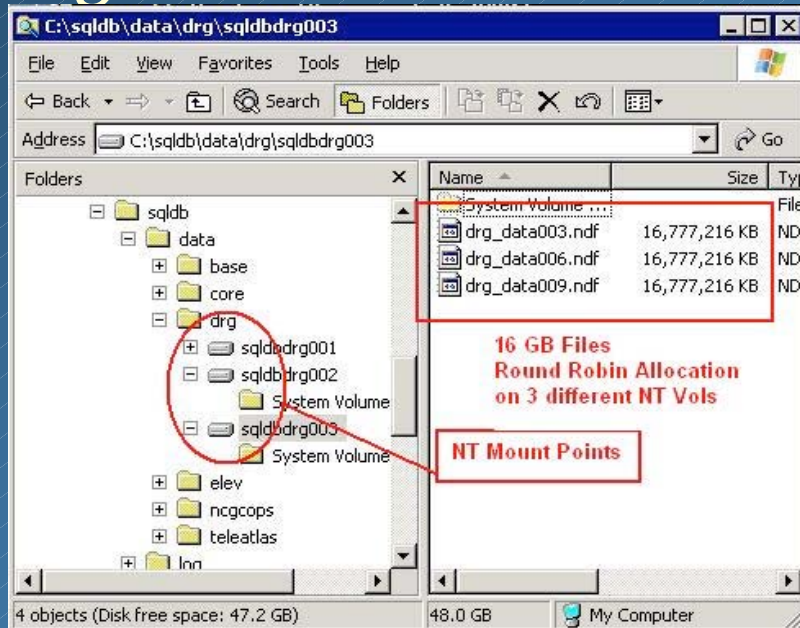



Hardware Infrastructure





Large Database / SAN Implementation

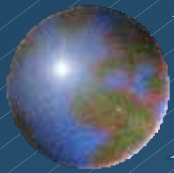


| General | | | |
|---|---------|------------|-----------|
| Table Info | | | |
| Wizards | | | |
| drg  | | | |
| Data: | | | |
| drg_sys | 64MB | 13.5MB | 50.5MB |
| drg_data001 | 16384MB | 15281.88MB | 1102.12MB |
| drg_data002 | 16384MB | 15281.94MB | 1102.06MB |
| drg_data003 | 16384MB | 15282.38MB | 1101.62MB |
| drg_data004 | 16384MB | 15281.69MB | 1102.31MB |
| drg_data005 | 16384MB | 15281.88MB | 1102.12MB |

Spread (round robin) Database Files out on multiple volume/mount points/LUNs on the SAN. NT Mount Points allow for LUNS and simplified management of space.

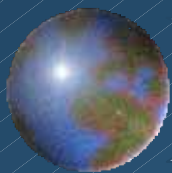
More spindles, more SAN Cache, better parallel processing for SQL Server queries and database creation.

Use AWE/PAE options in SQL Sever / Win2003 to support 16GB of memory.

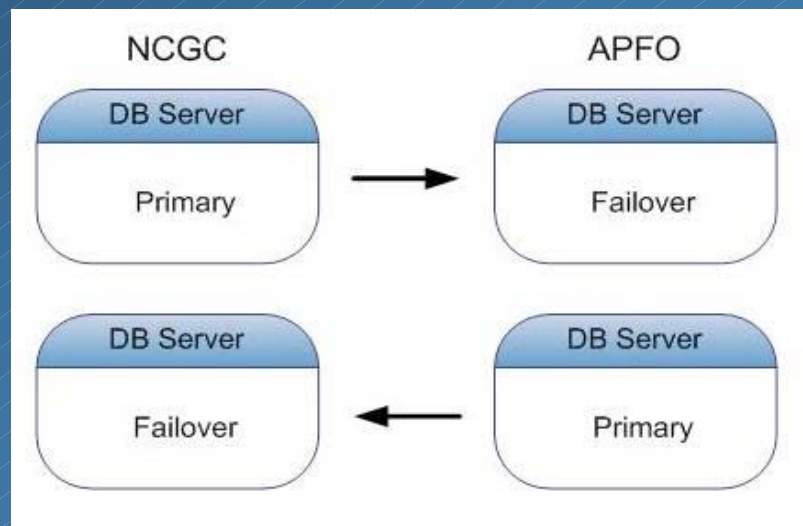


Agenda

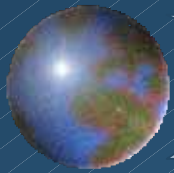
- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ **SQLServer Replication (Business Continuance)**
- ❖ ArcIMS Load Balancing using VMWare
- ❖ Web Services/Data Delivery



Database Replication



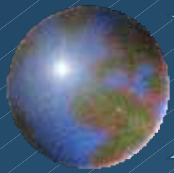
- ❖ Primary Database Servers at NCGC and APFO
- ❖ Failover Database Servers at NCGC and APFO (support failover for other site)
- ❖ Each site is an active data production/delivery site that acts as failover to the other site
- ❖ The GDW is a multi-site warehouse with redundancy and failover



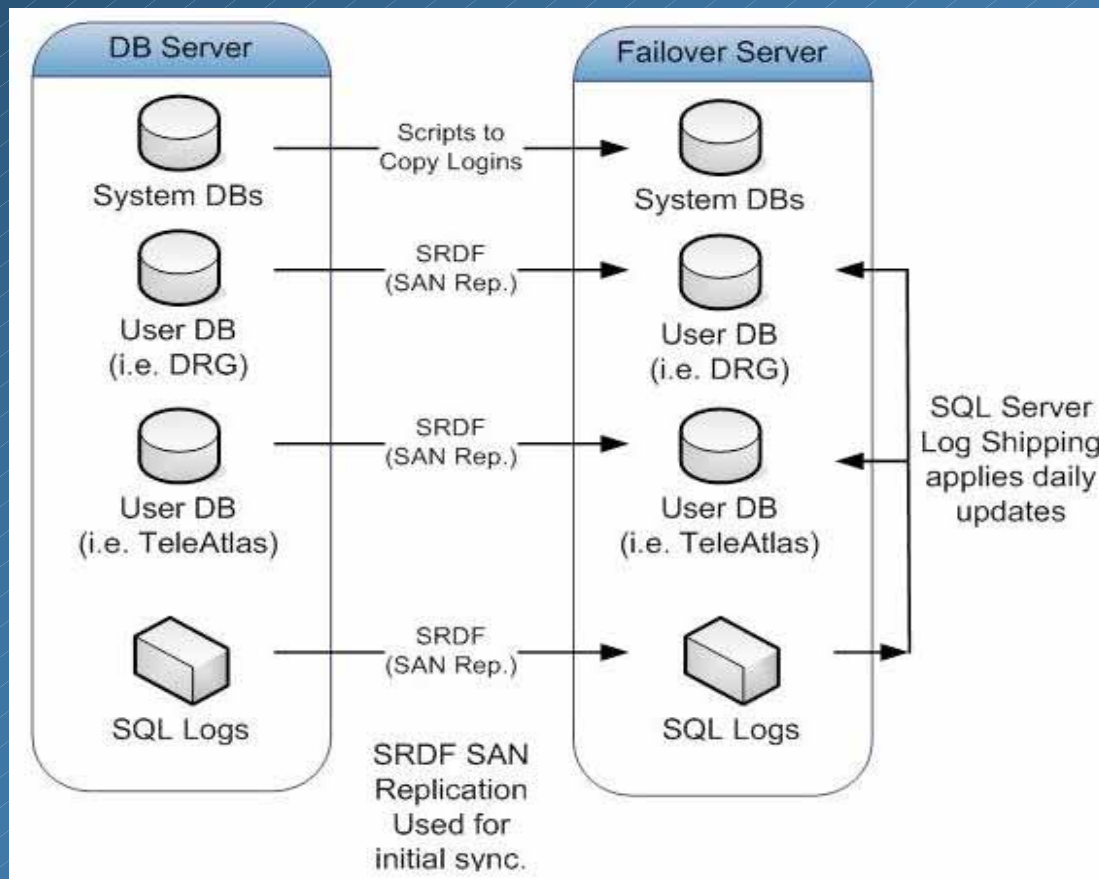
Replication Solution

Hybrid of SQL Server Log Shipping / SAN SRDF

- EMC SRDF Adaptive Copy technology provides very high speed SAN based volume replication (for synchronization and fail-back.)
- SQL Server Log Shipping provides a very flexible and efficient architecture for pushing database updates between sites.

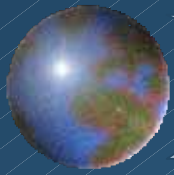


Replication Architecture



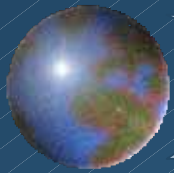
Steps -

- Initial Sync using SRDF (SAN Rep)
- Split Remote Mirror after Sync Complete
- Mount Failover Database in Recovery Mode (using EMC TSIM module)
- Push all updates using SQL Server Log Shipping



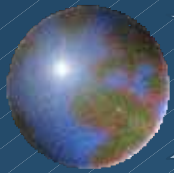
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ **ArcIMS Load Balancing using VMWare**
- ❖ Web Services/Data Delivery



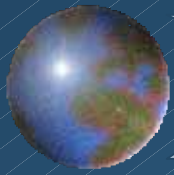
Network Load Balancing Overview

- ❖ Built into Windows 2003
- ❖ Load Balances IP Requests
- ❖ Supports scheduled or unscheduled node failure / shutdown
- ❖ Load Balances up to 32 Nodes
- ❖ Requires *stateless* applications (i.e. no session variables on web apps) to load balance requests
- ❖ Host Affinity can be set for *statefull* applications to support failover, but this will not provide load balancing

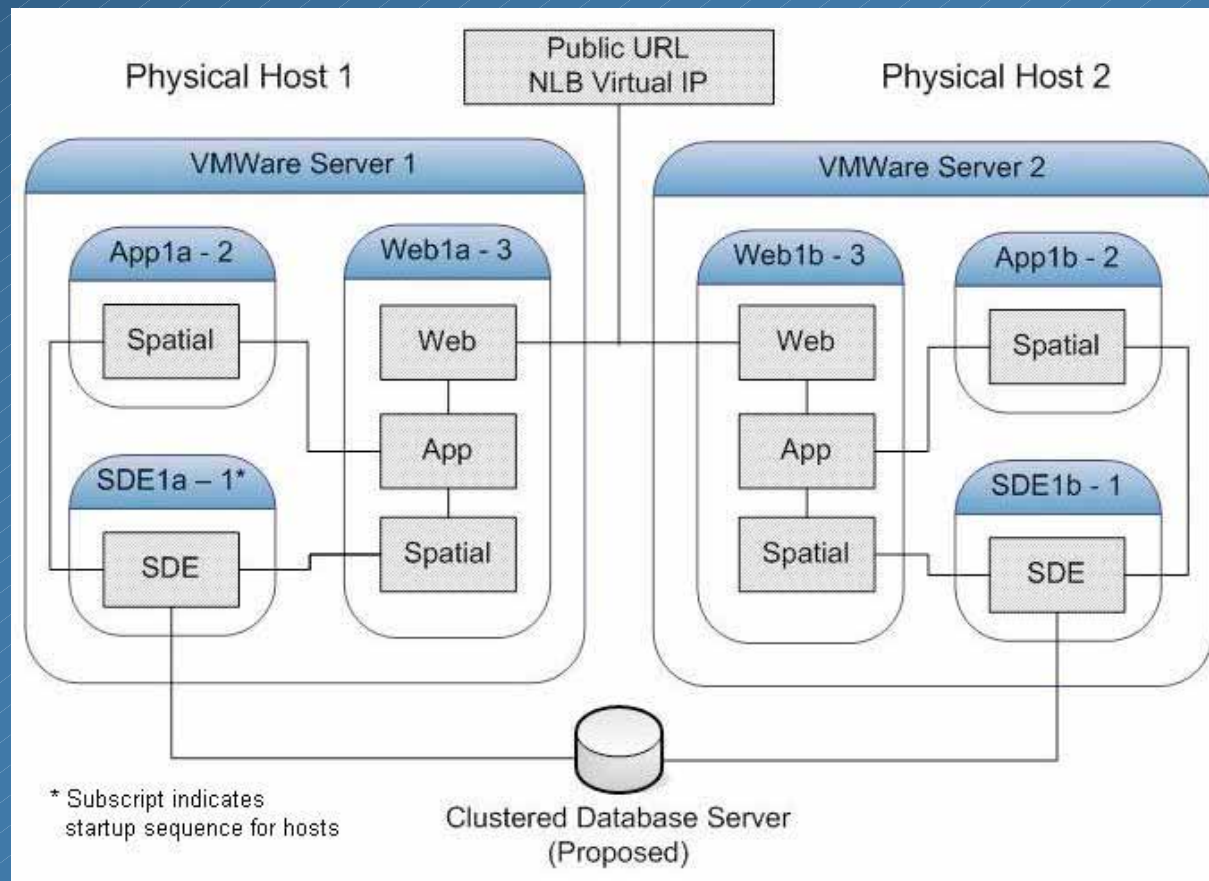


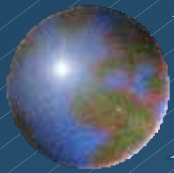
Virtual Server Implementation for ArcIMS Web Servers

- ❖ Server Virtualization using *VMWare ESX Server* for partitioning Window/Linux server on larger SMP servers
- ❖ Redundancy at a lower cost point
- ❖ Simplified Management
- ❖ Scalability at lower incremental cost
- ❖ Better utilization of resources (savings in HBAs alone pay for the software)
- ❖ Lower costs allow for higher end hardware (including SAN usage for all virtual hosts)



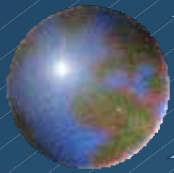
Map Services Load Balancing Architecture





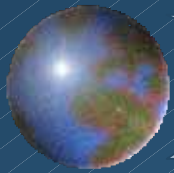
Agenda

- ❖ Overview of the USDA Organization
- ❖ Overview of the GDW
- ❖ ArcSDE Raster Tile Size Optimization
- ❖ Raster Data Management/Data Model
- ❖ SAN Architecture
- ❖ SQLServer Replication (Business Continuance)
- ❖ ArcIMS Load Balancing using VMWare
- ❖ **Web Services/Data Delivery**



GDW Data Access and Delivery

- ✦ Online Access to GDW Data
 - ✦ ArcIMS Map Services
 - ✦ Other WMS Compliant web services
- ✦ Offline Data Delivery System
 - ✦ USDA Geospatial Data Gateway
 - Public Data Delivery (FTP, CD, DVD)
 - Portal to GDW, USDA TerraServer and other USDA data sources
- ✦ Local Data Marts
 - ✦ Localized secure access to sub/superset of GDW data



Online Data Access

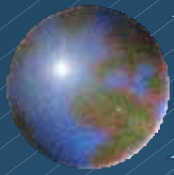
✦ ArcIMS Web Services

✦ Requirements:

- ✦ High available data access

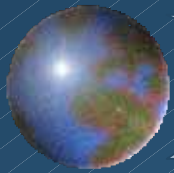
- ✦ Scalability

- Support for potentially all geospatial users within the USDA WAN
- Scale-out architecture – so that new servers can be added incrementally to support growth



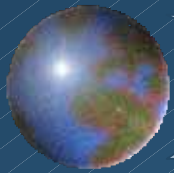
Offline Access and Data Marts

- ❖ Need for local copies of data
 - ❖ Local Data Marts
 - ❖ Offsite Partners
 - ❖ Un-dependable network connectivity
- ❖ Need to push data back to GDW
 - ❖ Updated Data Sets
 - ❖ Centralized Storage of new data themes from partner agencies and remote county offices
- ❖ Security and firewall requirements need to be addressed



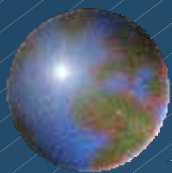
Data Distribution and Collection

- ✿ USDA Geospatial Gateway
 - ✿ Offline Provisioning / ordering through USDA Portal and public FTP site.
- ✿ OnCourse/Signiant Software for Data Transfers
 - ✿ Push/Pull technology that works well with firewalls
 - ✿ Multi-Streamed Data Delivery
 - ✿ Auto-restart for dropped connections
 - ✿ Encrypted communication for secure transfers to business partners outside USDA Network
 - ✿ Validated data transfers (guaranteed data delivery)
 - ✿ Web Management of data transfers



Data Push/Pull Technology

- ❖ EMC OnCourse/Signiant software pushes and pulls file changes between sites
- ❖ Backup SQL Server databases to multiple files (for multi-streamed data transfers), and push to partner sites
- ❖ Scanned images are pushed from remote sensing labs to data warehouse
- ❖ Data is encrypted and offsite hosts are validated using public key/private key technology



GDW Summary

Centralized Management of Geospatial data ensures:

Accuracy

- single authoritative data source

Redundancy

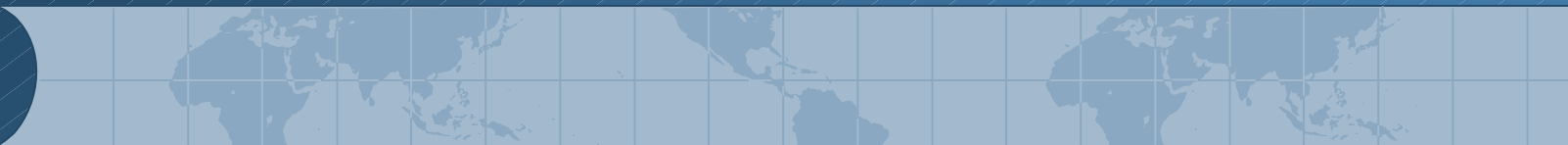
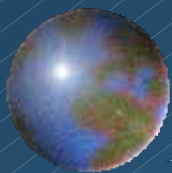
- replication technology

Availability

- load balanced web servers running multiple VMWare servers

Flexibility

- Customized data marts
- Gateway/Portal access for delivering offline downloads of data
- ArcIMS providing online web services



Contact Information

Kevin Clarke

USDA

Farm Service Agency

Salt Lake City, UT

kevin.clarke@apfo.usda.gov

<http://www.apfo.usda.gov>

Art Ullman

USDA

Natural Resources Conservation Service

Fort Worth, TX

art.ullman@ftw.nrcs.usda.gov

<http://www.ncgc.nrcs.usda.gov>

Visit the FSA and NRCS booths in the Exhibitor Area for additional information and a GDW demonstration