Designing and Deploying a Distributed Architecture for Research Data Management

Luke Sheneman, Ph.D

Technology and Data Services Manager
Northwest Knowledge Network (NKN)

Presentation to IS @ WSU
July 2014

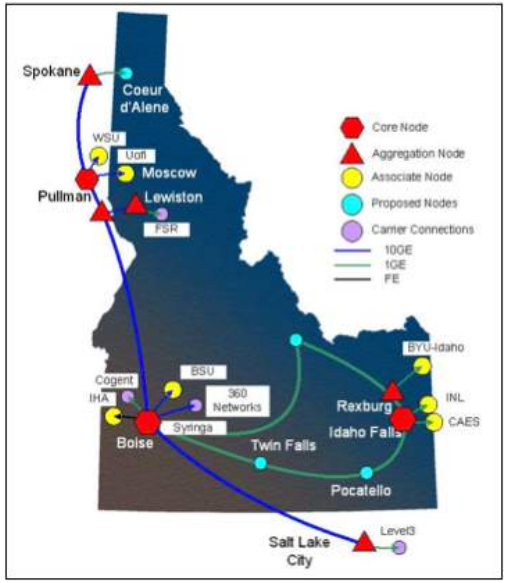
My Background &
Northwest Knowledge Network

- B.S. Computer Science from UofI (1990-1995)
- IT Work in Silicon Valley (1996-2002): Netscape, BigVine, Inktomi
- Ph.D Bioinformatics and Computational Biology (2002-2008)
  - Algorithm design for phylogenetics and molecular sequence alignments. Evolutionary Computation.
- IT Architect for Northwest Knowledge Network (2010-2014)
- Currently Technology and Data Services Manager for NKN
- NKN: Research Computing Support / Data Management
  - Support from UI Research Office, NSF, USGS, USDA, etc.
  - Repository / Catalog for Scientific Research Data Products
  - Scientific Metadata

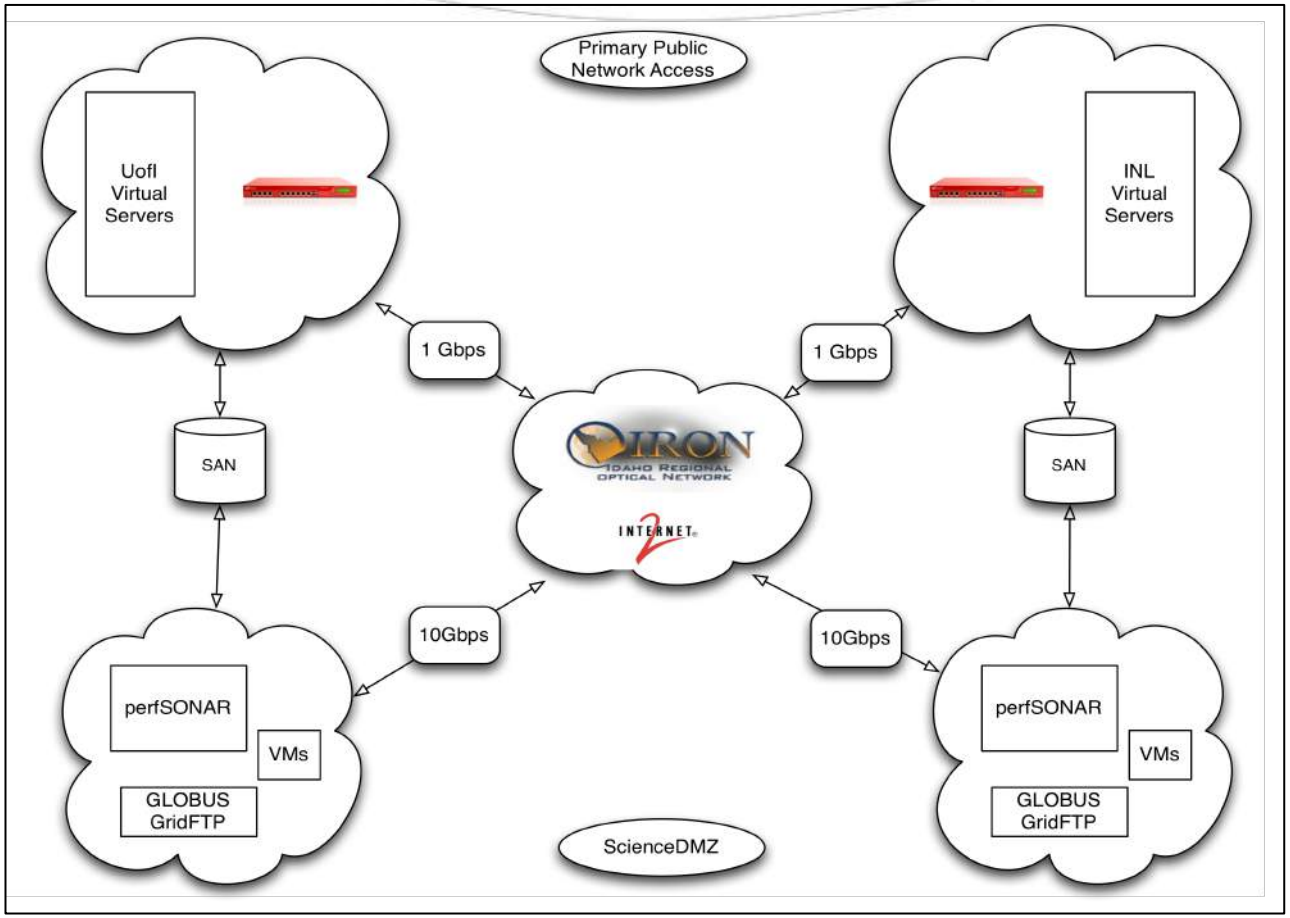
NKN Architecture Design Goals

- Distributed Datacenters
  - Backup, Recovery, Load Balancing, Failover
    - University of Idaho
    - Idaho National Laboratory
- Scalable Enterprise Storage
- Flexible Virtualized Server Environment
- Entirely Redundant Components
- Security: Tightly Controlled and Managed

IRON
IDAHO REGIONAL OPTICAL NETWORK



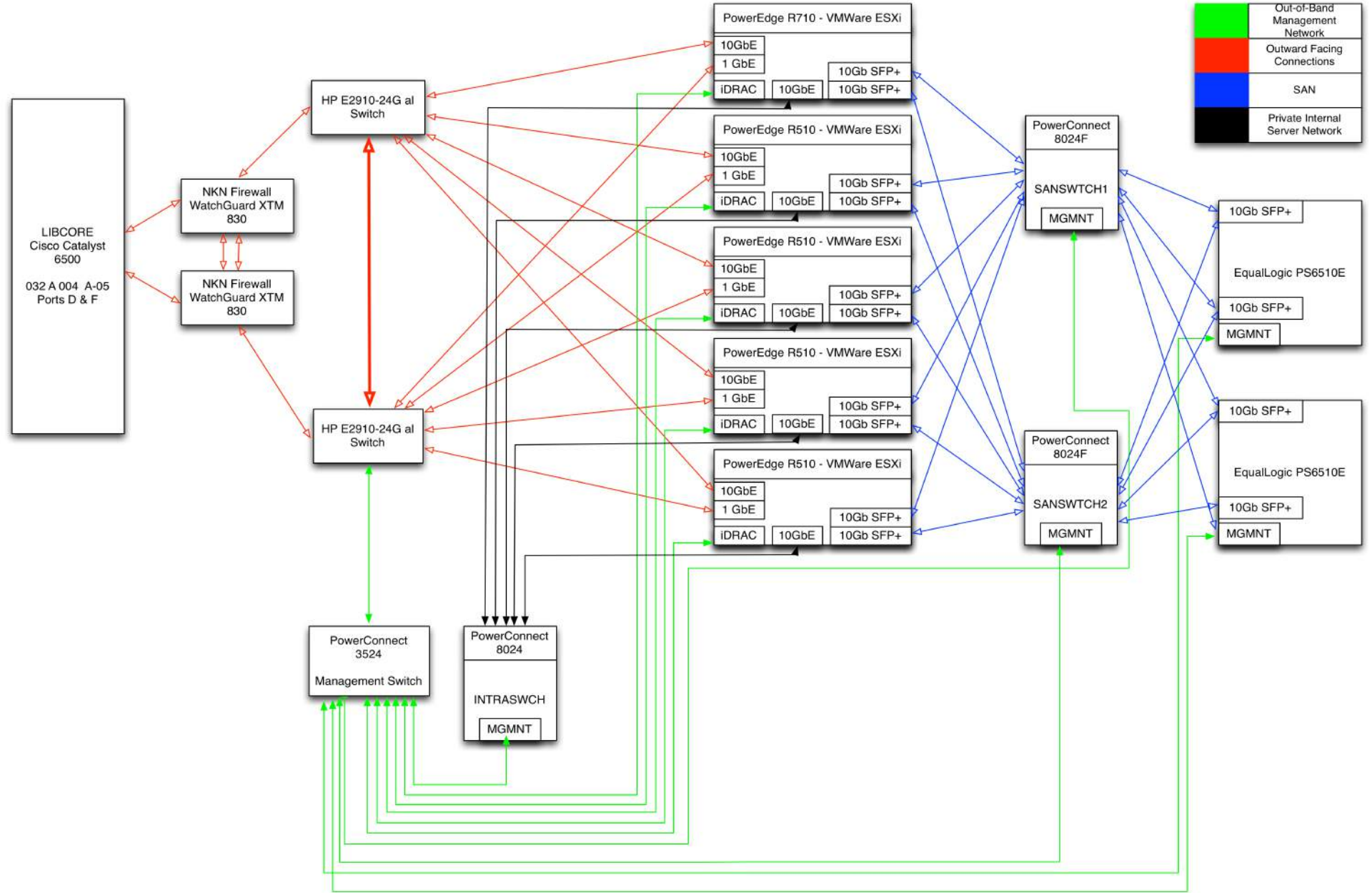
High Level NKN Network Architecture



Highlights:

- Dedicated Firewalls (Watchguard)
- Science DMZ
  - perfSONAR
  - Dedicated File Transfer
  - No Firewall
- Secure Shared SAN
- NSF CC-NIE

NKN Network Design



Network Address and Access Specifics

- Public /26 and /27 IP Address Space at UofI and INL
- Private /24 for SAN traffic across two 10Gbps PowerConnect 8024 Switches
- Private /24 for 10/100 Management Network
- Private /24 for dedicated 10Gbps server-to-server network
- VPN over SSL (port 443) Provided by WatchGuard XTM 830
- WatchGuard routes between all private/public networks. Packet filtering rules apply.
- WatchGuard provides Firewall authentication (via NKN LDAP)

NKN Enterprise Storage

- **Dell EqualLogic PS6510E**
  - High drive density (48 3TB disks in a 4RU Chassis)
  - 10Gbps NICs – Copper or SFP+
  - iSCSI – Initiators from ESXi Hypervisor or VM level
  - Simple Configuration and Management
  - Linear performance scalability – each chassis has multiple controllers
  - Each Chassis Configured RAID6 with 7200 RPM SATA
  - Scalable to several petabytes within one management group
  - Expansion requires no further networking (SAN Switch) investment
  - Same HW vendor (Dell) – support and integration tools
  - Good VMWare Support (host integration tools, EqualLogic multipath extension modules)
  - Snapshots, thin-provisioning, strong ACL and auth support, monitoring tools

Server and VMWare ESXi Environment

- Hardware is cookie cutter Dell PowerEdge R510, R710
- Maximum RAM Configuration, RAID Level 1 Boot into ESXi
- Run free/inexpensive versions of ESXi and vSphere (VMWare Essentials). No vMotion, etc. due to initial cost.
- Shared iSCSI datastores for all VMs and VMDKs – maximum flexibility for rapid VM recovery or migration
- iDRAC to dedicated private 10/100 management network

Rich Internal Linux Infrastructure

- Approx. 55 VMs between UI and INL
- Run primarily RHEL 6.5 – Academic licenses
- Some Windows Server 2008 and Ubuntu
- OpenLDAP on RHEL VM
  - NKN LDAP auth enables: Unix accounts, firewall, web applications (Drupal), databases, SAMBA SMB/CIFS, ArcGIS, ownCloud
- NFS & Samba, NTP, SSH, FTP, Jabber (openfire), rsync, Subversion, SMTP (postfix), IMAP (dovecot), OPeNDAP (THREDDS)
- Dozens of websites (Apache httpd, Tomcat, Drupal)
- Databases: MySQL, PostgreSQL

Data Replication

- **Synchronous**
  - GlusterFS "private cloud" distributed file system configured as a replicated volume using bricks at UI and INL.
  - Initially for main portal datastore (2011-2013).
    - WAN/IRON Bandwidth Sufficient
    - WAN/IRON Latency Problematic

- **Asynchronous**
  - Simple hourly or nightly rsync between VMs and single backup target at INL.
  - Bash scripts run out of cron with locking
  - Dump all databases to files prior to rsync

Handling Large Filesystems

- Single gridded climate model prediction dataset > 30TB
- EqualLogic limited to 15TB iSCSI targets
- ext4fs can handle 1024 TB volumes.
- But…e2fsprogs tools (mkfs.ext4) compiled as 16-bit on RHEL 6 and limited to 16TB.
- Looked at XFS and ZFS as alternatives to ext4.
- Used ZFS, simply adding multiple iSCSI target volumes to a single zpool.
  - Free and easy.
  - Performance and Reliability.
- RHEL 7 has native 64-bit e2fsprogs for full ext4fs support for > 16TB

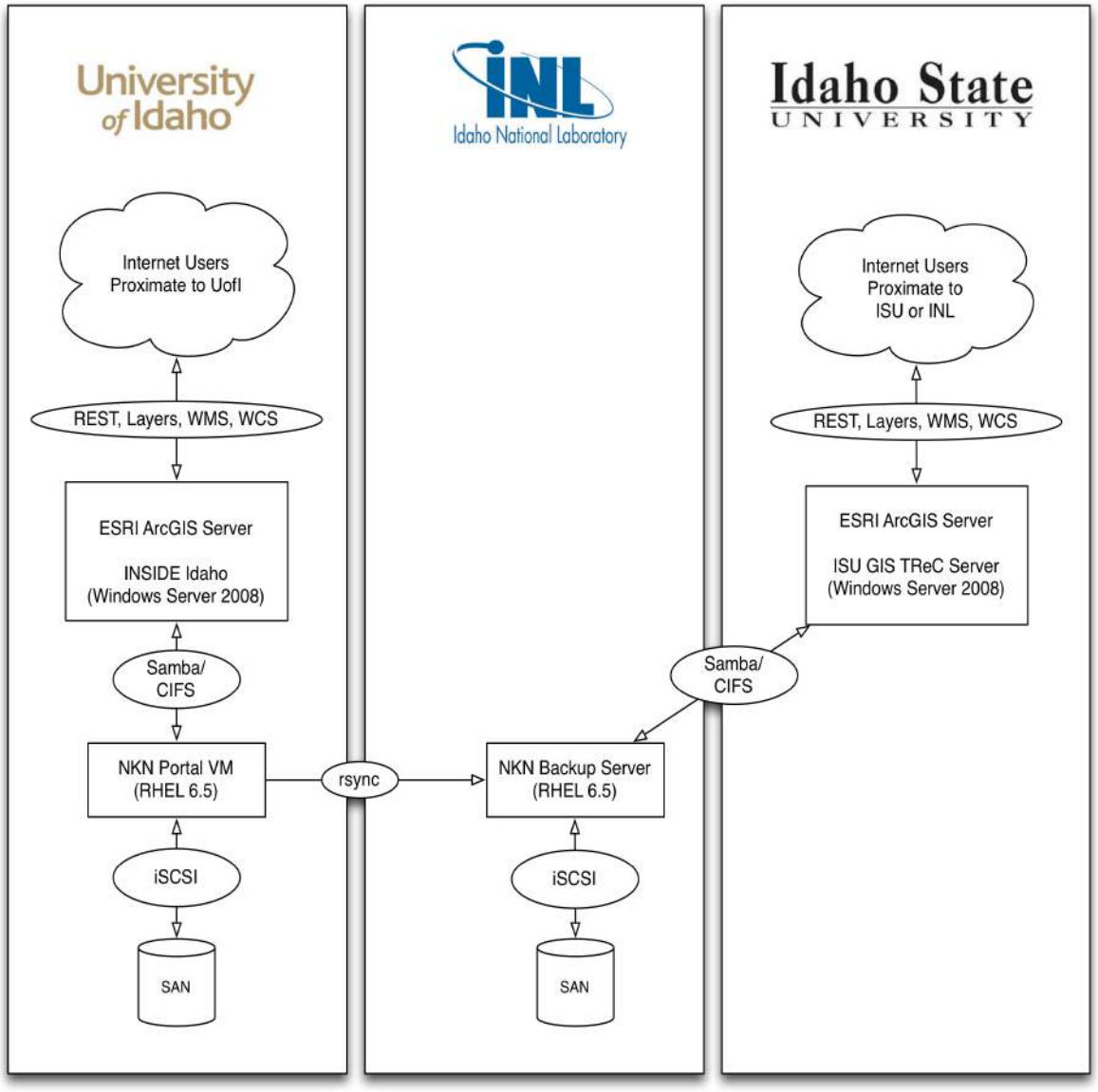
Handling Large Filesystems

```
[root@nknportal sheneman]# zpool list
NAME                SIZE  ALLOC   FREE    CAP  DEDUP  HEALTH  ALTROOT
preprod-datastore  29.8T  22.4T  7.32T    75%  1.00x  ONLINE  -
[root@nknportal sheneman]# zpool status
  pool: preprod-datastore
 state: ONLINE
  scan: none requested
config:

        NAME                                          STATE     READ WRITE CKSUM
        preprod-datastore                             ONLINE       0     0     0
          scsi-36090a0c800c35243643115265002 20cf     ONLINE       0     0     0
          scsi-36090a0c800c3a24b643155505002f0a3      ONLINE       0     0     0

errors: No known data errors
```

```
[root@nknportal sheneman]# df -k
Filesystem            1K-blocks        Used  Available Use% Mounted on
/dev/mapper/VolGroup-lv_root
                       51606140    25208200   23776500  52% /
tmpfs                   8166824           0    8166824   0% /dev/shm
/dev/sda1                495844      102157     368087  22% /boot
/dev/mapper/VolGroup-lv_home
                      402196392      708252  381057704   1% /home
/dev/sdd1           14189693800 11025472760 2431799204  82% /datastore
preprod-datastore   31444568448 24083549056 7361019392  77% /preprod-datastore
filecore.rocket.net:/homespace/sheneman
                     9612386304   704006144 8420099072   8% /nethome/sheneman
[root@nknportal sheneman]#
```

Using the Distributed Architecture

- Asynchronous replication of 15TB portal datastore from UI to INL on 5 minute interval.
- 8TB 2013 National Agriculture Imagery Program (NAIP) Orthoimagery
- Idaho State University (ISU) mounts NAIP data from INL over CIFS and exposes as value-added geospatial web services via their ESRI ArcGIS Server.
- INSIDE Idaho (geospatial clearinghouse) does similar in Moscow.
- Same live imagery data. Replicated across Idaho. Shared live among universities using NKN distributed servers and storage.



Thank You

Luke Sheneman
sheneman@hungry.com

www.northwestknowledge.net