Infrastructure and Services Enabling Academic Research

NITRD MAGIC Team Meeting, Feb-5-2014

John McGee, and the RENCI ACIS Team

renci

RESEARCH \ ENGAGEMENT \ INNOVATION

Today's Agenda

- RENCI architecture and support for High Performance Clusters, including, storage, data, compute, identity management,…
- Lots of details about hardware and software, with some lessons learned
- Discussion of a few research projects and how they build on top this CyberInfrastructure

About RENCI

- An applied research institute of UNC-CH
- We partner with UNC Research Computing on many initiatives
- Our board is comprised of the CROs and Provosts from NC State, Duke, and UNC-CH
- Build new collaborations and accelerate research activities spanning the three universities
- Develop new Cyberinfrastructure capabilities
- Host leading edge services and platforms for research

RENCI's ACIS Team

Casey Averill

VMWare, StorNext, AD, Exchange SQL Server; 90%

Mark Montazer

Data center facilities, cooling, UPS, generator, purchasing, end user support for RENCI; 100%

John McGee

Vision, team lead, team advocate, administrivia; 25%

Marcin Sliwowski

Sr. Linux admin, storage, clusters, 80%

Jonathan Mills

Sr. Linux admin, clusters, master puppeteer; 50%

We are hiring!!

Linux admin, cloud technology (ACIS, ExoGENI)
https://unc.peopleadmin.com/postings/35013

Everyone works on (almost) everything, topics listed are areas of leadership

Locations where RENCI operates equipment

EDC

GSB

GSC

MDC

EDC	Europa Data Center	RENCI Main location
MDC	Manning Data Center	UNC Central IT, Research Computing
GSB	Genome Sciences Building	UNC School of Medicine
GSC	Galapagos Science Center	UNC Center for Galapagos Studies

Europa Data Center: Facilities

- 2000 square feet of floor space on an 18 inch raised floor

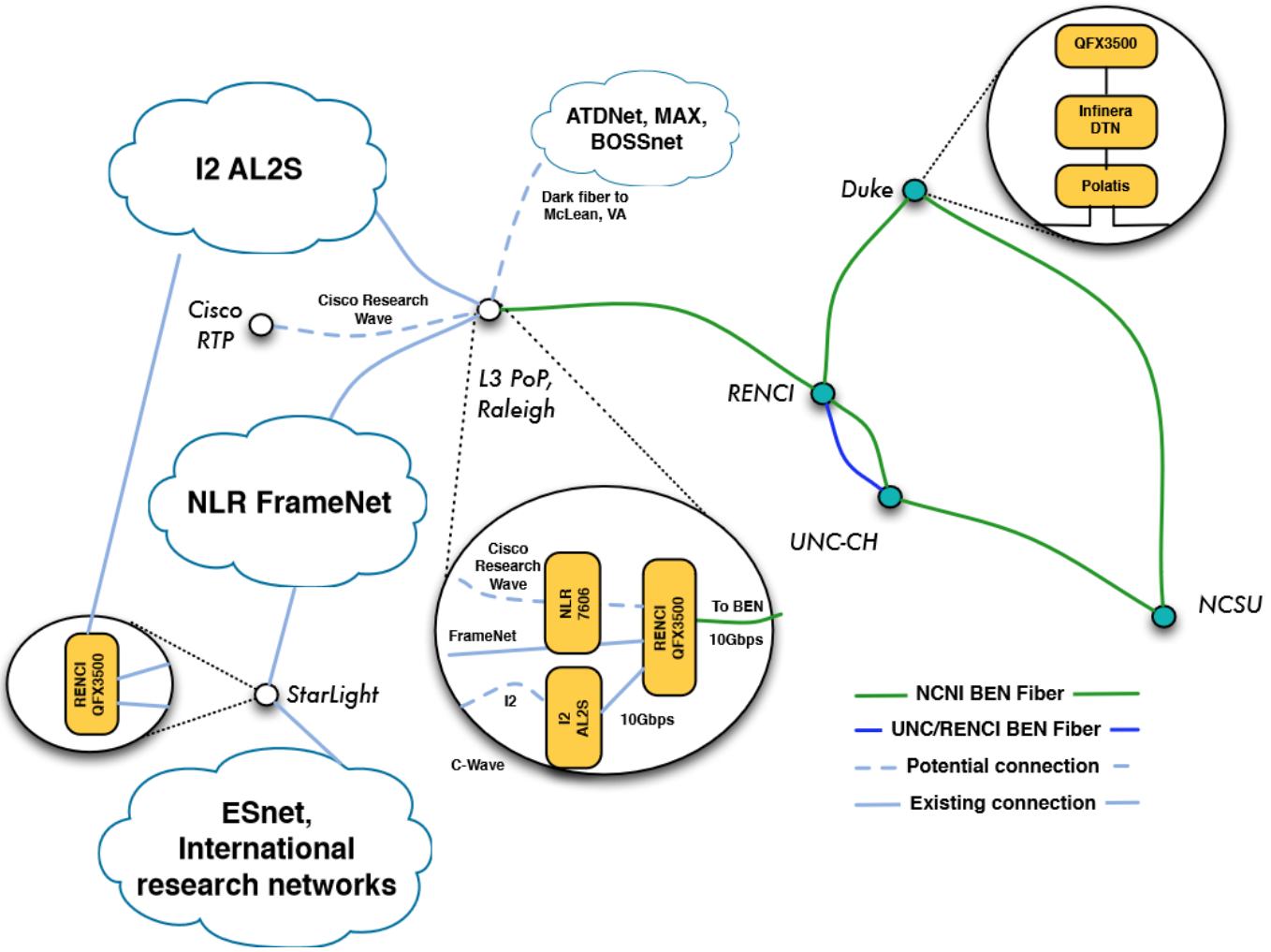
- 600 kva commercial power
- 375 kva UPS power
- 20 kva generator power

- 134 tons dedicated cooling

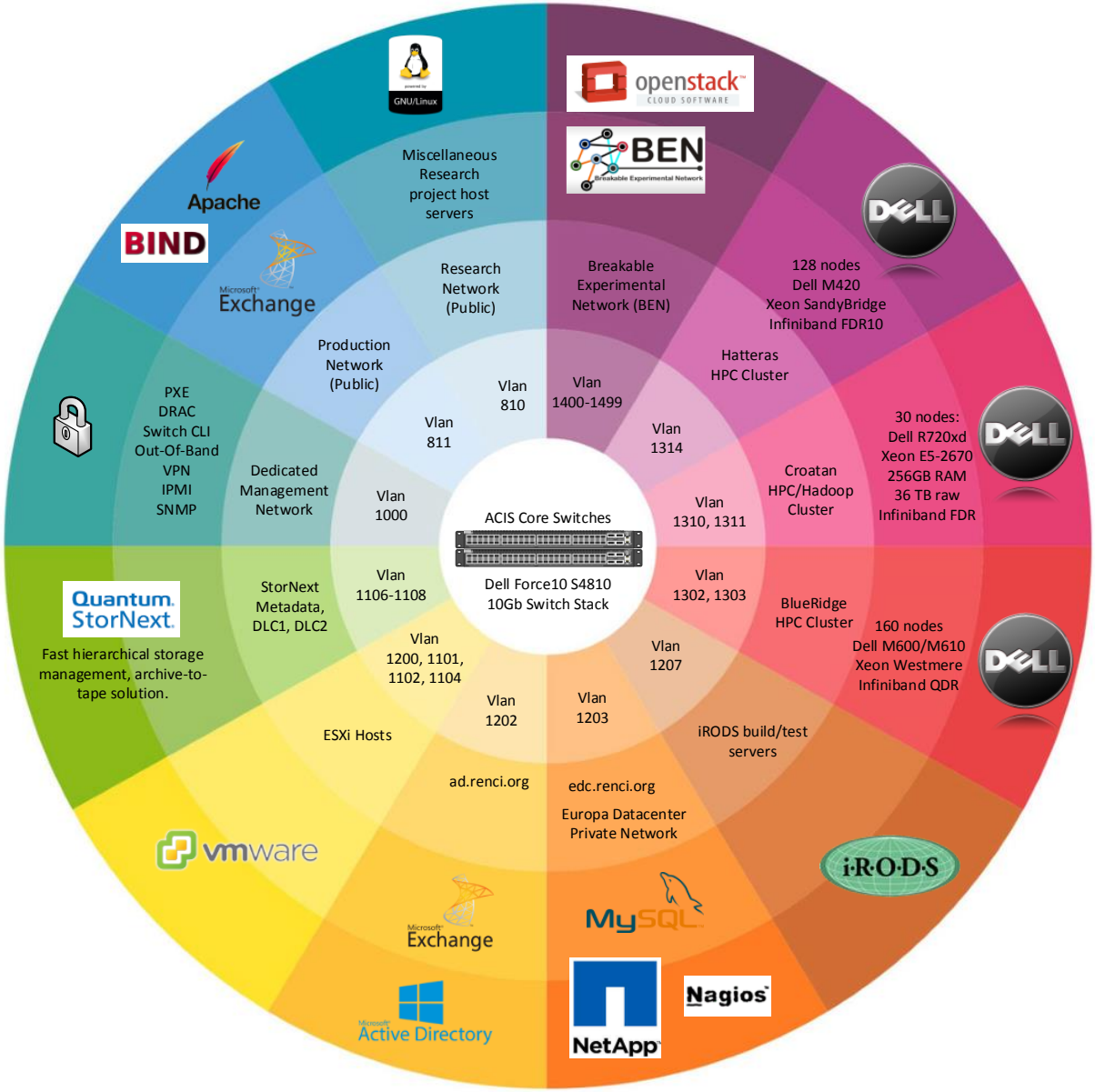
- Room for 40 Racks

EDC: RENCI's main location

Network Overview



Extensive use of vlans



Storage Strategy

- Highly diverse set of requirements, both in terms of performance characteristics, and in connectivity mechanics, eg: Fiber Channel, iSCSI, NAS

- Compute and data analytics clusters
- Bulk transfers to/from other data centers
- Underlying capabilities for iRODS data grids
- Archive, for some definition of the word
- Genomics, Environmental Science, National Collaborations

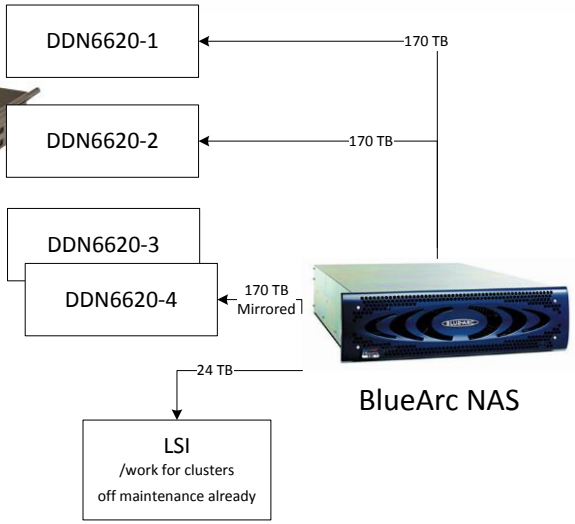
- We optimize for flexibility
- Minimize the human cost of enterprise storage management

Storage: Current Generation

1. NetApp FAS – research; 1.1PB
2. NatApp FAS – production; 50TB
3. NetApp FAS – MDC 1.5PB (March-2014)
4. Quantum StorNext; 2 - 3.5PB
5. Kaminario all flash array; 6TB

6. Croatan Data Analytics Cluster; 1PB
7. Individual Project Servers w/ Direct Attached Storage; varies
8. Winding down: DDN and BlueArc; 1.1PB

Storage: Prior Generation



BlueArc NAS

Network Attached Storage

DDN9900

Block storage only. Direct attached to various servers or VMs

Multi-vendor solution was complex
Numerous issues with DDN support and controller firmware
BlueArc was purchased by Hitachi during the lifetime of this deployment
Cannot operate these systems without vendor maintenance agreements
The geek in our soul would like to try low cost hardware with open source systems

Storage: Primary Research Storage (NAS, SAN)

NetApp Clustered Data ONTAP

2 x (FAS6620 Controller, 3TB FlashCache)

8 x (Disk shelves, each with 48 x 3TB drives)

8 x FC8 connections

8 x 10GbE connections

Hatteras Compute Cluster
Blue Ridge Compute Cluster
Croatan Data Analytics Cluster
VMs in the VMWare farm
Project Specific Servers

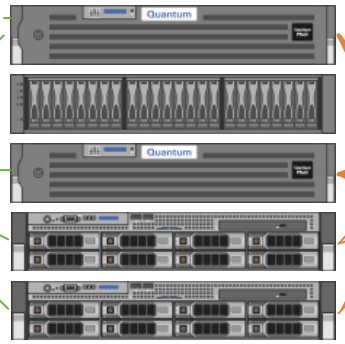
Storage: StorNext, secondary research storage + archive

528TB: (8) Dell MD3660f; each with 5 shelves
each shelf has: (1) 400GB SSD, (11) 1.2TB 10k disks

(2) Quantum M662
Metadata Controllers

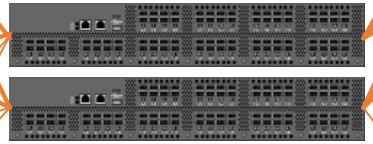
Metadata

LAN Client



(2) StorNext
G302 Gateways

FC8



Brocade 5300 (independent fabrics)

FC8

1.5 to 3 PB: Quantum i6000

- Hatteras Compute Cluster
- Blue Ridge Compute Cluster
- Croatan Data Analytics Cluster
- VMs in the VMWare farm
- Project Specific Servers

Storage: Experimental System

Kaminario 6TB all flash array

Evaluated performance with VMWare (see link below)
ExoGENI: tested as an iSCSI device for rapid VM provisioning
Relational Database for Genomics processing

http://www.hpcwire.com/off-the-wire/kaminario-renci-announce-report-storage-hpvcs/

CI-BER: CI for Billions of Electronic Records

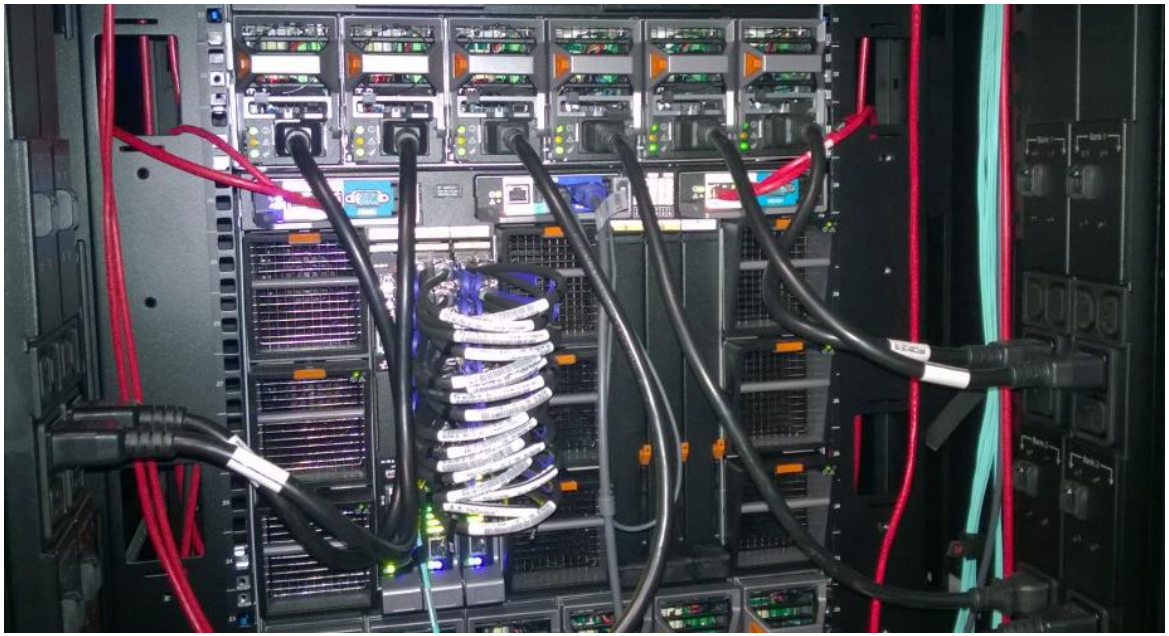
- Approximately: 105TB, 150 million files

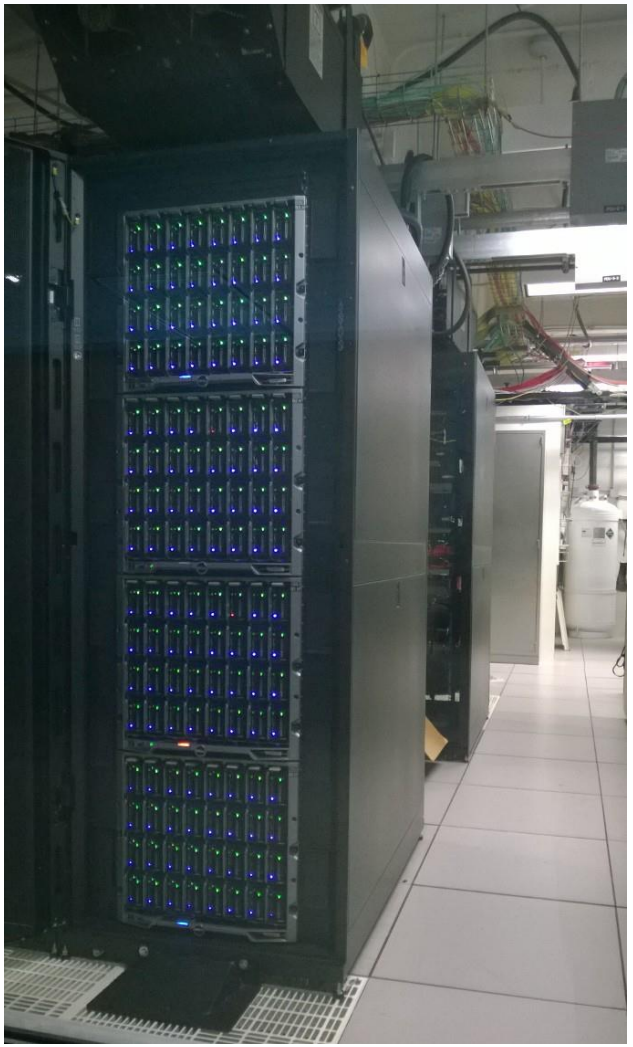
- Capacity is not unruly by our norm, however the file count has proven to be an interesting experience
- Migrating out of the shared infrastructure and onto a dedicated storage system
- Front-end services (iRODS) remain as VMs in the farm

http://sils.unc.edu/news/2012/ci-ber-big-data

Compute: Hatteras Cluster

- Designed for HPC ensembles
- 4 x (512 cores with 6GB per core; packaged in 32 nodes)
- 40Gb FDR-10 Infiniband Interconnect
- 20GbE uplink per M1000e to storage/rest of the world
- Chose Not to purchase and manage the interconnect to enable MPI across the boundaries of the four 512 core units
- Fits into a single rack
- First RENCI resource to use SLURM





Compute: Blue Ridge Cluster

- 160 nodes: each with 8 cores, 3GB per core
- 40gb Infiniband interconnect

- 2 GPGPU nodes: 96GB ram, nVidia Tesla S1070
- 2 LargeMem nodes: 32 cores, 1TB ram

- Includes nodes dedicated to a specific project (ADCIRC)

- 10 nodes of similar hardware configuration running Windows HPC

Data Analytics: Croatan Cluster

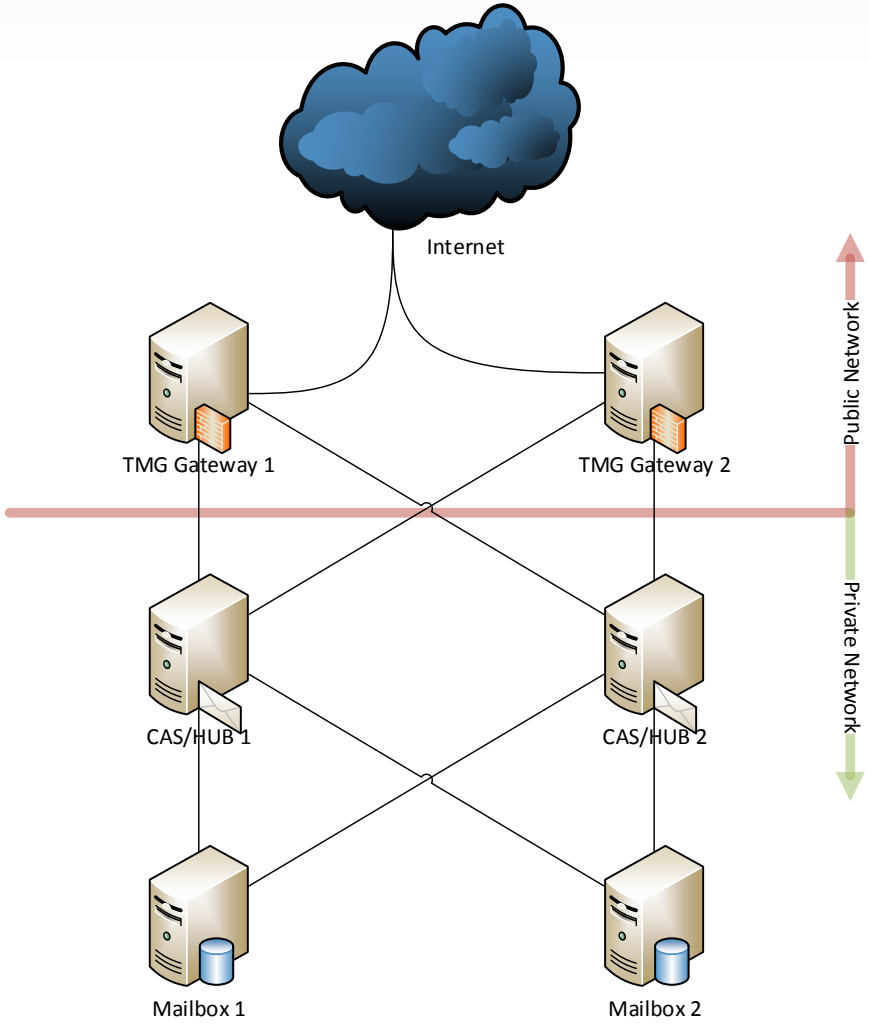
- 30 x (Dell R720xd), *each* with:
  - 16 cores at 3Ghz
  - 256GB memory
  - 36TB direct attached storage
  - 56Gbps FDR Infiniband and 40Gbps Ethernet interconnect
  - 10GbE Dedicated NAS Connectivity
  - 1GbE management network
- Aggregate: 1PB raw storage, 7.6TB ram

Virtual Machine demand has skyrocketed

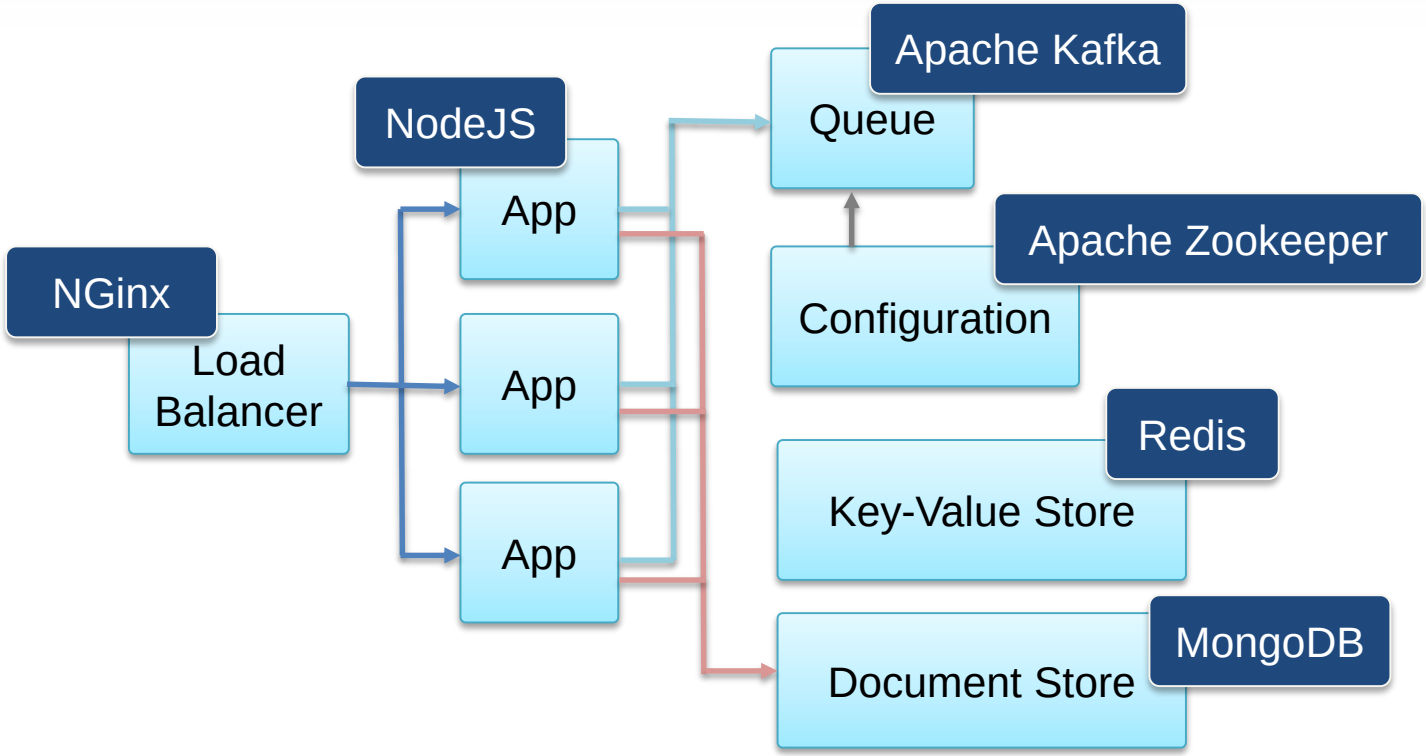
- More than 350 VMs in the server farm
- ACIS Core services such as:
  - AD controllers, LDAP, Exchange, Lync, Sharepoint
  - HA Clustered MSSQL, MySQL, PostgreSQL
  - DNS, DHCP, cluster login and service nodes
- Project based VMs
  - Software development, testing, code repos, continuous integration
  - iRODS servers, databases for iRODS catalog
  - ExoGENI has about 35 VMs, including
    - control.exogeni.net: exogeni's master ldap, puppet, and DNS
    - geni.renci.org: ORCA service manager which allows stitching vlans across racks
  - NetCDF data distribution
- We underestimated the value (demand)
  - Would prefer to have various levels of QOS, partition of services along differing SLA requirements

VM Farm Example: Exchange e-mail services

- 6 virtual machines to run the system
- We began hosting our own e-mail services before outsourcing was a viable option
- Community has grown accustomed to a level of service that can be difficult to achieve with the university central IT offerings, or commercial providers



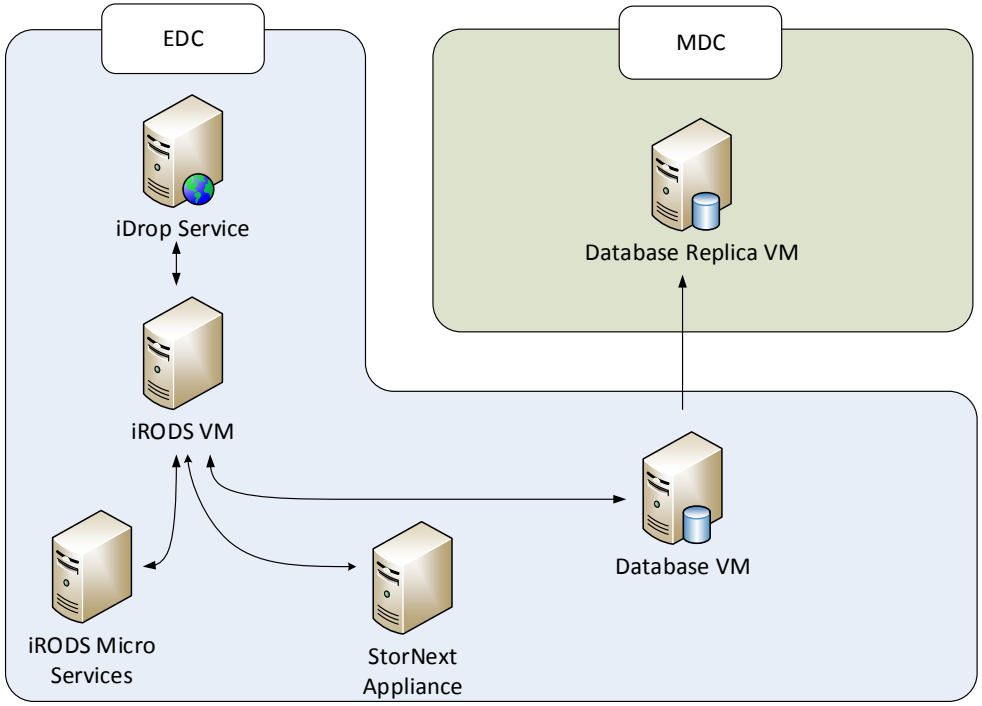
VM Farm Example: Skylr – Distributed Alpha



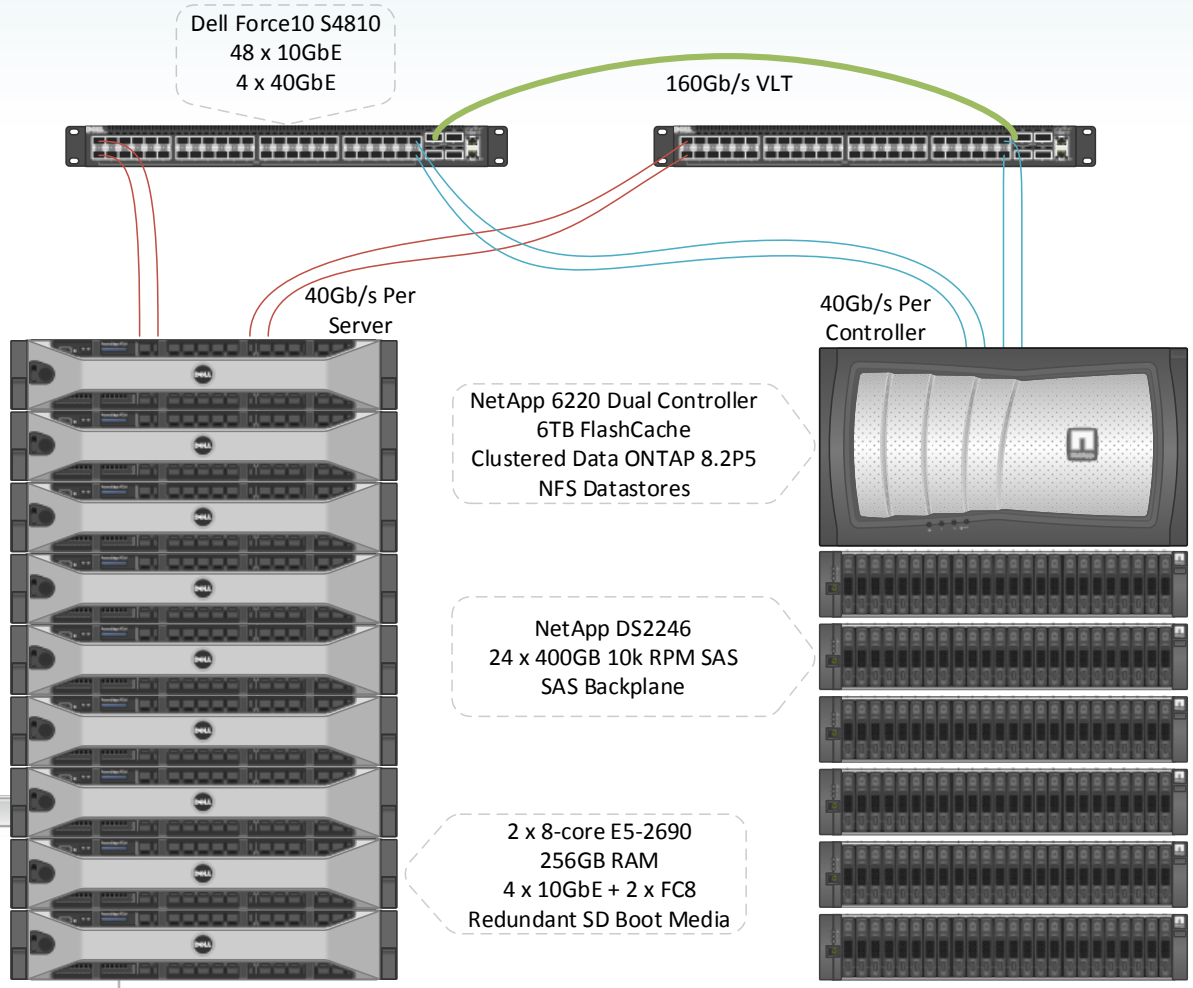
20 VMs total: two instances of this architecture, plus 4 dev/utility VMs

VM Farm Example: Production iRODS deployment

- iren2.renci.org
- Multisite catalog replication
- Hydroshare microservices
- Serves the following data grids:
  - DFC
  - UNC Lifetime Library
  - CAMERA
  - TDLC



Virtual Machine Services: VMWare



General

vSphere DRS:	On
vSphere HA:	On
VMware EVC Mode:	Disabled
Total CPU Resources:	417 GHz
Total Memory:	2.25 TB
Total Storage:	20.33 TB
Number of Hosts:	9
Total Processors:	144
Number of Datastore Clusters:	0
Total Datastores:	8
Virtual Machines and Templates:	366
Total Migrations using vMotion:	20942

Aggregate: 144 cores, 2.3TB RAM

VM Farm: Just because we can, doesn't mean we should …

We utilize external service providers wherever appropriate.
Websites on HostGator that we manage for more than 20 domains including:

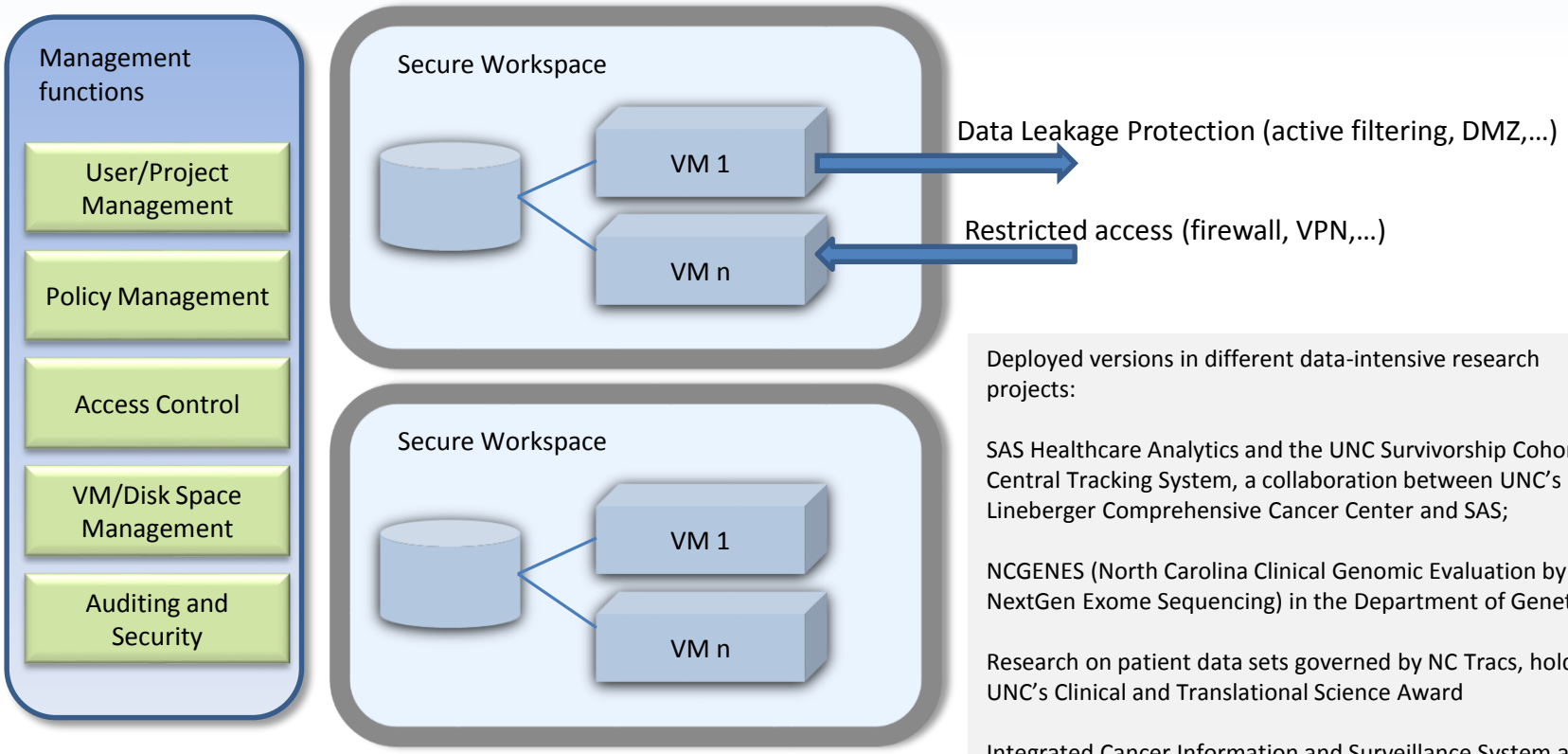
http://reachnc.org
http://coastalhazardscenter.org
http://mitigationguide.org
http://diph.renci.org
http://dice.renci.org
http://datafed.org
http://e-irods.org

http://ncdatascience.org
http://irods-consortium.org
http://www.ncgenes.org
http://www.exogeni.net
http://cuahsi.hydroshare.org
http://www.renci.org
. . .

RENCI web developer **Joe Hope** manages these in collaboration with our partners

Secure Medical Workspace

Managed Infrastructure



A secure workspace consist of one or more virtual machines (VMs) that are provisioned along with a common data space to a research team. Multiple workspaces are controlled through a unified management system

Deployed versions in different data-intensive research projects:

SAS Healthcare Analytics and the UNC Survivorship Cohort Central Tracking System, a collaboration between UNC's Lineberger Comprehensive Cancer Center and SAS;

NCGENES (North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing) in the Department of Genetics;

Research on patient data sets governed by NC Tracs, holder of UNC's Clinical and Translational Science Award

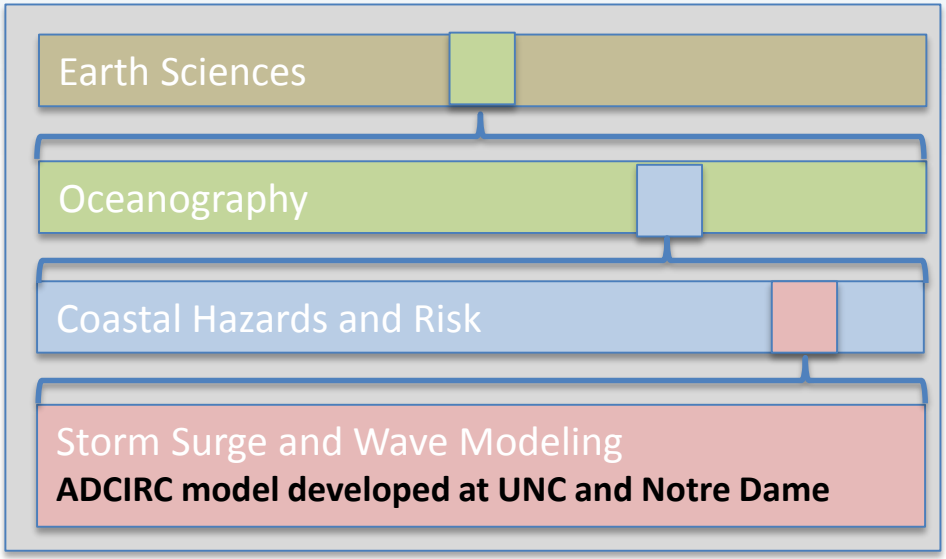
Integrated Cancer Information and Surveillance System at the Lineberger Comprehensive Cancer Center.

Evaluating the cost-effectiveness of the SMW for campus-wide adoption as a central component of its Information Security Plan following a comprehensive assessment of the technology.

Contact: Charles Schmitt, RENCI

http://www.ncbi.nlm.nih.gov/pubmed/23751029

Environmental Sciences Program: Brian Blanton et al



Funding sources

Public:
- NSF
- NOAA
- DHS & FEMA
- USACE
- NRC

Private:
- Engineering Firms
- Utilities
- Commercial

Insurance
- Consulting

Research, Applications, and Operations with the ADCIRC storm surge, tide, and wave model

Hindcasting of Historical Storms
Validation of model to Establish Uncertainty Estimates

Forecasting of Storm Surge and Waves

Real-time, operational
On-demand resource needs

Probabilistic Coastal Risk Assessment

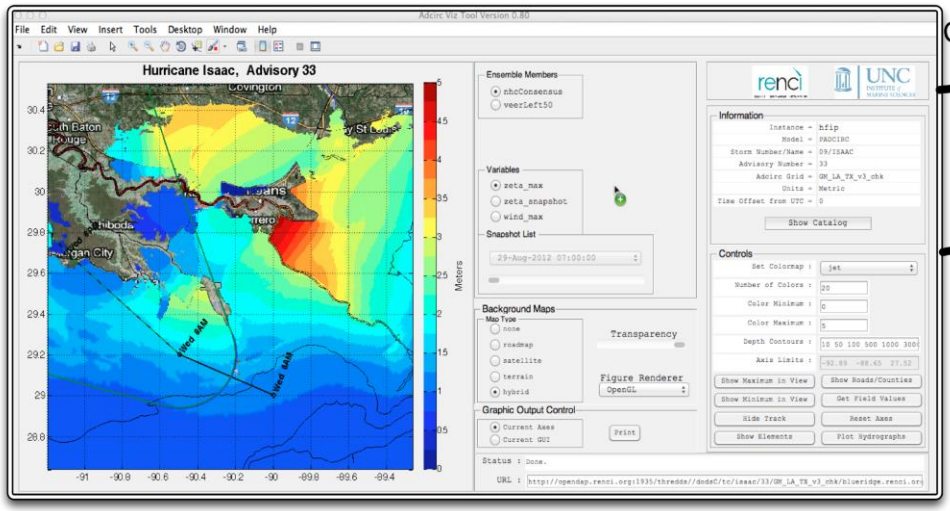
Floodplain Mapping
Very large datasets

Engineering and Design

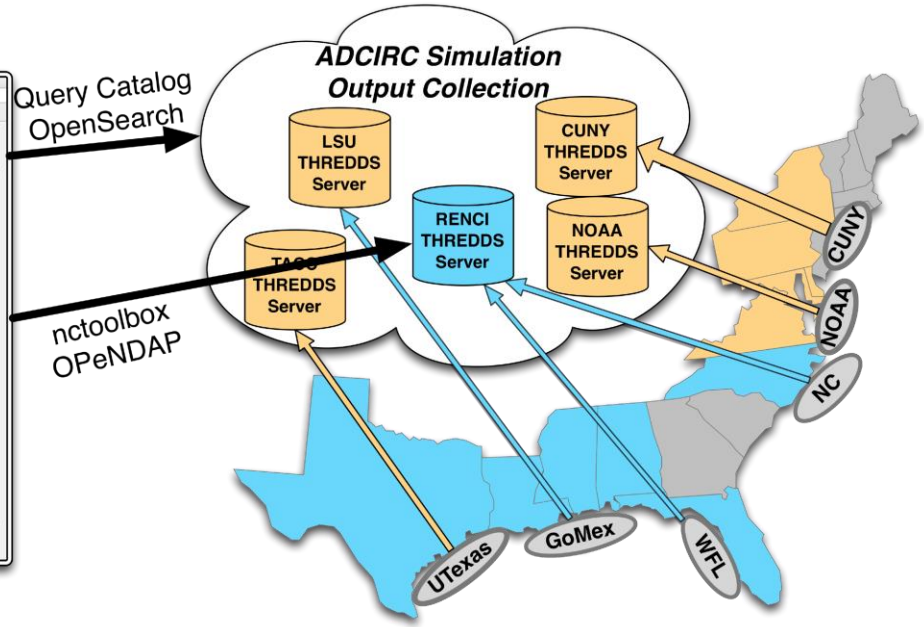
Flood protection systems (e.g., levees)
Coastal planning and development

Coastal Hazards Community of Practice

AdcircViz during Hurricane Isaac (2012)



Collection of ADCIRC forecasting systems



Time sensitive ensemble HPC, as storms approach
Would benefit greatly from HPC resource sharing

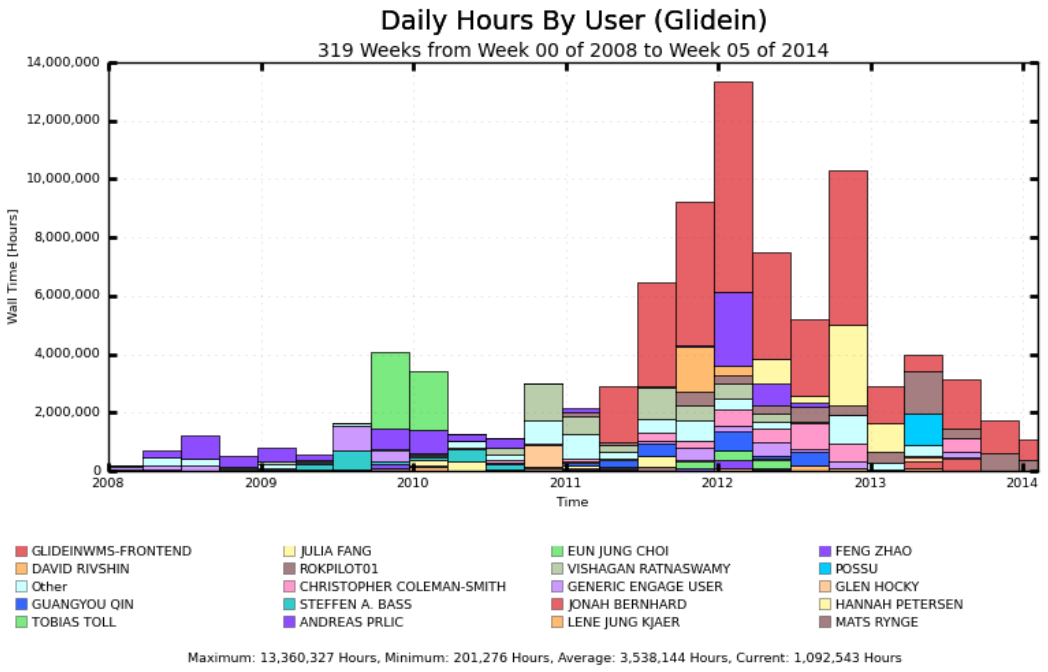
http://people.renci.org/~bblanton/files/AdcircDuringIrene.key.pdf

Some examples of Project Specific Hardware (PSH)

my apologies for the acronym, but too many letters otherwise

PSH: OSG EngageVO

- Assisting researchers with access to OSG since 2007
- engage-submit3.renci.org, physical machine
- 30 TB disk mounted to the submit host (re-purposed out of warranty)
- VMs for associated services (eg VOMS)
- RENCI Identity management for access to submit host, OSG credentials from there
- Newer methods available now: OSG Connect



PSH: Relational Database for Genomics

- Essentially a Croatan node
  - Dell R720xd, 256GB ram
  - 32TB DAS: 12 x 3TB NLSAS
  - Runs MS SQL Server

  - Supports genomics work of Dr. Kirk Wilhelmsen
  - Several Databases that are many TBs in size, with billions of rows
  - Very heavily loaded system

PSH: Coastal Emergency Risks Assessment, NC-CERA

- Visualize results of ensemble based HPC model runs
- Generates and serves map tile sets
- Tens of millions of small files generated regularly
- Storage latency is more important than throughput
- One small piece of the overall solution is an old c++ opengl code for Windows, and is a bottleneck which is driving hardware decisions

- Currently running on multiple blade servers with DAS from the out of warranty DDN equipment, planned upgrade to R720xd

http://nc-cera.renci.org/cera/_docs/CERA_NC_tutorial2013.pdf

http://www.ecu.edu/renci/_docs/2012HurricaneWorkshop/Blanton.pdf

MDC: Manning Data Center; UNC Central IT

- 11,000-square-foot data center;
- In partnership with UNC Research Computing
- We wish to bridge the capabilities of these two University assets (EDC, MDC) and facilitate research that moves between them, enabling cross site redundancy and building a culture of shared distributed systems.

- Genomics and medical informatics are the primary use cases thus far
- DR for EDC research and production storage tiers
- A few project specific servers

- RENCI has a few racks here
- Small VM Farm
- 1.5 PB raw storage

The Duke Social Computing Room (SCR)

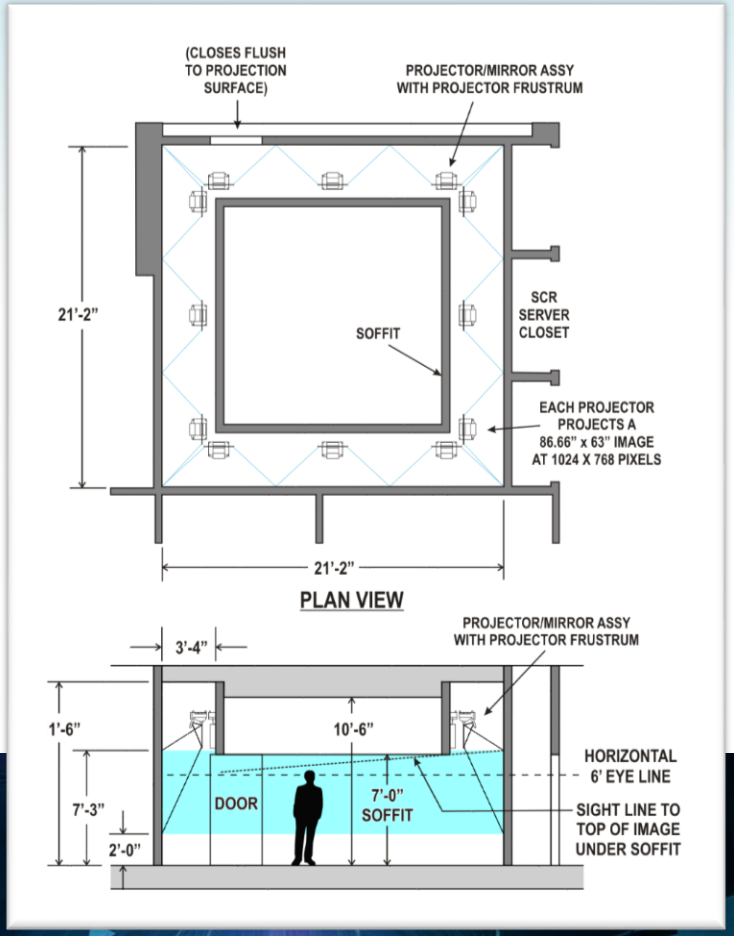
In Development

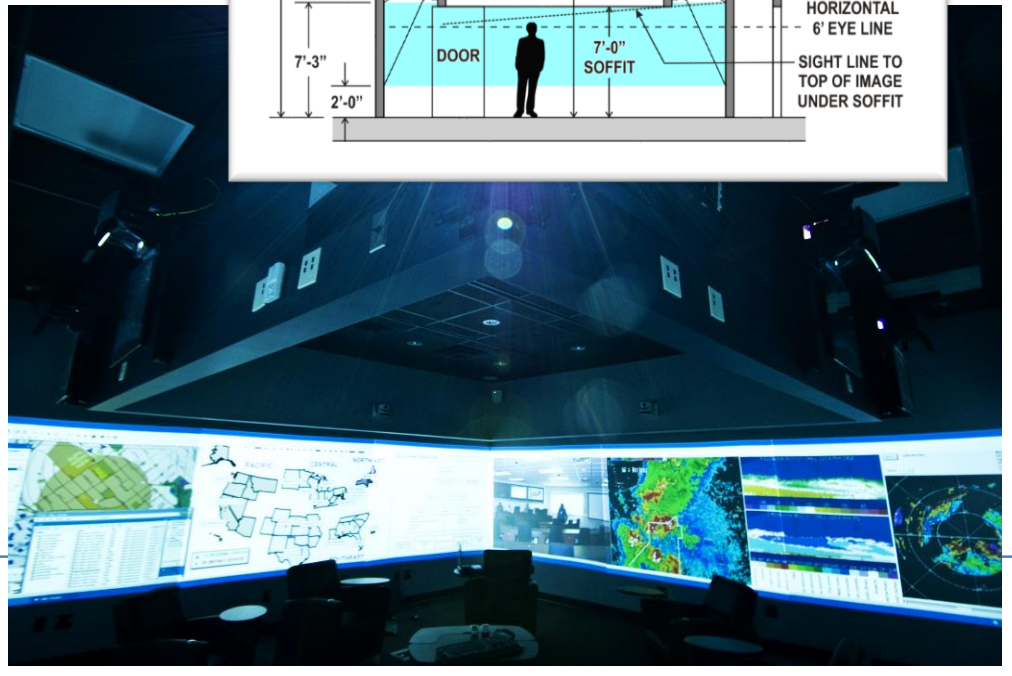
Currently deployed at RENCI's Engagement Center at UNC-Chapel Hill and NCSU, the SCR is …

- A collaborative visualization environment for researchers and students that provides a large physical real estate for arranging visual information[1]
- A Windows 7-based display that consists of a square room with 12 projectors (3 per wall) used to display a single 360-degree desktop environment spanning all four walls
- A tool to visualize and explore large amounts of heterogeneous data
- A user-friendly environment that has a low barrier to entry

In the SCR you can …

- Boost presentations to a new level
- Conduct classes and seminars while surrounded by images, documents, movies, and visualizations
- Visualize and explore large data and documents
- Collaboratively develop grants, review code, and analyze large images, maps, and engineering documents
- Host videoconferences immersed in virtual worlds
- Develop innovative media installations





SCR Applications: Scientific Visualization

"The Social Computing Room is an ideal location to perform the design reviews and voting for the visualization homeworks in the Visualization in the Sciences class. We can project the entries from each student on the walls all **at the same time**, along with a text description of what questions they were supposed to answer, and then **all sit and talk about the comparative strengths and weaknesses** of each approach. The ability to **walk right up to the images** lets us both see them at **high detail** and vote (by dropping pennies) for our favorite visualizations. The central seating area makes the **room comfortable for hour-long sessions and extended discussions**. The USB input lets us bring our own content to be displayed. The **auto-layout software** combined with the **ability to move images** around manually enables us to **rapidly set up a session**."

– Russ Taylor, UNC-CH CS

Multiple model runs

Time-series data

SCR Applications: Research Meetings

- Melanoma Image Analysis
  - Segmentation and feature extraction of stained biopsy slides
  - Lots of data
    - Original images
    - Feature images
    - Feature distribution scatter plots
  - Ability to spread data out extremely helpful
  - Used weekly

Matlab tool for sorting images based on feature strength

SCR Applications: Arts and Humanities

Spectacular Justice

Multimedia art installation by Joyce Rudinsky

Part of Carolina Performing Arts
Criminal/Justice: The Death Penalty Examined

Images, video, and audio respond to user locations

Extensions for this project:

- Ubisense tracking system
- 2 HoloPro rear-projection screens w/ auxiliary projectors
- 5 Directional speakers
- PosiTrack camera controller

SCR Take Home

- Easy to use
  - Single PC
  - Run practically any existing software
  - Standard interface
- Customizable
  - Hardware + software
  - Coding in general…
- Collaborative
  - Multiple people
  - Currently only single point of control…

GSB: Genome Sciences Building

TopSail: decommissioned from UNC ITS Research Computing, managed by Erik Scott

- 400 nodes, 3200 cores
- 400x1.5T (600 T) disk + 108 TB scratch
- 400 gbits/sec of usable Hadoop bandwidth

- 4 or 8 cores per disk drive
- 2 gigabit campus connectivity to datamover – less than 1 gb/s to Renci

- Apache Hadoop w/ Pig, Hive, Mahout, Zookeeper
- DB/2 parallel query edition
- HDFS FUSE Bridge on datamover
- Backups (slowly!) to RENCI tape

The National Consortium for Data Science

UNC Chapel Hill: Computer Science 790-042
Data Center Systems and Programming **Prof. Don Smith**

- One of the NCDS focus areas: Workforce Development

- Graduate Seminar on Large-Scale Computing Methods and Issues
- Key topics: Storage, Networking, Computing, Power and Cooling
- Basic introduction to Hadoop through directed readings and two Hadoop-based projects
- ~15 graduate students

www.data2discovery.org

Erik Scott

http://data2discovery.org/dev/wp-content/uploads/2014/02/NCDS-Summit-2013.pdf

GSC: Galapagos Science Center

- Remote Branch Office Scenario
- Extreme network limitations
- Supports Center operations and onsite research and teaching activity
- GIS Lab with approx. 15 workstations
- UNC and USFQ Campus Identity Management extended to the facility
- MSFT Hyper-V

print.gsc.edu.ec
share.gsc.edu.ec
file-01.gsc.edu.ec
deploy.gsc.edu.ec
update.gsc.edu.ec
…
…
license.gsc.edu.ec
multimedia.gsc.edu.ec
sql.gsc.edu.ec
wikipedia.gsc.edu.ec
dpm.gsc.edu.ec

Network Switch
HV-04
HV-03
HV-02
HV-01
Storage
Storage

Directory Services and Identity Management

- LDAP and Active Directory
- There are 12 times as many external collaborators in the directory as there are RENCI staff
- Consolidating on Active Directory, eliminating LDAP
- Cluster nodes are now using SSSD vs distributed passwd files
- Shibboleth is used in some cases (eg REACH-NC)
- iRODS based projects federated at the iRODS layer
- ExoGENI project maintains their own LDAP
- We wish to federate with UNC or adopt UNC credentials for institute staff, while maintaining a directory for external collaborators
- Very difficult to acquire the required time and service interruptions for a more appropriate solution

And a few things we did not talk about …

- Monitoring and notification: nagios, snmp, SCOM
- PXE boot: laptops, VMs, cluster nodes
- DNS, DHCP, VPN
- CommVault Backup
- HA Clustered relational databases
- FTP/HTTP servers and frameworks (PHP, Tomcat, etc.)
- Dell Open Manage Essentials
- Multisite data replication: block level, file level
- Compiler support: GNU gcc/gfortran, Intel icc/ifort and Portland Group (PGI) pgcc/pgfortran
- MPI Support: mvapich, openmpi and intel MPI
- Configure, compile and install: NetCDF, HDF, Nvidia CUDA, Intel MIC platform, hpctoolkit
- Manage licenses for Matlab, Intel and PGI compilers
- Special Engagements, eg experimental replacement for TCP, deployed at endpoints between EDC and MDC; Prof. Don Smith, UNC-CS

SSH.NET – .NET native open source SSH

- Open source .NET library for building solutions with SSH capabilities
- Makes use of advanced features of the Microsoft platform
- More than: 145k downloads, 1M page views
- Active discussion boards
- http://sshnet.codeplex.com

SELECTED PERIOD	CHANGE	DESCRIPTION
145907 downloads 0.61 downloads / visit 114.53 downloads / day	**+145907** +0.61 +114.53	Download counts are for all publicly available releases, source code changesets, and wiki attachments. Mouse over a data point to see download traffic for that specific date.

Orion Nebula by: Mark Montazer
Date Taken: 11/09/2013
Location: Pittsboro, NC
Software Used: Deep Sky Stacker & GIMP 2.9
30 3-minute exposures at ISO 800, 30 2-minute exposures at ISO 800, 40 20-second exposures at ISO 800 for a 2-hour & 43-minute integrated exposure
Equipment: Sony NEX-5 (modified for astrophotography) using a 731mm f/4.8 Maksutov-Newtonian telescope on a Celestron CG5-ASGT mount


END PRESENTATION
renci
RESEARCH ENGAGEMENT INNOVATION