Protected Environment at CHPC

Anita Orendt, anita.orendt@utah.edu
Wayne Bradford, wayne.bradford@utah.edu
Center for High Performance Computing

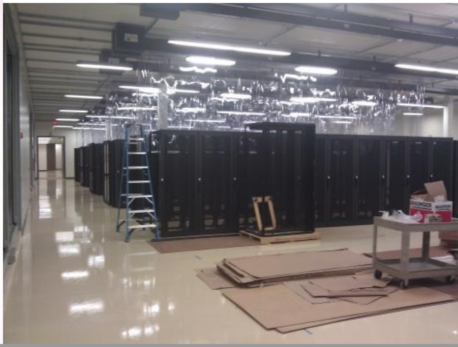
25 October 2016

CHPC Mission

In addition to deploying and operating high performance computational resources and providing advanced user support and training, CHPC serves as an expert team to broadly support the increasingly diverse research computing needs on campus. These needs include support for big data, big data movement, data analytics, security, virtual machines, Windows science application servers, protected environments for data mining and analysis of protected health information, and advanced networking.

Downtown Data Center

- Came online spring 2012
- Shared with enterprise (academic/hospital) groups (wall between rooms)
- 92 racks and 1.2MW of power with upgrade path to add capacity for research computing
- Metro optical ring connecting campus, data center, & internet2
- 24/7/365 facility



Overview

- Background on the protected environment (PE)
- How to get a PE account
- Description of PE resources
- How to access PE resources

Why do we have it?

- Researchers need a safe place to compute and work with restricted data
- Restricted data can be stolen from insecure places
  - insecure systems, laptops/phones and tablets/removable drives
- Required by law in order to comply with regulations such as HIPAA. PHI security breaches are serious. e.g., fines, potential lawsuits, loss of reputation/credibility/funding.

Safeguarding data is important for you and your institution

18 Personal Identifiers Under HIPAA

(any single or multiple identifiers that can identify a person)

1. Name
2. Address including city and zip code (except 1st 3 digits)*
3. Dates (birth, death, admission, discharge) except year*
4. Telephone number
5. Fax number
6. E-mail address
7. Social security number
8. Medical record number
9. Health plan ID number
10. Account number
11. Certificate/license number
12. Vehicle identifiers and serial number
13. Device identifiers and serial number
14. URL
15. IP address
16. Biometric identifiers including finger prints
17. Full face photo and other comparable image
18. Any other unique identifying number, characteristic, or code*

https://privacyruleandresearch.nih.gov/pr_08.asp

Use & Disclosure of PHI for Research

1. De-identify
2. Obtain written authorization from individuals to use data
3. Used without authorization IF:
   a. Have authorization requirement waived
   b. As part of a limited data set, with a data use agreement
   c. As needed in preparation for research
   d. Use for research on decedent's information

There are defined requirements and procedures for each of the above options

Two Methods for De-identifying Data

1. Removal of all 18 individual identifiers that could be used to identify the individual.
   - Can leave code that is not derived from any of the identifiers and cannot be translated back to the individual (randomly assigned with secure key)
2. A formal determination by a qualified expert who confirms that individual cannot be identified.

Other Potential Uses of PE

- While need for HIPAA compliance is the most common reason to use the PE, there are other uses, including:
  - ITAR (International Traffic in Arms Regulations) compliance
  - FERPA
  - FDA part 11 compliance
  - FISMA
  - Any other sensitive or restricted data and/or application
- These each come with their own regulations and requirements and must first be evaluated for PE suitability

NIH dbGaP

- https://www.ncbi.nlm.nih.gov/gap
- See Security Procedure section
- For "controlled-access human genomic and phenotypic data"
- Do not contain direct identifiers, but the data are sensitive and must be protected

CHPC developing plan to meet requirements

Restricted Data Handling & Compliance

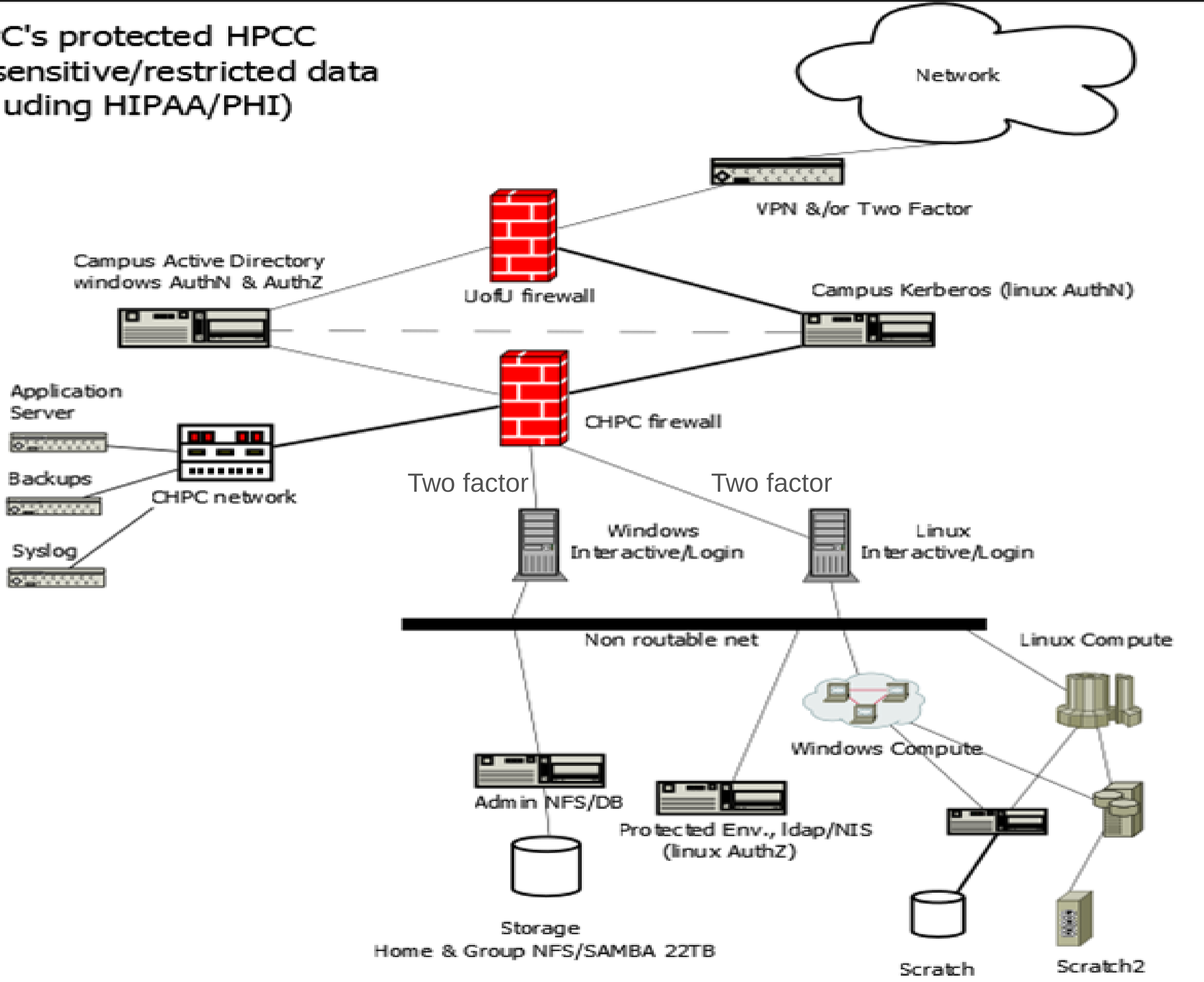
- U of U data classification and encryption policy http://regulations.utah.edu/it/rules/Rule4-004C.php
- PHI - must encrypt the data both when is it being stored (data at rest) on ANY device or transmitted (data in transit)
- dbGaP - data at rest encrypted if on mobile devices or on removable media

For details see article in CHPC Fall 2016 Newsletter that will be published in November

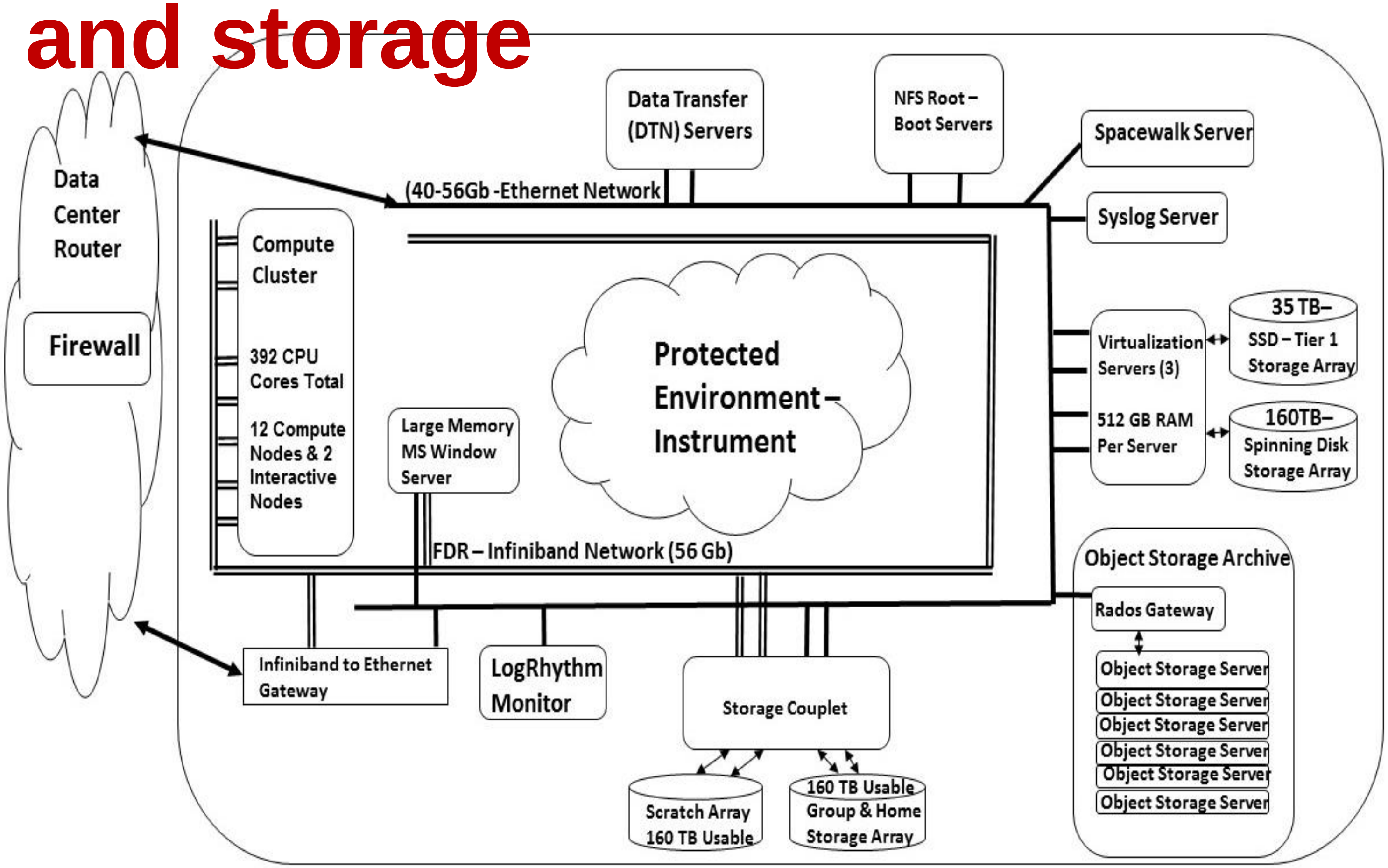
What is the CHPC Protected Environment (PE)?

- Developed in 2009 to strengthen the privacy and security protections for health information in scientific research
- Work closely with Security and Privacy office for consultations, security risk and compliance assessments, reviews, mitigation plans, and policy & regulation enforcement
- PE resource include:
    - HPC Cluster – Apexarch
    - Windows Server – Swasey – Windows software, Statistics packages, front end for web and database access
    - Storage – both home and scratch space; storage for VMs
    - VMs for project development and for use cases that do not fit in HPC world
        - RECAP/database services
        - Web proxy and WWW services

CHPC's protected HPCC for sensitive/restricted data (including HIPAA/PHI)



Proposed New PE – Condo model for ndes and storage


Data Center Router
Firewall
Data Transfer (DTN) Servers
NFS Root – Boot Servers
Spacewalk Server
(40-56Gb -Ethernet Network
Syslog Server
Compute Cluster
392 CPU Cores Total
12 Compute Nodes & 2 Interactive Nodes
Large Memory MS Window Server
Protected Environment – Instrument
Virtualization Servers (3)
512 GB RAM Per Server
35 TB– SSD – Tier 1 Storage Array
160TB– Spinning Disk Storage Array
FDR – Infiniband Network (56 Gb)
Object Storage Archive
Rados Gateway
Object Storage Server
Object Storage Server
Object Storage Server
Object Storage Server
Object Storage Server
Object Storage Server
Infiniband to Ethernet Gateway
LogRhythm Monitor
Storage Couplet
Scratch Array 160 TB Usable
160 TB Usable Group & Home Storage Array

Some of the Systems Controls in Place

- Isolated VLANS, separate VPN pools per project
- Standard baseline Build list
- Inventory assets & hardware POC
- Qualys scans, Center for Internet Security (CIS) scans, Nessus, nmap, security onion, traffic trending with cacti
- Central Syslog, logwatch reports, network flow reports
- The physical hardware in datacenter with controlled room access; hosts are racked in a locked cabinet and have locked server bezels
- Thorough Documentation!
- Needs assessment, training, MFA/VPN access, IRB certification

Requires constant review of technical & physical security controls

PE Need Assessment

https://www.chpc.utah.edu/resources/ProtectedEnvironment.php

Used to see if your project fits in the protected environment

- complete needs assessment
- talk to us

Requesting PE Access

- Provide HIPAA/CITI training certification
- Get a CHPC PE account
- Set up DUO two factor authentication
- If the resources you need already exist – you are ready to go
- If you need a new VM – complete VM request (jira issue)
  - Provide info on OS, number of cores, amount of memory, disk space and any additional software needs

Access Controls

- Login Access
  - General linux login nodes via ssh: apex.chpc.utah.edu (round robin of apex1 and apex2), farnsworth.chpc.utah.edu
    - Have FastX available (more later)
    - Farnsworth – has graphics card; no access to submit to batch from this node
  - Windows: swasey.chpc.utah.edu (preferred); drawbridge.chpc.utah.edu (VM, can be used for very lightweight needs) – connect via RDP
  - Access to all requires DUO 2 factor authentication; from non-UofU IP address must first use VPN (either @chpc.utah.edu or @utah.edu)
- Data access -- based on IRB number/project
  - We verify users' right to access the specified data
  - Use unix FACLs (File Access Control Lists)

Description of Resources

- HPC Cluster – Apexarch
  - https://www.chpc.utah.edu/documentation/guides/apexarch.php
  - 3 general login nodes
  - 16 compute nodes (148 total cores)
    - 11 nodes with 8 cores, all with 24G of memory
    - 5 nodes with 12 cores (one has 96G ram, two have 48G ram & two have 24G ram
  - Mellanox QDR Infiniband interconnect
  - 5.5 TB general scratch server
  - Slurm batch system

11 new nodes – 28 cores, 128GB memory on order

Description of Resources (2)

- Windows Server – Swasey
  - https://www.chpc.utah.edu/documentation/guides/swasey.php
  - 48 CPU cores, 512GB RAM, 1TB local space
  - SAS with text miner, AMOS, SPSS, R, STATA, Mathematica, Matlab, and Microsoft Office 2010
  - Can mount PE home and project space
- Storage
  - 5 TB file system for home and two 11 TB project space file systems
  - No quotas enforced at this time
  - Backed up to tape

New 100 TB home and project space on order

Description of Resources (3)

- VM farm (have 2 servers with fail over)
  - Currently support 81 VMs
  - 384 GB RAM
  - 15TB usable capacity with self encrypting drives
  - 32 cores (oversubscribe)

HPC login scripts

- CHPC provides login scripts (“dot” files) when creating account for both tcsh and bash shells
- These files set the environment so that applications are found, batch commands work – ***Do not remove!***
- Choose shell at account creation – can change at www.chpc.utah.edu (sign in, select edit profile)
- Four files: .bashrc, .tcshrc, .custom.sh, .custom.csh
  - **The first two should not be edited!**
  - **The second two is where to add custom module loads!**
- Will automatically execute a .aliases file if it exists

HPC Batch System -- SLURM

- Used to access compute nodes
  - https://www.chpc.utah.edu/documentation/software/slurm.php
- This site has example scripts, basic commands, information on SLURM environmental variables, table with correspondence between SLURM and PBS commands and variables

FastX2 – Tool for Remote X

- https://www.starnet.com/fastx and https://www.chpc.utah.edu/documentation/software/fastx2.php
- Used to interact with remote linux systems graphically in much more efficient and effective way then simple X forwarding
- Graphical sessions can be detached from without being closing, allowing users to reattach to the session from the same or different systems
- Server on apex1.chpc.utah.edu, apex2.chpc.utah.edu and Farnsworth.chpc.utah.edu
- Clients for windows, mac and linux; can be installed on both university and personal desktops.

Getting Help

- CHPC website and wiki
  - – www.chpc.utah.edu
    - Getting started guide, cluster usage guides, software manual pages, CHPC policies
- Jira Ticketing System
  - – Email: issues@chpc.utah.edu
- Help Desk: 405 INSCC, 581-6440 (9-5 M-F)
- We use chpc-hpc-users@lists.utah.edu for sending messages to users; also have Twitter accounts for announcements -- @CHPCOutages & @CHPCUpdates