Protected Environment at CHPC

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

6 December 2018

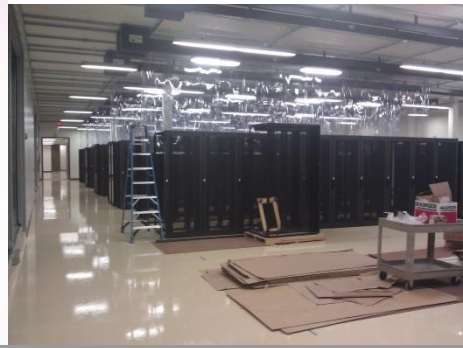
CHPC Mission

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)
- New PE Resources
- 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 Uses of PE

- While need for HIPAA compliance is the most common reason to use the PE, there are other uses, including:
  - FDA part 11 compliance
  - Any other sensitive or restricted data and/or application
  - Looking at other cases – ITAR, FERPA
- These each come with their own regulations and requirements

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

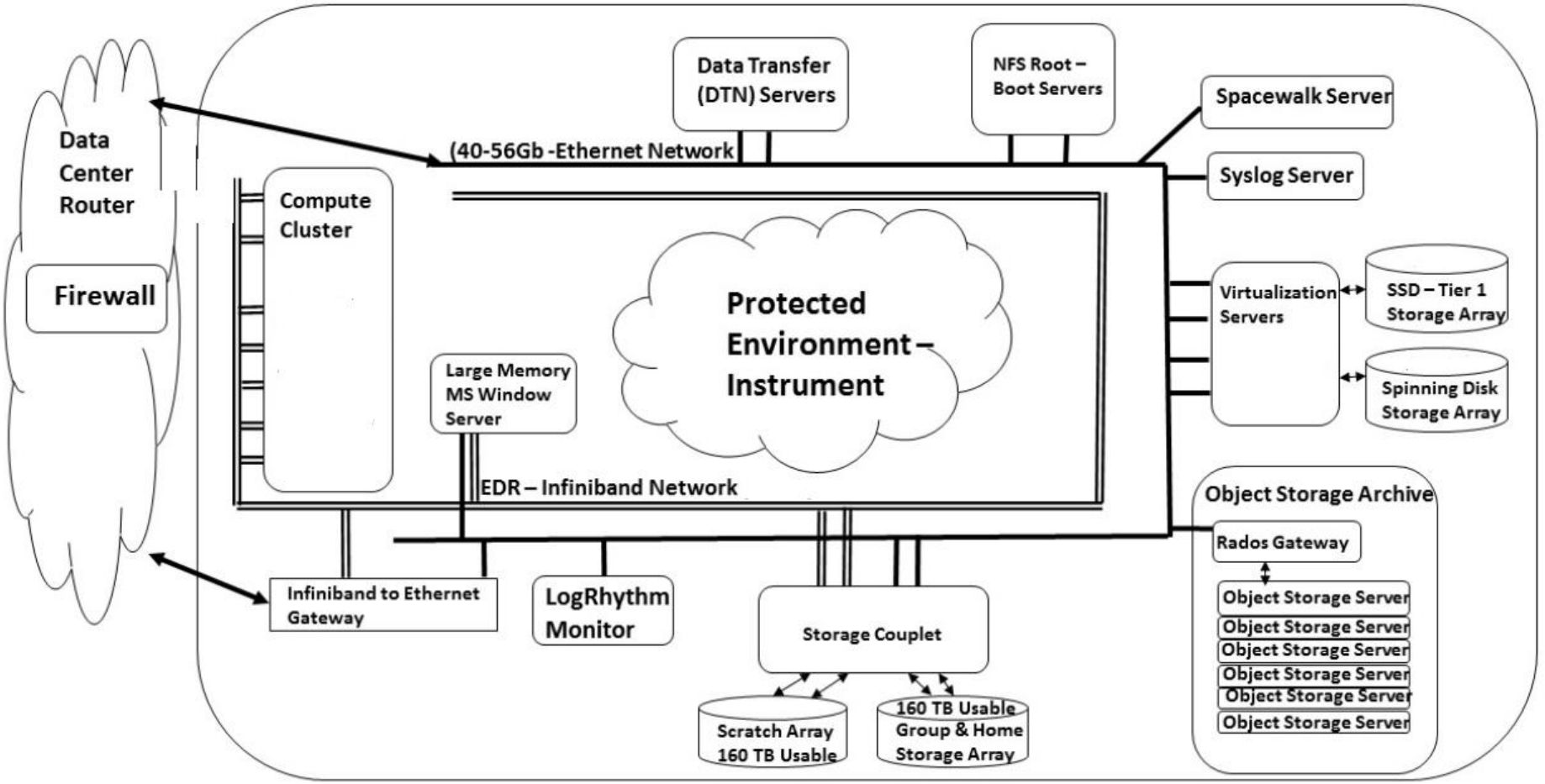
CHPC is promoting move of all human genomic work to PE

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
- Updated in late 2017/early 2018 with the assistance of a NIH Shared Instrumentation Grant.
- New PE will be more reliable and secure, have expanded capabilities, and will be scalable in a condominium fashion (similar to the general environment).
- Looking for users with NIH funding who can make use of new PE

PE Resources

- HPC Cluster – Redwood
- Home Directories – Mammoth, 50 GB per user
- Project Space – Mammoth, about 280 TB total space. Default 250GB per project, can purchase more
- Scratch Space – Mammoth, 160 TB
- Archive Storage – Elm, about 500 TB, can purchase as needed
- VM farm – Prismatic
  - Both SSD storage (25 TB) and SED spinning disk storage (16 TB)
- Windows Server – Narwhal


Data Center Router
Firewall
Data Transfer (DTN) Servers
NFS Root – Boot Servers
Spacewalk Server
(40-56Gb -Ethernet Network
Syslog Server
Compute Cluster
Protected Environment – Instrument
Virtualization Servers
SSD – Tier 1 Storage Array
Spinning Disk Storage Array
Large Memory MS Window Server
EDR – Infiniband Network
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

New Security Features

- Firewall (palo alto)
  - Classifies all traffic, including encrypted traffic, based on application, application function, user and content. You can create comprehensive, precise security policies, resulting in safe enablement of applications.
  - Innovative features reduce manual tasks and enhance your security posture, for example, by disseminating protections from previously unknown threats globally in near-real time, correlating a series of related threat events to indicate a likely attack on your network
  - Threat prevention feature WildFire identifies unknown malware, zero-day exploits, and advanced persistent threats (APTs) through static and dynamic analysis in a scalable, virtual environment, and automatically disseminates updated protections globally in near-real time

New Security Features

- SIEM – Security Information Event Management (LogRhythm)

  Due to growing need for a comprehensive log and event management; provides predefined reports to easily document evidence of compliance

  – enable better security of networks and optimize information technology operations with sophisticated log correlation and analytics.

  – automate collection, organization, analysis, archival, and recovery of log data that enables enterprises to comply with log data retention regulations.

  – ensure compliance with mandates for HIPAA and other government regulations and to protect patient confidentiality and safety.

Getting Started in the PE

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

- Step 1: Determining if your project fits in REDCAP
- Step 2: Needs Assessment
- Step 3: Requesting access to a PE resource

PE Need Assessment

https://www.chpc.utah.edu/role/user/needs_assessment_form.php

- Complete Form – one assessment needed for each project
  - Information about PI
  - Project funding information
  - IRB information
  - Project Computing Requirements
    - What do you intend to do in this environment?
    - Do you need a VM?
    - What services will the software/hardware provide?
    - To whom will these services be provided?
    - Who needs to have remote terminal (ssh/rdp) access?
    - Will there be any information sharing with third parties? – if yes, there are additional questions.
    - Brief description of the research with this project
    - How many people will use this system?
    - Estimate of how many people and records are anticipated to be stored
    - Amount of storage space (project space) needed?

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 (Service Now ticket)
  - 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: redwood.chpc.utah.edu (round robin of redwood1 and redwood2)
    - Have FastX available (more later)
  - Windows: narwhal.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 University VPN
  - Data access – based on IRB number/project
  - We verify users' right to access the specified data (check IRB)
  - Use unix FACLs (File Access Control Lists)

Description of Resources

- HPC Cluster – Redwood
  - General resources now on allocation
  - https://www.chpc.utah.edu/documentation/guides/redwood.php
  - 2 general login nodes (XeonSP, 32 cores, 192GB memory)
  - 2 general GPU compute nodes (32 cores, 4 GTX1080Ti, 192GB)
  - 15 general CPU compute nodes (308 total cores)
    - 4 XeonSP (skylake) nodes with 32 cores, 192GB of memory
    - 11 Broadwell nodes with 28 cores, 128GB memory
  - Owner nodes (both interactive/login and compute)
  - Mellanox EDR Infiniband interconnect (broadwell nodes connect at FDR)
  - 160 TB scratch server
  - Slurm batch system

Description of Resources (2)

- Storage - Mammoth
  - 50 GB/user home – backed up
  - Project space – will have quotas (250GB/project free; can purchase additional)
  - Archive space – can be used for user driven backup of project space

Description of Resources (3)

- Windows Server – Narwhal
  - https://www.chpc.utah.edu/documentation/guides/narwhal.php
  - Change from Swasey – separation of entry to PE versus compute needs
    - narwal.chpc.utah.edu will get you to one of the gateway nodes (1 physical box with VMs for when physical box not available)
    - From gateway can open session (window icon) on narwhal-c1 (compute node)
  - Narwhale compute --24 CPU cores @3GHz, 512GB RAM, 1TB SSD local space
    - Local space as scratch, with each user having a directory (unid); project directories created on request
  - SAS with text miner, AMOS, SPSS, R, STATA, Mathematica, Matlab, and Microsoft Office 2010
  - Can mount PE home and project space (mammoth)

Description of Resources (4)

- VM farm (have 4 servers with fail over for availability)
  - Have usable 72 cores, 1150GB RAM, 25TB SSD, 16TB SED spinning
- Sizing in incremental blocks (2 core, 4GB RAM, 50GB storage)
  - Storage SSD by default unless encryption needed
  - Can support 275 blocks
  - Can get additional space if needed (cost TBD)
  - Can also mount project space
- Working on costing model for VMs with block plus basic installation at $350/block/5years (hardware costs)
  - Total cost for external
- Customization billed at $75/hour

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 redwood1.chpc.utah.edu, redwood2.chpc.utah.edu (and other owner specific interactive nodes)
- 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
- Service Now
  - Email: helpdesk@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