CERIAS

The Center for Education and Research in Information Assurance and Security

Apache™ Hadoop® Implementation with CDH 5 and VMware® for the REU Data Spillage in Hadoop® Clouds Project

Advisers: Brandeis H. Marshall, Ph.D. & John A. Springer, Ph.D. Student: Niki Ierides

Problem Statement

- Data spills[1]
  - Classified data/metadata to unclassified systems
  - Most often by user error/negligence

Motivation & Significance

- Compromise of data confidentiality and integrity
- Threat to businesses and national security
- Costly & difficult clean-up and recovery

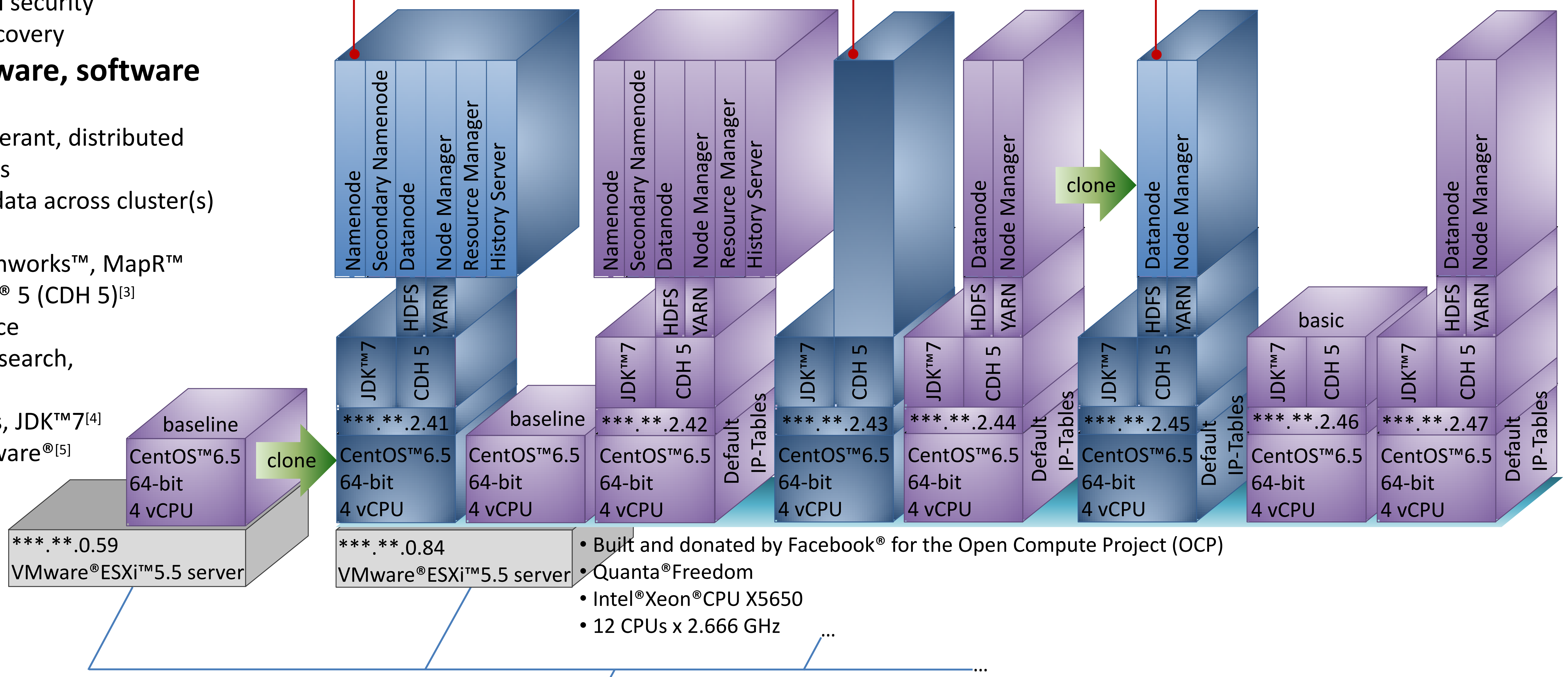
Scope – hardware, middleware, software

- Apache™ Hadoop®[2]
  - Open-source, scalable, fault-tolerant, distributed computing for Big Data analytics
  - Divides, distributes, replicates data across cluster(s)
    - –Complicates data spills
  - Top Vendors: Cloudera®, Hortonworks™, MapR™
  - Cloudera® Distribution Hadoop® 5 (CDH 5)[3]
    - –Apache™-licensed open-source
    - –Batch processing, interactive search, role-based access controls
    - –Requirements: 64-bit systems, JDK™7[4]
- Hadoop® Virtualization with VMware®[5]
  - Cost reduction
    - –datacenter efficiency
    - –server consolidation
  - At least 3 VMs for Hadoop® standard default replication
  - Multiple nodes per host for spill containment
  - Scalable number of hosts for simulating spills across multiple systems
  - Spill residue in VMDK[6]

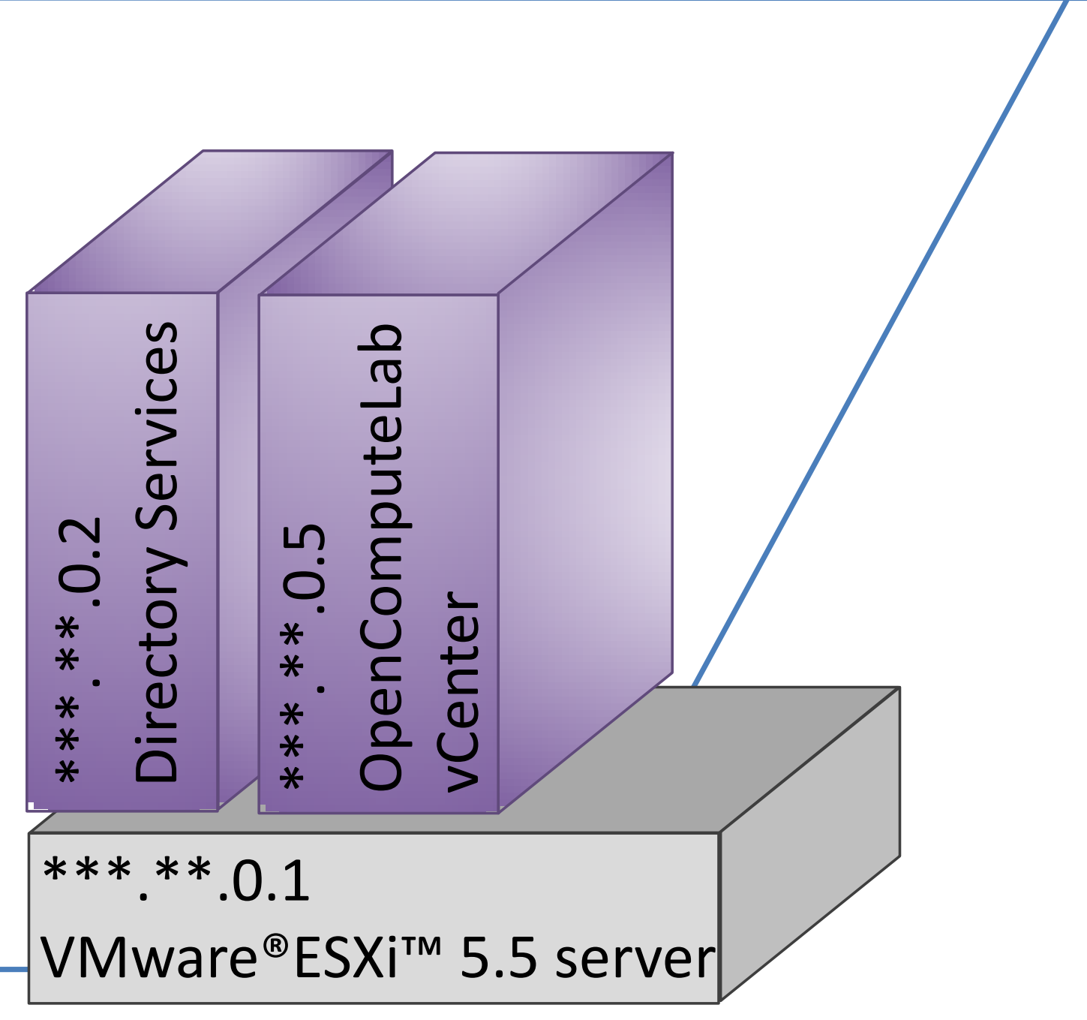
- Purdue campus power-outage
- Server power-box failure
- VM crash
- Unable to bring up daemons

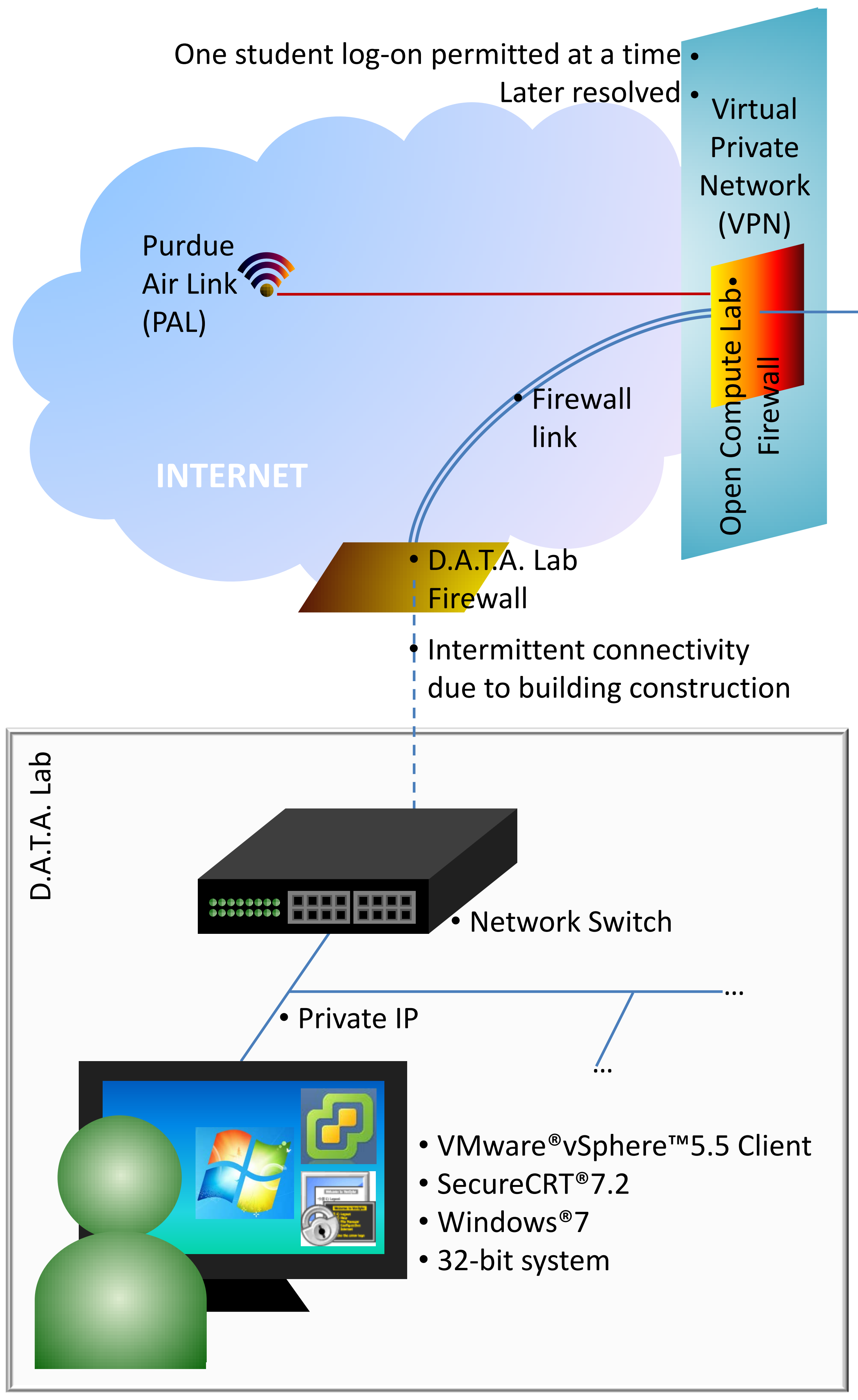
- Faulty CDH 5 installation
- Unable to install daemons
- Uninstallation & Re-installation of CDH 5 unsuccessful in resolving issue

- Cloning from formatted Datanode results in identical UUID received from IP address and port of ***.**.2.44
- Renders both Datanodes unable to establish connection to Namenode
- Uninstallation & Re-installation of CDH 5 unsuccessful in resolving issue



- Built and donated by Facebook® for the Open Compute Project (OCP)
- Quanta®Freedom
- Intel®Xeon®CPU X5650
- 12 CPUs x 2.666 GHz




One student log-on permitted at a time
Later resolved
Virtual Private Network (VPN)
Purdue Air Link (PAL)
INTERNET
Open Compute Lab Firewall
Firewall link
D.A.T.A. Lab Firewall
Intermittent connectivity due to building construction
D.A.T.A. Lab
Network Switch
Private IP
...
...
VMware®vSphere™5.5 Client
SecureCRT®7.2
Windows®7
32-bit system

Methodology

1. Assign server
2. Clone baseline VM to assigned server
3. Establish internet connection for each VM
   i. Assign IP address
   ii. Update Hardware address
   iii. Change Network Device Name
4. Download & Install JDK™ 7 for each VM
5. Download & Install CDH 5 with respective daemons
   i. VM 1: Namenode, Secondary Namenode, Datanode, Node Manager, Resource Manager
   ii. VM 2: Datanode, Node Manager
   iii. VM 3: Datanode, Node Manager
6. Verify inter-VM communication for each VM
   i. Update hostnames & FQDNs across multiple files
   ii. Set IP-Tables to default
   iii. Check Namenode web-console for registered Datanodes

Conclusions

- Results
  - Virtualized Hadoop® environment
  - 3 Datanodes for fault-tolerance
- Recommendations
  - Node fault-tolerance
    - –Independent Namenode
    - –VM snapshots for reversions
  - Server fault-tolerance[7]
    - –Backup generators
    - –Uninterruptable Power Supply (UPS)
  - Cloudera® Express[8]
    - –CDH
    - –Cloudera® Manager: cluster deployment, management, monitoring, performance diagnostics
  - Project Serengeti[9]
    - –Open-source VMware® project
    - –Automates deployment and management of clusters on vSphere™

References

[1] NSA Mitigations Group. (2012). *Securing Data and Handling Spillage Events* [White paper]. Retrieved from http://www.nsa.gov/ia/_files/factsheets/final_data_spill.pdf

[2] Apache™ Software Foundation. (2014). *Welcome to Apache™ Hadoop®!* Retrieved from: http://hadoop.apache.org/

[3] Cloudera®, Inc. (2014). *CDH*. Retrieved from: http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html

[4] Cloudera®, Inc. (2014). *CDH 5 Requirements and Supported Versions*. Retrieved from: http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH5/latest/CDH5-Requirements-and-Supported-Versions/CDH5-Requirements-and-Supported-Versions.html

[5] VMware®, Inc. (2012). *Benefits of Virtualizing Hadoop*. Retrieved from: http://www.vmware.com/files/pdf/Benefits-of-Virtualizing-Hadoop.pdf

[6] VMware® Community. (2007). Classified Spillage. Retrieved from: https://communities.vmware.com/thread/73150

[7] Sun Microsystems®, Inc. (2006). *Server Power and Cooling Requirements*. Retrieved from: http://docs.oracle.com/cd/E19088-01/v445.srvr/819-5730-10/powcool.html

[8] Cloudera®, Inc. (2014). *Cloudera Express*. Retrieved from: http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-express.html

[9] VMware®, Inc. (2014). Apache Hadoop on vSphere. Retrieved from: http://www.vmware.com/hadoop/serengeti.html

INSuRE is training students in information security research. Sponsored by the National Science Foundation with problems provided by the National Security Agency.

This material is based upon work supported by the National Science Foundation under grant #1062970. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

PURDUE UNIVERSITY