

BIG DATA INSTITUTE

Li Ka Shing Centre for Health Information and Discovery



Does data security rule out high performance?

Adam Huffman

2018-02-04 FOSDEM HPC & Big Data Dev Room

Agenda

The background brains of HPC

More ambitious science

HPC meets the “Real World”

Data security déjà vu

Modest Hopes

Context

New job, hence new Questions

...Answers may take longer

Some sites have always faced these problems

Biomedical focus, specifically in England

Secure https://www.bdi.ox.ac.uk

BIG DATA INSTITUTE


Li Ka Shing Centre for Health Information and Discovery

Site Map Accessibility Cookies Contact us Log in

BIG DATA INSTITUTE **UNIVERSITY OF OXFORD**


HOME ABOUT US NEWS EVENTS PUBLICATIONS STUDY TEAM VACANCIES

Search



WELCOME

The Big Data Institute (BDI) is a state-of-the-art building at Oxford University's Old Road Campus. This interdisciplinary research centre focuses on the analysis of large, complex, heterogeneous data sets for research into the causes and consequences, prevention and treatment of disease. To this end, BDI researchers develop, evaluate and deploy efficient methods for acquiring and analysing information for large clinical research studies. These approaches are invaluable in identifying the associations between lifestyle exposures, genetic variants, infections and health outcomes around the globe.



5dac89b5-95ff-4f8e-9...jpg

Show All

The Big Data Institute (BDI) is a new, interdisciplinary research centre that will focus on the analysis of *large, complex, heterogeneous data sets* for research into the causes and consequences, prevention and treatment of disease. Research will be conducted in 4 general themes: genomics, population health, infectious disease surveillance, and methodology (including informatics, statistics, and engineering). Big Data methods could transform the scale (breadth, depth and duration) and efficiency (data accumulation, storage, processing and dissemination) of large-scale clinical research. *The work of the BDI requires people and projects that span traditional departmental boundaries and scientific disciplines*, supported by technical resources to handle the vast quantities of data they generate.



Minerva & Me is a research project to help find people with rare diseases with an aim to understand these diseases better.

Doctors know that people's faces can sometimes have changes that might tell they have a certain disease.

We are teaching computers to look at photographs of peoples' faces to help find these diseases better and faster.



Download the participant information

Register Now!



Minerva & Me



Login

[Forget the password ?](#)

Sign in

☐ keep me logged-in

[About Minerva & Me](#)

[Advisory Board](#)

[FAQ](#)

[Privacy Statement](#)

[Terms of service](#)

The Background Brains of HPC



- *“Security is for someone else”*
- *“{Molecules, particles} don’t have rights”*
- *“Get out of my way”*
- *“Who’s going to check anyway?”*
- (there are exceptions...)



More ambitious science

- Pressure from hyper-scalers
- More capable instruments
- Working across domains
- Pressure from funders

More ambitious science



<https://www.genomicsengland.co.uk/the-100000-genomes-project/>

<https://allofus.nih.gov/>

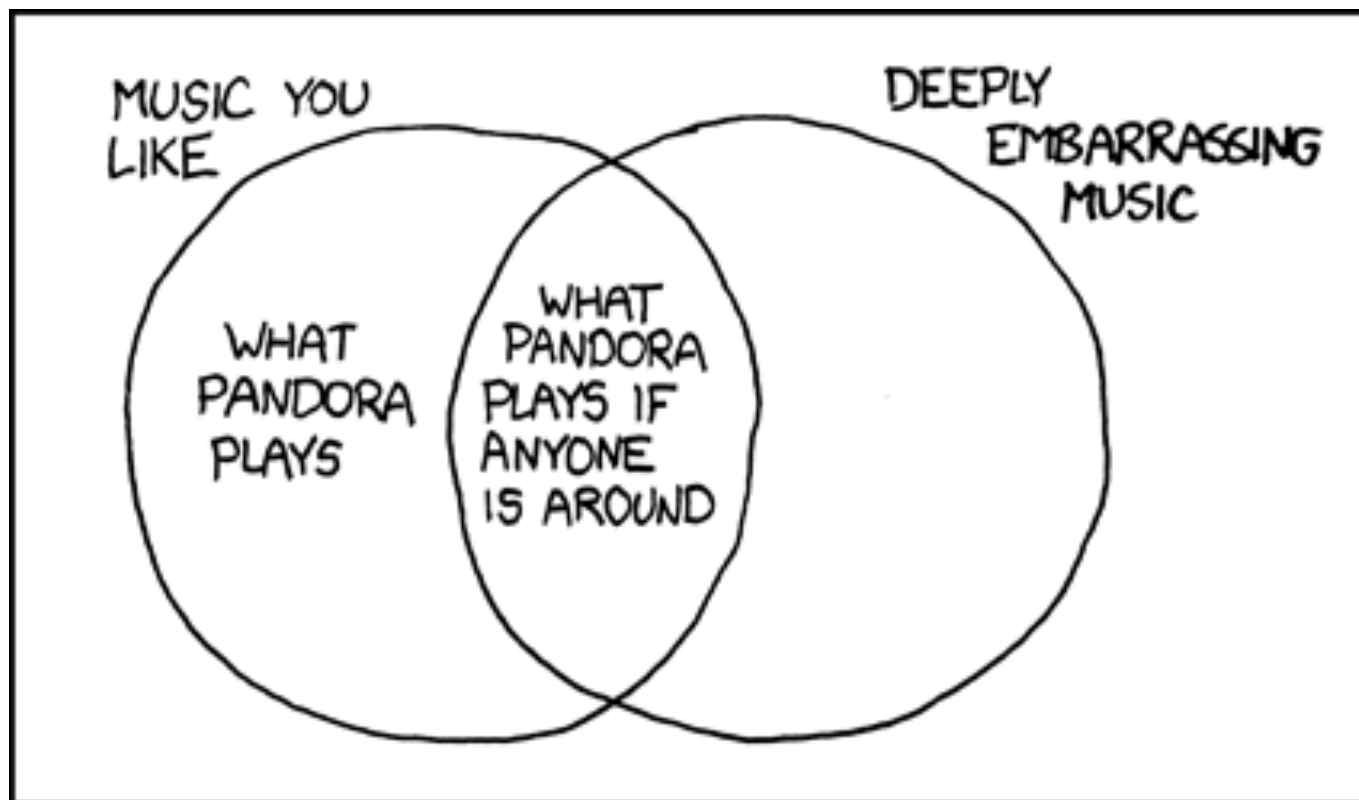
HPC meets the “Real World”

- Electronic Health Records (EHR)
- Hospital Episode Statistics (HES)
- Prescription Data
- <https://www.bigdata-heart.eu/>

HPC meets the “Real World”

- Protected data implies data sharing
- Data sharing implies agreements and audits
- Clashing requirements?

HPC meets the “Real World”



HPC meets the “Real World”

- Protected data implies data sharing
- Data sharing implies agreements and audits
- Clashing requirements?
- Take care of your reputation...

This is our old website. Most information can now be found on our [new NHS Digital website](#). Let us know what you think.

[Home](#) > [Services](#) > [Data Access Request Service \(DARS\)](#)

[Home](#)

[Services](#)

▼ [Data Access Request Service \(DARS\)](#)

[DARS Process](#)

[Information on type 2 opt-outs](#)

▶ [Data Sharing Audits](#)

[Local Authority HES Extract Service](#)

[Benefits case studies \(Data Extracts\)](#)

[Contact NHS Digital](#)

Data Sharing Audits


We carry out independent audits and where necessary post audit reviews to check that our customers are meeting the obligations in their [Data Sharing Contracts](#) and [Data Sharing Agreements](#). This helps to ensure that organisations abide by the terms and conditions set by NHS Digital and data is kept safe and secure.

To assist the data recipient with the audit process NHS Digital has produced an Audit Guide and an Action Plan template. Audits undertaken between October 2016 and December 2017 were against Version 1 of the Audit Guide. Audits conducted from 1 January 2018 will be against Version 2.

Each audit and post audit review carried out results in the publication of a formal audit report. From April 2016, audit reports are being published on monthly basis and are scheduled to appear every third Thursday. Note the style of the audit report varies according to version of the Audit Guide the audit was conducted against.

-  [NHS Digital Data Sharing Audit Guide Version 2 \[1Mb\]](#)
-  [NHS Digital Data Sharing Audit Guide \[618kb\]](#)
-  [Data Sharing Audit Action Plan - Template \[52kb\]](#)

January 2018

-  [Data Sharing Agreement Audit - City of York Council \[313kb\]](#)
-  [Data Sharing Agreement Audit - Cambridgeshire County Council \[389kb\]](#)
-  [Data Sharing Agreement Audit - London Borough of Enfield \[444kb\]](#)
-  [Data Sharing Post Audit Review - Moorfields Eye Hospital \[307kb\]](#)
-  [Data Sharing Post Audit Review - University of Leeds \[260kb\]](#)

November 2017

-  [Data Sharing Agreement Audit - Northamptonshire County Council \[294kb\]](#)
-  [Data Sharing Post Audit Review - NEW Devon CCG \[335kb\]](#)
-  [Data Sharing Agreement Audit - University of Oxford \[303kb\]](#)
-  [Data Sharing Post Audit Review - SCW CSU \[285kb\]](#)

October 2017

-  [Data Sharing Agreement Audit - Arden GEM \[296kb\]](#)



HPC meets the “Real World”

- Data sharing agreements and audits

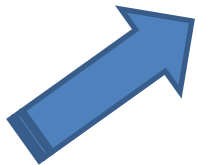
“There were three auditors – lead, support and trainee. All were friendly but well informed and looking hard at what we presented. As well as policies, They looked in almost forensic detail at the computer used to download the data, the drive where it was stored and the two machines in the secure computing room where it had been worked on.”

- Wulf Forrester-Barker - NDORMS, University of Oxford

HPC meets the “Real World”

- Data sharing agreements and audits

“There were three auditors – lead, support and trainee. All were friendly but well informed and looking hard at what we presented. As well as policies, They looked in almost forensic detail at the computer used to download the data, the drive where it was stored and the two machines in the secure computing room where it had been worked on.”



- Wulf Forrester-Barker - NDORMS, University of Oxford

HPC meets the “Real World”

- Audits on HPC systems conducted by external contractors when processing NIH data
- Shift in the burden of proof?
 - <https://deepmind.com/blog/trust-confidence-verifiable-data-audit/>

HPC (and Big Data) meets the “Real World”

“The ‘wow’ phase of big data appears to be coming to an end, and a more sober understanding of its power is replacing it.”

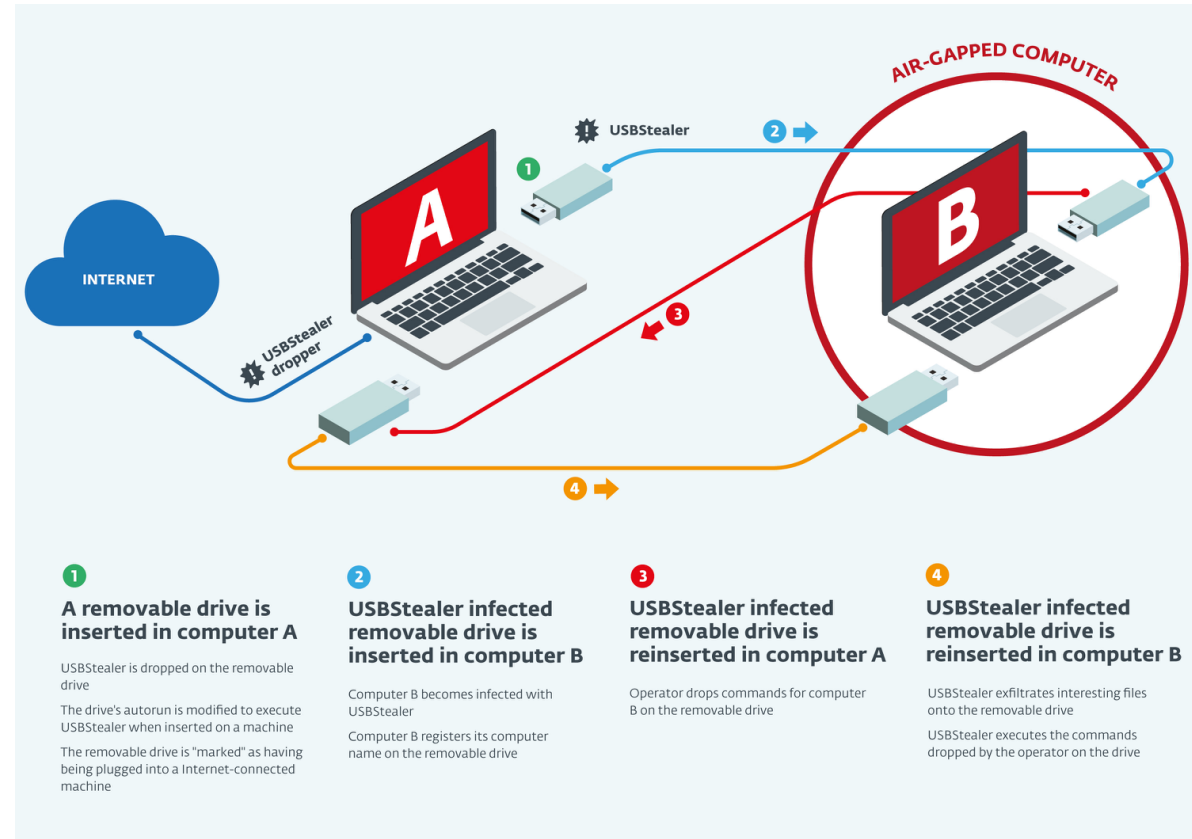
- Dr. Patrick Healy, University of Limerick



Big Data Déjà Vu?

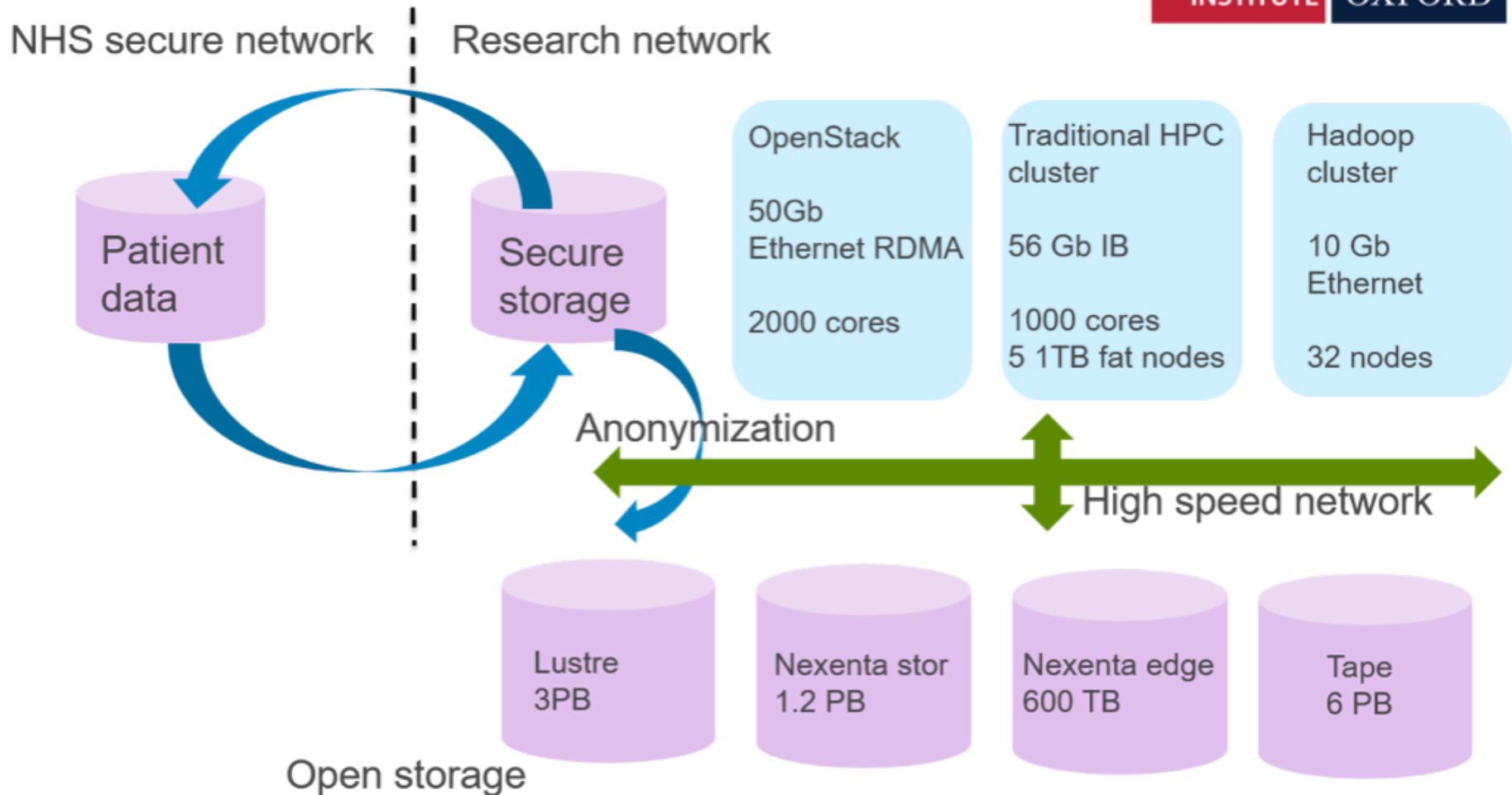
Can't we just use simple segregation of systems for this?

- Cf. traditional air-gap
- Affordability
- Flexibility



<https://www.welivesecurity.com/2014/11/11/sednit-espionage-group-attacking-air-gapped-networks/>

OpenStack Clinical Cloud



<https://www.linkedin.com/pulse/cambridge-university-transforms-medical-imaging-dell-openstack-eric/>



Exactly how anonymous?

- Process of anonymising data becoming harder?
 - http://knowledge.freshfields.com/m/Global/r/1640/can_clinical_trial_data_be_adequately_anonymised
- Correlating data sources becoming easier
- Can we safely process anonymised data on general purpose clusters?
- European Medicines Agency
 - Data anonymization workshop
 - http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2017/10/event_detail_001526.jsp&mid=WC0b01ac058004d5c3

Immutable Data Security Infrastructure?

OpenStack as data centre API



Move towards immutable infrastructure

Not just virtualisation - Ironic

Explicitly encode relationships between
networks, users, security policies

https://fosdem.org/2018/schedule/event/vai_openstack_gdpr_compliance/



Big Data Déjà Vu?

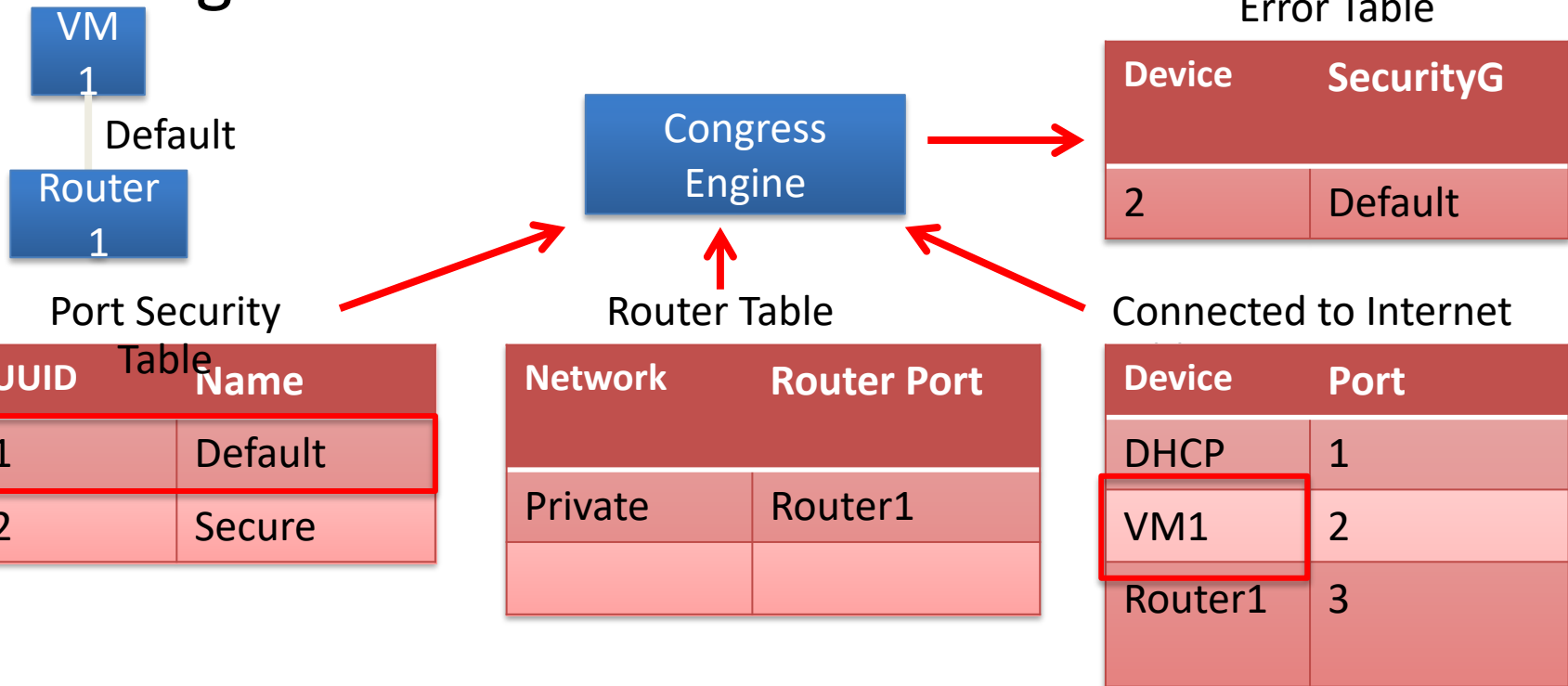
OpenStack Congress

“open policy framework for the cloud”

- Monitoring
- Proactive enforcement
- Reactive enforcement

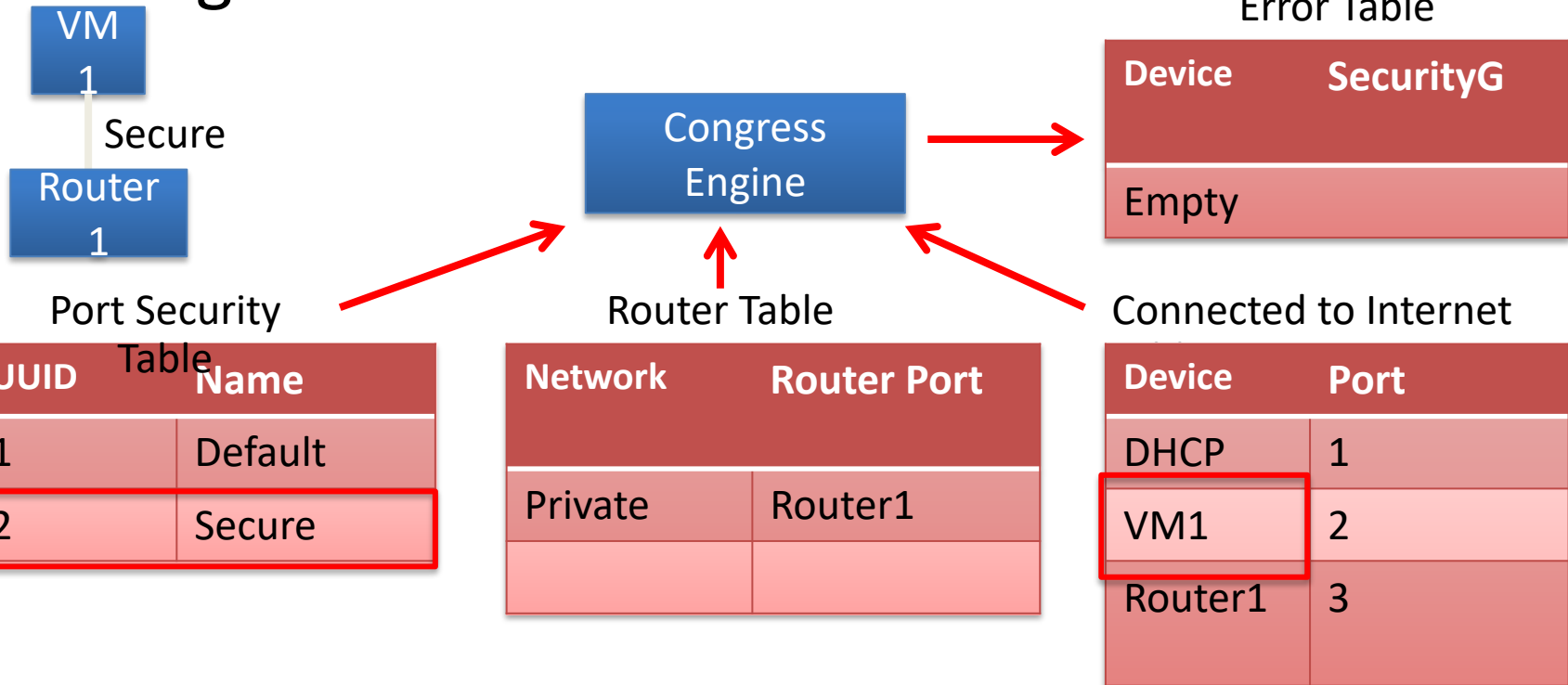
Part 1 – Error Table

Error if any VMs connected to Internet is not using Secure



Part 2 – Error Table

Error if any VMs connected to Internet is not using Secure



Big Data Déjà Vu?

OpenStack Congress

- Isn't this just what we do anyway?
- Things go wrong, and that's why we still have jobs
- Audit and proactive enforcement
- Delegate some admin rights to users, mistakes happen
- Effectively forces creation of documentation, essential for audit

Big Data Déjà Vu?

Containers help security?

Build on work on security in the container world

<https://github.com/cilium/cilium>

“API-aware Networking and Security for Containers based on BPF”



<https://github.com/coreos/clair>

“ [static analysis](#) of vulnerabilities in application containers”

Extend this to check for data privacy compliance?





Big Data Déjà Vu?

“The Cloud”

We need to find answers that work on infrastructures that we don't control
e.g. public clouds, owing to pressure to use them from funders

Can we have fast enough encryption, possibly via AVX512, to use it
ubiquitously?

Big Data Déjà Vu?

Hardware aspects of “The Cloud”

Meltdown/Spectre, VMs particularly badly affected

AMD Secure Encrypted Virtualization

<https://developer.amd.com/amd-secure-memory-encryption-sme-amd-secure-encrypted-virtualization-sev/>

“Secure Encrypted Virtualization is Unsecure”

<https://arxiv.org/pdf/1712.05090.pdf>

Modest Hopes, and a New Realism

Or, conclusions

- Data security needs to be considered **at the system design stage**
- The HPC community needs to engage **much more widely**
- ... and expect to be **challenged**, rather than left alone in the office with no windows
- Job time = computing time + I/O time + **data transfer time + anonymization time + data security negotiation time...**

Image credits

<http://spsswizard.com/assumptions-spss/>

<https://www.allmusic.com/album/things-have-changed-mw0002540390>

<https://blog.volkovlaw.com/2015/08/calculating-the-incalculable-reputational-damage-part-i-of-iii/>

<https://www.welivesecurity.com/2014/11/11/sednit-espionage-group-attacking-air-gapped-networks/>

<https://www.silicon.fr/shadow-cloud-menace-opportunit-e-les-dsi-97072.html>

<https://xkcd.com/668/>

OpenStack Congress presentation from the Vancouver Summit

Thank You

adam.huffman@bdi.ox.ac.uk

@adamhuffman