Setting Up Compute Infrastructure for Sensitive Data

Yohai Meiron, Scientific Applications Analyst
Shawn Winnington-Ball, Manager, Information Systems Security

SciNet, University of Toronto
13 Nov 2024

SciNet
ADVANCED RESEARCH COMPUTING at the UNIVERSITY OF TORONTO

Agenda for today

1. Intro to SciNet and S4H
2. Technical overview
3. Compliance overview

SciNet
ADVANCED RESEARCH COMPUTING at the UNIVERSITY OF TORONTO

What’s S4H?

- Secure enclave at SciNet for HPC jobs involving sensitive data
- Orphan initialism formerly meaning ‘SciNet4Health’
    - Not limited to the health space, and renaming talks are underway

What’s sensitive data?

- We use the UofT data classification standard for guidance
- Level 4 data:
  - *Highly sensitive research data, requiring stronger security controls, whose unauthorized access, disclosure, or loss poses significant financial, reputational, legal or physical risk to the data subject, researcher, University, etc.*
- Data custodians are rightly concerned with the distribution of this data to parties that can guarantee a secure and stable environment

With whom are we working on this?

- U of T Research Information Security
  - Michael Laurentius, and formerly Sue McGlashan
  - Giving us guidance on compliance and helping us navigate the overall landscape
- With the support of VP Research & Innovation, and Information Technology Services

Why this?

- Researchers indicated a need for hosting sensitive data that our main cluster Niagara (now Trillium) is not intended to host
- Other related projects include:
  - HPC4Health
  - Health Data Nexus
  - Ontario Health Data Platform (OHDP) (ret.)
  - T-CAIREM
  - SecureData4Health

S4H: THE TECHNOLOGY

Yohai Meiron

WHAT IS SciNet ?

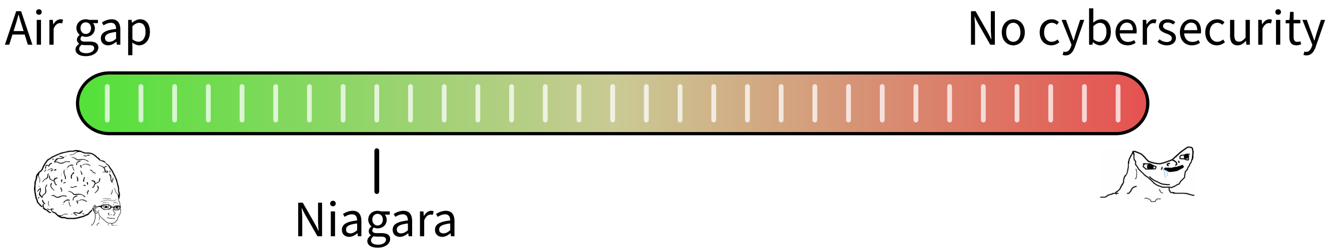
We are the supercomputing centre at the University of Toronto.

We provide Canadian researchers with computational resources and expertise necessary to perform their research at scale.

We operate the *Niagara* supercomputer, as well as several other systems.

(Soon to be replaced by the brand new *Trillium* system)

BALANCING SECURITY AND USABILITY



Air gap

No cybersecurity

Niagara

- Air-gapped system (“Fort Knox” approach)
  - Users have to go in person to a secure facility to do their work
  - Data go in an out on physical media under the supervision of the admins
- No cybersecurity precautions (“open house” approach)
  - Security through obscurity mostly
  - Unacceptable to data custodians
- Niagara already has some security measures in place
  - SSH keys only authentication (no passwords)
  - Mandatory 2-factor authentication (2FA)
  - Security patches applied to system in a timely manner
  - Admins monitoring for suspicious activity
  - Reasonable physical security measures

WHERE DOES NIAGARA FALL SHORT?

- Users
  - Come from a huge pool of potential users
  - Are minimally vetted
  - May connect from anywhere in the world
  - Can run arbitrary code
  - Make mistakes
    - `chmod -R 777 ~`
    - Rstudio server 🙄
- Data lifecycle
  - No encryption at rest
  - Backup in multiple copies (Home & Project)
  - No secure disposal mechanism

THE S4H SECURE ENCLAVE

Addresses *some* of the security concerns

- Hardened access
- Encryption at rest
- Group isolation
- Data egress control (select groups)

Striving to provide a *similar experience* to Niagara

- It’s an HPC environment, not cloud!
- Similar nodes
- Access to the same software
- Support from the same team at SciNet

THE S4H SECURE ENCLAVE (CONT.)

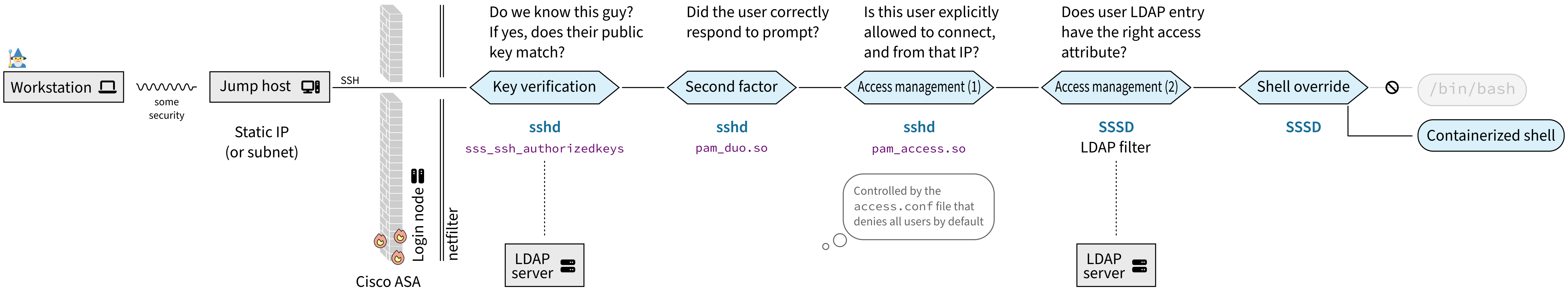
- Not a “Fort Knox” solution
  - Better than Niagara
  - We still make compromises for usability
  - Satisfactory level of security for our users

CURRENT STATUS

Pilot project at the University of Toronto

- Available for free for groups at UofT
- Online since July 2023
- 5 groups onboarded

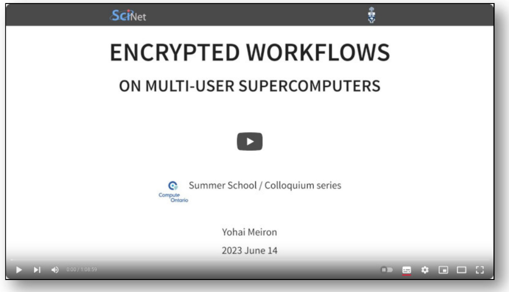
HARDENED ACCESS: FIREWALLING AND AUTHENTICATION



ENCRYPTION AT REST

- Parallel file system (IBM Spectrum Scale, a.k.a. GPFS)
- Key management server (for AES-128 keys)

Can I enjoy encryption on Niagara or the other systems?



SciNet

ENCRYPTED WORKFLOWS

ON MULTI-USER SUPERCOMPUTERS

Summer School / Colloquium series

Yohai Meiron

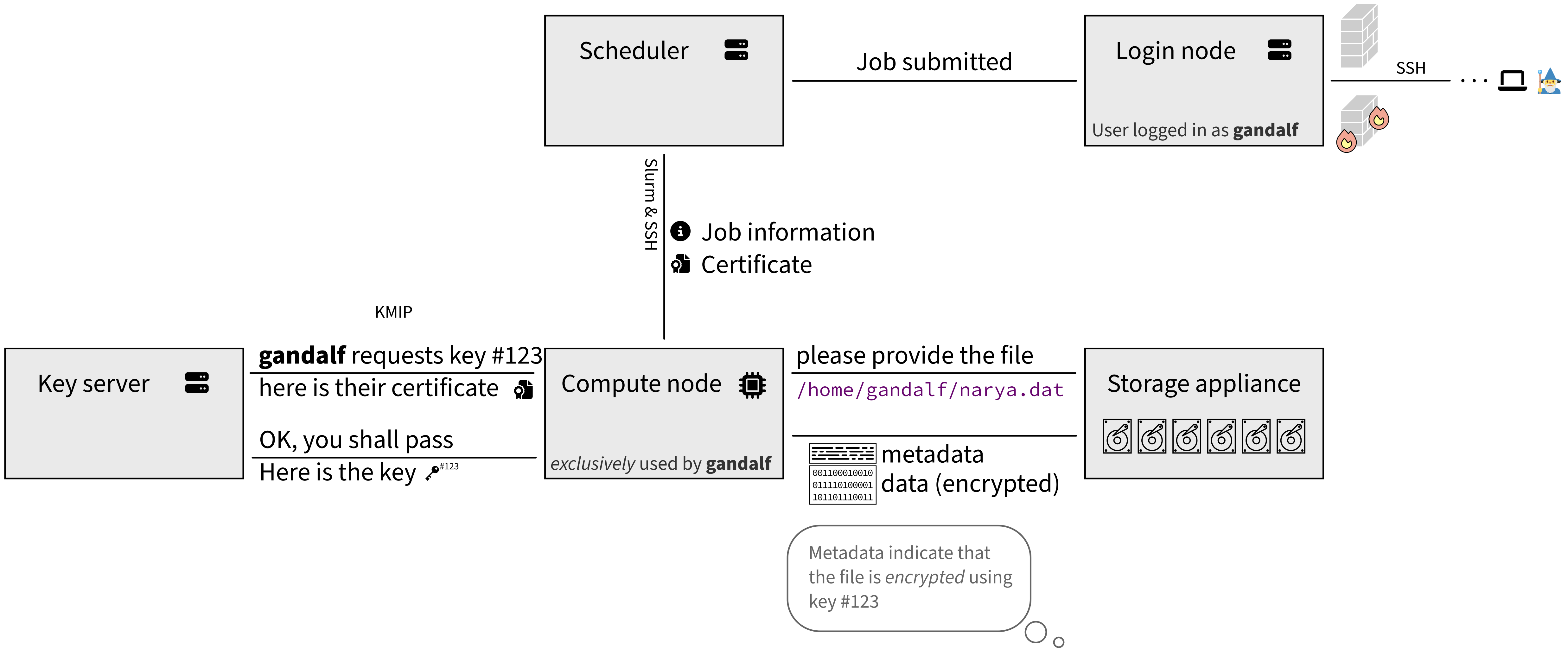
2023 June 14

Yes! But it’s not straightforward.

More details in my colloquium from last year https://youtu.be/xwXXknJbmmE

- You can use software encryption fully in user space
- That’s in contrast to “native” GPFS encryption in S4H
- Takes some time and expertise to set up
- Onus is on you to do it correctly
- Cybersecurity is more than just encryption at rest

ENCRYPTED JOB FLOW



Scheduler

Job submitted

Login node

User logged in as **gandalf**

SSH

Slurm & SSH

Job information

Certificate

KMIP

Key server

gandalf requests key #123

here is their certificate

OK, you shall pass

Here is the key #123

Compute node

exclusively used by **gandalf**

please provide the file

/home/gandalf/narya.dat

metadata

data (encrypted)

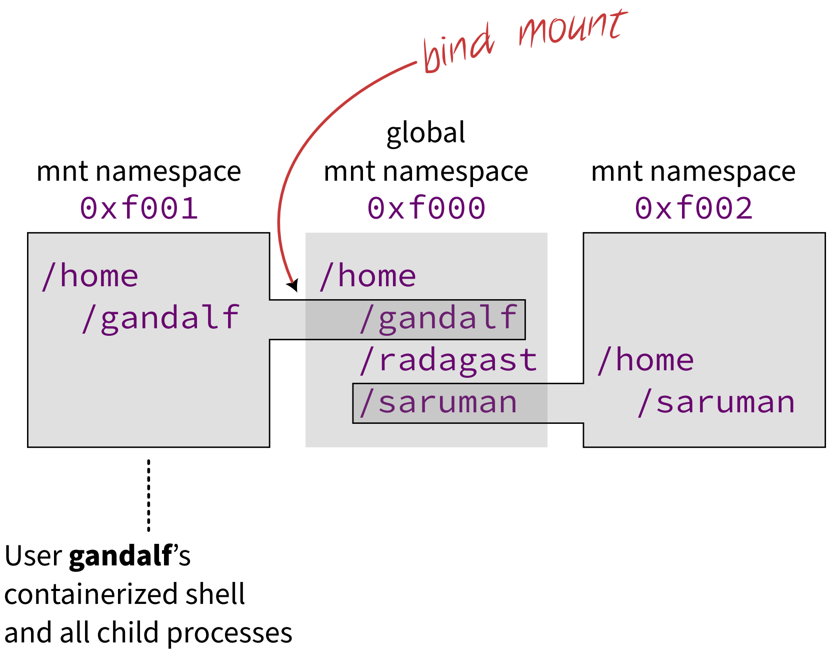
```
001100010010
011110100001
101101110011
```

Storage appliance

Metadata indicate that the file is *encrypted* using key #123

GROUP SEGREGATION ON THE LOGIN NODE

- The login node is an actual shared resource
- OS-level virtualization (container technology)
  - Containers are processes with reduced visibility to the host
  - System resources utilization control with cgroups
  - Each user gets their own container instance
  - Fully in userspace (no `setuid` or capabilities)



The Compliance: Shawn

Compliance

- How do we know how ‘secure’ we are? Is there a manual of sorts, a yardstick by which we can measure?
    - Yes, many: SOC2, ITSG-33, ISO 27001, PCI DSS, CMMC, etc.
- Our own opinion of our cybersecurity stance and effectiveness isn’t enough; we need to demonstrate to outside parties that we can be trusted to host sensitive data and not put UofT on the front page
- We follow UofT’s *Information Security Control Standard*, which is based on CMMC, which is based on NIST 800-171 rev. 2

NIST 800-171 rev. 2

*This publication provides federal agencies with **recommended security requirements** for protecting the confidentiality of CUI when the information is resident in nonfederal systems and organizations.* (https://csrc.nist.gov/pubs/sp/800/171/r2/upd1/final)

- CUI ~ Controlled Unclassified Information
  - *CUI is sensitive information that does not meet the criteria for classification but must still be protected* (https://www.dodcui.mil/)
- Set of 110 security controls in 14 categories that span all aspects of cybersecurity
  - Many technical, many documentation
  - Most organizations already have many of these controls to some degree

CMMC

- Cybersecurity Maturity Model Certification
- U.S. defense industry requires that all contractors achieve a certain level of cybersecurity maturity: makes sense given the sensitive nature of their industry
- Basically a compliance framework centered around the controls in NIST 800-171
- The standard itself is simply a set of security requirements; CMMC provides the structure:
  - Contract language, professional accreditation, assessment process, scoring methodology, etc.

CMMC con't

- Note that no organization or data custodian is requiring us to follow CMMC
  - In the absence of a directive or generally accepted practice, this was our choice
- NIST is fairly well-known in academic circles
  - https://www.regulatedresearch.org CoP
- In fall 2023, the Canadian government announced the Canadian Program for Cyber Security Certification (CP-CSC), which is essentially CMMC fitted to the Canadian landscape, intended for use by Canadian defence contractors working with their U.S. counterparts

SciNet compliance journey

- Familiarizing ourselves with the control set by reviewing it as a team
  - Some ambiguity in the requirements and for good reason (they have to be agnostic to cover all manner of environments and technologies)
- Added some management metadata to each control to track our notes, relative difficulty, predominantly technical or policy, responsibility, etc.
- Initial impressions: 800-171 is very thorough, covers all aspects of cybersecurity, and focuses a lot on documentation: *if it's not written down, it doesn't exist*
- Policies and corresponding procedures are important

SciNet compliance journey con’t

- System Security Plan (SSP)
    - Where all aspects of site security are documented directly or linked
        - Diagrams, technologies, per-control summaries
    - Think of it as a table of contents
    - Give this to an assessor and they will quickly get an idea of your posture
- Plans of Action and Milestones (POA&Ms)
    - Anything that isn’t an immediate ‘pass’ is something that needs add’l work and this effort needs to be documented

SciNet self-assessment against NIST 800-171

- Two templates, briefly
  - SciNet original
  - U of T Trusted Infrastructure Framework assessment

Next steps

- Complete internal assessment against NIST 800-171 performed by U of T Research Information Security
  - Michael Laurentius will wear his CMMC assessor’s hat
  - We need to meet some ‘acceptable’ level of compliance, TBD
- Continuing to onboard S4H clients with L3 data, for which we are approved
- Find clients willing to bring their L4 data and do what’s needed to gain their trust
  - Marketing effort required
  - S4H security statement has been drafted and approved

Next steps con’t

- Develop new user onboarding process
    - We know the rough steps already through onboarding with existing clients
    - Lots of back and forth with U of T departments, data custodians
    - How can streamline as much of this process as possible?
- Investigate cost, scope, necessity of privacy impact assessment, threat risk assessment, and penetration testing as external third-party reviews
    - Technical architecture may change as a result

Thank you!

- dgruner@scinet.utoronto.ca
- swball@scinet.utoronto.ca
- yohai.meiron@scinet.utoronto.ca