# The IT Infrastructure of the LUX-ZEPLIN Experiment at SURF

- CHEP 2023
- Norfolk, VA
- Steffen Luitz (SLAC National Accelerator Laboratory) luitz@slac.stanford.edu



# The LZ Dark Matter Experiment

#### LZ detector





S. Luitz

<u>NIM A 163047 (2020)</u>

- Liquid Xenon Time Projection Chamber (TPC)
  - 7 tonnes active LXe
  - Located 4850ft (~1mile) underground at Sanford Underground Research Facility (SURF) in Lead, South Dakota, USA
  - Currently operating and taking data
- Data rates
  - up to ~ 40MByte/s during routine Dark Matter search
  - $\circ$  up to ~ 350 MByte/s during calibration
- Dataflow deeply buffered
  - hold 2 months of Dark Matter search data at SURF
- ~200 active user accounts in system (shifters, experts, etc.)



# Designing the LZ Online IT system

# IT Environment at SURF

- Small lab in remote location, small IT org
- Very limited IT support for experiments
  - Firewall, VPN, basic user management
  - no services
- Site/underground network
  - hostile mine environment
  - too slow at design time, upgraded since
- Internet connectivity good
  - 10Gbit/s Internet2
  - 2 commercial ISP connections



### LZ requirements

- 10Gbit/s underground-surface-WAN
- Highly reliable access to underground
  - slow control & detector operations
- >> 100 physical devices on network
- ~ 500 user accounts over lifetime,
- "secure", including remote access
- "easy to manage", "cost-considerate"

#### Building our own

- Opportunity to **design** IT system
- Network
- Compute
  - Extensive use of virtualization
  - Centrally managed
    - Authentication, Directory
    - Configuration management
- Prototyping at SLAC to demonstrate performance and manageability
  - supporting LZ detector R&D and QA/QC

#### LZ System Test facility at SLAC





3



LAC S. Luitz

# Network

#### Physical

- Direct connection of LZ underground and surface
- 4 strands of "dark fiber" from SURF (+2 for GPS)
- 2 Netgear m4300 stacks (underground, surface)
  - 10Gbit/s, Layer 3, low cost, expansive features, redundant
  - 4 x 10Gbit/s, point-to-point, fully redundant between stacks + to SURF
  - 2 fiber pairs with different paths (via Yates and Ross mine shafts)
- All critical devices (VM hosts, PLC, etc.): connections to their switch stack
- Separate Emergency/Out-Of-Band (OOB) network.
- 10 access switches underground (48 ports, 1Gbit/s each) for devices

# Logical

Layer-2 (VLAN) and Layer-3 IPv4 design

S. Luitz

- VLANs/subnets and Layer 3 ACLs create zones (security/fault isolation)
- static routing, redundancy managed by stacking and Layer-2 (LACP)
- Connecting to SURF network via redundant vyOS Firewalls (VMs)
  - NAT to map internal services to (screened) public IPs
    - we use about 20 public IP addresses in a /27 network
    - once allowed by SURF firewalls, ports out on the open Internet
- OOB allows remote access with VM clusters (and LZ external firewalls) down
  - can use SURF network as alternative path for UG access





# Underground and Surface Clusters

Separate; underground cluster has no operational dependencies on surface

#### Hardware

- 2 VM host clusters (underground and surface), each:
  - o 3+ x Dell R530/R540, 256 GByte RAM, 5 x 4TB SSD
  - each server 2 x 10 Gbit/s Ethernet to switch stack
- 2 Supermicro JBOD servers at surface (128 GByte RAM, 40 x 6TB HDD each)
  - ZFS RAIDz2 (some storage set aside for backups and local data)
- Fully remotely manageable
  - built-in service processors, external KVM, power control via PDUs
- All equipment dual power (1 leg on UPS, one leg on house power, generator-backed in both locations)

# Virtualization Software (open source)

- Proxmox VE
  - We use full VMs, not containers
  - Cluster storage on Ceph
  - Clusters can each operate with 2 out of 3 nodes
    - Seamless routine maintenance using hot VM migration
- VM image backup with Proxmox Backup Server
- Running ca. 100 virtual machines total



S. Luitz



Surface Rack, front and back

Underground Infrastructure & Network Rack

(PROXMO)	× Virtua	I Environment	7.3-6 Sea	arch							Documentation	Create VM	😭 Create CT 🔒 Iu	itz-a@luxzeplin.org	
erver View		Datacenter												🚱 Help	
Datacenter (Underg	round)				ouroca										
withost2-pve		Q Search Summary Notes Cluster		CPU					Memory			Storage			
>   vmhost4-pve  >   vmhost5-pve															
		Cussel Coph Coph Cophons Storage Backup Replication Permissions Users API Tokens API Tokens				19%			31%			7%			
				of 80 CPU(s)				392.96 GiB of 1.23 TiB				6.67 TIB of 96.16 TIB			
				Noc	les									$\odot \odot$	
				Nan	пе			ID	Online	Support	Server Address	CPU usage	Memory usage	Uptime	
				vmh	iost1-pve			4	×		172.20.99.32	42%	43%	52 days 18	
				vmhost2-pve			5	×	-	172.20.99.33	23%	59%	52 days 19		
				vmh	iost3-pve			3	~		172.20.99.34	29%	52%	52 days 19	
		va, iwo⊧a	ctor	vmr	iosis-pve		1	÷	+	172.20.99.35	1%	1%	52 days 20 52 days 21		
Tasks Cluster log								~							
Start Time ↓	End Tim	10	Node		User name		Description	_			_		Status		
Apr 21 04:20:07	Apr 21 0	04:20:12	vmhost3	vmhost3-pve		root@p					ок				
Apr 21 03:37:16	Apr 21 0	03:37:21	vmhost2	vmhost2-pve		Proxmox dashboard			ard			ОК			
Apr 21 01:34:19	Apr 21 01:34:23		vmhost1	-pve	root@p (underground clu			clu	ster)				ОК		
Apr 21 00:58:18	Apr 21 0	Apr 21 00:58:24 vr		ost5-pve root@pa							ОК				
Apr 21 00:50:48	Apr 21 00:50:52		vmhost4	vmhost4-pve		ot@pam Update package da			labase	ase			ок		
00.00 44.05	• • • • • • • •				10		1 HOT 400						017		

# System Architecture, Management, and Services

# Infrastructure and Management

- CentOS-7
- Samba 4 Active Directory with sssd on clients
  - Kerberos, LDAP and DNS with replication
  - 2 domain controllers (VM) each underground and surface
- Saltstack for configuration management
  - authoritative configuration in YAML files
  - includes network and firewalls
- Storage
  - VM images on Ceph
  - Extensive use of ZFS (zfs on linux)
    - snapshots, send/receive file systems
- <u>https://luxzeplin.org</u> domain for access
- Multiple MySQL 8 database server VMs
  - largest is slow control history with 2.1TB
  - replication+snapshots of replicas for downtime-less backups
- Central syslog collection (searchable)
- Backups (leveraging ZFS)
  - snapshots, local copies, off-site to University of Rochester



### **Role Management**

- Developed role management system to manage fine-grained LDAP permissions (group memberships)
  - permissions granted/revoked based on 'role'
  - e.g. 'High Voltage Operator' role assigns all necessary access and slow control permissions to operate HV.
  - management via web interface, can be delegated
  - group and role definitions in YAML files



# Remote Access, Security

### **Remote Access**

- 2-factor authentication for **all** remote access
  - Bring-your-own-token (TOTP, TAN, Yubikey, App)
  - PrivacyIDEA, open source, locally hosted.
- Reverse web proxy with SAML login for internal web apps
  - all web apps under single tree (luxzeplin.org)
  - permissions from LDAP through SAML
- OpenVPN
  - automated certificate management
  - fine-grained IP target address restrictions (from LDAP)
  - Education institution license of Viscosity client
- SSH
  - fine-grained IP target address restrictions (LDAP)

# Security

- Defense in depth designed into the system
- Free/open source tools

#### PrivacyIDEA portal, mobile app, Yubikey



#### Security Architecture / Features





# **Operational Experience**

### Staged installation, Rochester to SURF

- Setting up production system at University of Rochester
  - first application was electronic logbook to support LZ construction activities
  - most system setup and configuration via remote management from California,
  - resulted in intrinsically remotely operable system
- shipping pre-configured network/clusters to SURF, installation at SURF in 2018/2019
- operating very reliably since
- Full/reliable remote access and management essential during Covid-19

# Availability

- Network/host monitoring system tracks downtime
- May 2022-now (1 year):
  - VM typical availability 99.995% ('4 ½ nines', ~30min downtime/year)
  - Physical host availability 99.95%-99.995% (slower boot times)
  - dominated by 2 scheduled LZ network outages (switch firmware upgrades)







- LZ network and computing at SURF is working reliably
- Experiment is taking data
- Personal perspective
  - In my 30 year career building and improving DAQ and Online systems, IT was always something already existing.
    - Organically grown, difficult to improve
    - In particular system management and Information Security aspects, retrofits usually collide with user expectations and convenience
  - I was always wondering if designing something from scratch is better
    - Both users and administrators
  - The answer is Yes!



# LUX-ZEPLIN (LZ) collaboration

#### • Black Hills State University

- Brookhaven National Laboratory
- Brown University
- Center for Underground Physics
- Edinburgh University
- Fermi National Accelerator Lab.
- Imperial College London
- King's College London
- Lawrence Berkeley National Lab.
- Lawrence Livermore National Lab.
- LIP Coimbra
- Northwestern University
- Pennsylvania State University
- Royal Holloway University of London
- SLAC National Accelerator Lab.
- South Dakota School of Mines & Tech
- South Dakota Science & Technology Authority
- STFC Rutherford Appleton Lab.
- Texas A&M University
- University of Albany, SUNY
- University of Alabama
- University of Bristol
- University College London
- University of California Berkeley
- University of California Davis
- University of California Los Angeles
- University of California Santa Barbara
- University of Liverpool
- University of Maryland
- University of Massachusetts, Amherst
- University of Michigan
- University of Oxford
- University of Rochester
- University of Sheffield
- University of Sydney
- University of Wisconsin, Madison







LZ Collaboration Meeting University Of Maryland 5<sup>th</sup>-7<sup>th</sup> January 2023





Science and Technology Facilities Council







BOLD PEOPLE VISIONARY SCIENCE REAL IMPACT BOLD PEOPLE VISIONARY SCIENCE REAL IMPACT

# **Backup Slides**



# List of Tools

#### Open Source

- CentOS-7, ZFS on Linux (OS)
- Saltstack (configuration management)
- Proxmox Virtual Environment (Virtual Data Center)
- Proxmox Backup Server (VM image backups)
- Sanoid/Syncoid for ZFS (snapshot+remote backups)
- Samba 4 AD (Directory, Kerberos, LDAP, DNS)
- MySQL (database)
- PrivacyIDEA (2-factor authentication)
- vyOS (router/firewall linux distribution)
- OpenVPN (VPN)
- simplesamlphp (SAML IdP and SP)
- PSI elog (electronic logbook)
- Drupal (web portal)
- LibreNMS and elastiflow (network monitoring)
- SSP (password portal)
- BestPractical RT (ticket tracker)
- Graylog (syslog indexing and access)
- Security Onion (intrusion detection)
- PerfSONAR (Internet performance measuring/monitoring)
- Commercial (unlimited users licenses)
  - Ignition for Slow Control
  - Viscosity VPN client (educational site license)



S. Luitz	
----------	--

2 Bill Cognet in an "Steller Laft"	
ner rent best, herr tomp soper allant my Managed by Teach M Taxons Teach Teleform - Triper	A-
KO Interest particular (a point 10 million)     Montania Interest (a point 10 million)     Montania     Montania     Montania     Montania     Montania     Montania     Montania     Montania	
<ol> <li>Our an answer of the second sec</li></ol>	
	Home 12 Online Solvers Becomentations - Link of article 12 arcrants - Link of disabled 12 arcrants - Ma arcrant
PSI elog	
SAML/LDAP	LZ Online System Resources
	Z Online System Resources
	IZ Electronic Lophonk
Construction	
16-16-29 Si tit ng insulining invé seres	
Cheered and work the transfer to be indeed to be been tauk.	Manage your account settings
	Change your luxzeplin.org user settings
17	Manage your luxzeplin.org 2-factor tokens
Weissme Steffen Luitz	Change your luxzeplin.org password
Use this fairs to be used update your an out of all of orders professions. To change your primary sends address, same or your reservance, phases cardial the LE Online Admin Team	
Administrative doublishers from C. (2) submits with to set a compact to our holds, the polytopole basis.  You will be upged and administrative in the second basis.  Execution of the compact to the comp	
Verified Television (Verified Television) (Verified Television)	
Visiogian         Audigiana andremi esta           Normania         (%,4%) features of non-construction (x)           Normania         (%,4%) features of non-construction (x)           La provide         (%,4%) features of non-construction (x)           La provide         (%,4%) features of non-construction (x)	VPN Host Instance User VPN IP Remote IP kB Received kB S
Account Switings Loge from [ tesh -	1 5-vpn1-1-vpnsrv s-vpn1-ubp patton 172.23.210.100 4 240 2 s-vpn1-1-vpnsrv s-vpn1-ubp harrynnelson 172.23.210.22 6,379 4,1 3 s-vpn1-1-vpnsrv s-vpn1-ubp mannino 172.23.210.54 4 155,553 100,
LZ VVM extrange vom dysenite anterlander Vom 2.000 knows whet is anterlande to inter	4 s-vpn1-1-vpnsrv s-vpn1-udp afan 172.23.210.222 5161 140,851 112, 5 s-vpn1-1-vpnsrv s-vpn1-top tjw 172.23.212.22 116 133,956 110, 6 s-vpn1-1-vpnsrv s-vpn1-top tjw 172.23.212.46 77804 2.507 11.
Processor of the interactional and the interaction of the CE Contine phonestock) Proceed Number(1) (Section 2) (Se	7 s-vpni-1-vpnsrv s-vpni-top dwoodwardi 172.23.212.98 51945 347,676 486, 8 s-vpni-1-vpnsrv s-vpni-top wolfs-a 172.23.213.34 149,399 169.1
Account Info and	<ul> <li>5 -vpn2-1-vpnsrv</li> <li>5 -vpn3-1-vpnsrv</li> <li>5 -vpn3-1-vpnsrv</li></ul>
self service	
NetFoldmon Band Addressed TealCost Units 1. 000000000	CLIMING A more House Queen Ann Weat Hap Manage and
National Data 2	NETGEAR memoriak
NetFoction Subscriptions NetFoction Sector Primary Deal Sector Sector Deal NetFoction NetFoction NetFoct NetFoction NetFoct Ne	
The very transport of concernment of the concernmen	Annual Montha Auron Auron Montha (2014) and 10578 Aurona Montha M
	an Lin App (methy) Faurties Line Tates Line Tates Line App
	Network Monitoring
Heres v Saacch v Reports v Articles v Assets v Tools v Admin v More v Starch Titlett Descentes Start >>> REQUEST	LibreNIMS and electiflow (old)
RT at a glance	
	ChAAn
10 highest priority tickets I own (Found 2 tickets)     # Subject     Priority Ousue     Status	Deterfies Points
1042 Password reset	A DECEMBER OF A
1117 Rec (EXT) Change of 2-step authentication settings for your	
keeneelin oog account of admin System	En la la ce El la la ce el la
10 newest unowned tickets (Found 14 tickets)     # Subject Source Status Created	
1165 New release (s) available for lazzeplin.org	
adrin	
	Construction     C
e → σ O ∂ tracijetovi kaopinepitoki/	the test of t
PerTS WNAK Toolid on performar Jucceptin.org	Log Int - Consequences - Tendo
Perfsonar.luxzeplin.org at 151.150.227.167     ✓ce	interfaces Details v
Sites: L2 at Santosi Underground Research Facility Addenses: Levis, 35 3774-V35 https: Addentisational: L2 Control Sites Status Address editored Research Research	Printyry Interface #80 NTP Spriced Nes
	Gobuly Jugitizend Nes Todo Non Method
Standods New Minkey Dig C Standod Status Velsion Polis	Access Folicy Public Access Folicy Public
sensed - Running 4.46.1.87	Wisai Aucline Nec
10000 1000 1000 1000 1000 1000 1000 10	NM M GB
publicadar - publicadar -	All Privacy Policy (2
PerfSONAR	emand centre tools
Test Results 12 Reputs	across Cf
Sends	
* SOLINCE I DESTINATION THROUGHPUT LATENCY (MS) LOSS	R Other services Global mode directory/07
performativezepin.org Klgot0-bergrid-hep-phila.ac.uk +3.87 Geps + #Aa + #Aa	

- 55.3 - 6.019% - 5/4 - 6.025%

# LZ Network Layer 3 sketch

