Hydra Infrastructure Management Guide

Student Container Platform Administration

Computer Science Department
SUNY New Paltz

Last Updated: January 2025

Contents

1 System O er iew

Hydra is a contain riz d d v lopm nt platform providing p rsist nt d v lopm nt nvironm nts for Comput r Sci nc stud nts and faculty at SUNY N w Paltz. Th syst m us s SAML 2.0 Singl Sign-On via Azur AD and Dock r for contain r orch stration across a 3-nod clust r.

1.1 Key Features

- **SSO Authentication:** Azur AD SAML 2.0 with automatic us r provisioning
- **Persistent Containers:** On d v lopm nt nvironm nt p r stud nt with data p rsist nc
- **Built-in Services:** VS Cod (cod -s rv r), Jupyt r Not book, Dock r-in-Dock r
- **SSH Access:** Dir ct SSH acc ss to contain rs via assign d ports
- **GPU Computing:** Acc ss to NVIDIA GPUs on Chim ra and C rb rus nod s
- **Dynamic Routing:** Tra fik-bas d routing for custom w b applications
- **Resource Management:** Tim -limit d r sourc allocations with automatic xpiry
- **Integration:** Op nW bUI (GPT) and n8n account manag m nt

1.2 Access URLs

Service	URL	Description
Dashboard	`https://hydr .newp ltz.edu/d shbo rd`	Main us r int rfac
Op nW bUI	`https://gpt.hydr .newp ltz.edu/`	AI chat int rfac
VS Cod	`https://hydr .newp ltz.edu/students/{user}/vscode`	Brows r IDE
Jupyt r	`https://hydr .newp ltz.edu/students/{user}/jupyter`	Not books
S rv rs	`https://hydr .newp ltz.edu/servers`	Clust r status

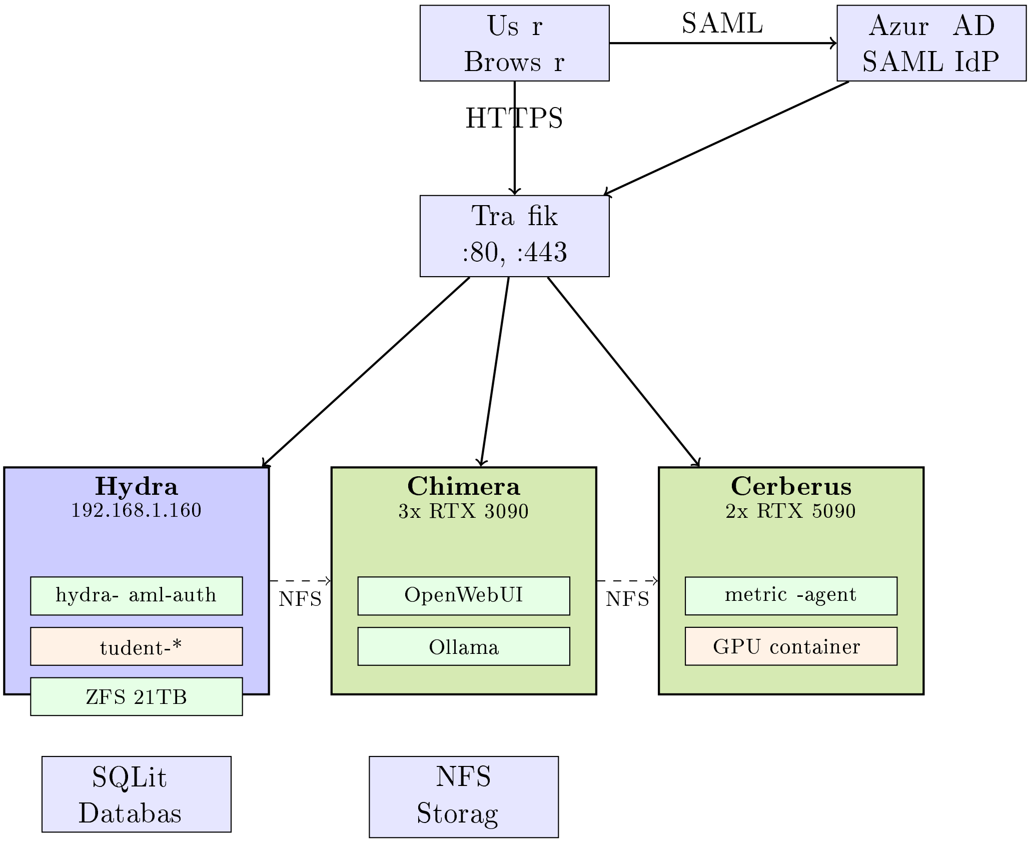
2 Cluster Architecture

Th Hydra platform op rat s across a 3-nod clust r, ach with sp cializ d rol s.

2.1 Cluster Nodes

Node	IP	Role	GPU	Description
Hydra	192.168.1.160	Control Plan	Non	Main s rv r, ZFS storag , stud nt contain rs
Chim ra	192.168.1.150	Inf r nc	3x RTX 3090	Op nW bUI, inf r nc workloads
C rb rus	192.168.1.242	Training	2x RTX 5090	Stud nt GPU training

2.2 Architecture Diagram



2.3 Storage Configuration

Node	Storage	Capacity	Purpose
Hydra	ZFS RAID-10	21 TB	Stud nt volum s, primary data
Hydra	S agat SSD	1.1 TB	Daily backups
Chim ra	NFS mount	–	Stud nt data acc ss
C rb rus	NFS mount	–	Stud nt data acc ss

2.4 Component Overview

Component	Port	Description
Tra fik	80, 443	R v rs proxy, TLS t rmination, routing
hydra-saml-auth	6969	SAML auth, dashboard, contain r manag m nt
Op nW bUI	3000	AI chat int rfac (Ollama front nd)
Ollama	11434	LLM inf r nc ngin
m trics-ag nt	9100	Nod m trics coll ction (GPU nod s)
Stud nt Contain rs	Dynamic	P r-us r d v lopm nt nvironm nts

2.5 Network Architecture

Stud nt contain rs op rat on an isolat d Dock r n twork (`hydr _students_net`) with:

- No dir ct int rn t acc ss (configurabl)
- Int rnal DNS r solution
- Tra fik-m diat d xt rnal acc ss via ForwardAuth
- Cross-nod NFS acc ss for GPU workloads

3 Authentication System

3.1 SAML 2.0 SSO Flow

1. User visits https://hydra.newpaltz.edu/login

2. Hydra redirects to Azure AD with SAML AuthnRequest

3. User authenticates with New Paltz credentials

4. Azure AD returns signed SAML assertion

5. Hydra validates signature, extracts: email, groups, displayName

6. Session created, JWT cookie issued

7. User redirected to /dashboard

3.2 Session Management

S ssions ar manag d via:

- **Express Session:** S rv r-sid s ssion storag in SQLit
- **JWT Cookie:** Sit -wid auth ntication cooki for cross-s rvic SSO
- **JWKS Endpoint:** Public k y ndpoint for JWT v rification by oth r s rvic s

JWT Configuration:

- TTL: Configurabl via `JWT_TTL_SECONDS` (d fault: 86400)
- Algorithm: RS256

- Cooki Domain: `.newp ltz.edu`

4 Container System

4.1 Student Container Features

Each stud nt r c iv s a singl p rsist nt contain r with:

Feature	Details
Nod .js	Lat st LTS via nvm
Python	3.11+ with pip, v nv, Jupyt r
Java	Op nJDK 21
Dock r	Full Dock r-in-Dock r support (privil g d mod)
VS Cod	cod -s rv r brows r IDE with xt nsions
Jupyt r	Not book and Jupyt rLab
SSH	Dir ct SSH acc ss via assign d port
Tools	Git, curl, wg t, build- ss ntial, tc.

4.2 SSH Access

Stud nts can acc ss th ir contain rs via SSH from any t rminal:

```
# Connect to your container
ssh -p <assigned_port> student@hydra.newpaltz.edu

# Example with port 2222
ssh -p 2222 student@hydra.newpaltz.edu
```

- SSH ports ar dynamically assign d from rang 2200-2299
- Password is display d in th dashboard aft r contain r cr ation
- SSH supports k y-bas d auth ntication (add k ys to `~/.ssh/ uthorized_keys`)

SSH Setup: Each contain r runs an SSH s rv r via sup rvisord. Th SSH port and password ar shown in th dashboard und r "SSH Acc ss" s ction.

4.3 Resource Presets

Preset	RAM	CPU	Storage	GPU	Approval
Minimal	256 MB	0.5	5 GB	0	Auto
Cons rvativ	512 MB	1	10 GB	0	Auto
Standard	1 GB	1	20 GB	0	Auto
Enhanc d	2 GB	2	40 GB	0	R quir d
GPU Inf r nc	32 GB	8	100 GB	1	R quir d
GPU Training	48 GB	16	200 GB	2	R quir d

GPU Access: GPU pr s ts run on Chim ra (inf r nc) or C rb rus (training). Chim ra GPUs ar shar d with Op nW bUI; us C rb rus for d dicat d GPU work.

5 Resource Management

5.1 Time-Limited Allocations

R sourc allocations ar tim -limit d to nsur fair sharing among stud nts:

Duration	Approval	Use Case
1 Day (D fault)	Auto	Quick t sting
3 Days	Auto	Short assignm nt
1 W k	Auto	Short proj cts
2 W ks	Auto	Standard proj cts
1 Month	Auto	S m st r proj ct
2 Months	R quir d	Ext nd d proj ct
3 Months	R quir d	Full s m st r

5.2 Resource Expiry

Wh n a r sourc allocation xpir s:

1. Configuration r s ts to **minimal** pr s t
2. Contain r mov s back to **Hydra** nod
3. Contain r automatically **restarts** to apply n w limits
4. Stud nt r c iv s notification (if mail configur d)

Th xpiry ch ck r runs hourly via th `resource-expiry` s rvic .

5.3 Requesting Additional Resources

1. Navigat to Dashboard ! Configur R sourc s
2. S l ct d sir d pr s t, nod , and duration
3. Submit r qu st (auto-approv d if within thr sholds)
4. If approval r quir d, admin r vi ws within 7 days

Auto-Approval Thresholds (Hydra only):

- M mory: up to 2 GB
- CPU: up to 2 cor s
- Storag : up to 20 GB

6 GPU Computing

6.1 GPU Node Configuration

Node	GPUs	Model	VRAM	Primary Use
Chim ra	3	RTX 3090	72 GB total	Op nW bUI/Inf r nc
C rb rus	2	RTX 5090	64 GB total	Stud nt training

6.2 GPU Access Guidelines

- **Cerberus (Recommended):** Us for GPU training and stud nt proj cts. RTX 5090 off rs n w r archit ctur with 32GB VRAM p r card.
- **Chimera:** R s rv d for Op nW bUI inf r nc . 1 GPU is r s rv d for Ollama. Only us if C rb rus is unavailabl .

6.3 Requesting GPU Access

1. S l ct "GPU Training" pr s t in Configur R sourc s
2. Choos C rb rus as targ t nod
3. Provid justification for GPU acc ss
4. Wait for admin approval

7 Backup System

7.1 Daily Cluster Backups

All clust r nod s ar back d up daily at 1:00 AM to th S agat driv on Hydra.

Setting	Value
Backup Location	`/mnt/sdh4/b ckups/`
Sch dul	Daily at 1:00 AM
M thod	rsync with compr ssion
Log Fil	`/v r/log/cluster-b ckup.log`

7.2 What Gets Backed Up

- **Hydra:** Full OS, configuration, application cod (xclud s Dock r volum s)
- **Chimera:** Full OS, NVIDIA driv rs, Op nW bUI config
- **Cerberus:** Full OS, NVIDIA driv rs, m trics ag nt

7.3 Backup Exclusions

Th following ar   xclud d from backups:

- `/dev/*`, `/proc/*`, `/sys/*`, `/run/*`
- `/tmp/*`, `/v r/tmp/*`, `/v r/c che/*`

- /mnt/*, /medi /*, /lost+found
- /v r/lib/docker/* (Dock r data)

7.4 Manual Backup

```
# Run backup manually
sudo /home/infra/backup-cluster.sh

# Check backup status
cat /var/log/cluster-backup.log

# View backup sizes
du -sh /mnt/sdh4/backups/*
```

8 File Structure

```
hydra-saml-auth/
|-- index.js              # Main entry: SAML, JWT/JWKS, routes,
   WebSocket
|-- routes/
| ||-Wa|-toWa-49(s)n-11(.)-js122(j)-0(s)-982(#)-6(R)i-9(e)-9(s)-94(r)816(t)-9(r)-94(c)-592(u)-9r(i)-9e(u)-
|   |-- resource-requests.js  # Resource allocation requests
|   |-- webui-api.js      # OpenWebUI account proxy
|   |-- n8n-api.js        # n8n account management
|   |-- servers-api.js    # Cluster status endpoints
|   |-- admin.js          # Admin panel routes
|-- services/
|   |-- db-init.js        # Database initialization and migrations
|   |-- resource-expiry.js    # Time-limited resource expiry checker
|   |-- activity-logger.js    # Activity tracking
|   |-- email-notifications.js # Email alerts
|   |-- 45(a)-5(i)0.925u2-(e)-81(r)-82(s)-49(a)-9oe a n t-49(s)-146(.)-12(j)-0(s)-1105N(e)-84(s)-
```

```
docker ps --filter "name=student-"
```

9.2 Access Container Shell

```
docker exec -it student-<username> /bin/bash
```

9.3 View Container Logs

```
docker logs -f student-<username> --tail=100
```

9.4 Restart a Container

```
docker restart student-<username>
```

9.5 Remove a Stuck Container

```
docker rm -f student-<username>
```

9.6 Rebuild Student Container Image

```
cd student-container
docker build -t hydra-student-container:latest .
```

Note: Stud nts with xisting contain rs must r cr at th m to us updat d imag s.

9.7 Check Cluster Node Status

```
# Check metrics from Chimera
curl http://192.168.1.150:9100/metrics

# Check metrics from Cerberus
curl http://192.168.1.242:9100/metrics
```

9.8 Trigger Resource Expiry Check

```
# From within the application
curl http://localhost:6969/api/admin/resource-expiry/check
```

10 Ser ice Management

10.1 Restart Main Service

```
docker compose restart hydra-saml-auth
```

10.2 Rebuild and Redeploy

```
docker compose build hydra-saml-auth
docker compose up -d hydra-saml-auth
```

10.3 View Service Logs

```
docker compose logs -f hydra-saml-auth
```

10.4 Check Traefik Routing

```
docker compose logs traefik | grep -i error
curl -I https://hydra.newpaltz.edu/
```

10.5 Manage Metrics Agent (GPU Nodes)

```
# On Chimera or Cerberus
sudo systemctl status metrics-agent
sudo systemctl restart metrics-agent
sudo journalctl -u metrics-agent -f
```

11 Troubleshooting

11.1 Authentication Issues

Symptom	**Solution**
SAML ass rtion invalid	V rify `METADATA_URL` and `SAML_SP_ENTITY_ID` match Azur config xactly
Cooki not s t	Ch ck `COOKIE_DOMAIN`, nsur HTTPS, ch ck brows r s ttings
JWT v rification fails	V rify JWKS ndpoint acc ssibl , ch ck k y rotation

11.2 Container Issues

Symptom	**Solution**
Contain r won't initializ	V rify `hydr -student-cont iner:l test` imag xists
Contain r 404	Ch ck contain r is on `hydr _students_net`, Tra fik running
S rvic won't start	Ch ck sup rvisord logs insid contain r
Port routing fails	V rify port not r s rv d (8443, 8888) and not in us
SSH not working	Ch ck sshd proc ss in contain r, v rify port assignm nt

11.3 GPU Issues

Symptom	Solution
GPU not d t ct d	Run `nvidi -smi` on host, ch ck NVIDIA driv rs
GPU contain r fails	V rify nvidia-contain r-toolkit install d
M trics not showing	Ch ck m trics-ag nt s rvic , fir wall port 9100

11.4 Service-Specific Issues

- **VS Code not loading:** Ch ck cod -s rv r proc ss, ForwardAuth working
- **Jupyter issues:** V rify `NotebookApp.b se_url` s tting
- **Docker-in-Docker fails:** Contain r must hav privil g d mod
- **Files not persisting:** Only `/home/student/` is p rsist d
- **Resource expiry not working:** Ch ck r sourc - xpiry s rvic logs

12 Ansible Deployment

12.1 Cluster Setup Overview

Th clust r can b d ploy d using Ansibl playbooks in `nsible/` dir ctory:

```
# Full cluster deployment
cd ansible
ansible-playbook -i inventory.yml playbooks/site.yml
```

12.2 Playbook Execution Order

1. `00-preflight-b ckup.yml` - Cr at backups b for chang s
2. `01-prep re-nodes.yml` - Install packag s, configur k rn l
3. `02-rke2-server.yml` - S tup RKE2 control plan
4. `03-rke2- gents.yml` - Join GPU nod s to clust r
5. `04-gpu-setup.yml` - Configur NVIDIA driv rs and GPU Op rator
6. `05-deploy-hydr .yml` - D ploy Hydra application stack

12.3 Inventory Configuration

Th clust r inv ntory is d fin d in `nsible/inventory.yml`:

```
# Key variables
rke2_version: "v1.28.4+rke2r1"
cluster_domain: hydra.newpaltz.edu
nfs_server: "192.168.1.160"
nfs_path: "/srv/hydra-nfs"
```

13 En ironment Configuration

13.1 Required Variables

Variable	Description
BASE_URL	Ext rnal URL (https://hydra.n wpaltz. du)
METADATA_URL	Azur AD f d ration m tadata URL
SAML_SP_ENTITY_ID	SP Entity ID (must match Azur xactly)
COOKIE_DOMAIN	Cooki scop (.n wpaltz. du)
PORT	S rvic port (d fault: 6969)
DB_PATH	Databas path (/app/data/w bui.db)

13.2 Optional Variables

Variable	Description
PUBLIC_STUDENTS_BASE	Stud nt URL bas
JWT_TTL_SECONDS	JWT tok n lif tim
CHIMERA_HOST	Chim ra IP (d fault: 192.168.1.150)
CERBERUS_HOST	C rb rus IP (d fault: 192.168.1.242)
STUDENT_IMAGE	D fault contain r imag
GPU_STUDENT_IMAGE	GPU contain r imag

14 Monitoring

14.1 Servers Dashboard

Th `/servers` pag displays r al-tim m trics for all clust r nod s:

- CPU usag and load av rag
- M mory usag
- Disk usag and ZFS pool status
- Contain r count
- GPU utilization (Chim ra/C rb rus)
- GPU t mp ratur and VRAM usag

14.2 Metrics Collection

Node	Agent	Port
Hydra	m trics-coll ctor.js (int rnal)	N/A
Chim ra	m trics-ag nt.py	9100
C rb rus	m trics-ag nt.py	9100

15 References

- Dock r Docum ntation: https://docs.docker.com/

- Tra fik Docum ntation: https://doc.tr efik.io/tr efik/
- SAML 2.0 Sp cification: https://docs.o sis-open.org/security/s ml/v2.0/
- Azur AD SAML: https://docs.microsoft.com/en-us/ zure/ ctive-directory/develop/single-sign-on-s ml-protocol
- cod -s rv r: https://coder.com/docs/code-server/l test
- Jupyt r: https://jupyter.org/document tion
- RKE2 Docum ntation: https://docs.rke2.io/
- NVIDIA GPU Op rator: https://docs.nvidi .com/d t center/cloud-n tive/gpu-oper tor/