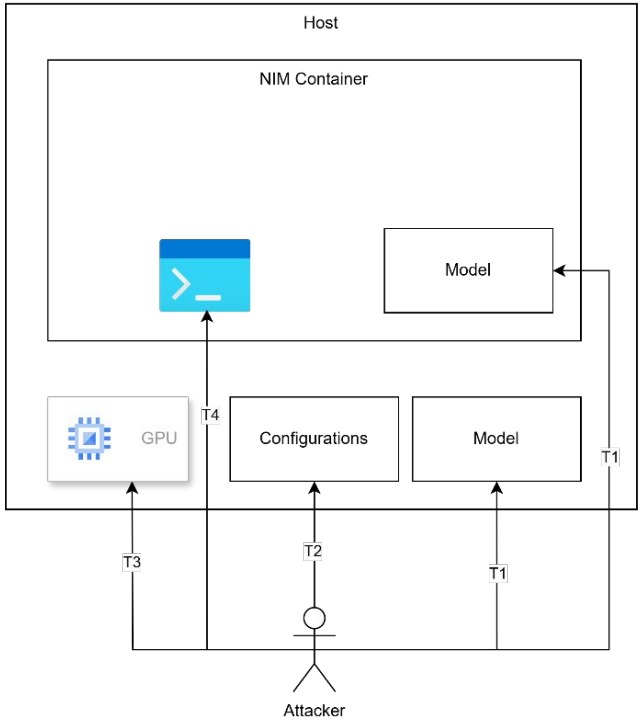
Securing NVIDIA NIM

AccuKnox Security

Agenda

- Risks with NIM microservices
  - Exposed models
  - Exposed configuration
  - GPU/CUDA compromise
- Threat Model
- Observability
  - Process/Network
- Security Hardening
  - Container Hardening
  - Host Hardening

Risks with NIM Microservices



Threats
T1: Exposed Models Unauthorized access to machine learning models that can lead to data leakage or model theft.
T2: Exposed Configurations Misconfiguration or exposure of sensitive configuration files that may contain API keys or credentials.
T3: GPU/CUDA Compromise Unauthorized access or exploitation of GPU resources for cryptojacking or malicious computations.
T4: Arbitrary Code Execution Attackers gaining the ability to execute unauthorized commands within the container or environment.

Assets
A1: Machine Learning Models
A2: Configuration Files
A3: GPU Resources

Use Cases

- Securing NIM Microservices
  - Zero-Trust
  - Restrict execution of NVIDIA Tools
  - Audit access to HPC/CUDA

- Host Security- Audit access to HPC/CUDA
- Cryptojacking
- Further Hardening

Securing NIM microservices - Process Observability

ACCUKNOX

Home > Runtime Security > App Behavior

Search anything...

Accuknox

Rahul

Search

- Dashboard
- Inventory
- Issues
- Compliance
- Runtime Protection
  - CWPP Dashboard
  - App Behavior
  - Policies
- Remediation
- Monitors / Alerts
- Identity
- Reports
- Notifications
- Settings

nvidia-nim-demo | container_namespace | Workloads

LIST | GRAPH | All

Container Name eq dazzling_benz,ecstatic_jones,focused_allen,intelligent_galileo,lucid_rubin,nifty_dijkstra

Add filter

File Observability | **Process Observability** | Network Observability

Last Updated Time	Process Accessed	Process	Cluster	Namespace	Workload	Action	Occu...	
10/04/2024 10:31 AM	/sbin/ldconfig.real	/usr/bin/bash	nvidia-nim-demo	container_namespace	ecstatic_jones	Allow	1	Det
10/04/2024 10:31 AM	/usr/bin/ls	/usr/bin/bash	nvidia-nim-demo	container_namespace	ecstatic_jones	Allow	12	Det
10/04/2024 10:31 AM	/usr/sbin/ldconfig	/usr/bin/bash	nvidia-nim-demo	container_namespace	ecstatic_jones	Allow	1	Det
10/04/2024 10:31 AM	/opt/nim/llm/.venv/bin/nim-llm-set-cache-env	/usr/bin/bash	nvidia-nim-demo	container_namespace	ecstatic_jones	Allow	1	Det
10/04/2024 10:29 AM	/usr/bin/nvidia-smi	/usr/bin/containerd-shim-runc-	nvidia-nim-demo	container_namespace	intelligent_galileo	Allow	1	Det
10/04/2024 10:27 AM	/usr/bin/ls	/usr/bin/bash	nvidia-nim-demo	container_namespace	nifty_dijkstra	Allow	12	Det
10/04/2024 10:27 AM	/opt/nim/llm/.venv/bin/nim-llm-set-cache-env	/usr/bin/bash	nvidia-nim-demo	container_namespace	nifty_dijkstra	Allow	1	Det
10/04/2024 10:27 AM	/usr/bin/sed	/usr/bin/bash	nvidia-nim-demo	container_namespace	nifty_dijkstra	Allow	2	Det
10/04/2024 10:27 AM	/usr/sbin/ldconfig	/usr/bin/bash	nvidia-nim-demo	container_namespace	nifty_dijkstra	Allow	1	Det

21 - 30 of 44

1 2 3 4 5

Securing NIM microservices - Network Observability

ACCUKNOX

Home > Runtime Security > App Behavior

Search anything...

Accuknox

Rudraksh

- Dashboard
- Inventory
- Issues
- Compliance
- Runtime Protection
  - CWPP Dashboard
  - App Behavior
  - Policies
- Remediation
- Monitors / Alerts
- Identity
- Reports
- Notifications
- Settings

nvidia-nim-demo | container_namespace | deployment/nvidia-nim

LIST | GRAPH | All

Add filter

Select Connections

File Observability | Process Observability | **Network Observability**

Last Updated Time	Network Flow	Source Command	Dest. POD/SVC/IP	Port	Cluster	Namespace	Workload	Action	Occu...	
10/07/2024 19:41 PM	egress	/opt/nim/llm/.venv/lib/p	127.0.0.1	49010	nvidia-nim-demo	container_namespace	nvidia-nim	Allow	13	Detc
10/07/2024 19:41 PM	egress	/usr/bin/python3.10	127.0.0.1	49010	nvidia-nim-demo	container_namespace	nvidia-nim	Allow	238	Detc
10/07/2024 19:41 PM	ingress	/usr/bin/python3.10	172.17.0.1	8000	nvidia-nim-demo	container_namespace	nvidia-nim	Allow	7	Detc
10/07/2024 19:40 PM	egress	/opt/nim/llm/.venv/lib/p	127.0.0.1	49010	nvidia-nim-demo	container_namespace	nvidia-nim	Allow	13	Detc

1 - 4 of 4

1

Securing NIM microservices - Zero Trust

ACCUKNOX

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorPolicy
metadata:
  name: autopol-system-1123564177
  namespace: container_namespace
spec:
  action: Allow
  selector:
        matchLabels:
          app: nvidia-nim
  file:
        matchDirectories:
        - dir: /opt/nim/llm/vllm_nvext/
          fromSource:
          - path: /opt/nim/llm/.venv/bin/nim-llm-set-cache-env
          - path: /opt/nim/llm/.venv/bin/nim-llm-check-cache-env
          - path: /opt/nim/llm/.venv/bin/python3
  network:
        matchProtocols:
        - fromSource:
          - path: /usr/bin/python3.10
            protocol: raw
        - fromSource:
          - path: /usr/bin/python3.10
            protocol: tcp
  process:
        matchDirectories:
        - dir: /opt/nim/llm/.venv/bin/
          recursive: true
        matchPaths:
        - path: /opt/nim/start-server.sh
        - path: /opt/nvidia/nvidia_entrypoint.sh
        - path: /usr/bin/python3.10
  severity: 1
```

Only allow Python and NIM binaries to access LLMs

Only allow Python applications to access network

Allow execution of NIM binaries only

*Only important paths shown for ref. Actual policy would be more granular

Securing NIM microservices - Zero Trust

ACCUKNOX

```
[ec2-user@ip-172-31-29-79 ~]$ docker exec -it nvidia-nim bash
exec /usr/bin/bash: permission denied
[ec2-user@ip-172-31-29-79 ~]$
```

Shell access denied

Informational October 7, 2024 at 21:36 a minute ago

kubearmor syslog

Policy Name	Resource	Source	Action	Result
DefaultPosture	/usr/bin/bash	/usr/sbin/runc	Block	Permission denied

Operation	Cluster Name	Pod Name	Workload Name	Workload Type
Process	nvidia-nim-demo	nvidia-nim	nvidia-nim	Deployment

Detailed information in Alerts

Raw Logs


```
{
  "Action" : "Block"
  "ClusterName" : "nvidia-nim-demo"
  "ContainerID" : "f16a8fa13422eb68ac83552fee793bb19da6c5e93a71d7f5226e6c4c70e303e3"
  "ContainerImage" : "nvcr.io/nim/mistralai/mistral-7b-instruct-v0.3:latest"
  "ContainerName" : "nvidia-nim"
  "Cwd" : "/"
  "Data" : "lsm=SECURITY_BPRM_CHECK"
  "Enforcer" : "BPFLSM"
  "HostName" : "ip-172-31-29-79.ec2.internal"
  "HostPID" : 40996
  "HostPPID" : 40990
  "Labels" : "maintainer=NVIDIA CORPORATION <cudatools@nvidia.com>,com.nvidia.nim.mo..."
  "NamespaceName" : "container_namespace"
  "Operation" : "Process"
  "Owner" : {
    "Name" : "nvidia-nim"
    "Namespace" : "container_namespace"
    "Ref" : "Deployment"
  }
  "PID" : 1380
  "PPID" : 40990
  "ParentProcessName" : "/usr/sbin/runc"
  "PodName" : "nvidia-nim"
  "PolicyName" : "DefaultPosture"
  "ProcessName" : "/usr/bin/bash"
  "Resource" : "/usr/bin/bash"
  "Result" : "Permission denied"
```


Inference continues as expected

```
[ec2-user@ip-172-31-29-79 ~]$ curl --silent -X 'POST' 'http://0.0.0.0:8000/v1/chat/completions' -H 'accept: application/json' -H 'Content-Type: applica
tion/json' -d '{
    "model": "mistralai/mistral-7b-instruct-v0.3",
    "messages": [{"role":"user", "content":"Write a limerick about the wonders of GPU computing."}],
    "max_tokens": 64
}' | jq
{
  "id": "chat-dcb568242eda404e913093bfe39f4825",
  "object": "chat.completion",
  "created": 1728317219,
  "model": "mistralai/mistral-7b-instruct-v0.3",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": " In halls where circuits hum and gleam,\n\nGPU dreams awake in every stream.\nWith powers vast to render so bright,\nIt brings the
impossible to light,\nIn digital art or games, it's the scheme.\n\nWith billions of tiny minds at play,\nIt learn"
      },
      "logprobs": null,
      "finish_reason": "length",
      "stop_reason": null
    }
  ],
```

Securing NIM microservices - NVIDIA Tools

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorPolicy
metadata:
  name: ksp-block-nvidia-tools-exec
  namespace: container_namespace
spec:
  action: Block
  selector:
        matchLabels:
          app: nvidia-nim
  process:
        matchPaths:
        - path: /usr/bin/nvidia-ctk
        - path: /usr/bin/nvidia-modprobe
        - path: /usr/bin/nvidia-smi
        - path: /usr/bin/nvidia-cuda-mps-control
        - path: /usr/bin/nvidia-cuda-mps-server
        - path: /usr/bin/nvidia-debugdump
        - path: /usr/bin/nvidia-persistenced
        - path: /usr/bin/nvidia-container-cli
        - path: /usr/bin/nvidia-container-runtime-hook
  message: "ALERT! Blocked execution of NVIDIA container runtime
tools inside container."
  severity: 1
```

Tools in container which are not used during inference

```
nim@f16a8fa13422:/$ nvidia-
nvidia-cuda-mps-control  nvidia-cuda-mps-server   nvidia-debugdump         nvidia-persistenced      nvidia-smi
nim@f16a8fa13422:/$ nvidia-
```

Securing NIM microservices - NVIDIA Tools

Critical October 7, 2024 at 20:40 9 minutes ago

ALERT! Blocked execution of NVIDIA container runtime tools inside container.

Policy Name	Resource	Source	Action	Severity
ksp-block-nvidia-tools-ex...	/usr/bin/nvidia-cuda-mps...	/usr/bin/bash	Block	1

Operation	Cluster Name	Pod Name	Workload Name	Workload Type
Process	nvidia-nim-demo	nvidia-nim	nvidia-nim	Deployment

Raw Logs Copy


```
{
    "Action" : "Block"
    "ClusterName" : "nvidia-nim-demo"
    "ContainerID" : "f16a8fa13422eb68ac83552fee793bb19da6c5e93a71d7f5226e6c4c70e303e3"
    "ContainerImage" : "nvcr.io/nim/mistralai/mistral-7b-instruct-v0.3:latest"
    "ContainerName" : "nvidia-nim"
    "Data" : "lsm=SECURITY_BPRM_CHECK"
    "Enforcer" : "BPFLSM"
    "HashID" : "77bce3626d485dda07245ec402c6c04a1857b8ad8de33e3bc9ae7b66a849a337"
    "HostName" : "ip-172-31-29-79.ec2.internal"
    "HostPID" : 35346
    "HostPPID" : 33988
    "Labels" : "app=nvidia-nim com.nvidia.nim.model-nspect=NSPECT-YDAW-FMDD com.nvidia..."
    "Message" : "ALERT! Blocked execution of NVIDIA container runtime tools inside cont..."
    "NamespaceName" : "container_namespace"
    "Operation" : "Process"
    "Owner" : {
        "Name" : "nvidia-nim"
        "Namespace" : "container_namespace"
        "Ref" : "Deployment"
    }
    "PID" : 1358
    "PPID" : 1347
```


```
nim@f16a8fa13422:/$ nvidia-cuda-mps-control --help
bash: /usr/bin/nvidia-cuda-mps-control: Permission denied
nim@f16a8fa13422:/$
```

Securing NIM microservices - HPC/CUDA

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorPolicy
metadata:
  name: ksp-audit-hpc-libs-access
  namespace: container_namespace
spec:
  action: Audit
  selector:
        matchLabels:
          app: nvidia-nim
  file:
        matchDirectories:
         - dir: /opt/hpcx/
           recursive: true
         - dir: /usr/local/cuda/
           recursive: true
         - dir: /usr/local/cuda-12.3
           recursive: true
        matchPaths:
         - path: /usr/lib/x86_64-linux-gnu/libcuda.so.560.35.03
         - path: /usr/lib/x86_64-linux-gnu/libcudadebugger.so.560.35.03
         - path: /usr/lib/x86_64-linux-gnu/libcuda.so.545.23.08
         - path: /usr/lib/x86_64-linux-gnu/libcudadebugger.so.545.23.08
  message: WARN! Access to NVIDIA HPC/CUDA libs detected.
  severity: 1
```

Medium October 7, 2024 at 21:00 3 minutes ago

WARN! Access to NVIDIA HPC/CUDA libs detected.

Policy Name	Resource	Source	Action	Severity
ksp-audit-hpc-libs-access	/opt/hpcx/ucx/lib/libuct.so.0.0.0	/opt/hpcx/ucx/bin/ucx_perftest	Audit	5

Operation	Cluster Name	Pod Name	Workload Name	Workload Type
File	nvidia-nim-demo	nvidia-nim	nvidia-nim	Deployment

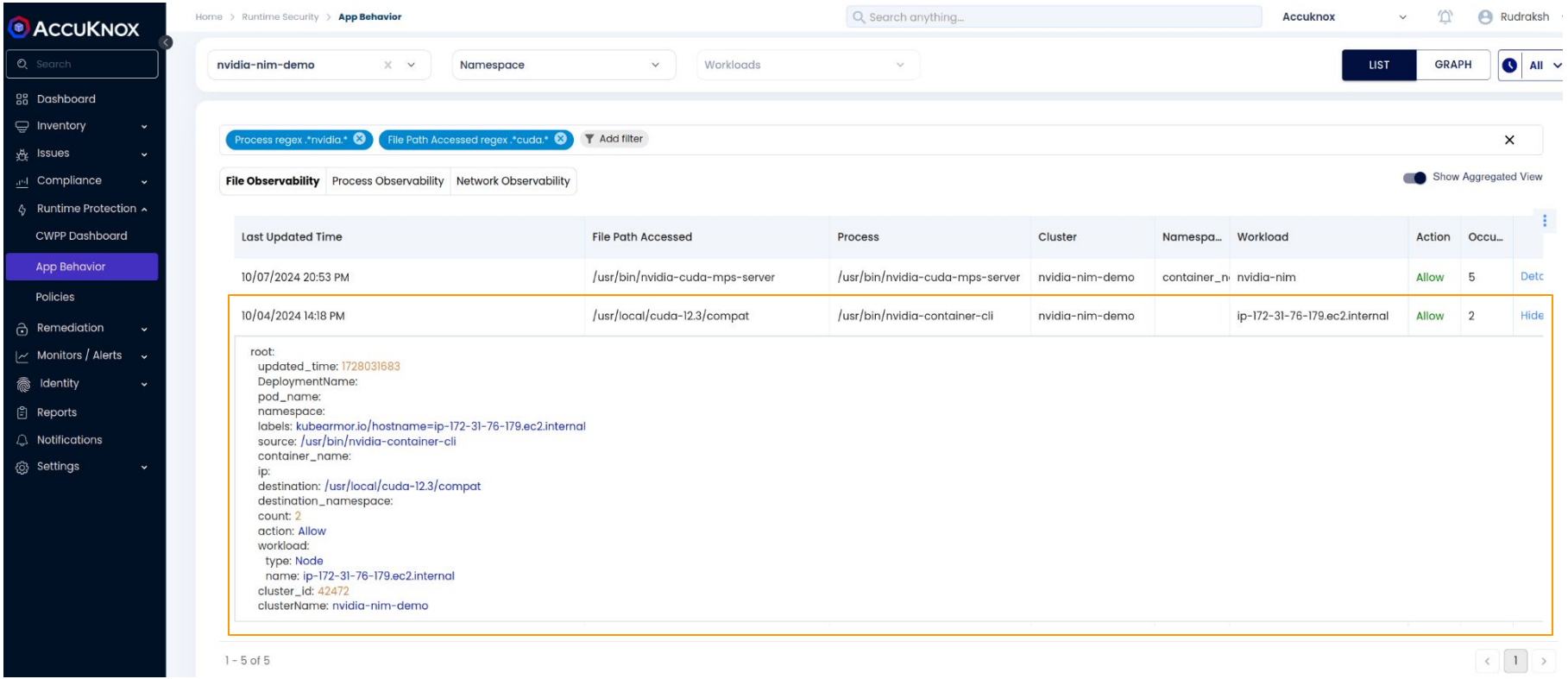
Raw Logs Copy


```
{
  "Action" : "Audit"
  "ClusterName" : "nvidia-nim-demo"
  "ContainerID" : "f16a8fa13422eb68ac83552fee793bb19da6c5e93a71d7f5226e6c4c70e303e3"
  "ContainerImage" : "nvcr.io/nim/mistralai/mistral-7b-instruct-v0.3:latest"
  "ContainerName" : "nvidia-nim"
  "Data" : "syscall=SYS_OPENAT fd=-100 flags=O_RDONLY|O_CLOEXEC"
  "Enforcer" : "eBPF Monitor"
  "HashID" : "d3921da5b98eb3f510af3240254ddb0d0dd8c40516c872e7e0afceec3a37490a"
  "HostName" : "ip-172-31-29-79.ec2.internal"
  "HostPID" : 37299
  "HostPPID" : 33988
  "Labels" : "app=nvidia-nim com.nvidia.nim.model-nspect=NSPECT-YDAW-FMDD com.nvidia..."
  "Message" : "WARN! Access to NVIDIA HPC/CUDA libs detected."
  "NamespaceName" : "container_namespace"
  "Operation" : "File"
  "Owner" : {
    "Name" : "nvidia-nim"
    "Namespace" : "container_namespace"
    "Ref" : "Deployment"
  }
  "PID" : 1365
  "PPID" : 1347
```


Audit HPC/CUDA library access in container

Host Security - Observability

```
docker run --gpus all -d --shm-size=16GB --labels app=nvidia-nim --name nvidia-nim -e NGC_API_KEY -v
"$LOCAL_NIM_CACHE:/opt/nim/.cache" -u $(id -u) -p 8000:8000 nvcr.io/nim/mistralai/mistral-7b-instruct-v0.3:latest
```



Host Security - HPC/CUDA

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorHostPolicy
metadata:
  name: hsp-protect-hpc-libs-access
spec:
  action: Allow
  nodeSelector:
        matchLabels:
          kubearmor.io/hostname: '*'
  file:
        matchDirectories:
         - dir: /
           recursive: true
         - dir: /usr/lib/firmware/nvidia/
           recursive: true
           fromSource:
           - path: /usr/bin/nvidia-container-cli
         - dir: /usr/local/cuda-12.6/
           recursive: true
           fromSource:
           - path: /usr/bin/nvidia-container-cli
         - dir: /usr/src/nvidia-560.35.03/
           recursive: true
           action: Audit
        matchPaths:
         - path: /etc/nvidia-container-runtime/config.toml
           fromSouce:
         - path: /usr/bin/nvidia-container-runtime-hook
  message: "ALERT! Blocked unexpected access to NVIDIA libs on host."
  severity: 6
```

Only nvidia-container-cli can access CUDA Libs

```
[ec2-user@ip-172-31-29-79 ~]$ head /usr/local/cuda-12.6/src/fortran.h
head: cannot open '/usr/local/cuda-12.6/src/fortran.h' for reading: Permission denied
[ec2-user@ip-172-31-29-79 ~]$ nvidia-container-cli --help
Usage: nvidia-container-cli [OPTION...] COMMAND [ARG...]
Command line utility for manipulating NVIDIA GPU containers.

 Commands:
  configure                  Configure a container with GPU support
  info                       Report information about the driver and devices
  list                       List driver components
```

Medium October 7, 2024 at 21:19 a minute ago

ALERT! Blocked unexpected access to NVIDIA libs on host.

Policy Name	Resource	Source	Action	Severity
hsp-protect-hpc-libs-access	/usr/local/cuda-12.6/src/fortran.h	/usr/bin/head /usr/local/cuda-1...	Block	6

Operation	Cluster Name	Pod Name	Workload Name	Workload Type
File	nvidia-nim-demo	--	ip-172-31-29-79.ec2.internal	Node

Cryptojacking - Prominent Attack On GPUs

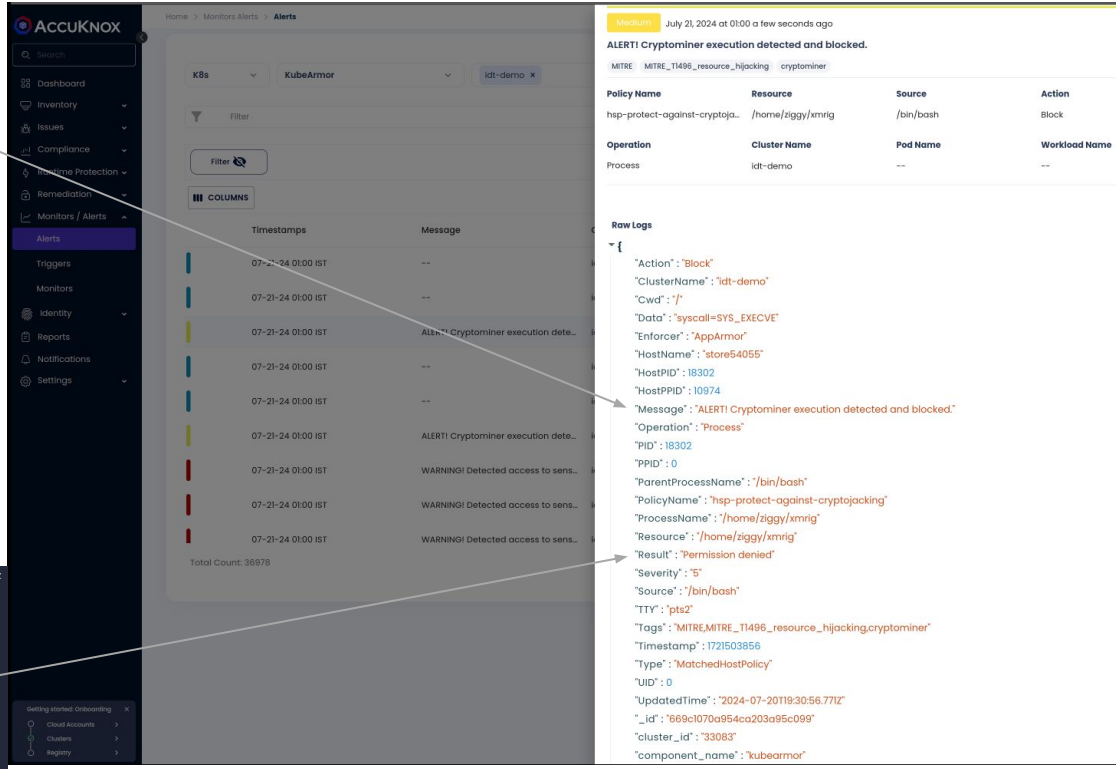
ACCUKNOX

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorHostPolicy
metadata:
  name: hsp-protect-against-cryptojacking
spec:
  action: Block
  message: "ALERT! Cryptominer execution detected and blocked."
  process:
    matchDirectories:
    - dir: /tmp/
      recursive: true
    matchPaths:
    - execname: apk
    - execname: apt
    - execname: dero-miner-linux-amd64
    - execname: dero-wallet-cli-linux-amd64
    - execname: dero
    - execname: derod-linux-amd64
    - execname: masscan
    - execname: nmap
    - execname: ntpdate
    - execname: xmrig
    - execname: zgrab2
  nodeSelector:
    matchLabels:
      kubearmor.io/hostname:   '*'
  severity: 5
  tags:
  - MITRE
  - MITRE_T1496_resource_hijacking
  - cryptominer
```

```
root@store54055:/tmp# curl -LO https://github.com/xmrig/xmrig/releases/download/v6.21.3/xmrig-6.21.3-linux-static-x64.tar.gz

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 3423k  100 3423k    0     0   9.9M      0 --:--:-- --:--:-- --:--:--  9.9M
root@store54055:/tmp# tar -xvzf xmrig-6.21.3-linux-static-x64.tar.gz

xmrig-6.21.3/
xmrig-6.21.3/config.json
xmrig-6.21.3/SHA256SUMS
xmrig-6.21.3/xmrig
root@store54055:/tmp# ./xmrig-6.21.3/xmrig
-bash: ./xmrig-6.21.3/xmrig: Permission denied
root@store54055:/tmp# cp ./xmrig-6.21.3/xmrig /home/ziggy
root@store54055:/tmp# /home/ziggy/xmrig
-bash: /home/ziggy/xmrig: Permission denied
root@store54055:/tmp#
```



Host Security - File Integrity Monitoring

ACCUKNOX

```yaml
apiVersion: security.kubearmor.com/v1
kind: KubeArmorHostPolicy
metadata:
  name: hsp-file-integrity-monitoring
spec:
  action: Block
  file:
          matchDirectories:
          - dir: /bin/
            readOnly: true
            recursive: true
          - dir: /sbin/
            readOnly: true
            recursive: true
          - dir: /usr/bin/
            readOnly: true
            recursive: true
          - dir: /usr/sbin/
            readOnly: true
            recursive: true
          - dir: /usr/lib/
            readOnly: true
            recursive: true
          - dir: /boot/
            readOnly: true
            recursive: true
  message: "ALERT! Detected and prevented compromise to file integrity"
  nodeSelector:
          matchLabels:
            kubearmor.io/hostname: '*'
  severity: 1
  tags:
  - MITRE_T1036_masquerading
  - MITRE_T1565_data_manipulation
  - NIST_800-53_AU-2
  - NIST_800-53_SI-4
```

Adversaries can't modify system binaries

Critical July 20, 2024 at 01:15 3 minutes ago

ALERT! Detected and prevented compromise to file integrity

MITRE MITRE_T1036_masquerading MITRE_T1565_data_manipulation NIST NIST_800-53_AU-2 NIST_800-53_SI-4

Policy Name	Resource	Source	Action	Result	Severity
hsp-file-integrity-monitoring	/bin/exploit-bin	/bin/cp /home/ziggy/exploit-bin /bi...	Block	Permission denied	1

Operation	Cluster Name	Pod Name	Workload Name	Workload Type
File	idt-demo	--	--	--

```
root@store54055:/home/ziggy# cp /home/ziggy/exploit-bin /bin/
cp: cannot create regular file '/bin/exploit-bin': Permission denied
root@store54055:/home/ziggy#
```

Further Hardening

ACCUKNOX

Search

Activate | Create Policy +

All (23) | Discovered (3) | Hardening (15) | Custom (5)

Ignore | Delete | Make Inactive

Policy Name	Category	Status	Clusters	Na...	Selector Labels	Aler	Tags
Cryptojacking, Crypto mining, Malware protection KubeArmor	Hardening	Inactive	nvidia-nim-der	-	app=nvidia-nim +9	0	cryptominer +2
File Integrity Monitoring/Protection KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.mod... +9	0	NIST +5
Prevent execution of container administration tools within a container KubeArmor	Hardening	Inactive	nvidia-nim-der	-	maintainer=NVIDIA ... +9	0	MITRE_T1609_container_adm +7
Limit adversaries from gathering system information KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.mod... +9	0	MITRE +1
Audit execution of network service scanning tools KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.mod... +9	0	MITRE +4
Restrict adversaries from writing malicious code under the shm folder KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.mod... +9	0	MITRE_TA0002_Execution +1
Audit access to cronjob files as a part of system monitoring for better integrity KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.mod... +9	0	NIST +5
Restrict or limit maintenance tool usage KubeArmor	Hardening	Inactive	nvidia-nim-der	-	app=nvidia-nim +9	0	PCI_DSS +2
Audit defense control points to detect defense impairments KubeArmor	Hardening	Inactive	nvidia-nim-der	-	com.nvidia.nim.ver... +9	0	MITRE +4
Prevent data exfiltration attempts using utility tooling							

1 - 10 of 15

Rows per page: 10 | 1 | 2