Centrum spoločných činností SAV, v. v. i.

Výpočtové stredisko SAV

Cyber Security in the HPC World in the Era of Artificial Intelligence

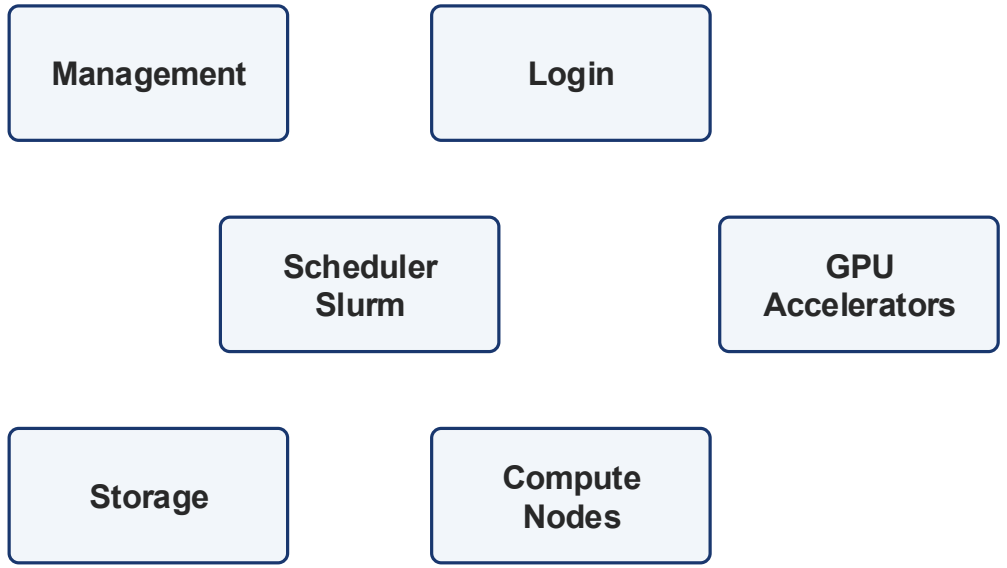
VSSAV, Bratislava, 2026
(HPC tím)

PRESENTATION ROADMAP

- Why HPC security is different from traditional IT security
- Why GPU-rich clusters are very valuable targets
- How artificial intelligence changes offensive and defensive security
- How we monitor, classify and remediate vulnerabilities
- Recent vulnerabilities and events analyzed in our HPC environment
- What this means for users and future operation

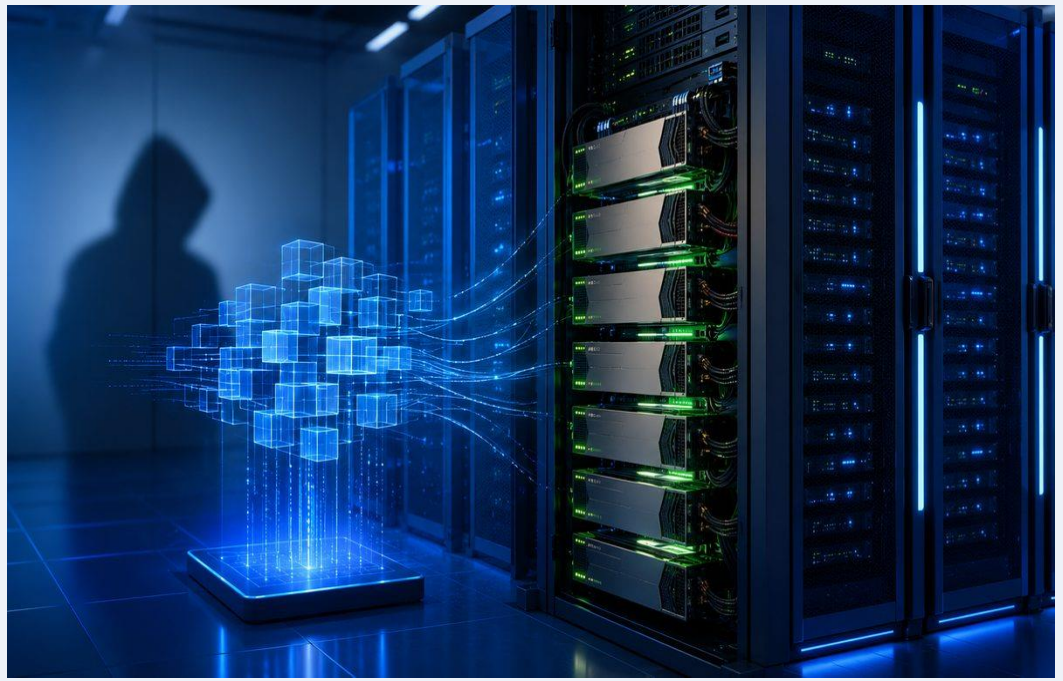
WHAT MAKES HPC DIFFERENT

- Users log in directly and submit their own code
- Schedulers such as Slurm execute arbitrary workloads
- Clusters combine login, management, compute, storage and network layers
- Performance and scientific productivity are part of the security equation
- One weak component can affect a large shared environment



THE VALUE OF HPC FOR ATTACKERS

- Ten years ago: mainly raw computing power
- Today: GPUs are the most valuable resource
- Access alone can be monetized through cryptomining, AI training or inference
- Attackers do not always need to steal data



A compromised cluster with 100 GPUs can represent hundreds of euros per hour in compute value.

APPROXIMATE VALUE OF GPU RESOURCES

- GPU resources are scarce and expensive
- AI workloads increase demand for accelerators
- Value of compute capacity creates motivation even without data theft

GPU	Approx. rental cost	Security relevance
NVIDIA A100 40GB	1–3 €/hour	Common AI/HPC accelerator
NVIDIA H100	3–6 €/hour	High-value AI training GPU
NVIDIA GH200	Higher	Premium GPU/CPU superchip
100 GPUs	Hundreds €/hour	Attractive target by itself

Compute power is an asset.
It must be protected like data.

OPEN SOURCE AS THE FOUNDATION OF HPC

- Linux, Slurm, OpenMPI, Kubernetes and scientific libraries form the stack
- Source code transparency helps defenders analyze and verify software
- The same transparency also helps attackers search for weaknesses
- The security of the whole environment is limited by the weakest dependency

AI AS A TOOL FOR ATTACKERS

AI

Source code analysis

AI

information gathering

AI

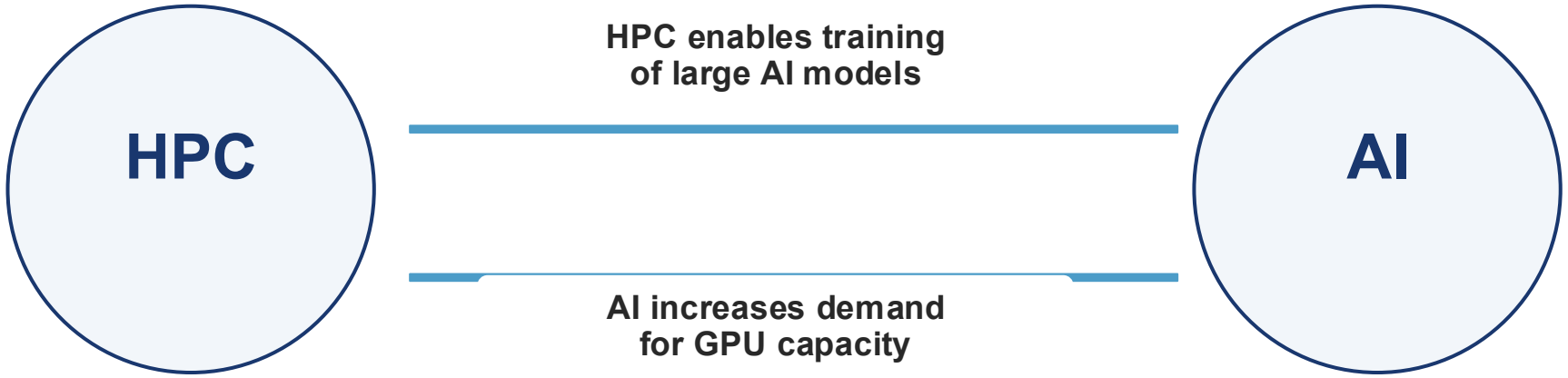
PoC exploit generation

AI

Leaked data analysis

The most important change is lower entry barrier: tasks that previously required specialized knowledge can be partially automated.

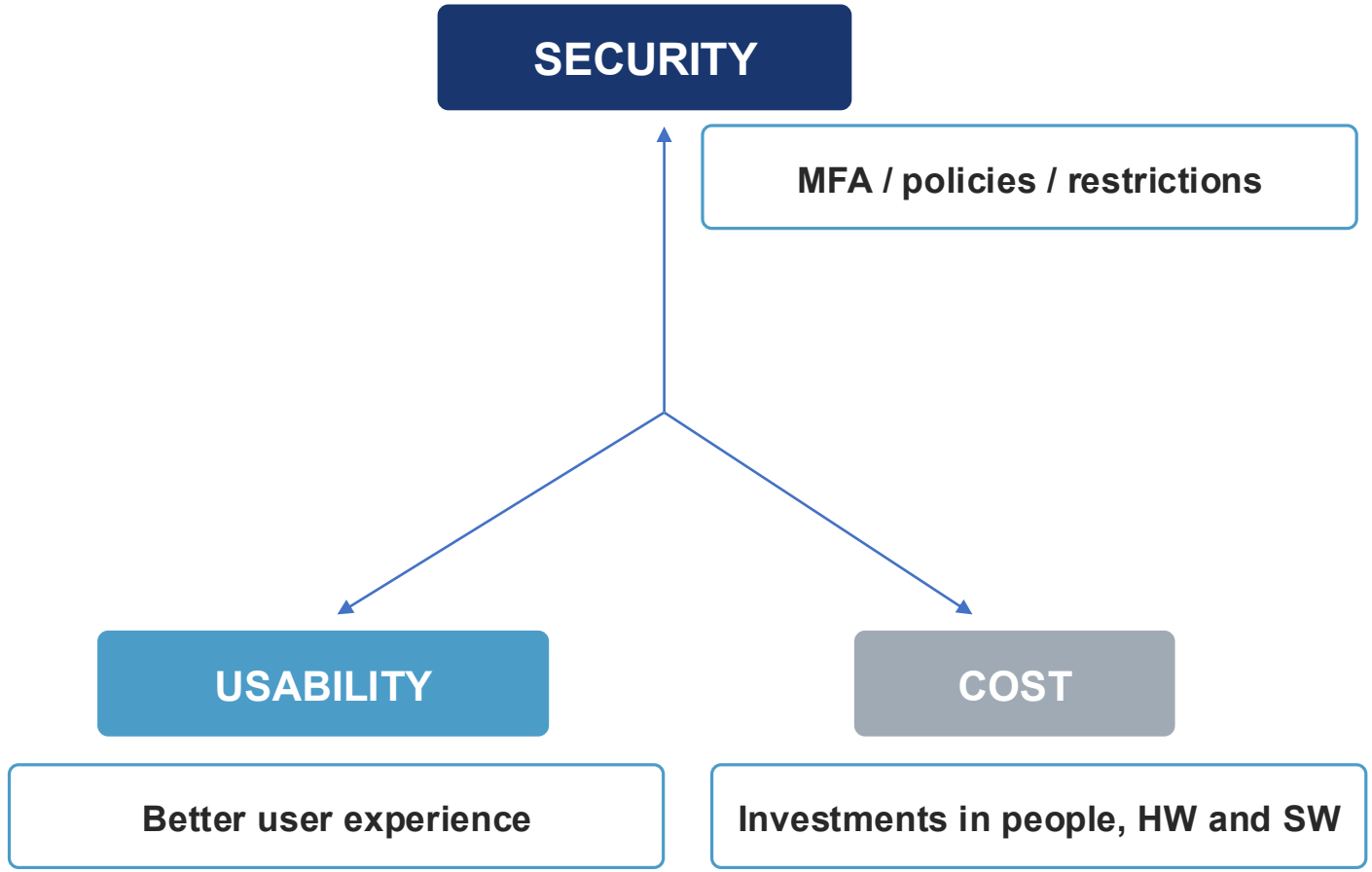
THE IMPACT OF AI ON HPC



- More GPUs in data centers
- More machine-learning workloads
- Higher value of compromised access

IMPROVING HPC SECURITY: BALANCE

- Security is all about trade-offs
- Absolute security does not exist
- The goal is to reduce risks to an acceptable level (while maintaining reasonable cost and usability)



CLASSIFICATION OF VULNERABILITIES

- Unknown vulnerabilities — not yet discovered or publicly disclosed
- Zero-day vulnerabilities — known to attackers before defenders or vendors
- Known vulnerabilities — publicly disclosed and often assigned a CVE
- KEV vulnerabilities — known to be actively exploited in the wild
- In operations, KEV and local exploitability (LPE) receive the highest priority

The most urgent issue is not always the newest CVE, but the one that is exploitable in our real environment.

SOURCES OF VULNERABILITY INFORMATION

- CISA (CVE,KEV) — vulnerabilities known to be actively exploited
- Red Hat Security Advisories — relevant for RHEL-compatible Linux infrastructure
- NVIDIA Security Bulletins — GPU drivers, CUDA and accelerator stack
- HPC Community Security Advisories – vulnerabilities, incidents and mitigations shared by HPC operators

Priority is based on exploitability and environment impact, not only on the existence of a CVE identifier.

TYPES OF VULNERABILITIES ANALYZED

Type	Examples
Linux Kernel	DirtyFrag, Dirty Pipe, CIFSwitch
Open Source Supply Chain	Shai-Hulud npm malware
GPU / AI stack	NVIDIA Security Bulletins
Web Services	BADHOST, FastAPI Host Header Injection
Configuration / Operations	cgroup release_agent, authentication, kernel modules

In HPC, local privilege escalation is especially important because users normally run code on shared systems.

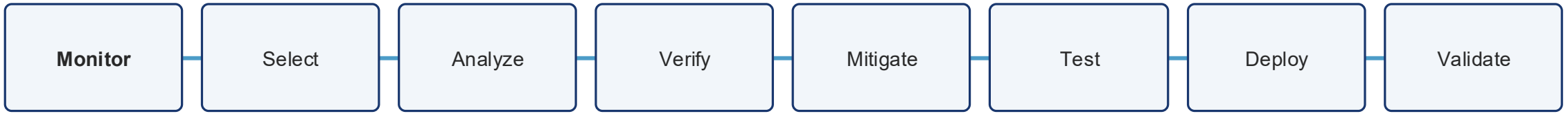
RECENT VULNERABILITIES AND SECURITY EVENTS — I

Approx. date	CVE	Name(s)	Action
2026-06-02	CVE-2022-0492	Linux cgroup-v1 release_agent, KEV 2026	Rocky 8/9 kernel verification; Red Hat backport confirmed
2026-05-20	CVE-2026-46333	CIFSwitch, CIFS LPE	CIFS audit; cifs blacklist; install cifs /bin/false
2026-05-21–23	N/A	PinTheft, DirtyFrag, ESP4/ESP6 chain, RDS/io_uring LPE	Kernel analysis; module review; blacklist esp4/esp6
2026-05-22	CVE-2022-0847	Dirty Pipe	Reference analysis and comparison with newer Linux LPEs

RECENT VULNERABILITIES AND SECURITY EVENTS — II

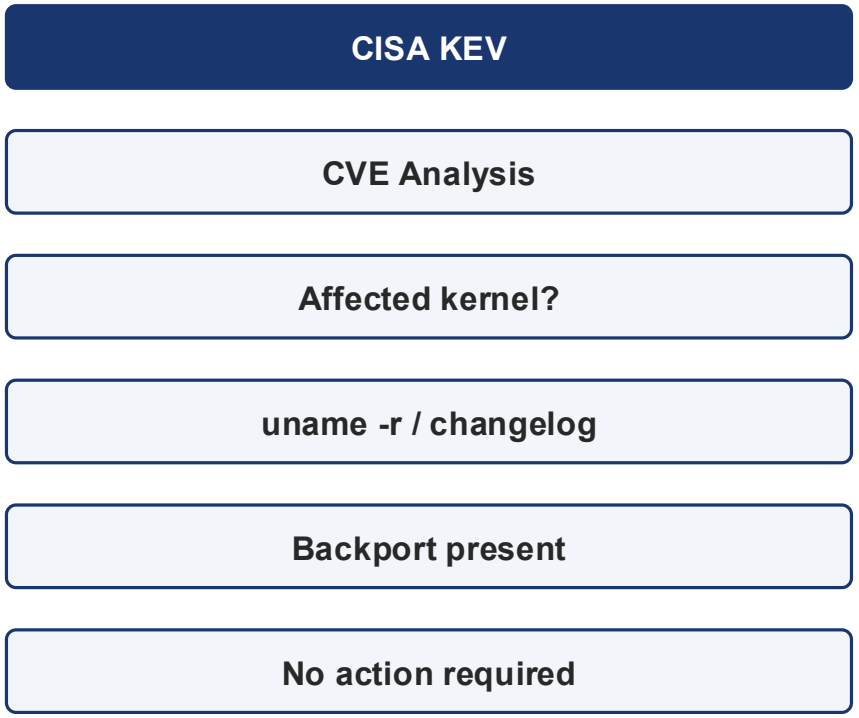
Approx. date	CVE	Name(s)	Action
2026-05-22	N/A	Copy Fail, AF_ALG exploit	Impact assessment for multi-user HPC environment
2026-05-24	N/A	Shai-Hulud, npm supply-chain malware	Node.js user audit; npm package/build environment inspection
2026-05-25	CVE-2026-31431, CVE-2026-43284, CVE-2026-46300	Linux Kernel LPE group	Kernel version verification and Rocky Linux impact assessment
2026-05-27	N/A	NVIDIA Security Bulletins	Impact analysis for A100/GH200 GPUs and R570/R580 drivers
2026-05-28	N/A	BADHOST, FastAPI Host Header Injection	Assessment of impact on AI and web services

VULNERABILITY HANDLING PROCESS



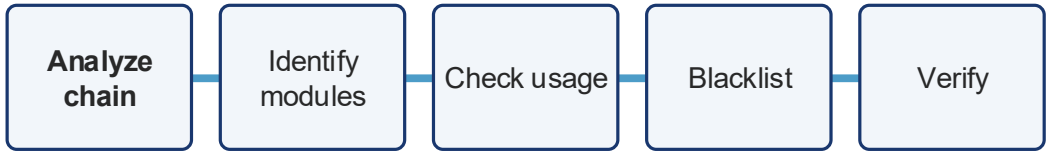
- Installing an update is often only one part of the process
- HPC requires compatibility checks with drivers, kernels, storage, MPI, Slurm and user workloads
- Mitigation may be safer than upgrade when functionality is at risk

CASE STUDY: CVE-2022-0492



- The vulnerability appeared in the CISA KEV catalog
- Analysis showed that our Rocky Linux kernels already contained the fix
- Result: documented verification, no emergency production change

CASE STUDY: DIRTYFRAG / ESP4 / ESP6



- Exploit-chain analysis identified ESP4 and ESP6 as relevant modules
- These IPsec ESP modules were not used in the HPC environment
- Risk was reduced by preventing module loading through modprobe policy

Mitigation example:
blacklist esp4 / esp6
install esp4 /bin/false
install esp6 /bin/false

AI IS DUAL USE

- Defenders use AI for log analysis, code review and prioritization
- Attackers use AI to analyze, discover and exploit development
- The advantage belongs to teams that integrate AI into process

AI does not remove the need for expertise. It amplifies both good and bad decisions.

IMPACT ON HPC USERS

Temporary unavailability

Password changes

MFA deployment

Access restrictions

Stricter policies

- Security measures may reduce usability in the short term
- The purpose is long-term reliability, trust and protection of shared resources

CURRENT SITUATION AND FUTURE OUTLOOK

- More vulnerabilities and more security events are being reported
- Incidents are becoming more complex and more automated
- GPU-rich HPC systems are increasingly attractive
- AI will accelerate both attacks and defense
- Quantum computing may later challenge current cryptographic systems

PRACTICAL PRIORITIES FOR HPC SECURITY

Identity and access control

Vulnerability monitoring and prioritization

Kernel / driver / module hardening

Logging, auditability and incident response

User communication and policy

CONCLUSION

- Cyber security is no longer limited to commercial IT environments
- HPC contains valuable compute power, GPU resources and scientific data
- AI transforms both offensive and defensive security capabilities
- Open source remains the foundation of HPC and a supply-chain risk (xz utils)
- Security is a continuous process, not a one-time project

Thank you for your attention

Questions?