Breaking Into The AI-protected Realm

A Practical Session on (pen)testing Embedded Targets with AI-based EDR

> Whoarewe

Lalie:

> PhD Student at CEA-Leti
> Studies energetic networks' cybersecurity using HW/SW related signals

> Nix Lover 
> ENSIMAG alumni

> Uses AI to filter and identify threats in OS signals trace

Ulysse:

> PhD Student at CEA-Leti
> Studies fuzzing with side-channel
> Has studied the subject of interest in the past

> Parisian bobo exiled to the mountains

> Uses statistical tools to perform signal processing

> Whoarewe

Lalie:

> PhD Student at CEA-Leti

> Studies energetic networks' cybersecurity using HW/SW related signals

> Nix Lover

> ENSIMAG alumni

> Use IA to filter and identify threats in OS signals trace

Ulysse:

> PhD Student at CEA-Leti
> Studies fuzzing with side-channel
> Has studied the subject of interest in the past

> Parisian bobo exiled to the mountains

> Uses statistical tools to perform signal processing

> What will u do today?

> Understand why Network-based Detection Systems (NIDS) are not always up to the task

> Understand what a Host-based Intrusion Detection System (HIDS) is

> Introduction to some simple statistical tools to perform anomaly detection

> Try to break into the target without triggering the detection mechanism

> What are IRL current targets?

IoT

> What are IRL current targets?

IoT

PCs

> What are IRL current targets?

IoT

PCs

Uncrewed vehicles

> What are IRL current targets?

Uncrewed vehicles

> What are IRL current targets?

Uncrewed vehicles

Servers

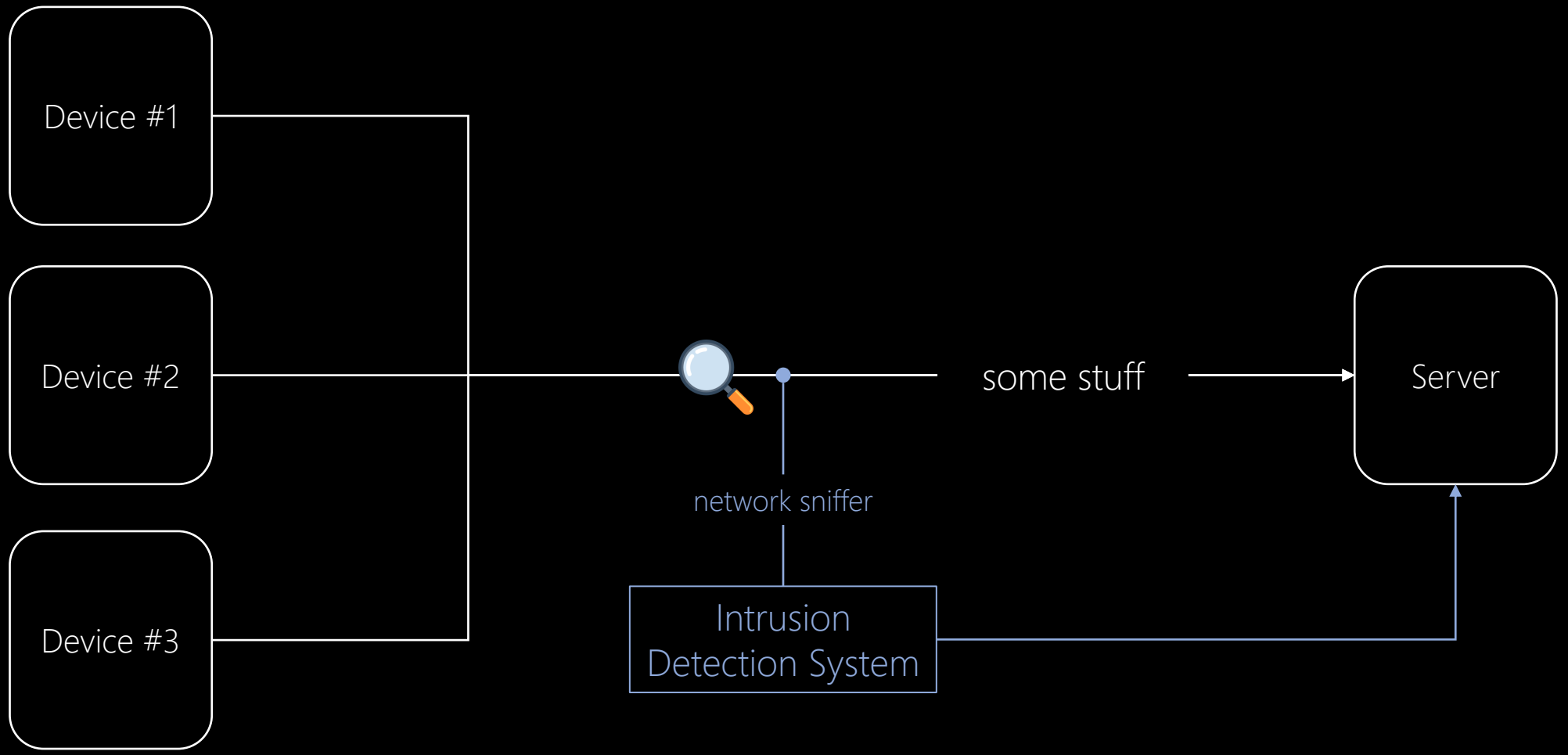
> What are IRL current targets?

Uncrewed vehicles

Servers

IIoT

> What is an NIDS?

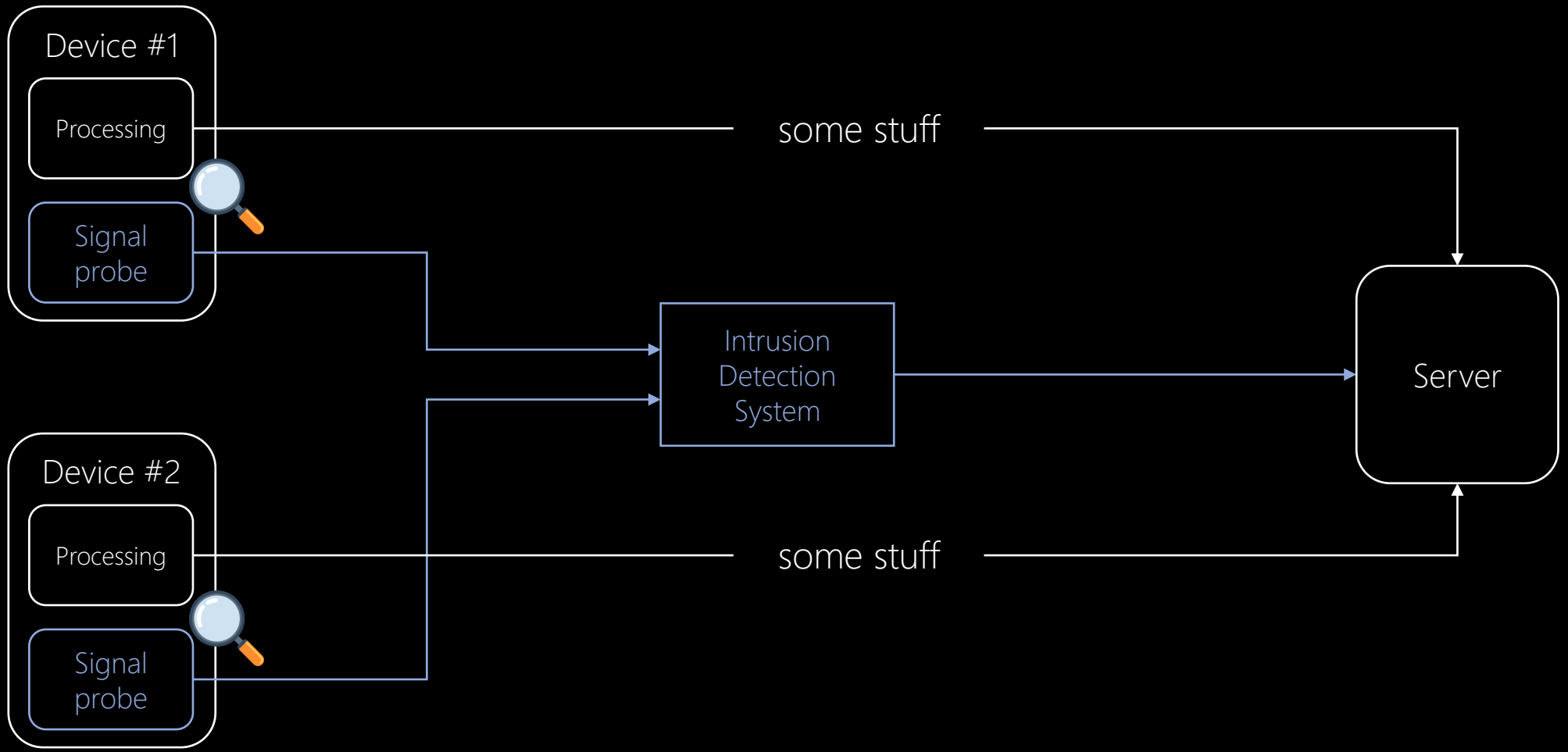


> What is an NIDS?

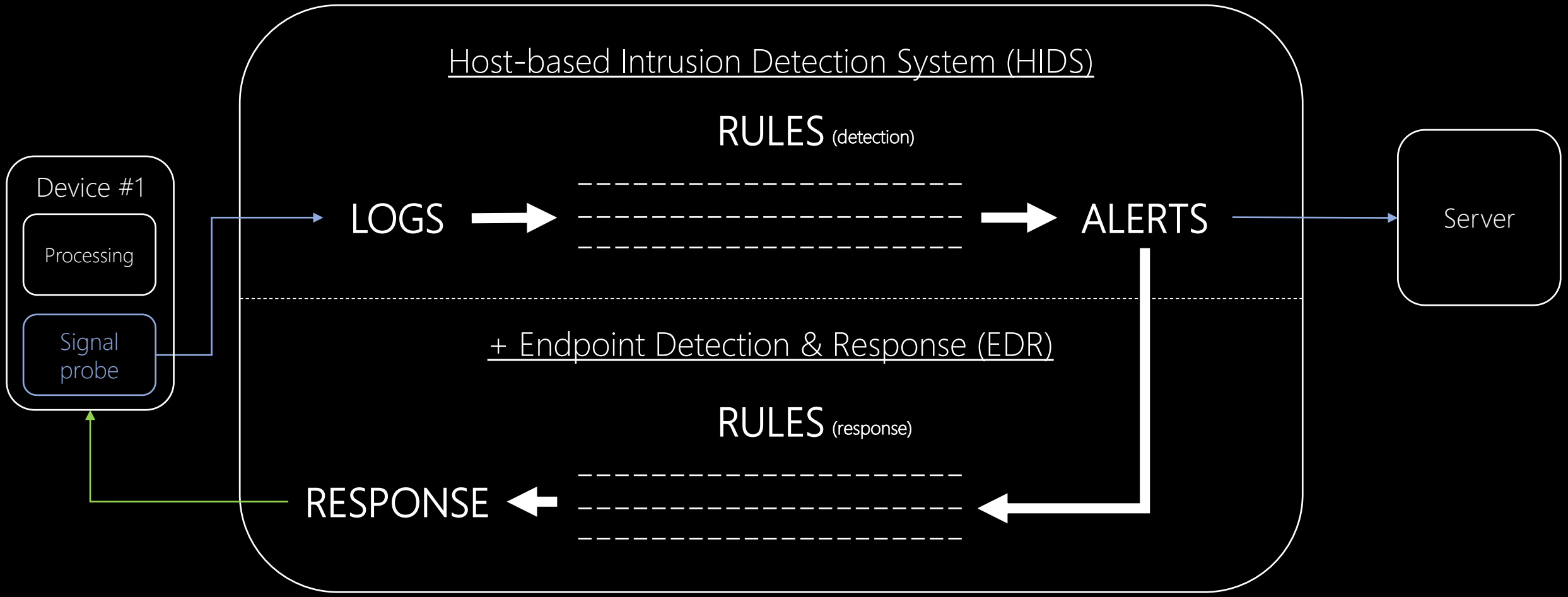
Device #1

« It enough detects anything in whole world! »

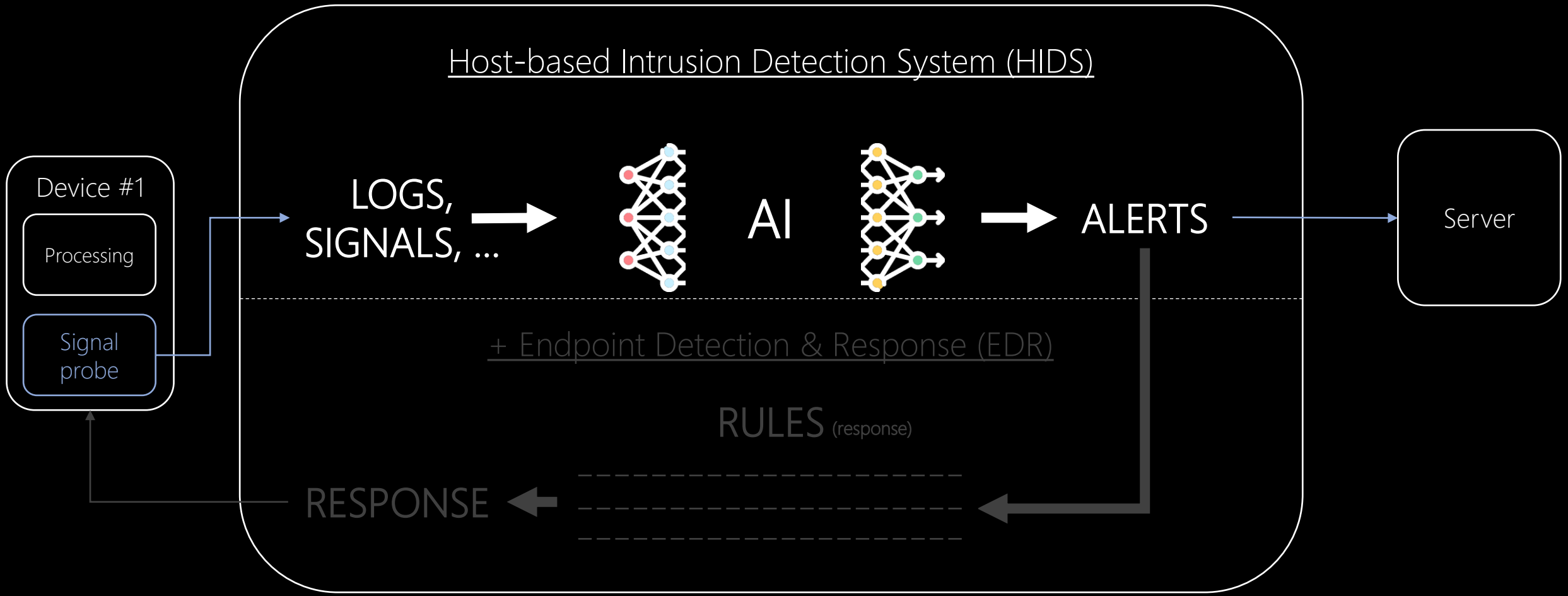
> What is an HIDS?



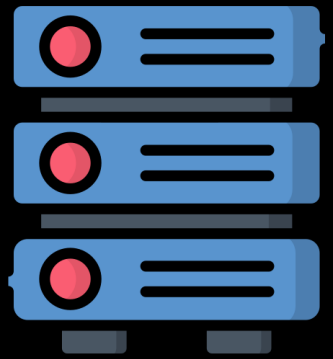
> What is an HIDS currently?

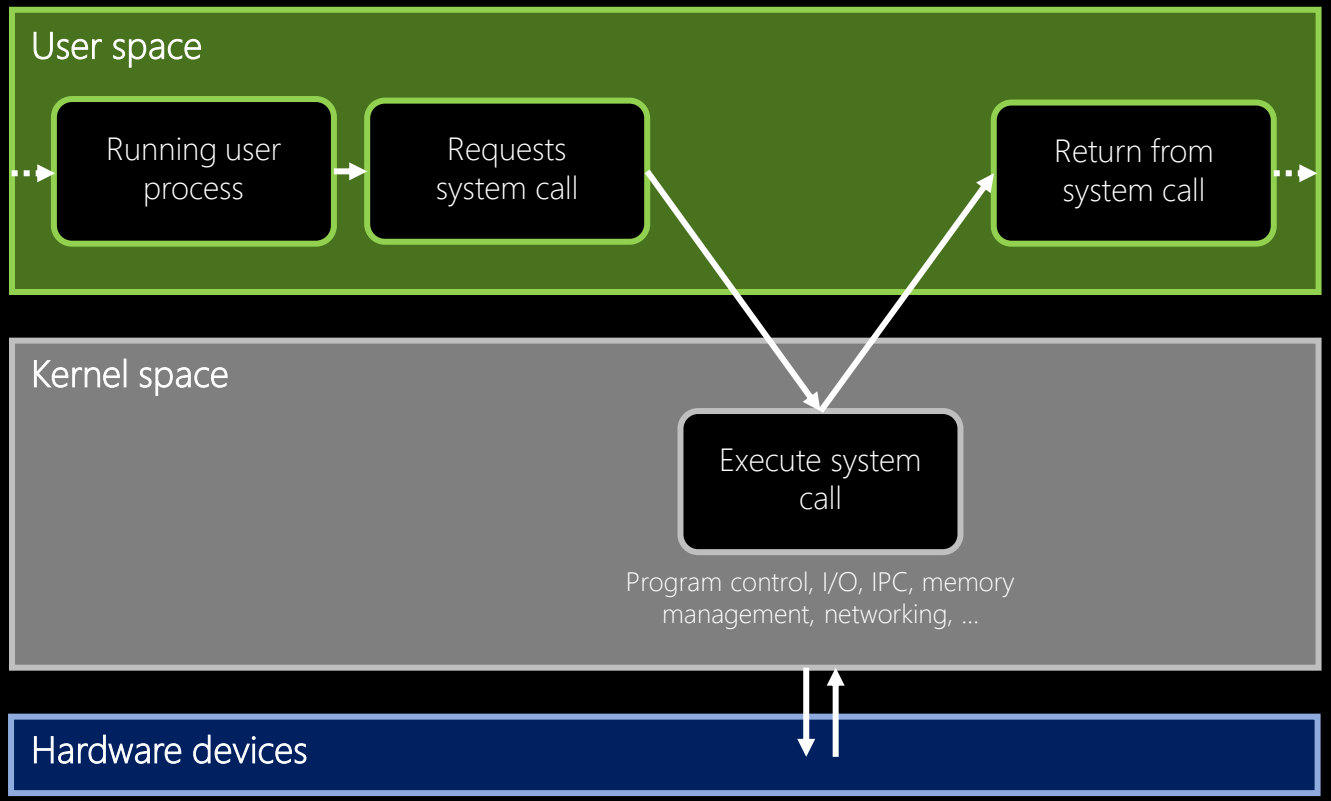


> What can be an HIDS soon?

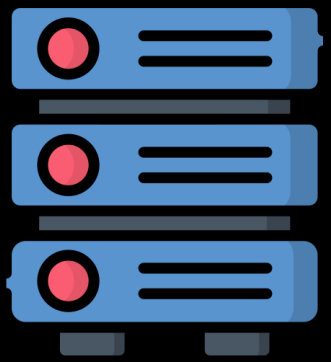


> System calls

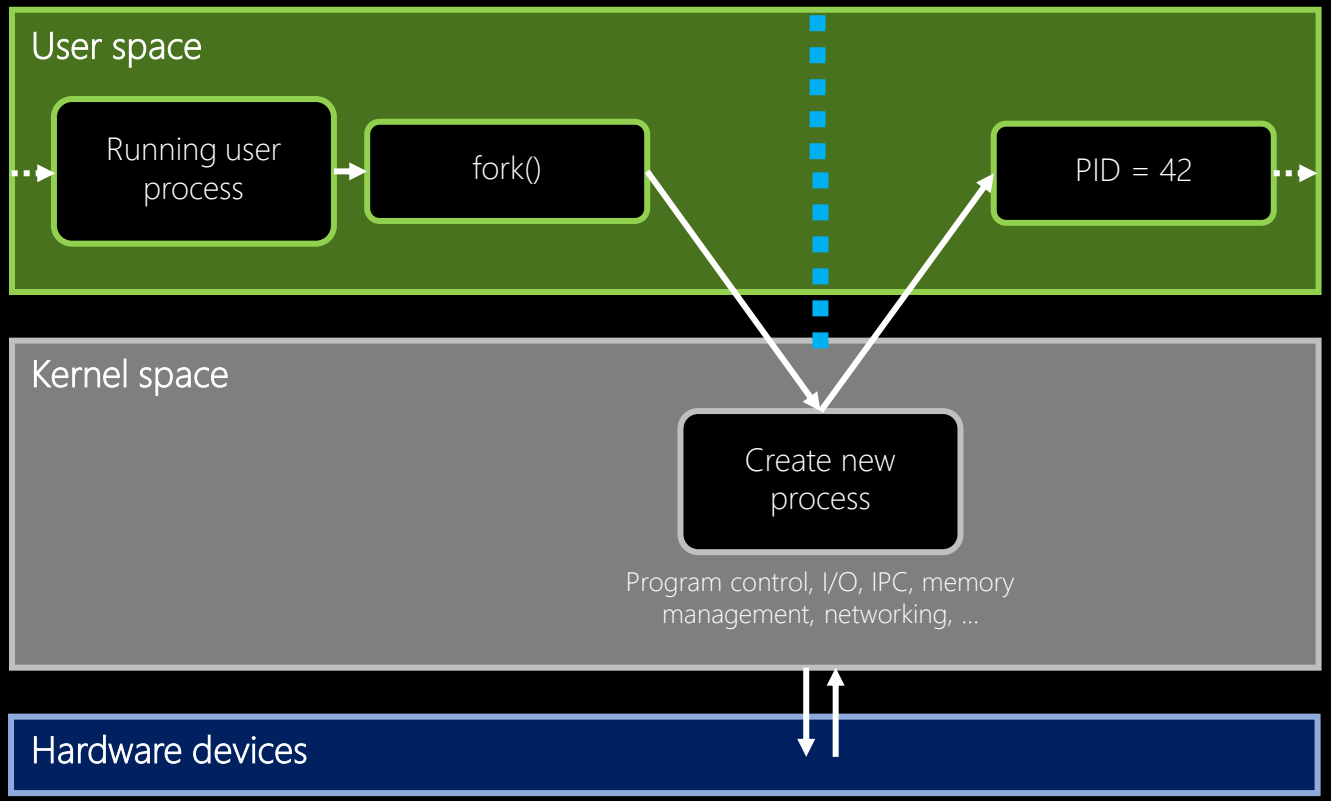




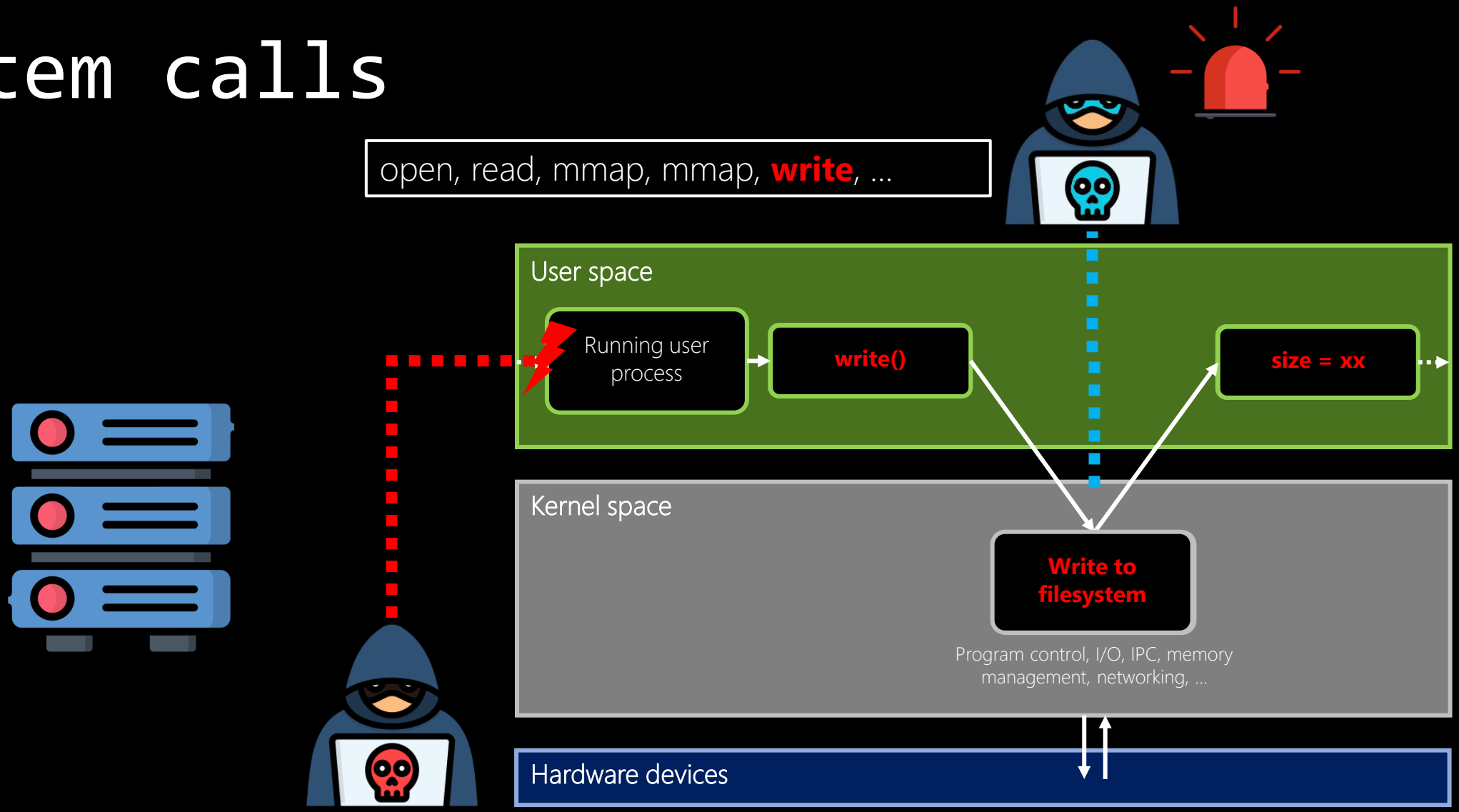
> System calls



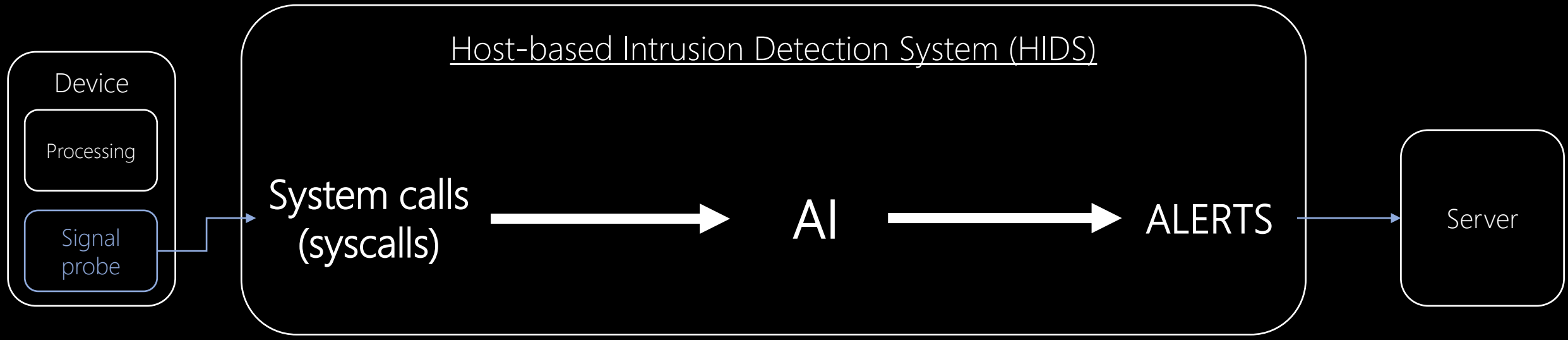
open, read, mmap, mmap, **fork**, ...



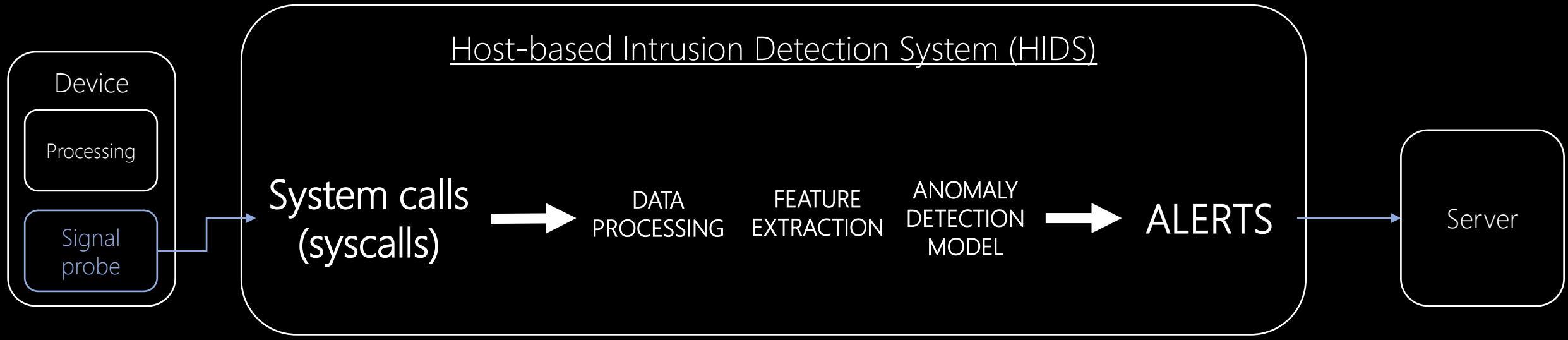
> System calls



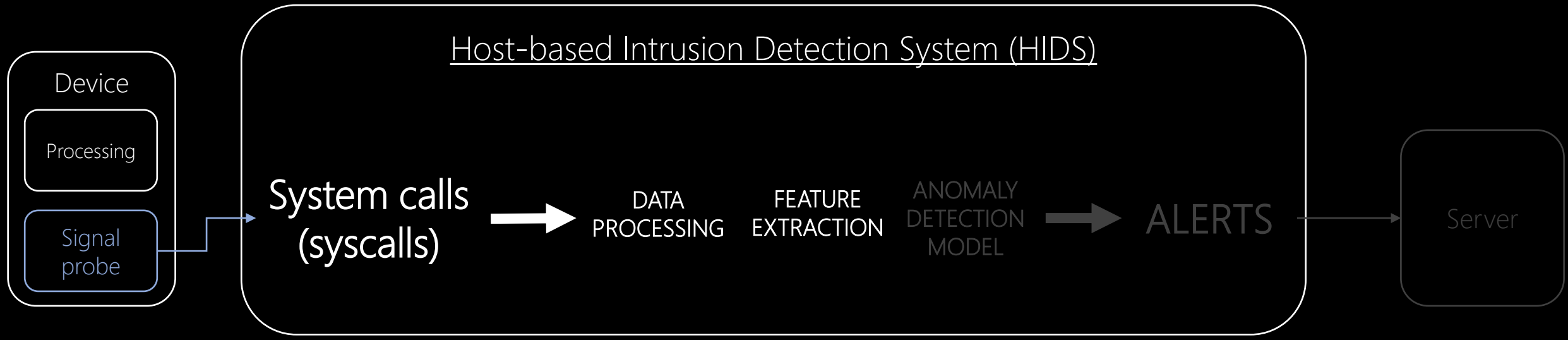
> What’s the HIDS you’ll deal with ?



> What’s the HIDS you’ll deal with ?



> What’s the HIDS you’ll deal with ?

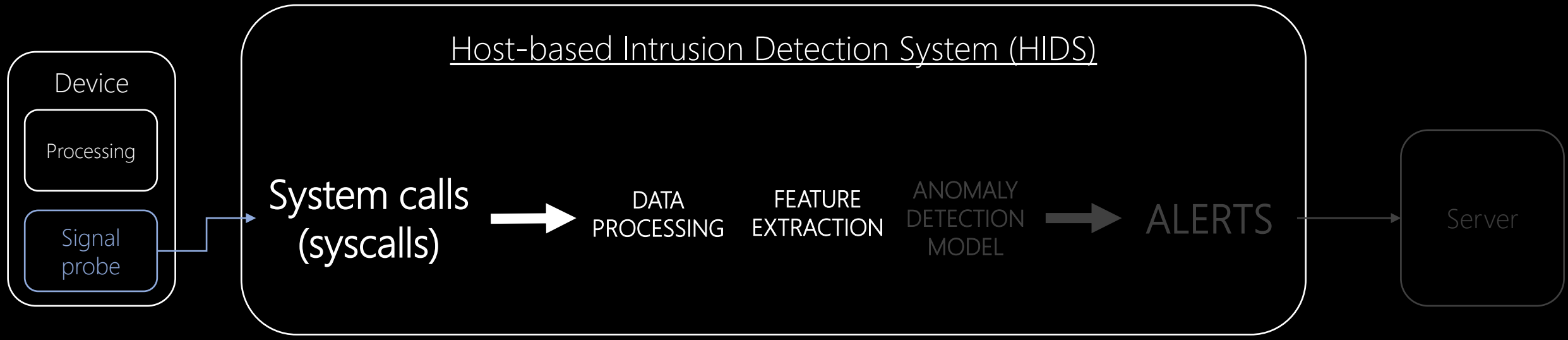


RAW SYSCALLS

Trace #10

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, ‘hi’)	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

> What’s the HIDS you’ll deal with ?



RAW SYSCALLS

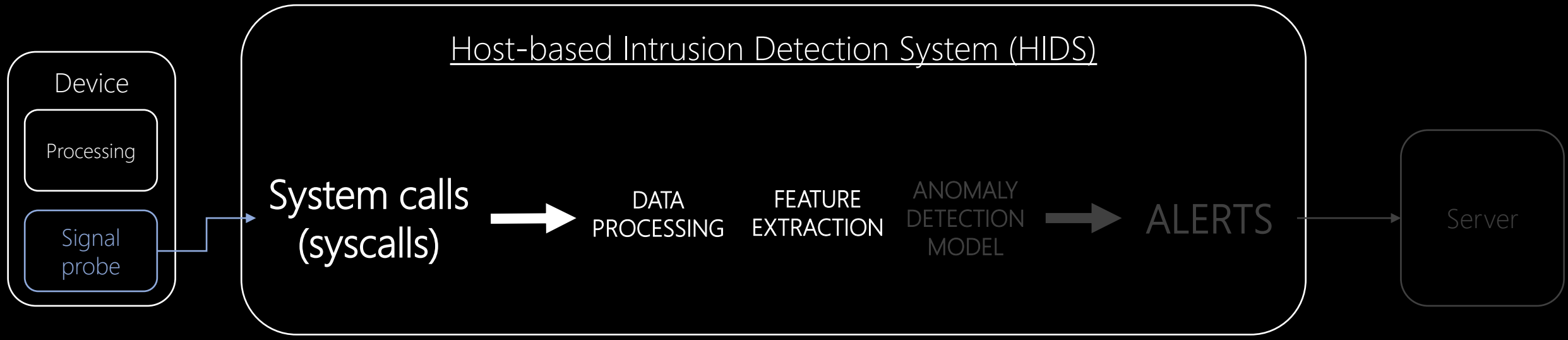
DATA PROCESSING

Trace #10

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, ‘hi’)	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

ID
#2
#102
#10
#102
#138
...

> What’s the HIDS you’ll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, ‘hi’)	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

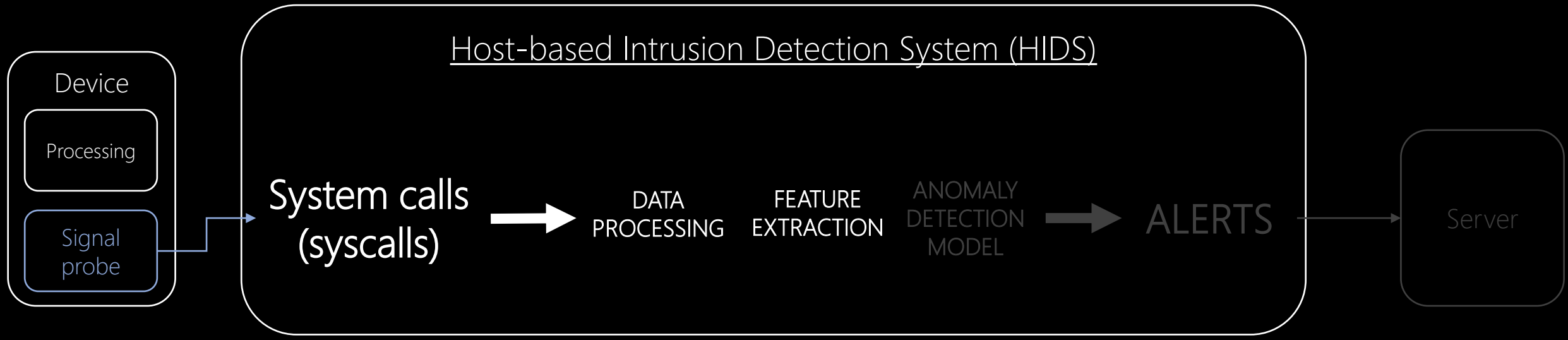
DATA PROCESSING

ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

3gram-ID
(#2, #102, #10)
(#102, #10, #102)
(#10, #102, #138)
(#102, #138, ...)
(#138,..., ...)
...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

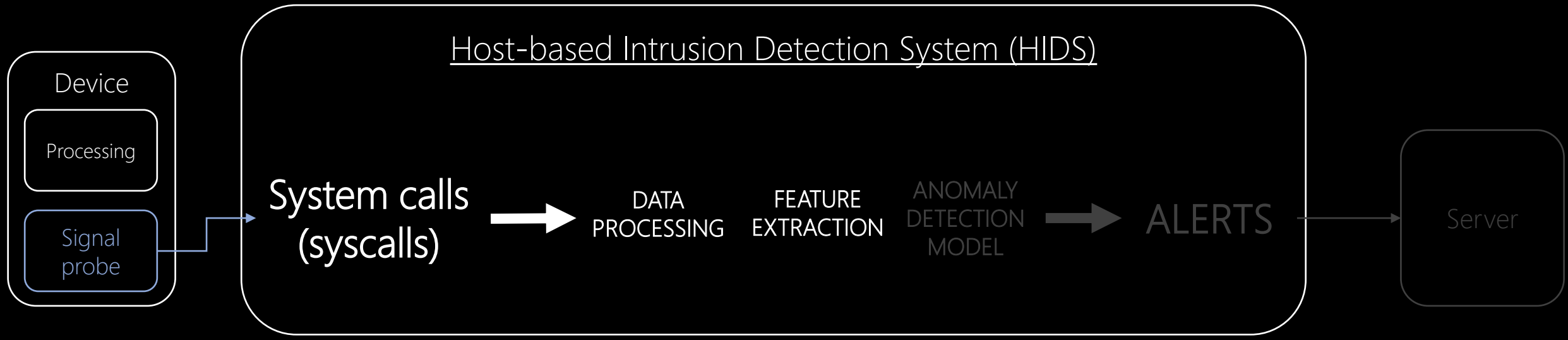
DATA PROCESSING

ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

3gram-ID
(#2, #102, #10)
(#102, #10, #102)
(#10, #102, #138)
(#102, #138, ...)
(#138,..., ...)
...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

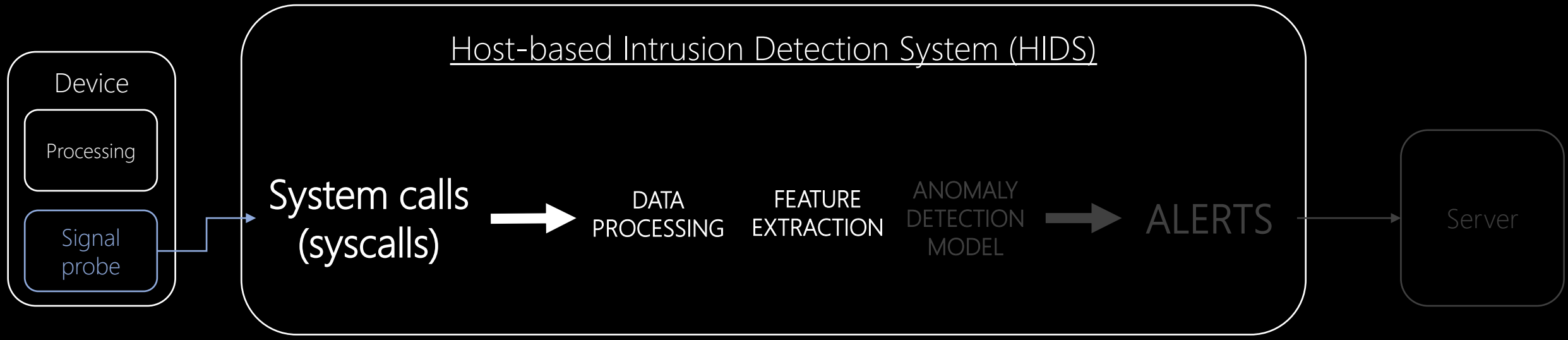
DATA PROCESSING

ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

3gram-ID
(#2, #102, #10)
(#102, #10, #102)
(#10, #102, #138)
(#102, #138, ...)
(#138,..., ...)
...

> What’s the HIDS you’ll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, ‘hi’)	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

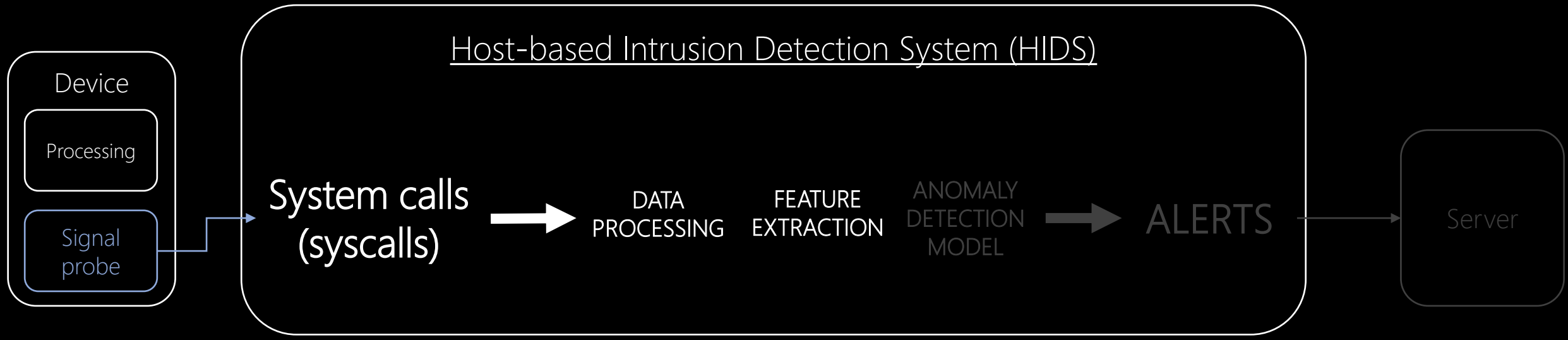
DATA PROCESSING

ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

3gram-ID
(#2, #102, #10)
(#102, #10, #102)
(#10, #102, #138)
(#102, #138, ...)
(#138,..., ...)
...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

DATA PROCESSING

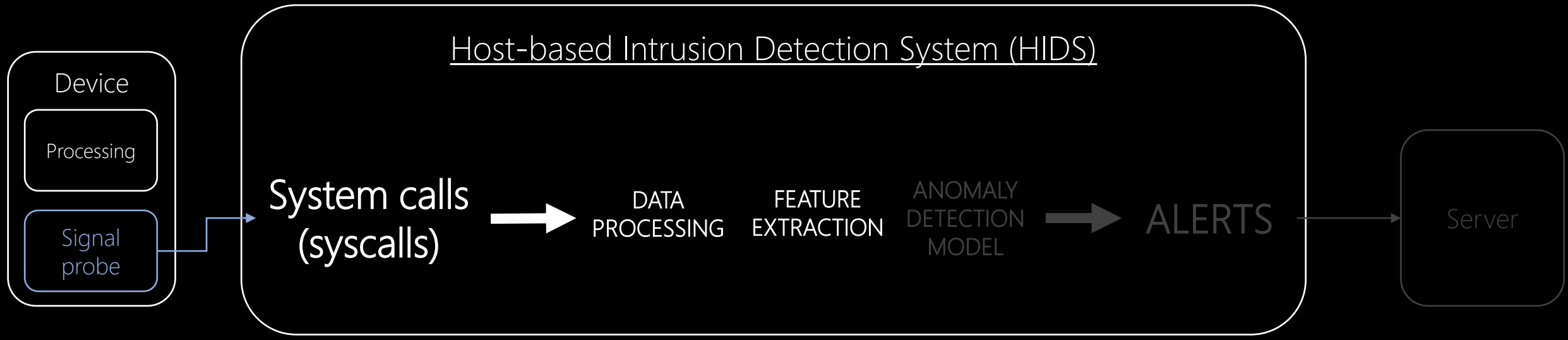
ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

*TF-IDF = N-gram frequency in trace * log(number of past traces / number of past traces with N-gram)*

3gram-ID	TF-IDF
(#2, #102, #10)	0.3
(#102, #10, #102)	0.5
(#10, #102, #138)	0.02
(#102, #138, ...)	0.9
(#138,..., ...)	0.4
...	...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

DATA PROCESSING

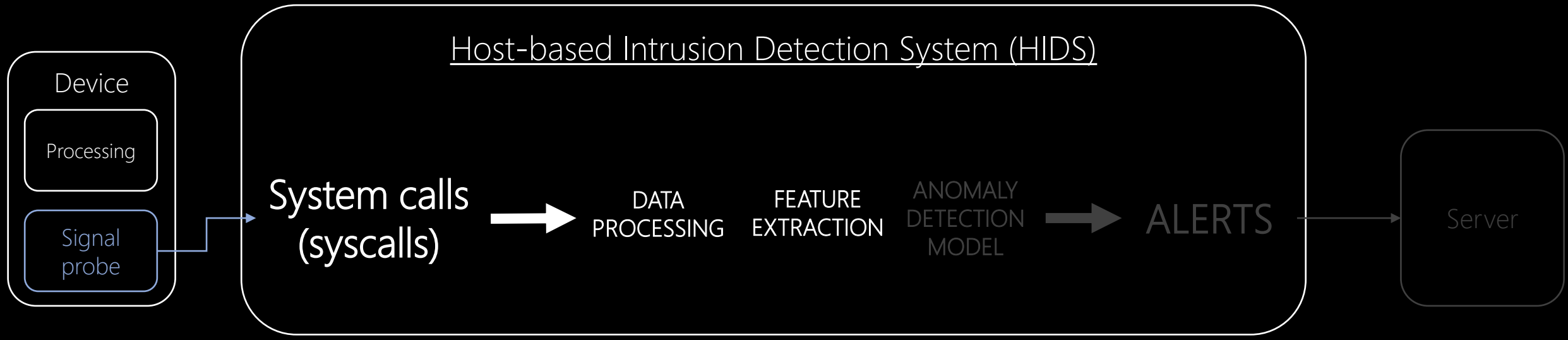
ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

*TF-IDF = N-gram frequency in trace * log(number of past traces / number of past traces with N-gram)*

3gram-ID	TF-IDF
(#2, #102, #10)	0.3
(#102, #10, #102)	0.5
(#10, #102, #138)	0.02
(#102, #138, ...)	0.9
(#138,..., ...)	0.4
...	...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

DATA PROCESSING

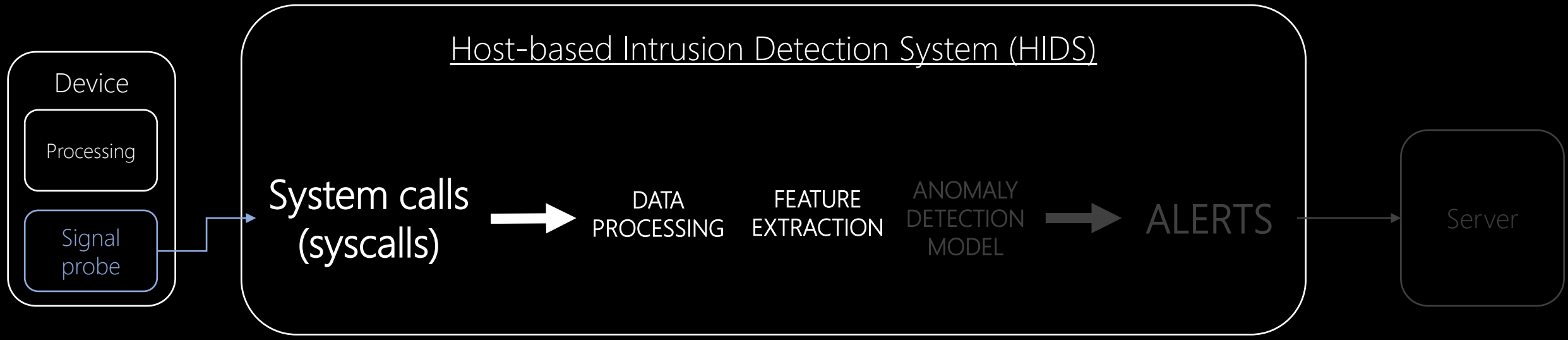
ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

*TF-IDF = N-gram frequency in trace * log(number of past traces / number of past traces with N-gram)*

3gram-ID	TF-IDF
(#2, #102, #10)	0.3
(#102, #10, #102)	0.5
(#10, #102, #138)	0.02
(#102, #138, ...)	0.9
(#138,..., ...)	0.4
...	...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

DATA PROCESSING

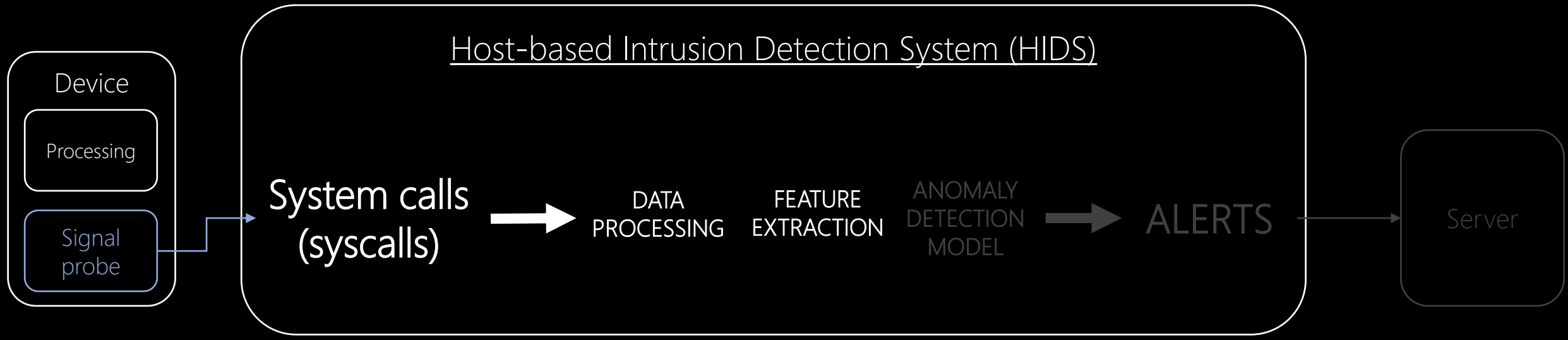
ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

*TF-IDF = N-gram frequency in trace * log(number of past traces / number of past traces with N-gram)*

3gram-ID	TF-IDF
(#2, #102, #10)	0.3
(#102, #10, #102)	0.5
(#10, #102, #138)	0.02
(#102, #138, ...)	0.9
(#138,..., ...)	0.4
...	...

> What's the HIDS you'll deal with ?



Trace #10

RAW SYSCALLS

Timestamp	ID	ARGS	RETVAL
00:01	#2	(2, 'hi')	2
00:02	#102	(3, {...})	3
00:04	#10	(39489)	&...
00:05	#102	({...}, {...})	{...}
00:07	#138	(409)	4
00:10	...	...	...

DATA PROCESSING

ID
#2
#102
#10
#102
#138
...

FEATURE EXTRACTION

*TF-IDF = N-gram frequency in trace * log(number of past traces / number of past traces with N-gram)*

3gram-ID	TF-IDF
(#2, #102, #10)	0.3
(#102, #10, #102)	0.5
(#10, #102, #138)	0.02
(#102, #138, ...)	0.9
(#138,..., ...)	0.4
...	...

> What's the HIDS you'll deal with ?

Device

Processing

Signal probe

Host-based Intrusion Detection System (HIDS)

System calls (syscalls)

DATA PROCESSING

FEATURE EXTRACTION

ANOMALY DETECTION MODEL

ALERTS

Server

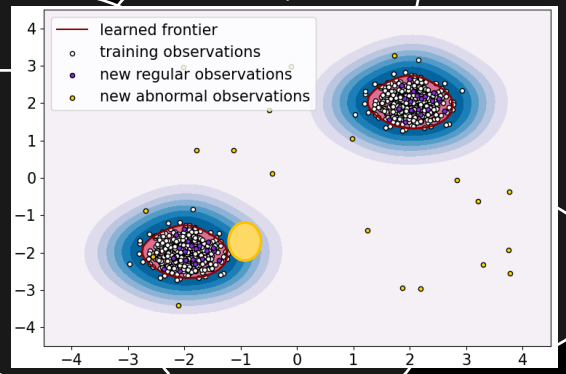
FEATURE EXTRACTION

Trace #10

TF-IDF
0.3
0.5
0.02
0.9
0.4
...

ANOMALY DETECTION MODEL

Anomaly detection algorithm



threshold

Anomaly score

> What's the HIDS you'll deal with ?

Device

Processing

Signal probe

Host-based Intrusion Detection System (HIDS)

System calls (syscalls)

DATA PROCESSING

FEATURE EXTRACTION

ANOMALY DETECTION MODEL

ALERTS

Server

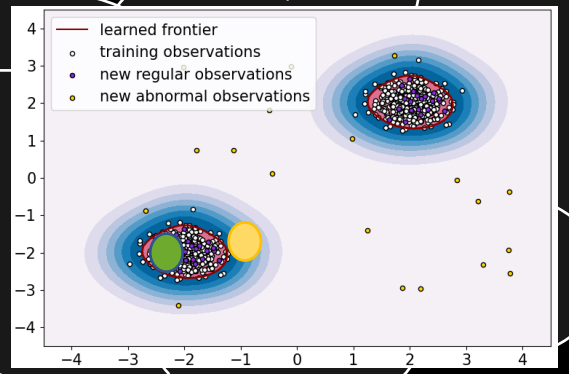
FEATURE EXTRACTION

Trace #11

TF-IDF
1
0.1
0.005
0.85
0.2
...

ANOMALY DETECTION MODEL

Anomaly detection algorithm



threshold

Anomaly score

> What's the HIDS you'll deal with ?

Device

Processing

Signal probe

Host-based Intrusion Detection System (HIDS)

System calls (syscalls)

DATA PROCESSING

FEATURE EXTRACTION

ANOMALY DETECTION MODEL

ALERTS

Server

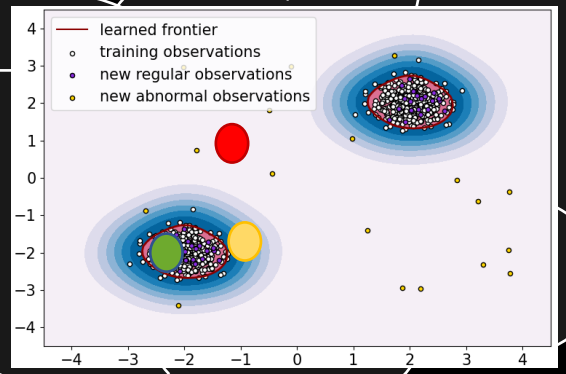
FEATURE EXTRACTION

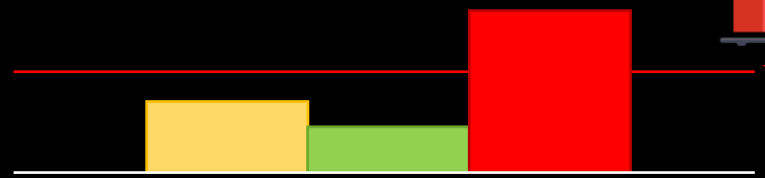
Trace #12

TF-IDF
0.048
0.2
0.13
0.2
0.9
...

ANOMALY DETECTION MODEL

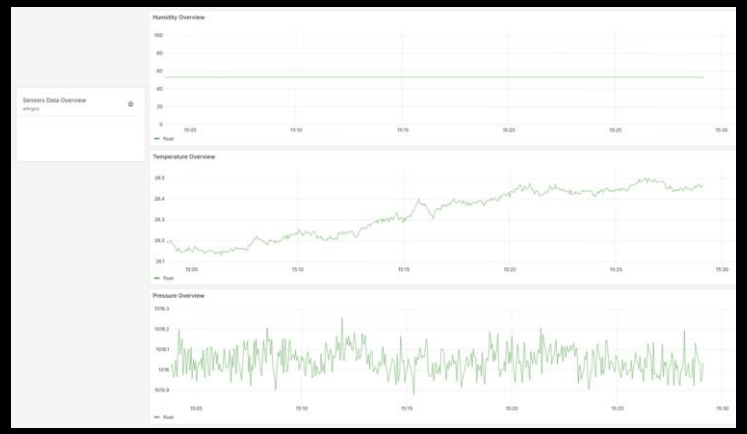
Anomaly detection algorithm

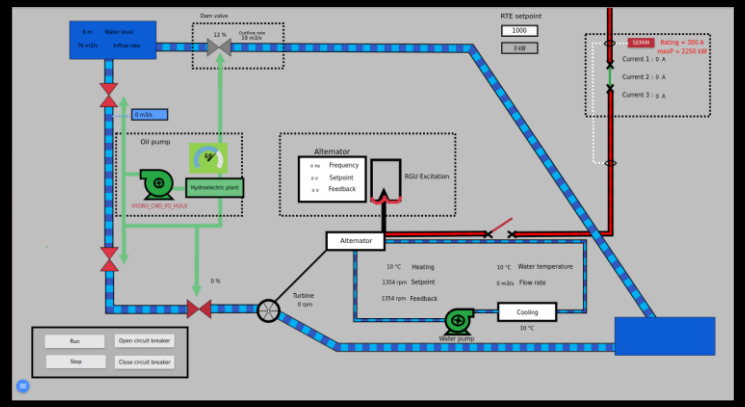


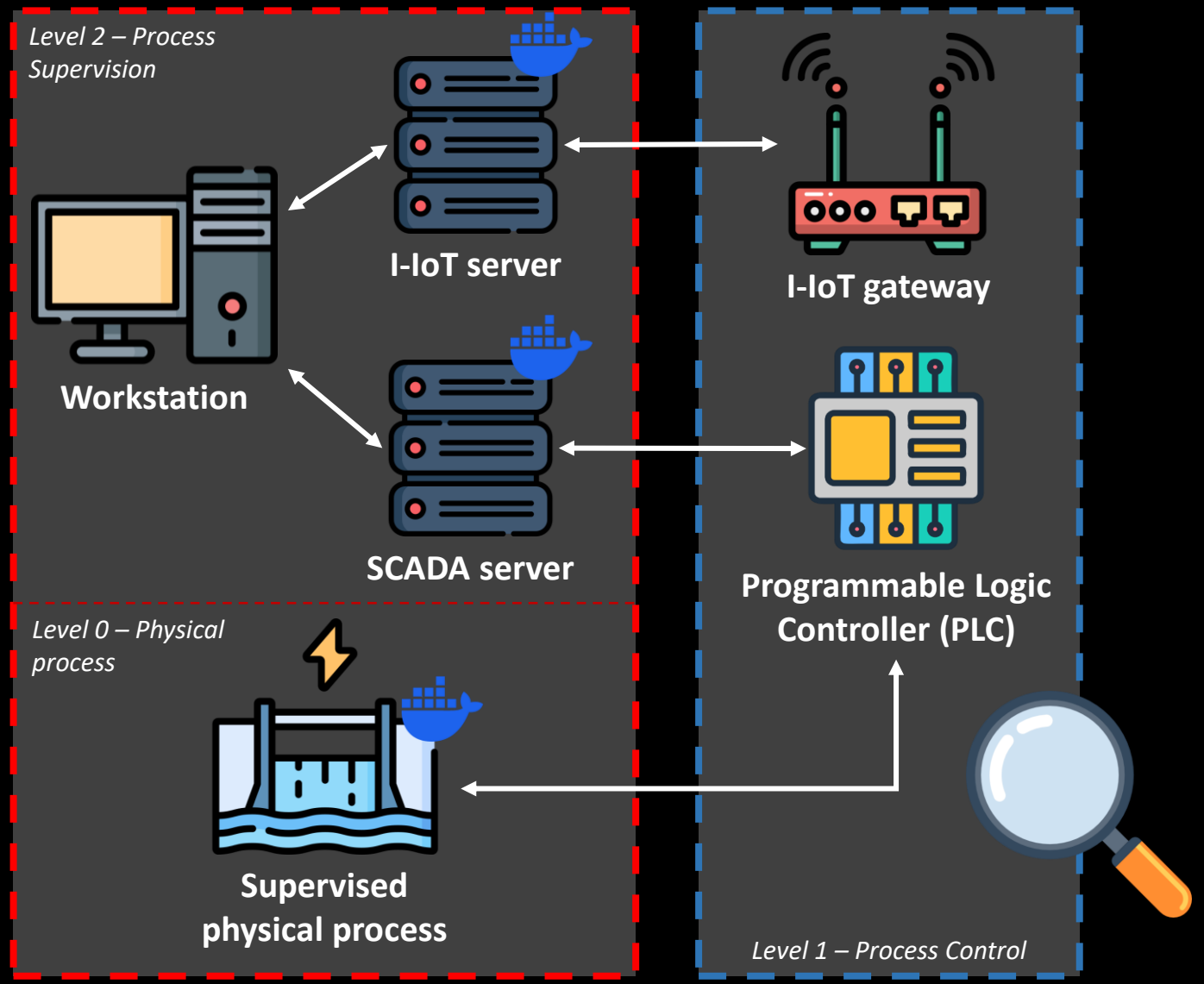


Anomaly score

> Hands-on: use case

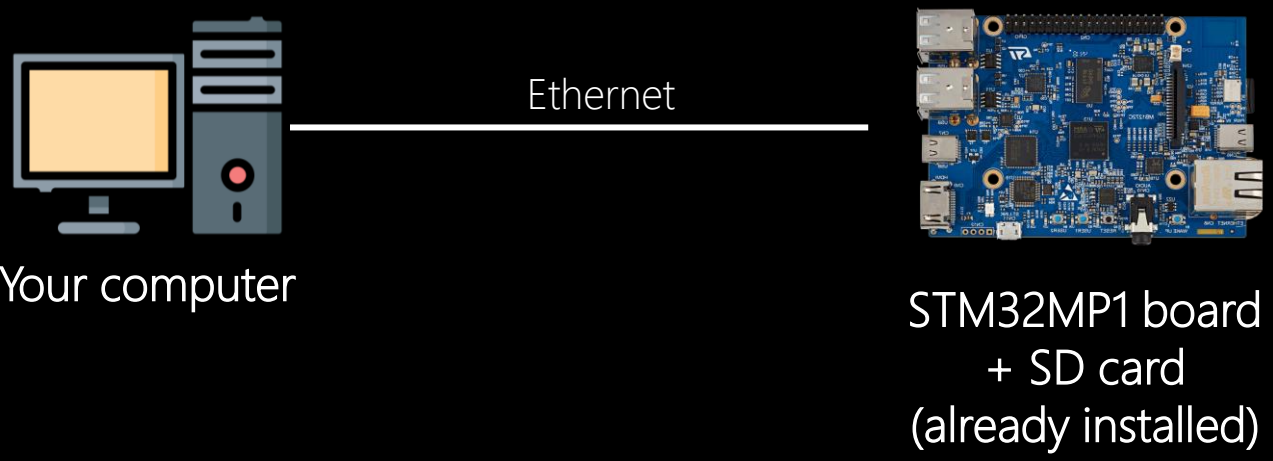




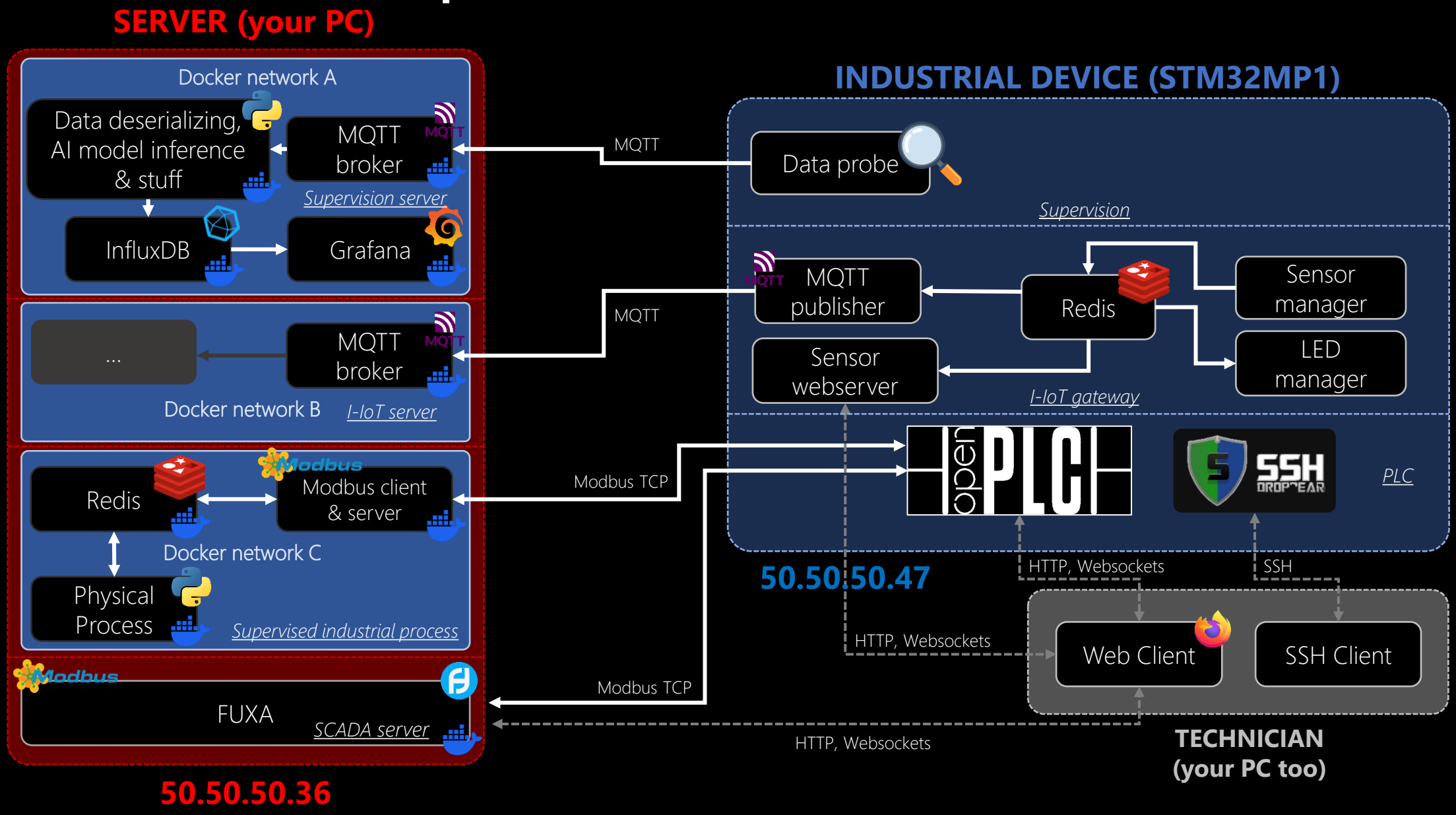


https://github.com/CEA-Leti/HENDRICS

> Hands-on: platform (hardware)



> Hands-on: platform (running services)



> Hands-on: threat model

An attacker (you) is in the same network as the PLC

(this is a critical scenario for a real-life facility, but it has already been seen multiple times…)

WHAT ARE THE RULES?

1. Suppose you do not have physical access to the device

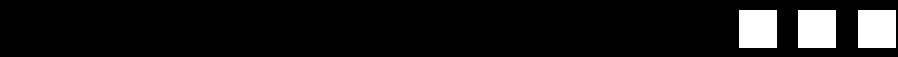
2. Do not try to attack things on your computer (industrial process simulator, AI & supervision stuff); that's not the goal of the workshop.

X. There are hidden flags on the target… Try to find 'em, without being caught by the IDS ☺

> Installation + Demo

30min

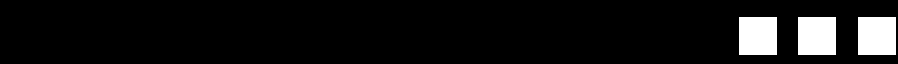
Everything that needs to run on your computer: **github.com/LalieA/Workshop_AI_Protected_Realm**



For further information after GreHack, don’t hesitate to reach us:
lalie.arnoud@cea.fr
ulysse.vincenti@cea.fr

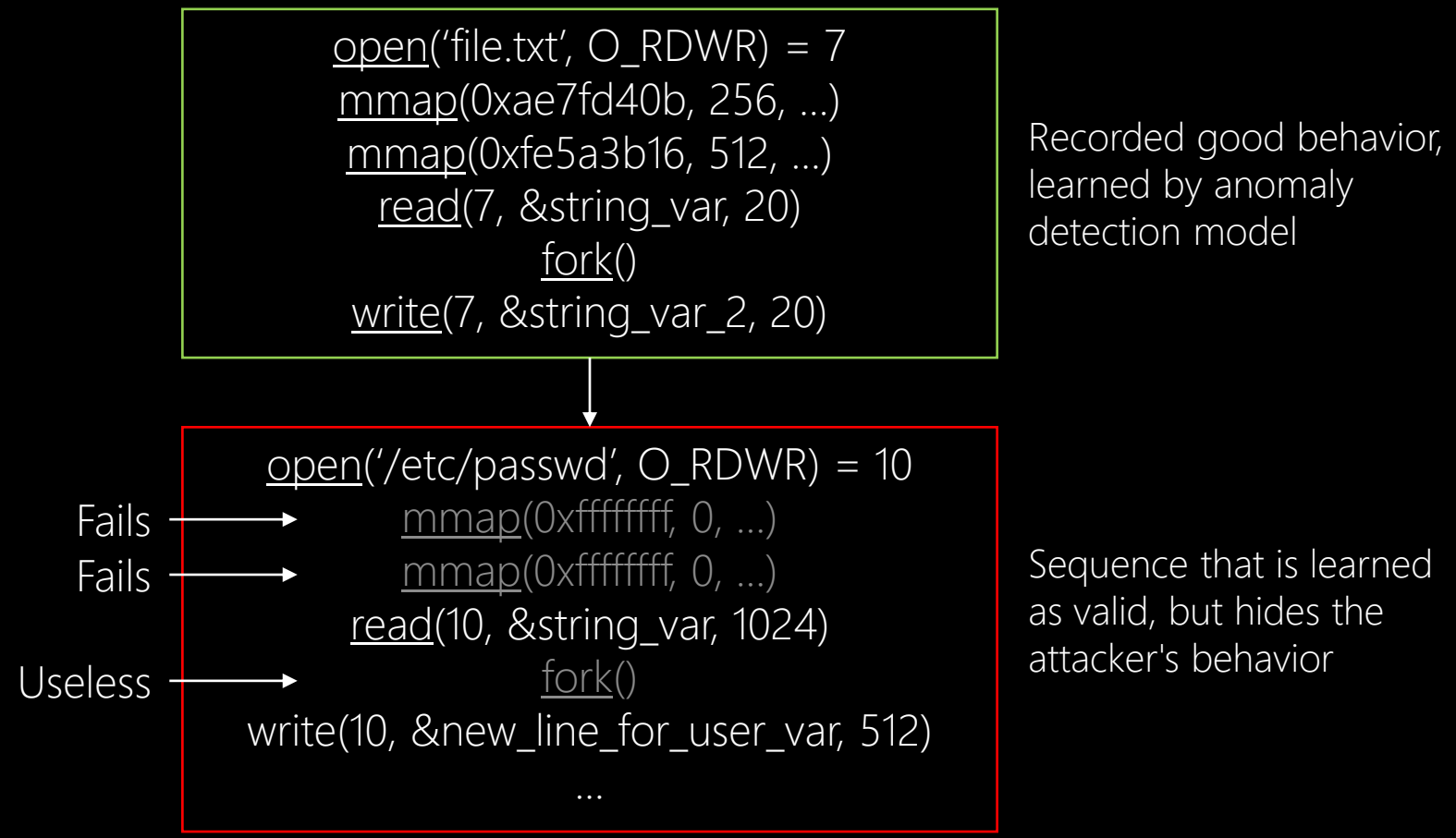
> Now, it’s up to you!

45min



> Closing remarks

This intrusion detection system is flawed (mimicry attacks)



> Closing remarks

AI helps solving a problem, but brings others

- It is difficult to explain anomaly scores (see **Explainable AI**)
- Need other mechanisms for forensics after an alert is raised
- AI models can be resource consuming and hardly integratable as is, especially in embedded systems (see **Pruning**, **Quantization**, **Distillation**)
- AI models are vulnerable to **data poisoning** and **adversarial attacks**
- **LLMs are not magic** 🤓☝

> Closing remarks

This EDR, as others, is still vulnerable to EDR evasion techniques

- Payloads can **behave too similarly** to known benign processes
- System calls' **hooking mechanism** can be altered
- **Indirect system calls** can be made

A 6-months internship is available in CEA Grenoble on bluetooth fuzzing
→ ulysse.vincenti@cea.fr

A 6-months internship is available in CEA Grenoble on the hardening of an EDR for embedded systems, similar to what has been seen in this workshop.
→ lalie.arnoud@cea.fr

> Thank you for participating in this workshop!

If you have feedbacks, questions, or just want to drink a beer, don’t hesitate to reach us:

lalie.arnoud@cea.fr
ulysse.vincenti@cea.fr