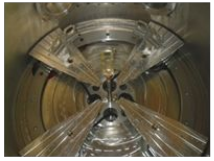


SINAPSE

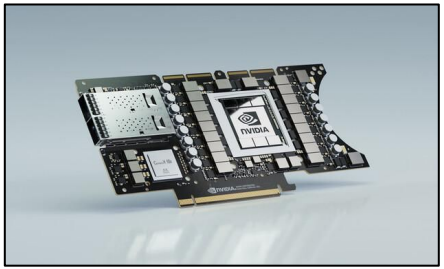
SImulateur **N**umérique pour l'**A**pprentissage **P**rofond sur **S**ystème **E**mbarqué

Guillaume Cubéro (AI), David Etasse (IR),
Adrien Matta (CR), Isabelle Yvon (AI),
Thomas Carreau (CDD-IR spécialiste en IA)

etasse@lpccaen.in2p3.fr

Funded by the European Union

FASTERv3

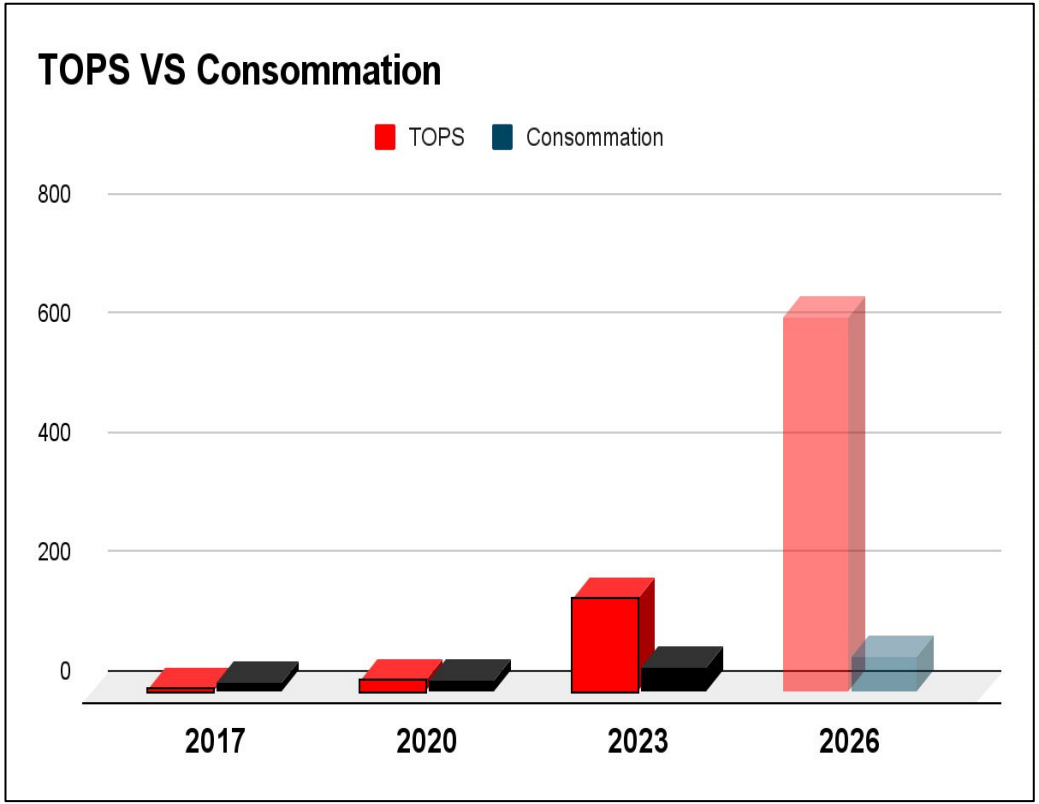


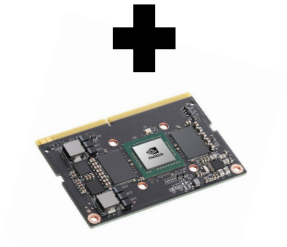
Ubuntu 22.04
Cuda 10
PyTorch
TensorRT
C/C++
TCP/IP
DPDK
RoCEv2

Jetson Orin Nx 16GB

- 1 CPU 8 cores Arm 2,2 GHz
- 1 GPU 1024 Cores Ampere 1 GHz
- 16 GB memory (103 GB/s)

FASTERv3

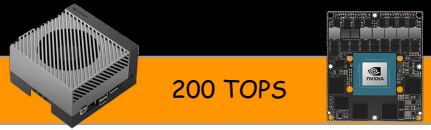


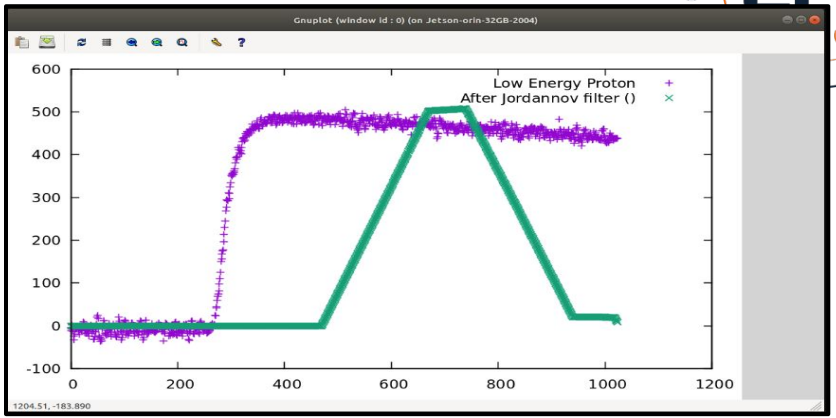


Jetson Orin Nx 16GB

- 1 CPU 8 cores Arm 2,2 GHz
- 1 GPU 1024 Cores Ampere 1 GHz
- 16 GB memory (103 GB/s)

FASTERv3 → CPU vs GPU processing

	Jetson AGX Orin 32GB 
AI Performance	200 TOPS
GPU	2048-core@1.3GHz NVIDIA Ampere Architecture 16 MP, 128 CUDA Cores/MP Max number of threads per MP: 1536 1536 * 16 = 24 576
GPU	48 Tensor Cores
CPU	12-core Arm® Cortex®-A78AE
Memory	32GB 256-bits LPDDR5 204.8GB/s
PCIe	2 x8 + 1 x4 + 2 *1 (PCIe Gen4, Root Port, & Endpoint)
Power	10 W – 40 W
Cost (VAT)	1238 €



100 000 Frames of 1024 samples
In double precision 819 MB
Processing a BLR, Moving Average
And Jordanov filter
On jetson Orin AGX 32 GB

	12 CPUs	GPU (Shared memory)	GPU (Dedicated memory)
CPP	1330 ms		
C	400 ms (50us)	330 ms	300 us (75 us)

FASTERv3

FASTERv3

High-throughput,
Embedded ML processing,
Centralized deployment

Data acquisition,
Slow control (EPICs),
synchronization (IJCLab WR)

daniel.charlet@ijclab.in2p3.fr
antoine.back@ijclab.in2p3.fr

User Algorithms

SINAPSE
2025 → 2027

SImulateur **N**umérique pour l'**A**pprentissage **P**rofond sur **S**ystème **E**mbarqué
Digital SImulator for Deep Learning on Embedded Systems

→ 4 LPC Caen, 1 CDD [2 years], 2 years, 150K€

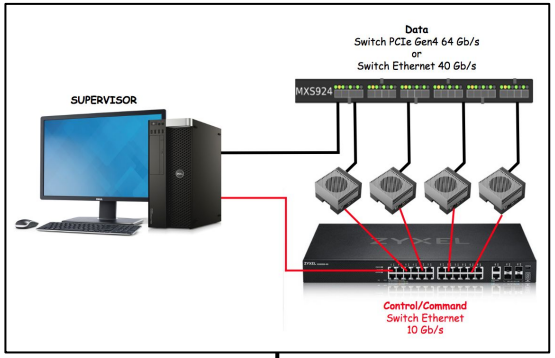
Board Design, Firmware,
VHDL Algorithms, Synchronization

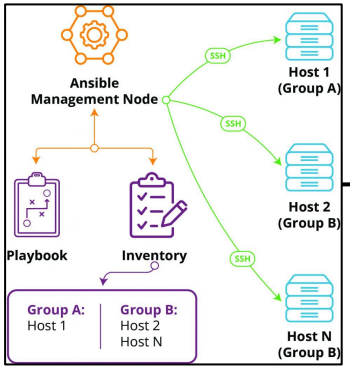
→ 5 LPC Caen, 1 CDD [2 years], X years, FASTERv2 funding

ML or Traditional Algorithms

→ LPC Caen or other laboratories

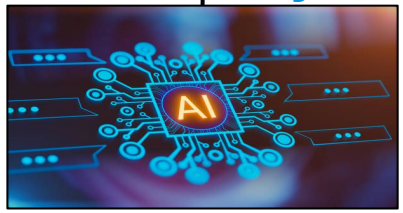
SINAPSE





SINAPSE

DVC mlflow



RÉGION NORMANDIE

Funded by the European Union

Guillaume Cubéro (AI), David Etasse (IR), Adrien Matta (CR), Isabelle Yvon (AI), Thomas Carreau CDD-IR

SINAPSE: Deep neural networks in FASTER v3 for online data analysis

T. Carreau [1] A. Matta [1] G. Cubero [1] D. Etasse [1]

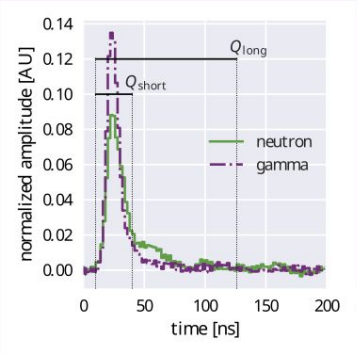
[1]LPC Caen, CNRS/IN2P3, Caen, 14000, France

November 19, 2025

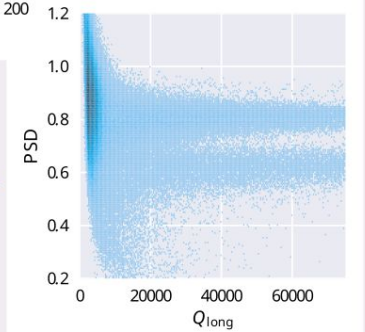
Context

Neutron-γ discrimination

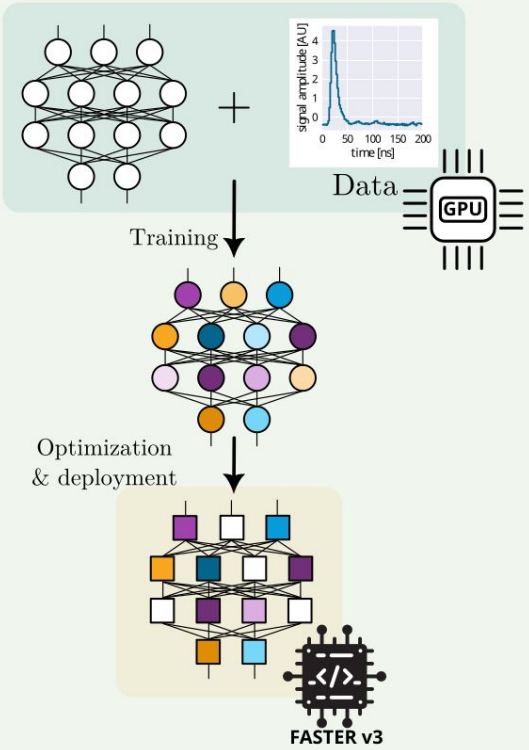
- Fission fragments emit both neutrons and γ rays.



$$\mathrm{PSD} = \frac{Q_{\mathrm{short}}}{Q_{\mathrm{long}}}$$



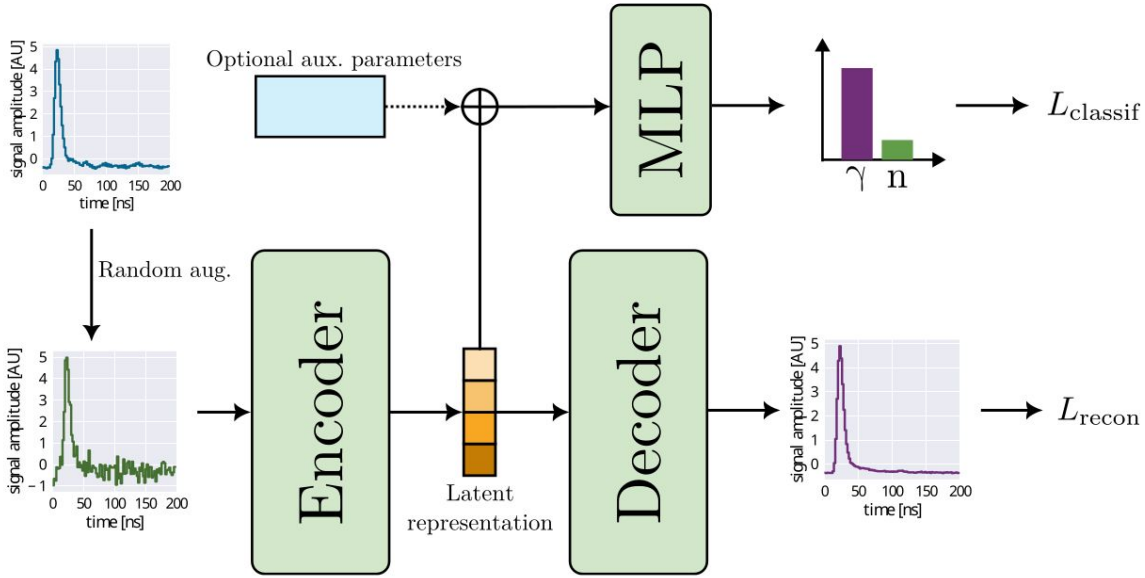
Neural networks in FASTER



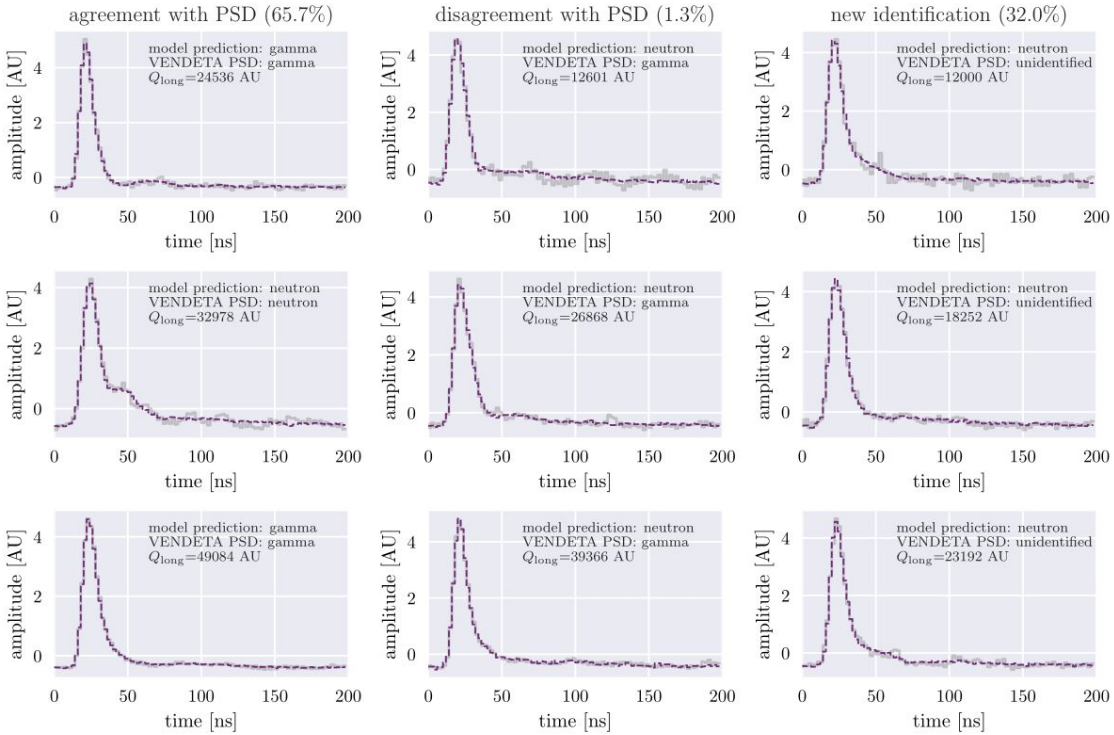
SINAPSE framework

Semi-supervised learning

Semi-supervised learning consists in training machine learning models using both labeled and unlabeled data.



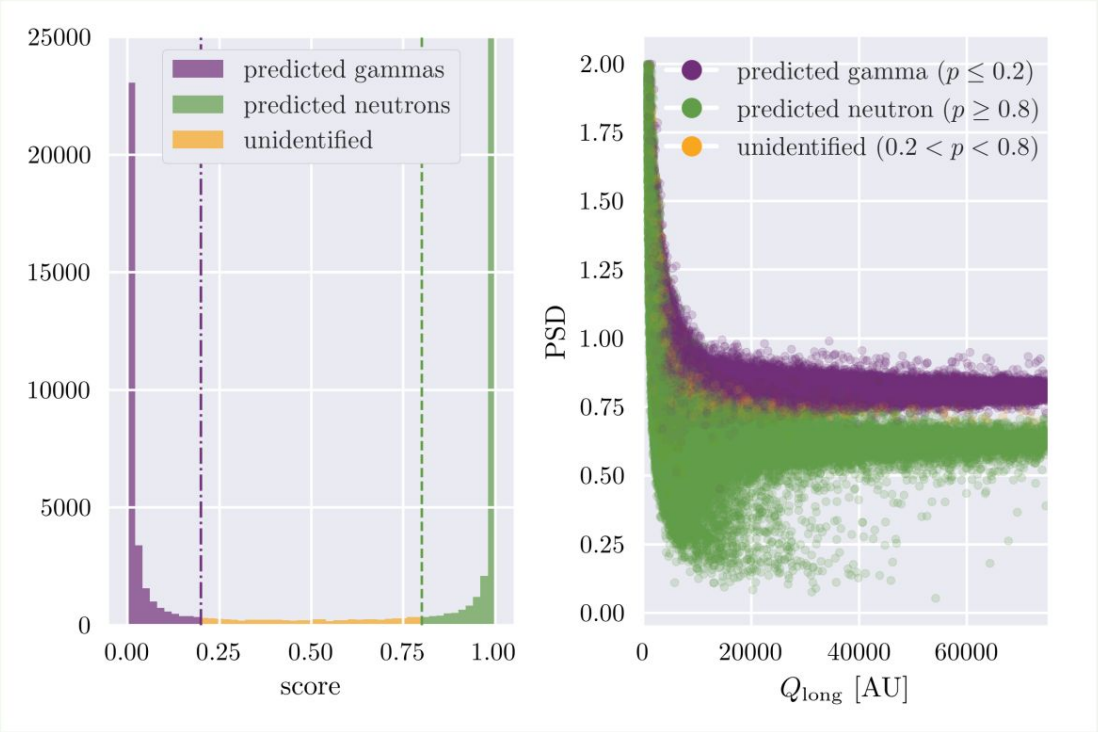
Signal denoising



$\longrightarrow$ Compression ratio = 25

Classification

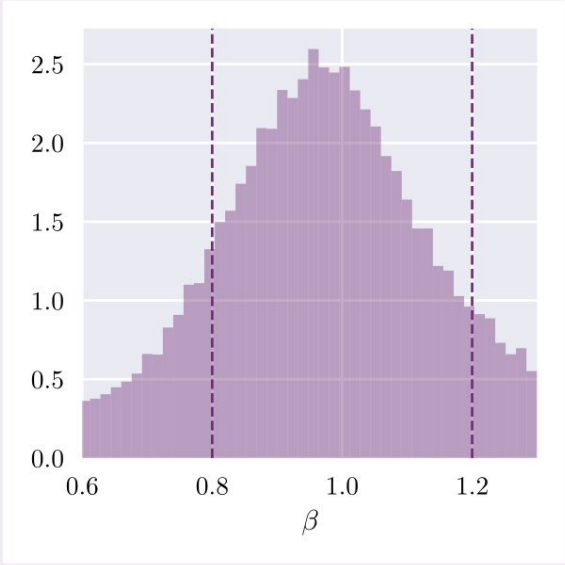
Results for ^{240}Pu (test set)



Validating the results

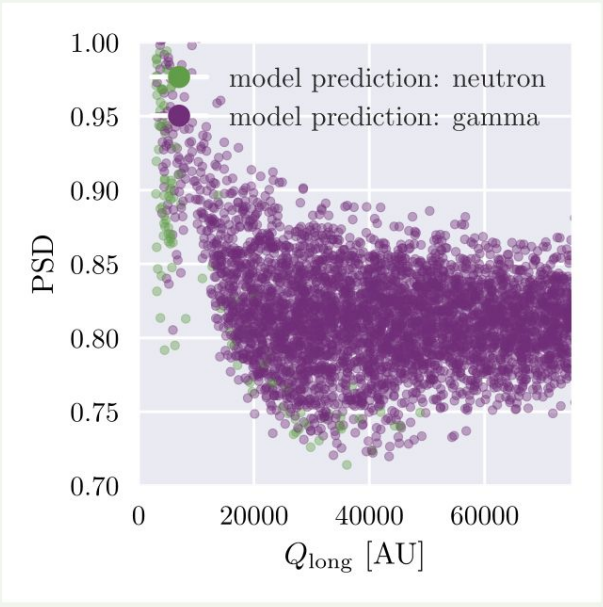
Prompt γ selection

$$\beta = \frac{1}{c}\frac{d_{\text{det}}}{\text{ToF}} \qquad (1)$$



Performance on low-energy γ

- Accuracy = 98.3% $\rightarrow$ very few FN



Model interpretability

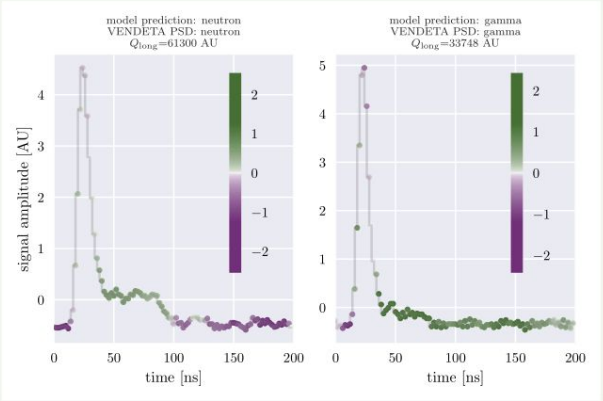
Saliency maps

Gradient of the score with respect to the input samples:

$$S(\mathbf{x}_0) = \left.\frac{\delta y}{\delta \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0} \tag{2}$$

Results

- Elevated baseline shifts the saliency score toward the γ class.
- Large time window is crucial for accurate neutron identification.
- Increasing the tail amplitude shifts the score toward the neutron class.



Conclusions & next steps

Conclusions

- Semi-supervised learning is a promising approach for neutron-γ discrimination.
- Using random augmentations allows for reliable extrapolation at low-energy and robust signal denoising.
- Model interpretability is permitted by saliency methods.

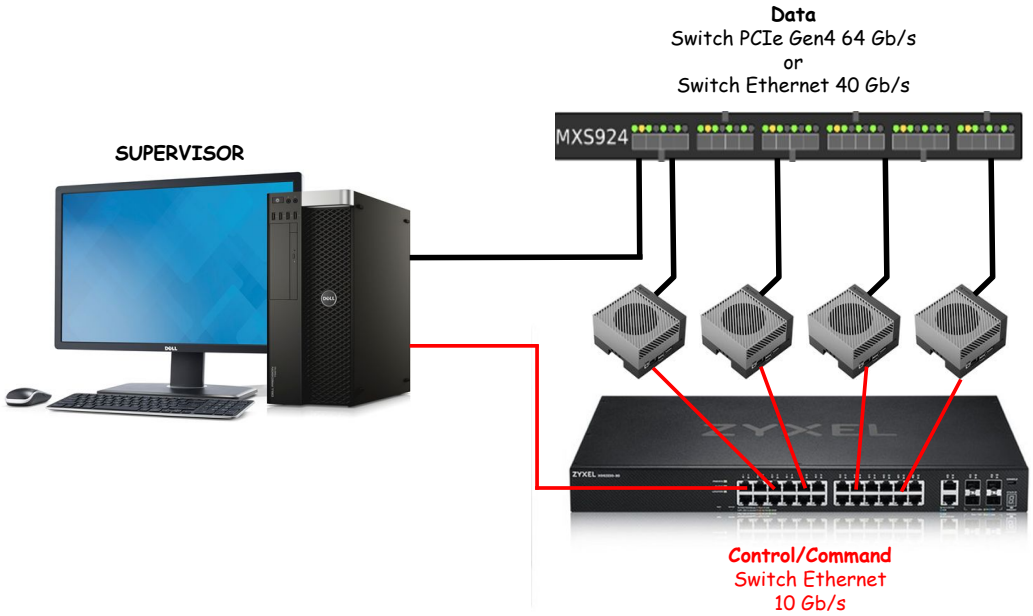
Next steps

- Enrich training data:
  - Low-energy non-annotated samples (semi-supervised learning)
  - Low-energy γ signals
- Optimize model inference for deployment on FASTER v3 (TensorRT, model quantization).
- PSA based particle discrimination with GRIT.

Bibliography I

- Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).
- Kingma, Diederik P et al. "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems* 27 (2014).
- Syrett, O et al. "VENDETA: VErsatile Neutron DETector Array, a new high-resolution neutron time-of-flight measurement array". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* (2025), p. 170824.
- *FASTER*. https://faster.in2p3.fr/.
- NVIDIA. *TensorRT*. https://github.com/NVIDIA/TensorRT.

SINAPSE → High throughput packet processing platform



Low-latency, high-throughput packet processing platform

- DPDK →**D**ata **P**lane **D**evelopment **K**it
- RoCEv2→**R**dma **o**ver **C**onverged **E**thernet

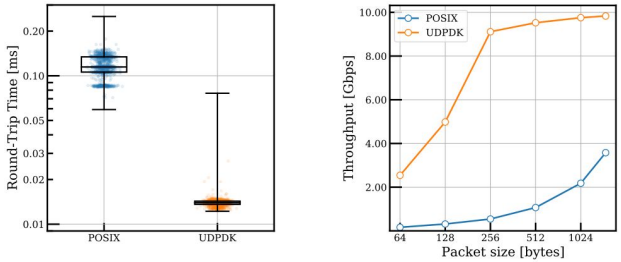
- Linux Kernel Networking (Ssh, Web, TCP/IP …)

SINAPSE →DPDK → Summary

Data Plane Developement Kit → https://www.dpdk.org/

How does DPDK improve packet processing ?

- **Bypasses the kernel, Enables fast packet processing in userspace**
- **Poll Mode Driver** (PMD) → uses polling instead and avoids the overhead associated with interrupts.
- **Lockless Sync** → Queues are managed with the ring library. Enqueue and dequeue operations are lockless
- **Zero copy**
- **Processor Affinity** →Ties specific processing to specific cores
- **Huge Pages, I/O Batch**
- **NUMA Aware, Cache Alignment,**
- **No networking stack** implementation



(a) (b)

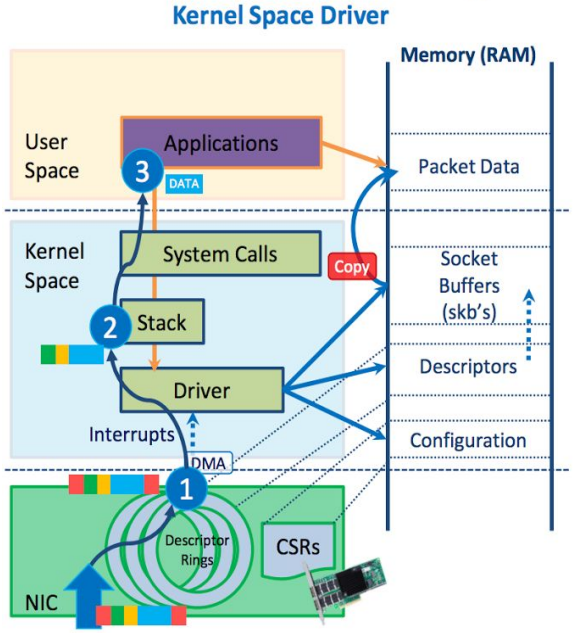
DPDK versus POSIX performance: (a) latency; (b) throughput. Source: Lai et al. 2021, fig. 4.

Drawback !

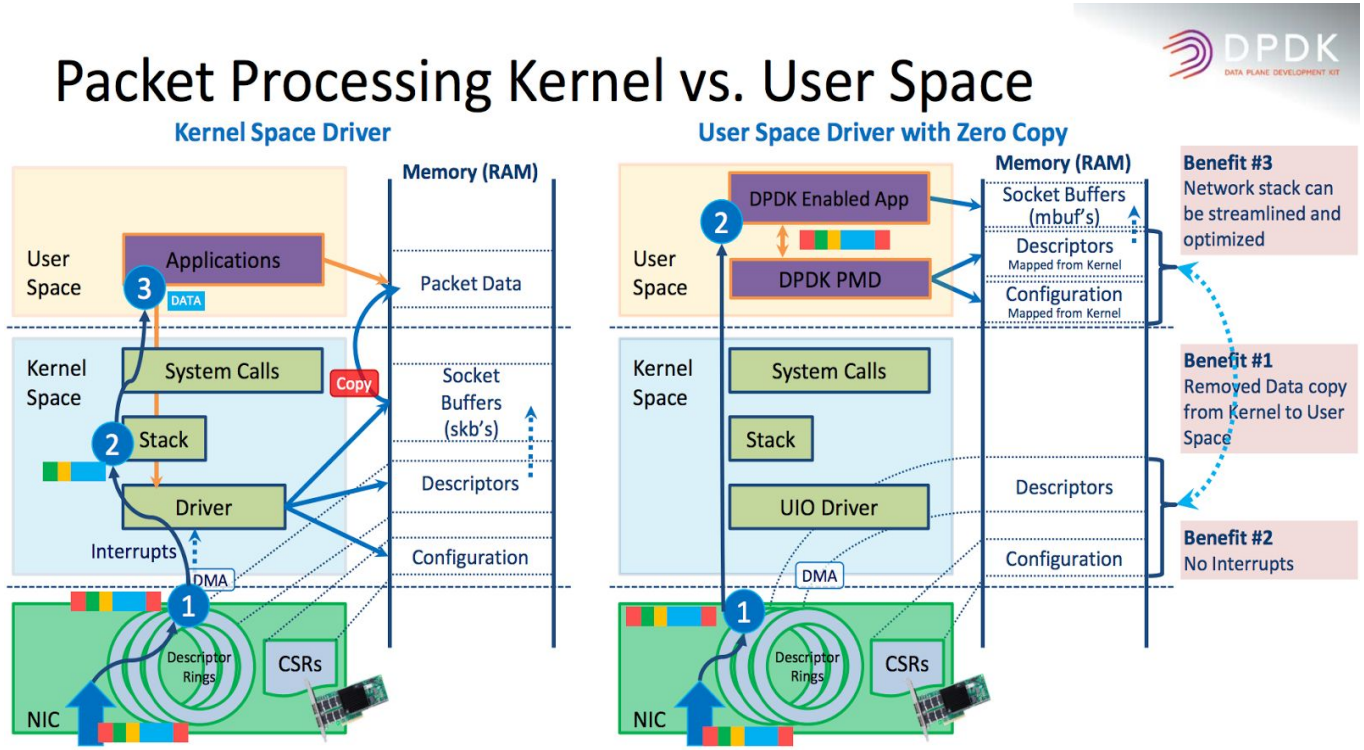
- DPDK causes 100% CPU usage even when only a few packets are being processed.

Polling is used instead of interrupts

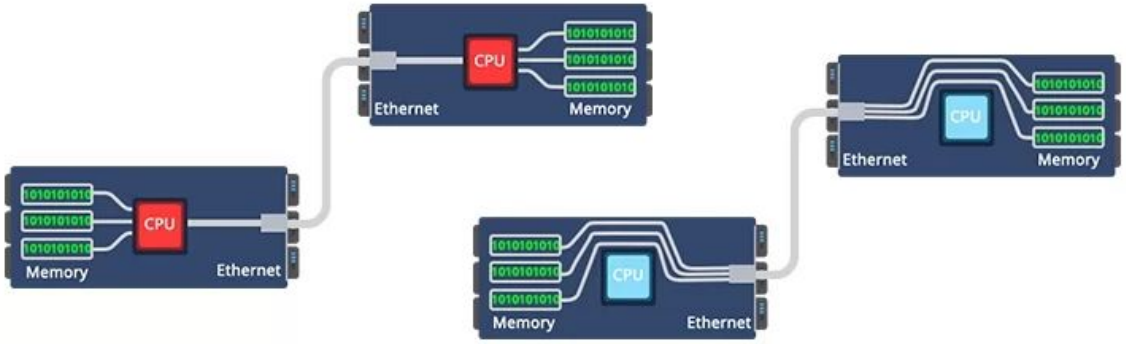
Packet Processing Kernel vs. User Space



SINAPSE →DPDK → Kernel vs User space packet processing



Remote Direct Memory Access



Before Vs. After RoCE

Remote Direct Memory Access

RDMA is a hardware mechanism through which the network card (NIC) can directly access all or parts of the main memory of a remote node without involving the processor.

- **Remote** → data is transferred between nodes in a network
- **Direct** → no CPU or OS kernel is involved in the data transfer
- **Memory** → data transferred between two apps and their virtual address spaces
- **Access** → One sided operations (read, write, and do atomic operations)
  → Two sided operations (send, receive)

RDMA communication operation is based on a set of three queues :

- **Send queue (SQ)**
- **Receive queue (RQ)**

} Used to schedule the work to be done (Queue Pair →QP) .

- **Completion queue (CQ)**→ Used to notify when the work has been completed (Optional).

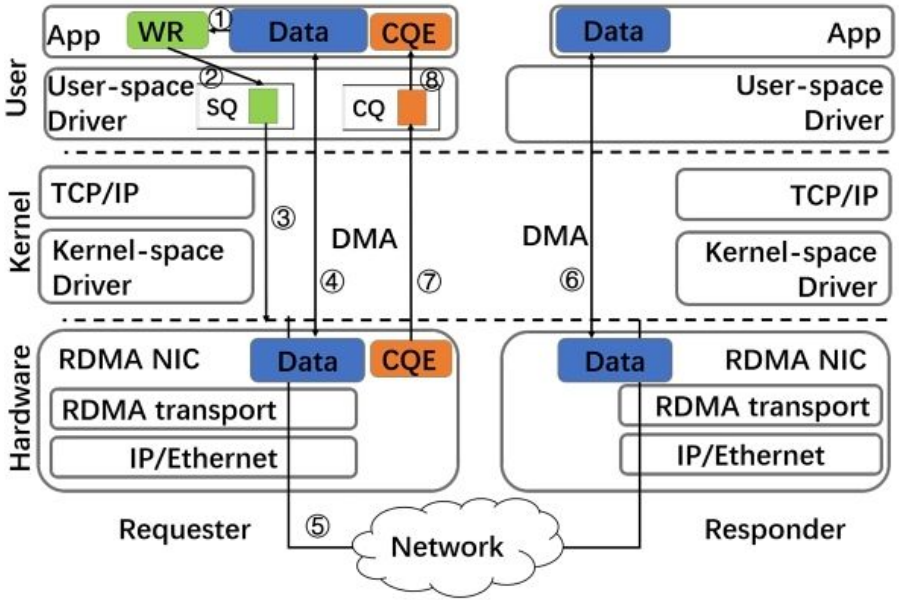
RDMA transport layer → Reliable Connection (RC), Unreliable Connection (UC), Unreliable Datagram (UD)

RDMA network layer→ IB, RoCEv1, RoCEv2, iWARP

https://www.doc.ic.ac.uk/~jgiceva/teaching/ssc18-rdma.pdf

SINAPSE → RDMA → Example

The workflow of RDMA transport

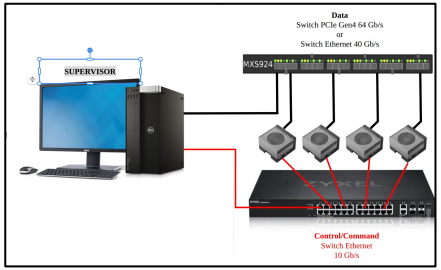


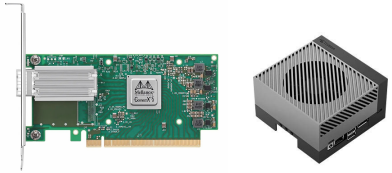
WR → Work Request
SQ → Send Queue
CQ → Completion Queue
CQE → Completion Queue Element

https://dl.acm.org/doi/fullHtml/10.1145/3600061.3600082

SINAPSE → From Software to Hardware

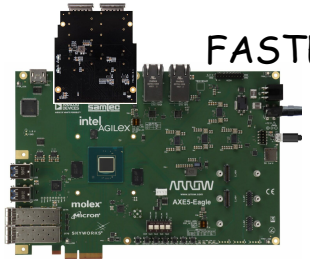
SINAPSE





NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE



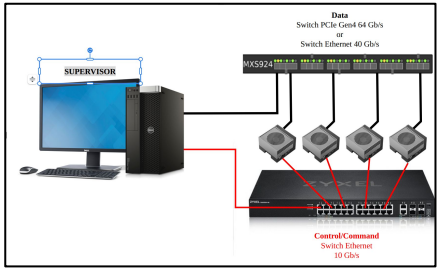
- Linux Kernel Networking (Ssh, Web, TCP/IP …)
  1 →1/10Gbe Bridge PCIe/Ethernet
  2 →40Gbe Bridge PCIe/Ethernet
  https://docs.corundum.io/en/latest/

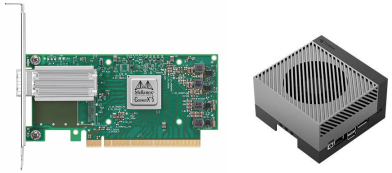
NAT-MCH-G4
Up-link → 2x 1/10/25G
DownLink → 12x 1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x 100G or 8x 25G
Downlink → 12x 40G or 12x 4x 1/10G

SINAPSE → From Software to Hardware

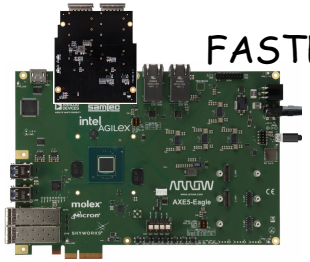
SINAPSE

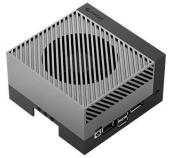




NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE





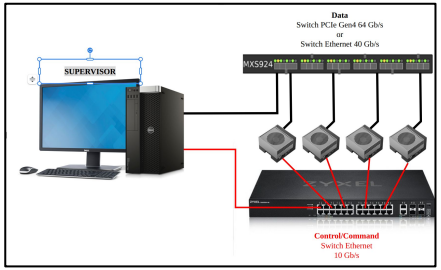
- Linux Kernel Networking (Ssh, Web, TCP/IP …)
  1 →1/10Gbe Bridge PCIe/Ethernet
  2 →40Gbe Bridge PCIe/Ethernet
  https://docs.corundum.io/en/latest/
- DPDK →**D**ata **P**lane **D**evelopment **K**it
  3 →Added **DPDK** feature to 1/10/40Gbe Bridge
  https://github.com/liuw666-bruce/DPDK-with-Corundum
  4 →Added **SR-IOV** feature to 1/10/40Gbe Bridge
  https://github.com/UIUC-ChenLab/OS4C

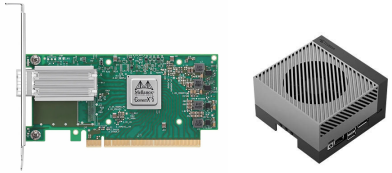
NAT-MCH-G4
Up-link → 2x 1/10/25G
DownLink → 12x 1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x 100G or 8x 25G
Downlink → 12x 40G or 12x 4x 1/10G

SINAPSE → From Software to Hardware

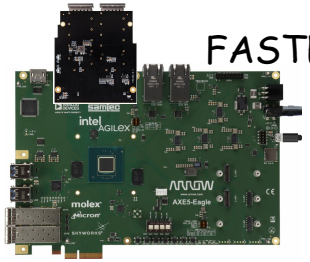
SINAPSE

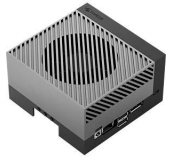




NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE





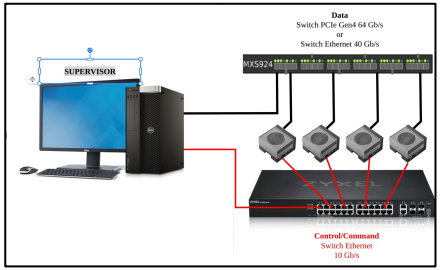
- Linux Kernel Networking (Ssh, Web, TCP/IP …)
  1 →1/10Gbe Bridge PCIe/Ethernet
  2 →40Gbe Bridge PCIe/Ethernet
  https://docs.corundum.io/en/latest/
- DPDK →**D**ata **P**lane **D**evelopment **K**it
  3 →Added **DPDK** feature to 1/10/40Gbe Bridge
  https://github.com/liuw666-bruce/DPDK-with-Corundum
  4 →Added **SR-IOV** feature to 1/10/40Gbe Bridge
  https://github.com/UIUC-ChenLab/OS4C
- RoCEv2 →**R**dma **o**ver **C**onverged **E**thernet
  5 →Added **RoCEv2** feature 1/10/40Gbe Bridge
  https://www.bittware.com/partners/grovf-rdma/
  https://github.com/ETH-PLUS/Jingzhao

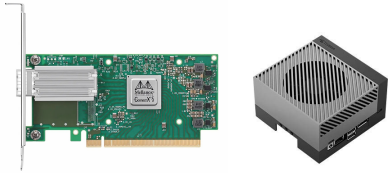
NAT-MCH-G4
Up-link → 2x 1/10/25G
DownLink → 12x 1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x 100G or 8x 25G
Downlink → 12x 40G or 12x 4x 1/10G

SINAPSE → From Software to Hardware

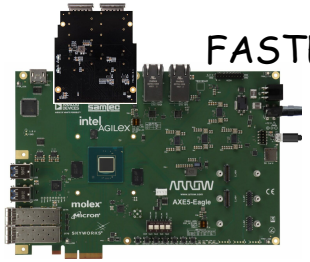
SINAPSE

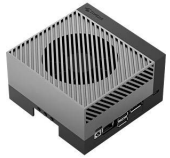


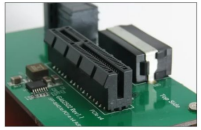


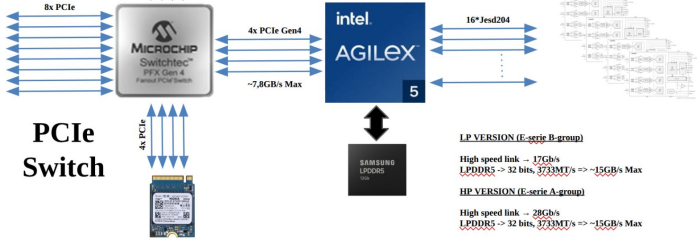
NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE

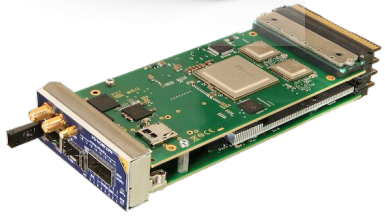








2 or 4 TB

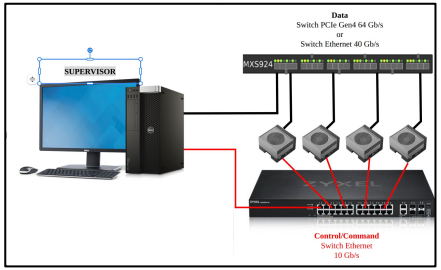


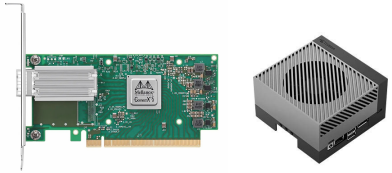
NAT-MCH-G4
Up-link → 2x1/10G
DownLink → 12x1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x100G or 8x25G
Downlink → 12x40G or 12x4x1/10G

SINAPSE → From Software to Hardware

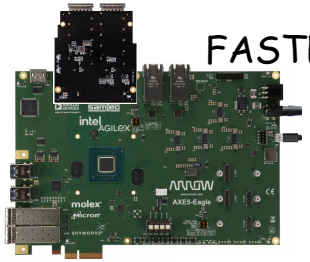
SINAPSE

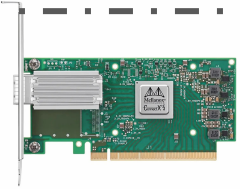


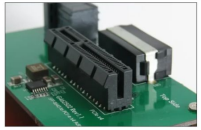


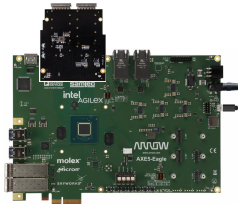
NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

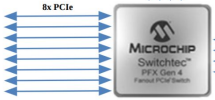
FASTERV3-HARDWARE











PCIe Switch

2 or 4 TB

Data Read out

Digitizers

LP VERSION (E-serie B-group)

High speed link → 17Gb/s
LPDDR5 -> 32 bits, 3733MT/s => ~15GB/s Max

HP VERSION (E-serie A-group)

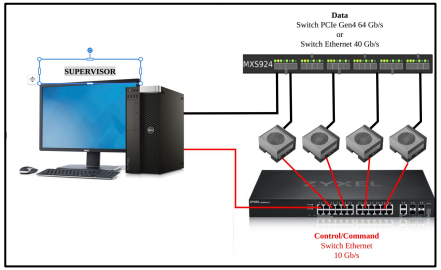
High speed link → 28Gb/s
LPDDR5 -> 32 bits, 3733MT/s => ~15GB/s Max

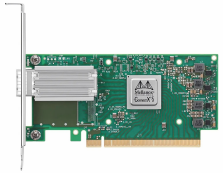
NAT-MCH-G4
Up-link → 2x1/10G
DownLink → 12x1/10G

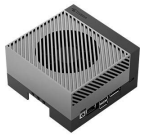
NAT-MCH-G4-HUB-EX
Up-link → 2x100G or 8x25G
Downlink → 12x40G or 12x4x1/10G

SINAPSE → From Software to Hardware

SINAPSE

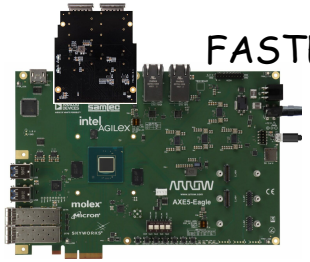






NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE

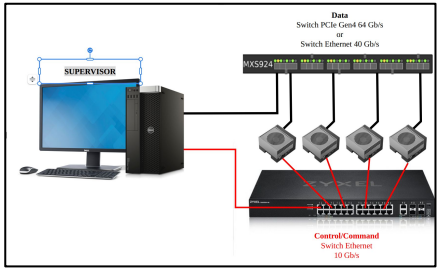


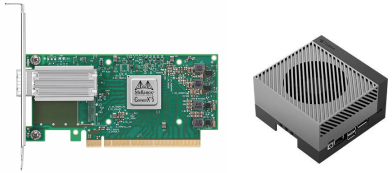
NAT-MCH-G4
Up-link → 2x1/10G
DownLink → 12x1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x100G or 8x25G
Downlink → 12x40G or 12x4x1/10G

SINAPSE → From Software to Hardware

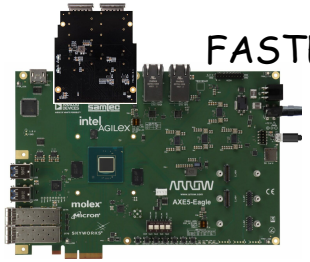
SINAPSE

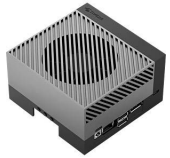


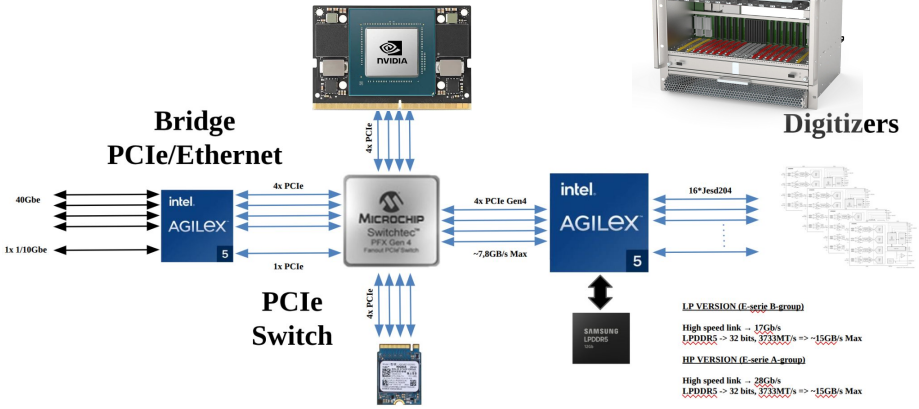


NVIDIA Mellanox MCX515A-CCAT ConnectX®-5
Jetson AGX Orin 64GB

FASTERV3-HARDWARE







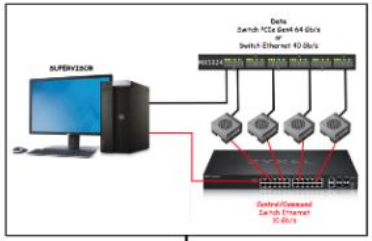
Digitizers

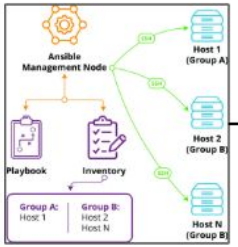
NAT-MCH-G4
Up-link → 2x1/10G
DownLink → 12x1/10G

NAT-MCH-G4-HUB-EX
Up-link → 2x100G or 8x25G
Downlink → 12x40G or 12x4x1/10G

SINAPSE → It's an opportunity

1. Break new ground in real-time data analysis in nuclear physics,
2. Attract an ML specialist to Normandy,
3. Develop ML skills within a Normandy laboratory,
4. Develop the software part of a high-throughput packet processing platform,
5. Add emerging techniques to an existing project,
6. Present the results in order to raise public awareness of these new technologies.





RÉGION NORMANDIE

Guillaume Cubéro (AI), David Etasse (IR),
Adrien Matta (CR), Isabelle Yvon (AI),
Thomas Carreau CDD-IR