# GPUs usage at L2IT

*Sylvain Caillou, Charline Rougier*

On behalf of L2IT

Réunion du CC-IN2P3 avec les expériences (virtuelle) : les GPUs

April 30th 2021

# A new lab’ in Toulouse

**« Le laboratoire des deux infinis – Toulouse » (L2IT) »**

- Created as a « FRE » on  1st September 2019
  - 4 members (scientists)
  - *Tutelles*: CNRS (IN2P3) and Université Paul Sabatier

- It is accepted as an « UMR » on January 1st 2020

- As of today: 9 scientific members

- We don’t have (and don’t plan to have in the near future) any infrastructure locally and we rely in the CC-IN2P3 for all of our needs (thanks to them).

# The research teams at L2IT

**Particle physics**

. Higgs boson (CERN)

. Now ->2040:
precise study of Higgs boson

. Commitment in ITk track reconstruction

5 scientific members

**Gravitationnal waves**

. New windows on the Universe

. Virgo, LISA

1 -> 6 scientific members this automn

**Nuclear physics (GANIL)**

. Equation of state of nuclear matter

. Study of nuclear interaction in laboratory

1 -> 2 scientific members this automn

**Key aspect of L2IT:**

Innovative algorithm and simulation

Computing and software

**‘Calcul, algorithmes et données’ (CAD)**

2 engineers + 1 doc ATLAS/CAD

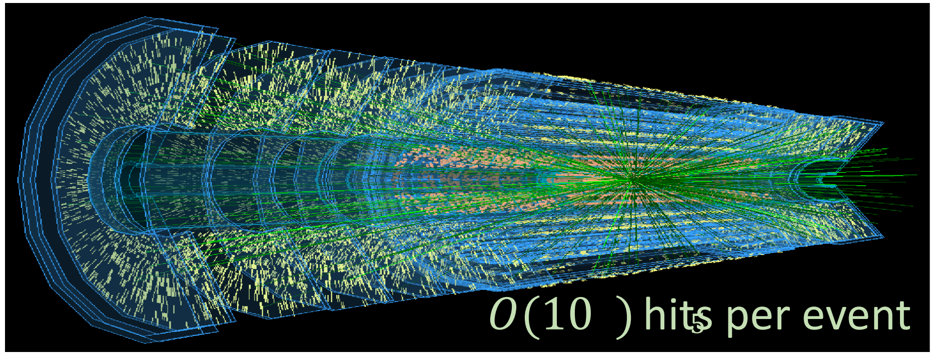
First project: track reconstruction in ATLAS (HL-LHC)

To be defined: LISA commitment

# GNN for track pattern recognition at HL-LHC

- Physics reach during HL-LHC will be limited by affordable software and computing and by how efficiently these resources can be used.

$\Rightarrow$ Study new track finding algorithm.

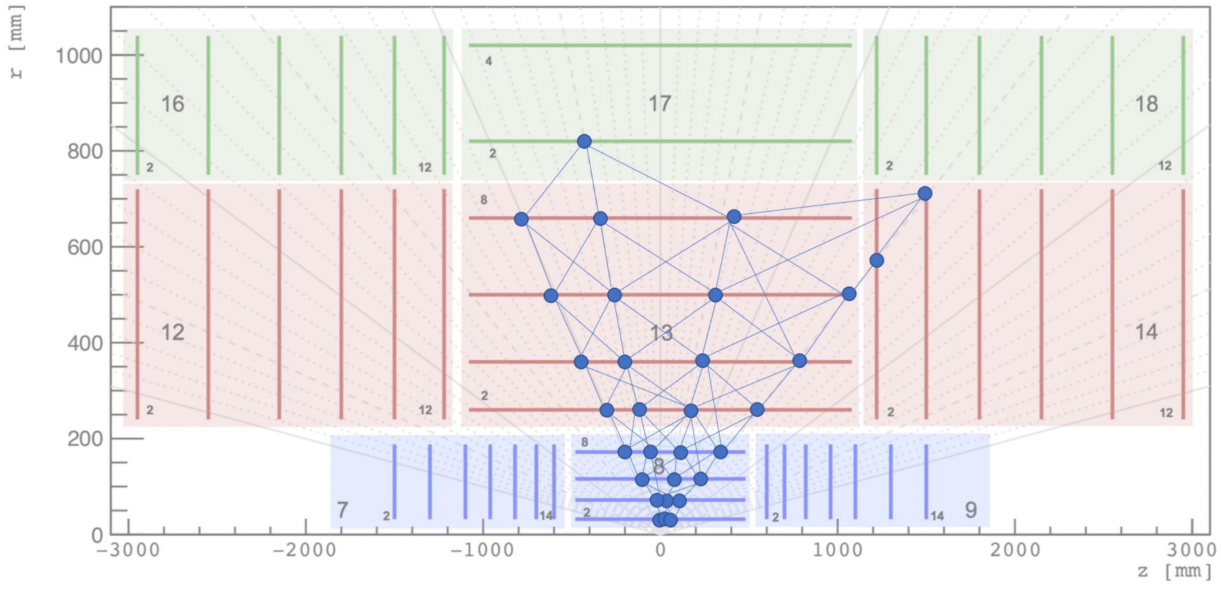


$O(10^5)$ hits per event

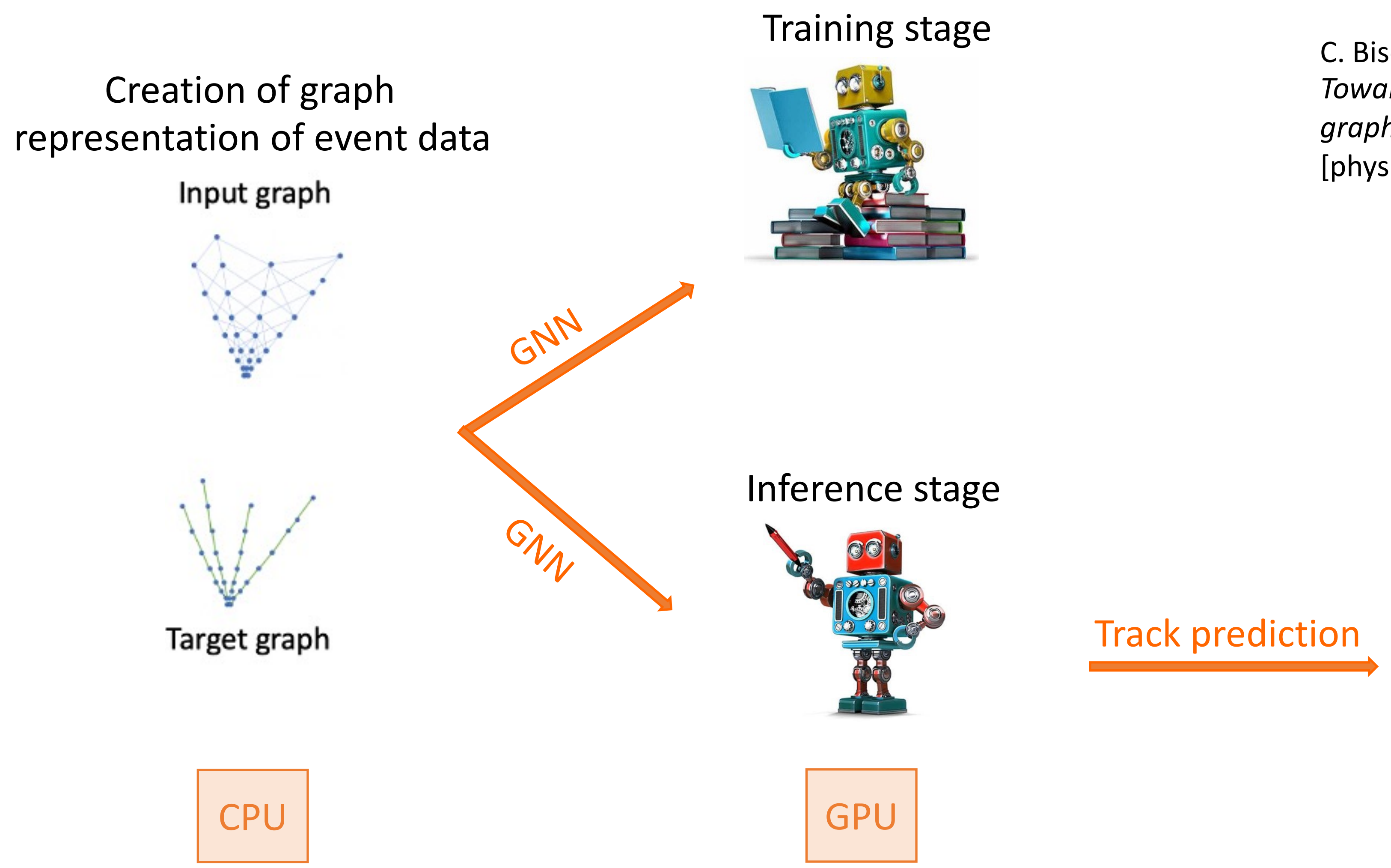
- New effort at L2IT to implement a realistic GNN-based algorithm that can be deployed in an HL-LHC experiment.
- Proof of principle from Exa.Trkx (arXiv:2003.11603)

# GNN for track pattern recognition at L2IT

- Graphs representation of data:
  - Nodes = hits
  - Edges = two potential successive hits on a track



# GNN for track pattern recognition at L2IT


Creation of graph representation of event data

Input graph

Target graph

GNN

GNN

Training stage

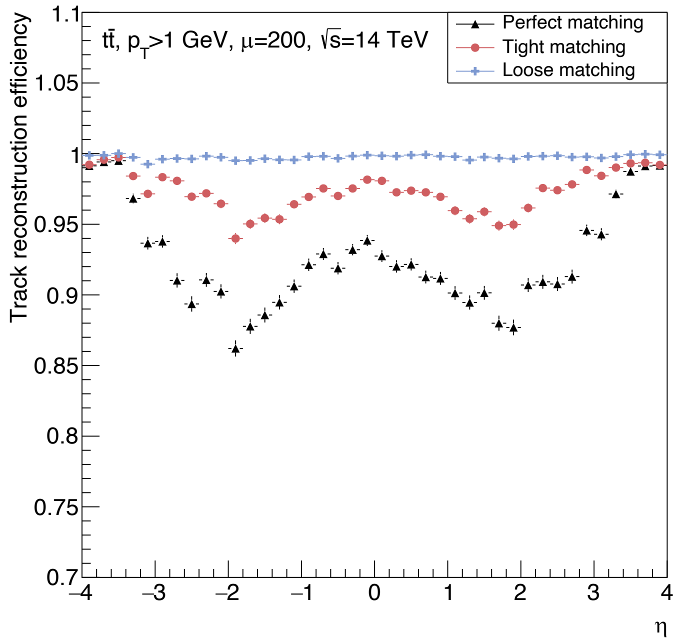
Inference stage

Track prediction

CPU

GPU

C. Biscarat, S. Caillou, C. Rougier, J. Stark, J. Zahreddine, *Towards a realistic track reconstruction algorithm based on graph neural networks for the HL-LHC,* arXiv:2103.00916 [physics.ins-det] (**vCHEP2021**).



Track reconstruction efficiency

# Specific memory need

1 graph: $10^5$ nodes and $10^6$ edges

Automatic differentiation

Complexity of the model

With simple model architecture:

- Training stage: $\sim 100\ \mathrm{GB}$
- Inference stage: $\sim 6\ \mathrm{GB}$

Currently, **training** doesn’t fit on usual GPUs:

- Use of IBM TFLMS (ref) (tensor swapping : GPU ⟷ CPU host)

  => Need of large memory GPUs with large CPU host memory e.g. one single tensor could be O(10) GB
- Use of 2 Nvidia Quadro RTX 8000 with 48 GB memory and 1 TB CPU host (thanks to the CC-IN2P3 ✓ )
  - The commissioning of this new type of GPUs required efforts (CC-IN3P3 and L2IT sides)

# Training stage

- Training with a simple model architecture:
  - Memory Peak = 115 GB (48 GB on RTX 8000 & 67 GB on host)
  - Batch size = 1
  - Runtime = a week
- More complex model: runtime ~ one month
- Benchmark of different kinds of GPU on reduced detector graphs
- Thanks to Huma-Num for A100 access ✓

| | Tesla V100 | Quadro RTX 8000 | Ampere A100 |
|---|---|---|---|
| GPU Memory (GB) | 32 | 48 | 40 |
| Runtime precision 32 | 1 min 20 s / epoch | 1 min 20 s / epoch | 1 min 20 s / epoch |
| Runtime precision 64 | 2 min 54 s / epoch | 9 min 30 s / epoch * | 2 min 27 s / epoch |
| Performance | same | | |

* as expected

Results are not public, please do not share

# Inference stage

- Inference with a simple model architecture:
  - Memory peak = 5.4 GB
  - Runtime = 0.3 s /event

- Test on different GPU on full detector
- Thanks to CPPM for GeForce RTX 2080Ti access ✓

| | Quadro RTX 8000 | GeForce RTX 2080 Ti |
|---|---|---|
| GPU memory capacity (GB) | 48 | 11 |
| Runtime mixed precision (16/32) | 0.3 s / event | |
| Memory peak | 5.4 GB | |
| Physics performance | same | |

- **Possibility to run inference on cheaper GPUs**

Results are not public, please do not share

# What we learn on the way

- Memory limitation: it is possible that one computation tensor exceeds the memory of the RTX.
  - Cut it into multiple subsets
  - Now even the GPU memory is no longer a conceptual problem
  - More subtensors slow the training
  - Limitation: time computation

- Multi-GPUs: to speed-up the training stage:
  - Use of Horovod
  - On 2 GPUs: double the memory swap on the host

# L2IT needs

### ATLAS – well advanced

- Training stage: we will take advantages of more complex model (DL)
  ⇒increase model architecture complexity
  ⇒increase memory needs
  ⇒ swapping/cutting computation tensors slow down the training
- Inference stage: this stage is not bound to memory and we can run it on more casual ("gamer") GPUs
- What would be our needs for preparing the HL-LHC:
  - A few GPUs with high memory like the new Nvidia A100 80 GB mem. with large host dedicated to training stage
  - Casual GPUs integrated in a farm dedicated to inference in production

### Gravitational waves & Nuclear physics – to come

- The teams are being assembled today
- Discussions with the LISA project on a future commitment