Graph Neural Networks for charged-particle track reconstruction

Jan Stark
Laboratoire des 2 Infinis - Toulouse (L2IT)

European AI for Fundamental Physics Conference (EuCAIFCon 2024)
Amsterdam, April 30th – May 3rd 2024

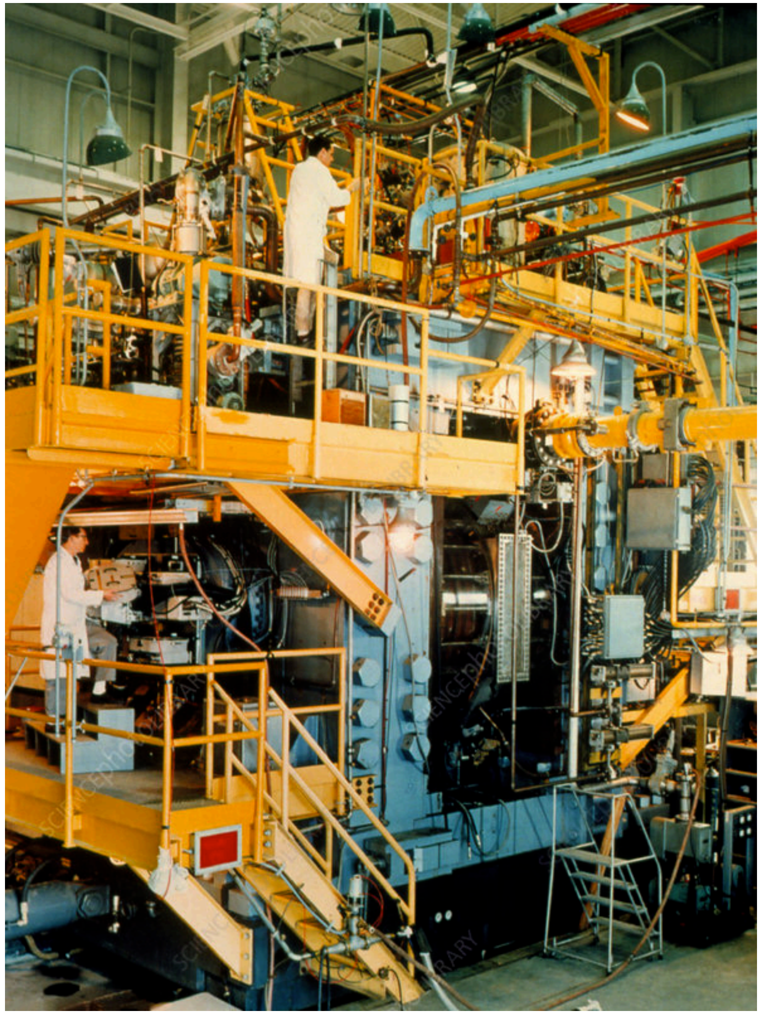
UNIVERSITE TOULOUSE III PAUL SABATIER

cnrs

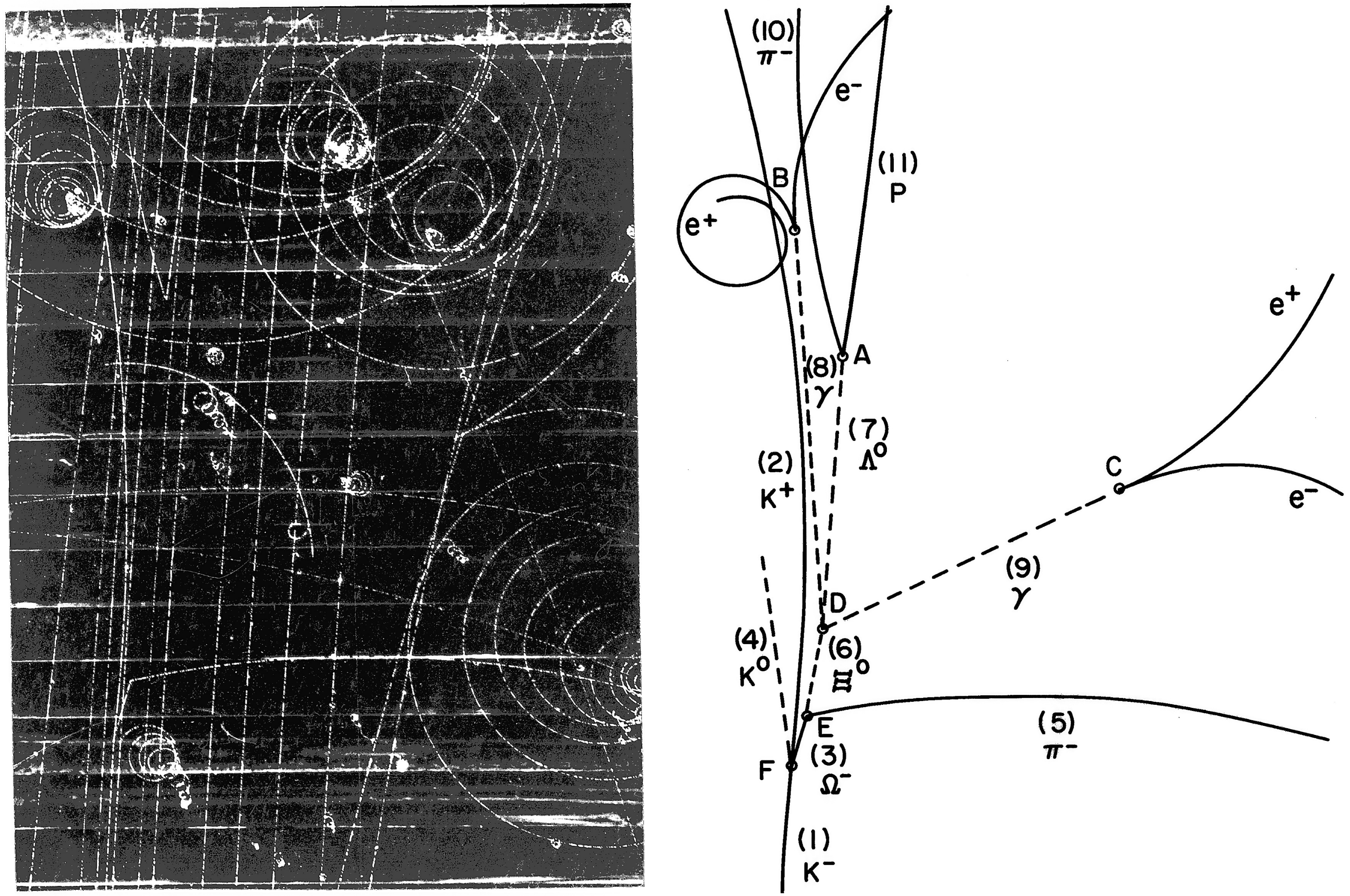
IN2P3
Institut national de physique nucléaire et de physique des particules

L2IT

Charged particle tracking – 1960s style



The 80 inch (2.0 m) bubble chamber at BNL

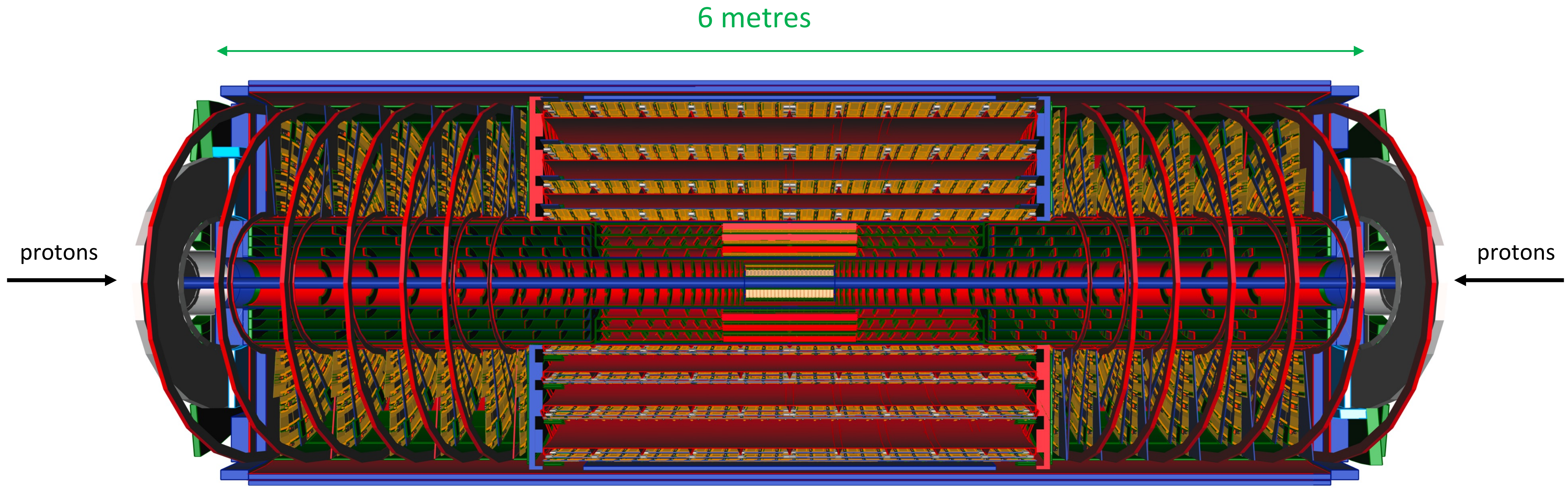


Discovery of the Omega-minus baryon in 1964

Today: silicon trackers

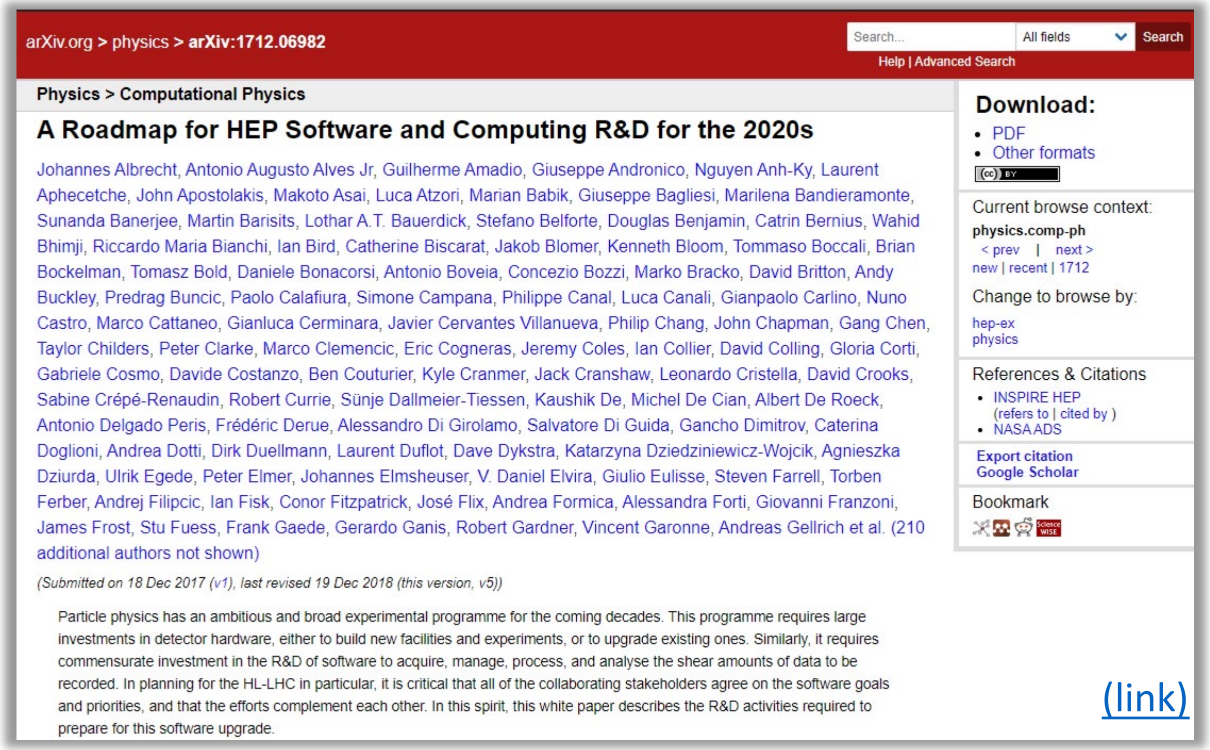
For example ATLAS ITk: new all-silicon tracker for the HL-LHC (data taking starting 2029)

27 thousand silicon modules
total area : ~180 m^2
~5 billon (5*10^9) readout channels



“10 years to prepare ourselves” for HL-LHC *(statement from 2017)*

- *Community white paper* (2017)
  - Algorithms, infrastructure, data access…
- Specific actions:
  - HEP Software Foundation (HSF)
  - Software Institute for Data-Intensive Sciences (SIDIS)
  - Creation of the Journal “Computing and software for big Science” (Springer)
  - IRIS-HEP (NSF project, US)
  - International project “Data Organization, Management and Access” (DOMA)
- The 2020 update of the EU strategy for particle physics


arXiv.org > physics > arXiv:1712.06982

Search... All fields Search
Help | Advanced Search

Physics > Computational Physics

A Roadmap for HEP Software and Computing R&D for the 2020s

Johannes Albrecht, Antonio Augusto Alves Jr, Guilherme Amadio, Giuseppe Andronico, Nguyen Anh-Ky, Laurent Aphecetche, John Apostolakis, Makoto Asai, Luca Atzori, Marian Babik, Giuseppe Bagliesi, Marilena Bandieramonte, Sunanda Banerjee, Martin Barisits, Lothar A.T. Bauerdick, Stefano Belforte, Douglas Benjamin, Catrin Bernius, Wahid Bhimji, Riccardo Maria Bianchi, Ian Bird, Catherine Biscarat, Jakob Blomer, Kenneth Bloom, Tommaso Boccali, Brian Bockelman, Tomasz Bold, Daniele Bonacorsi, Antonio Boveia, Concezio Bozzi, Marko Bracko, David Britton, Andy Buckley, Predrag Buncic, Paolo Calafiura, Simone Campana, Philippe Canal, Luca Canali, Gianpaolo Carlino, Nuno Castro, Marco Cattaneo, Gianluca Cerminara, Javier Cervantes Villanueva, Philip Chang, John Chapman, Gang Chen, Taylor Childers, Peter Clarke, Marco Clemencic, Eric Cogneras, Jeremy Coles, Ian Collier, David Colling, Gloria Corti, Gabriele Cosmo, Davide Costanzo, Ben Couturier, Kyle Cranmer, Jack Cranshaw, Leonardo Cristella, David Crooks, Sabine Crépé-Renaudin, Robert Currie, Sünje Dallmeier-Tiessen, Kaushik De, Michel De Cian, Albert De Roeck, Antonio Delgado Peris, Frédéric Derue, Alessandro Di Girolamo, Salvatore Di Guida, Gancho Dimitrov, Caterina Doglioni, Andrea Dotti, Dirk Duellmann, Laurent Duflot, Dave Dykstra, Katarzyna Dziedziniewicz-Wojcik, Agnieszka Dziurda, Ulrik Egede, Peter Elmer, Johannes Elmsheuser, V. Daniel Elvira, Giulio Eulisse, Steven Farrell, Torben Ferber, Andrej Filipcic, Ian Fisk, Conor Fitzpatrick, José Flix, Andrea Formica, Alessandra Forti, Giovanni Franzoni, James Frost, Stu Fuess, Frank Gaede, Gerardo Ganis, Robert Gardner, Vincent Garonne, Andreas Gellrich et al. (210 additional authors not shown)

(Submitted on 18 Dec 2017 (v1), last revised 19 Dec 2018 (this version, v5))

Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

Download:
- PDF
- Other formats

Current browse context:
physics.comp-ph
< prev | next >
new | recent | 1712

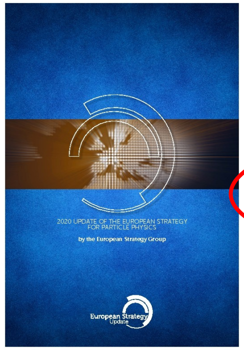
Change to browse by:
hep-ex
physics

References & Citations
- INSPIRE HEP (refers to | cited by)
- NASA ADS

Export citation
Google Scholar

Bookmark

(link)



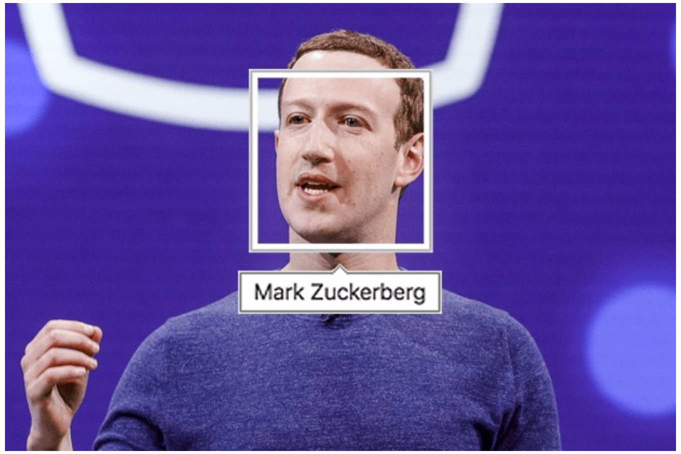
D. Large-scale data-intensive software and computing infrastructures are an essential ingredient to particle physics research programmes. The community faces major challenges in this area, notably with a view to the HL-LHC. As a result, the software and computing models used in particle physics research must evolve to meet the future needs of the field. ***The community must vigorously pursue common, coordinated R&D efforts in collaboration with other fields of science and industry, to develop software and computing infrastructures that exploit recent advances in information technology and data science. Further development of internal policies on open data and data preservation should be encouraged, and an adequate level of resources invested in their implementation.***

(link)

Machine learning for track pattern recognition ?

Challenge on Kaggle platform (in 2018): (link)

Article in proceedings of CHEP 2018: (link)

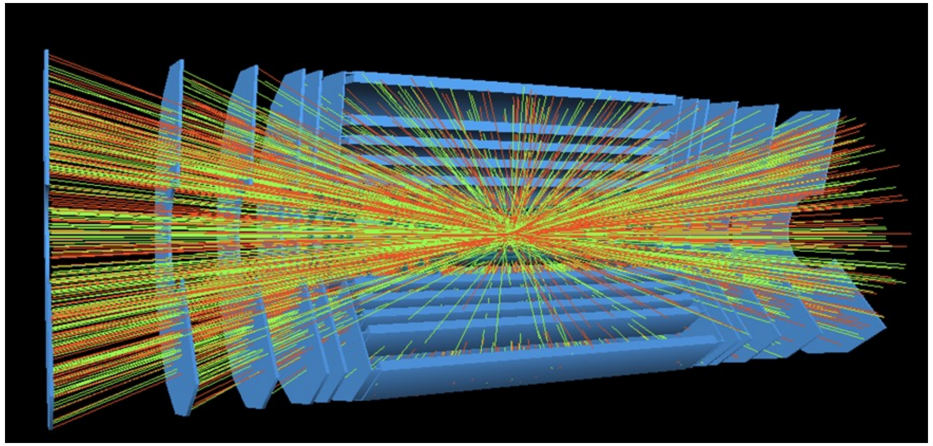


622 * 415 pixels

a large fraction carries information about the person

Can't use the same tools

How to present tracking data to a neural network ?



ATLAS tracker for HL-LHC:
$5 * 10^9$ readout channels
~$3 * 10^5$ 3D space-points per event

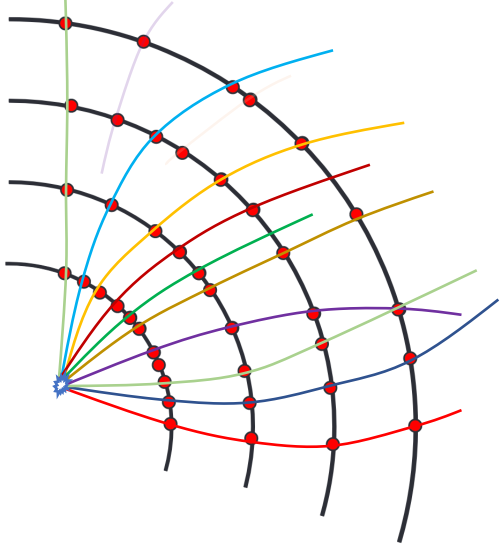
=> data are *sparse*

Representing tracking data using graphs

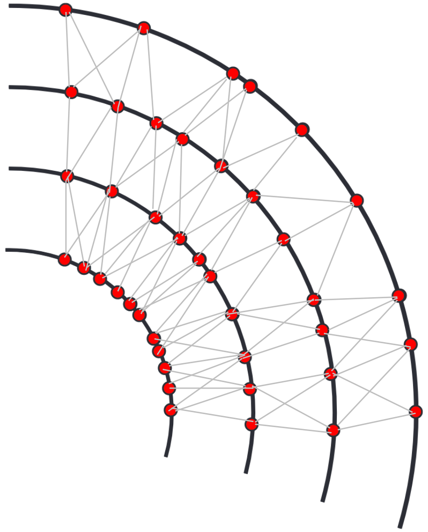
F. Siklér, "Combination of various data analysis techniques for efficient track reconstruction in very high multiplicity events", *Connecting the Dots* conference 2017 (link)

S. Farrell *et al.*, "Novel deep learning methods for track reconstruction", proceedings of *Connecting the Dots* conference 2018 (link)

Charged particles leave hits in the detector



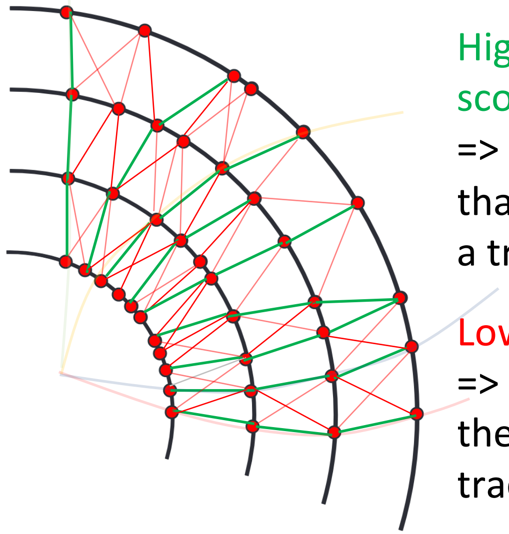
Represent the data using a graph



One node of the graph = one hit in the detector

Connect two nodes using an edge
if "it seems possible" that the two hits
are two (consecutive) hits on a track

Goal:
classify the edges of the graph



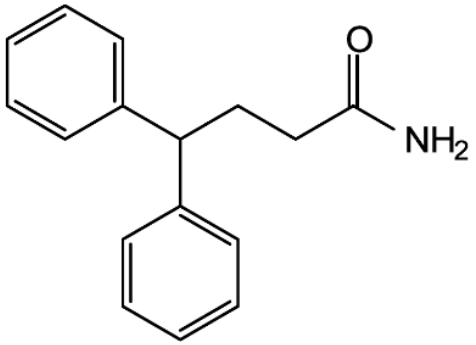
High classification score
=> **high probability** that the edge is part of a track

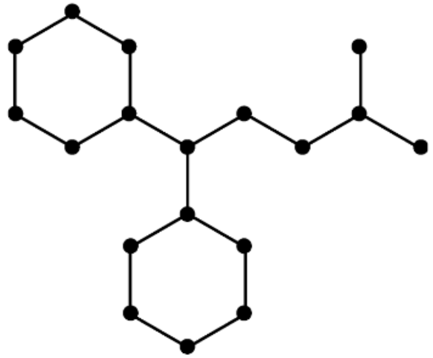
Low classification score
=> **low probability** that the edge is part of a track

Graph creation

A classic use case for graph neural networks:

Study molecules and their chemical bonds





"I suppose I'll be the one to mention the elephant in the room."

andrewgenn, iStock

For particle tracking, e.g. using ATLAS ITk, we have O(300k) hits per event.

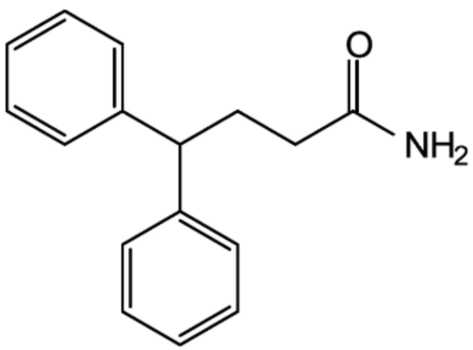
$\Rightarrow$ A fully connected graph would have O(300k) nodes and O(10^{11}) edges. This is not going to fly.

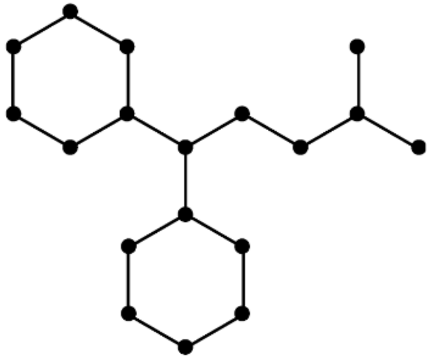
Keep in mind that we want to run this at high throughput.
Efficient graph creation becomes an area of study on its own.

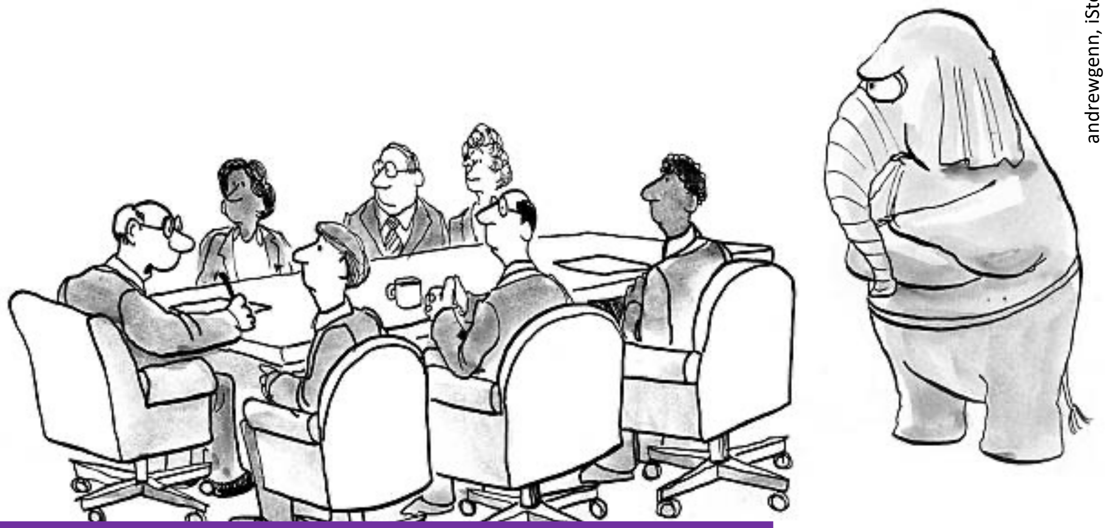
Graph creation

A classic use case for graph neural networks:

Study molecules and their chemical bonds







andrewgenn, iStock

And we need to run this (event reconstruction) on hundreds of billions of events expected to be recorded at the HL-LHC

[For comparison: O(millions) of molecules in modern databases]

For particle tracking, e.g. using ATLAS ITk, we have O(300k) hits per event.

$\Rightarrow$ A fully connected graph would have O(300k) nodes and O(10^{11}) edges. This is not going to fly.

Keep in mind that we want to run this at high throughput.
Efficient graph creation becomes an area of study on its own.

Graph creation: “module map”

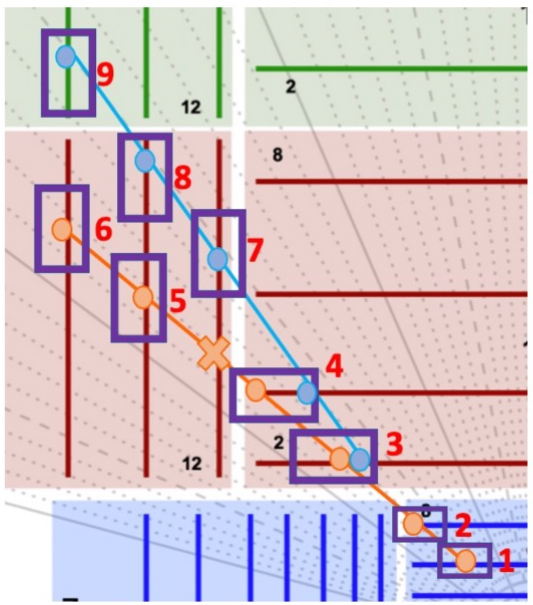
C. Biscarat *et al.*, “Towards a realistic track reconstruction algorithm based on graph neural networks for the HL-LHC”, proceedings of the *vCHEP2021* conference (link)

Refined version using module triplets:
C. Rougier, PhD thesis, Université de Toulouse, defended September 2023 (link)

New data-driven graph construction method:

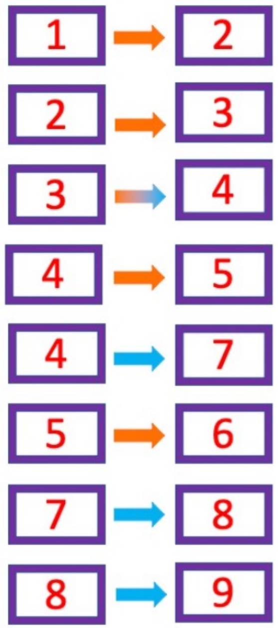
- build graphs starting from a list of possible connections from a *zone* to another *zone*: the *module map*
- done using 90k simulated $t\bar{t}$ events at $<\mu>$ = 200, considering particles with p_T > 1 GeV and leaving at least 3 hits

Particles leaving hits



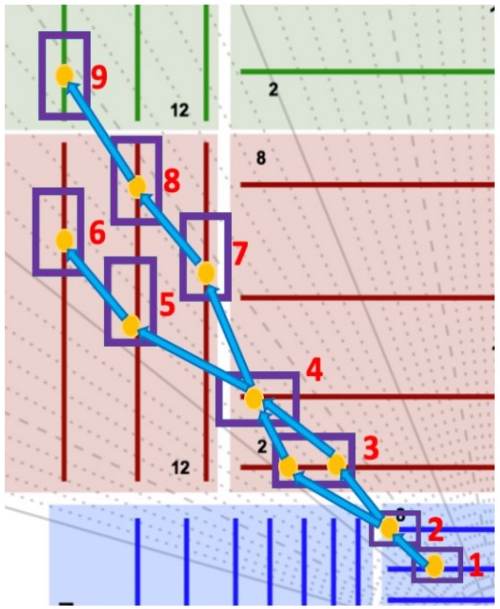
Done once

***Module map* creation**



For event reconstruction

Graph creation

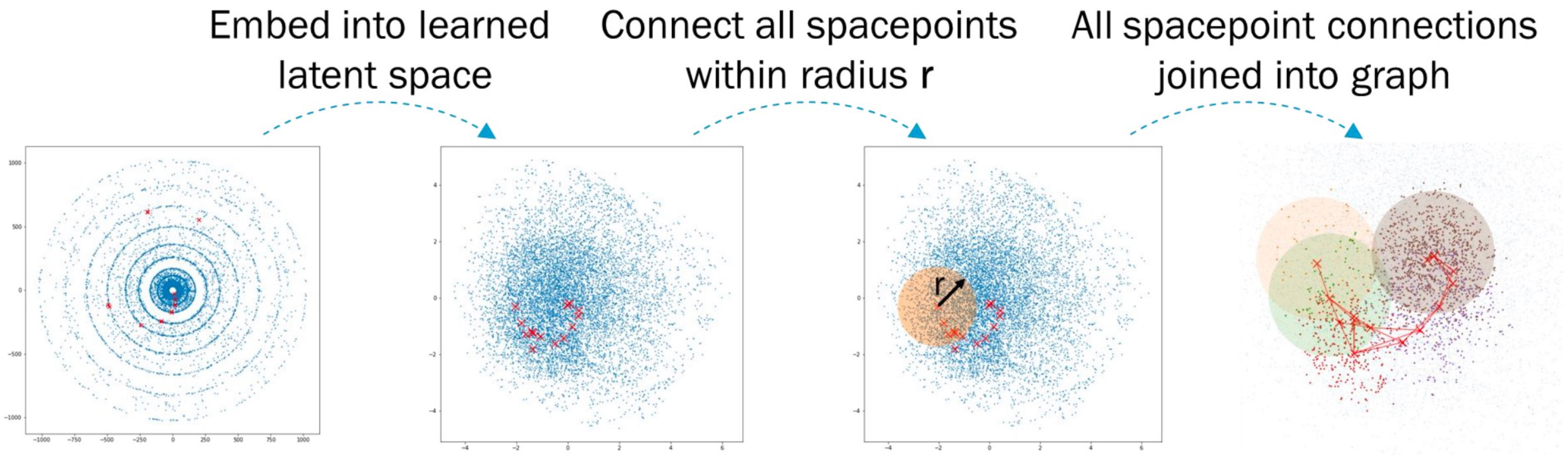


Graph creation: metric learning

X. Yu *et al.*, “Performance of a geometric deep learning pipeline for HL-LHC particle tracking”, Eur. Phys. J. C 81 (2021) 876 [(link)]

First Step: metric learning

- ➢ For all hits, embed features (coordinates, cell direction, ...) with multi-layer perceptron (MLP) into N-dimensional space
- ➢ Associate hits on same track as close as N-dimensional distance
- ➢ Score each neighbour hit within embedding neighbourhood against the “source” hit at centre
- ➢ Create edges between the source hit at centre and the neighbouring hits above a given threshold on the score.



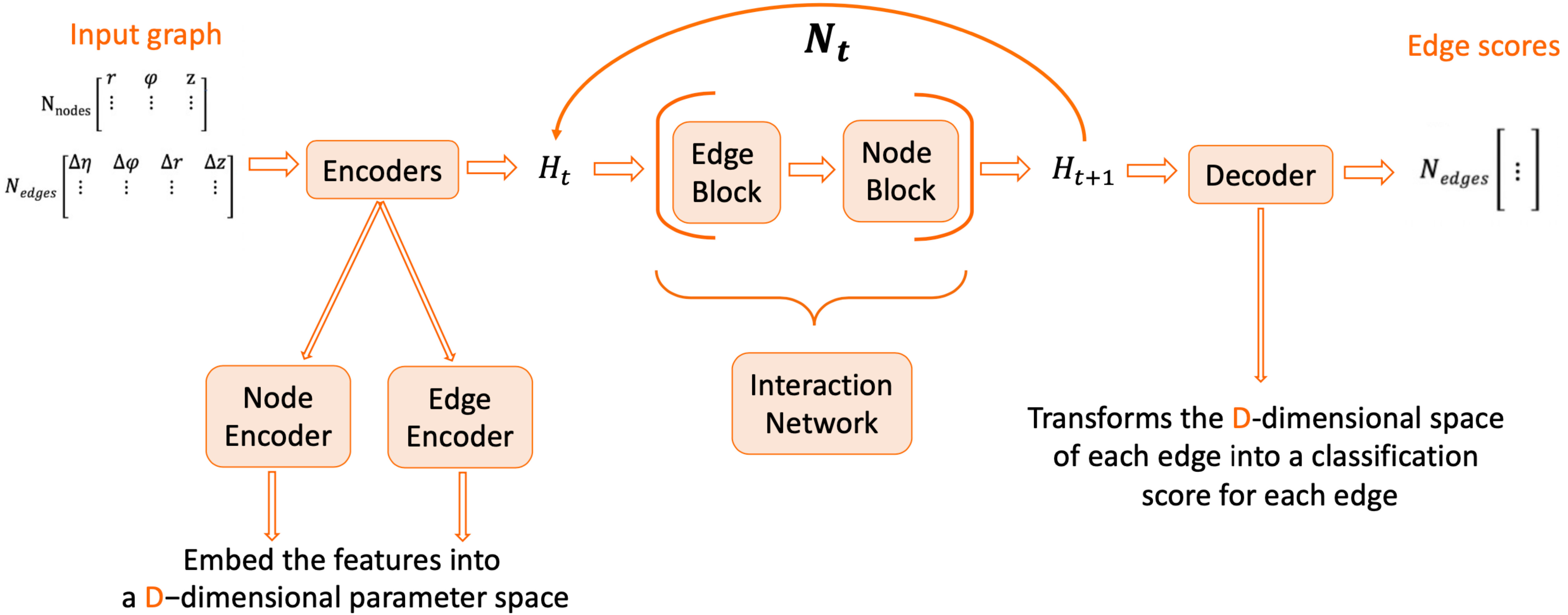
Second step: filtering

Reduce the number of edges using an MLP that looks separately at each edge (the features of the two nodes).

GNN architectures

S. Farrell *et al.*, "Novel deep learning methods for track reconstruction", proceedings of *Connecting the Dots* conference 2018 (link)

Also used in Biscarat *et al.* (vCHEP2021, link) and in ATLAS Collaboration, IDTR-2023-06 (link).

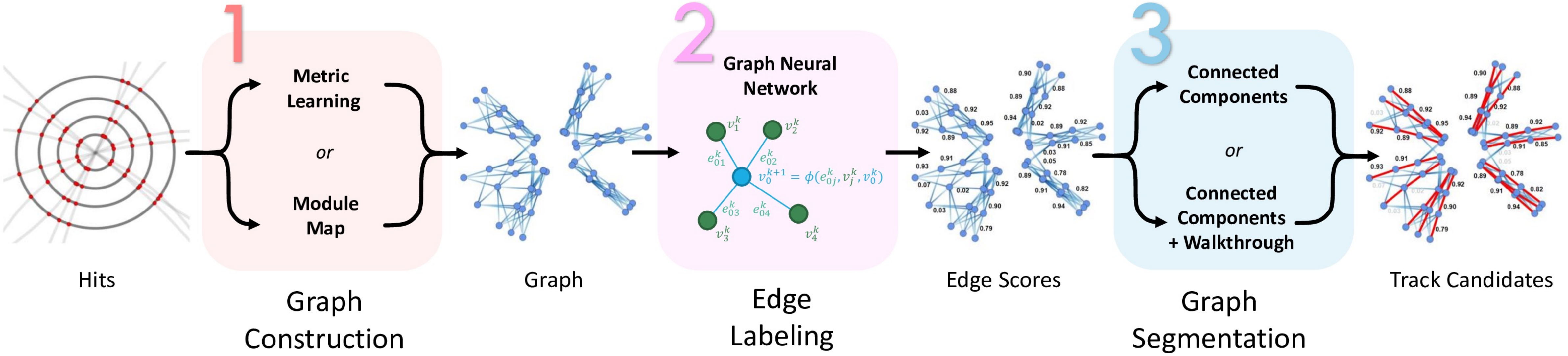


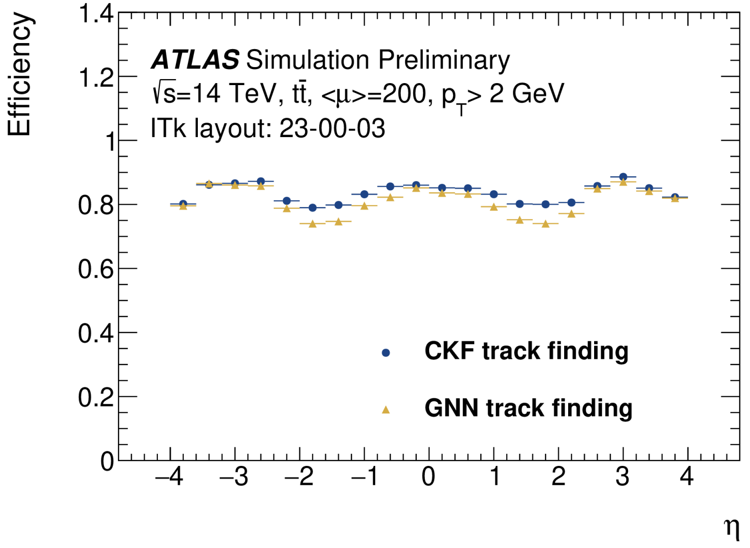
An alternative GNN architecture ("Recurrent Attention Message Passing") is presented in N. Choma *et al.* (CTD 2020) (link)

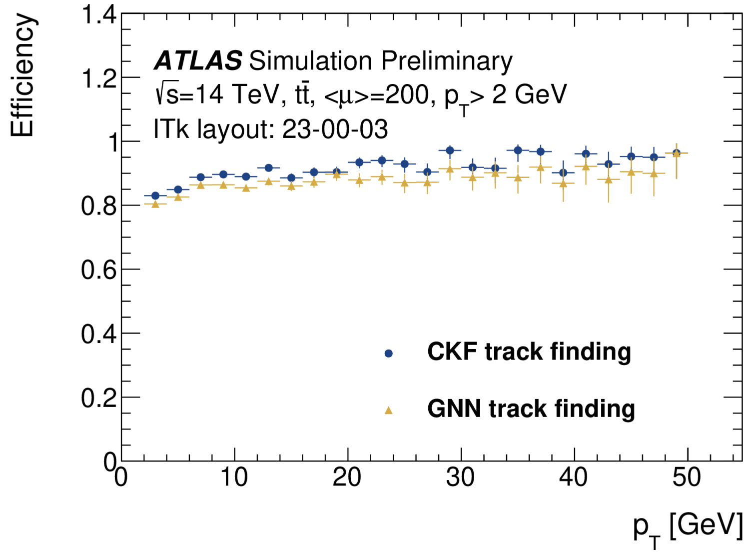
Full pipe-line; latest results for ATLAS ITk

ATLAS Collaboration, IDTR-2023-06, October 2023 (link)

H. Torres of behalf of the ATLAS Collaboration, Proceeding of Connecting the Dots 2023 (link)







Inference speed

Currently available literature: in the vast majority of studies, no attempt is made to optimise execution speed (demonstrate feasibility first).

Constraints imposed by the need to run the final algorithm at high throughput must be kept in mind, cf. slides 6 and 7.

For the pipeline presented in this talk:
initial goal of 0.5 seconds per event (for ITk data at $<\mu> = 200$) on a low-end GPU is within reach.
[correspondingly faster/ more parallel on a high-end GPU]

Significant gains expected from future implementations with custom CUDA kernels.
(so far, have initial implementation of custom CUDA kernel for module map – will be presented at the CHEP 2024 conference)

Wish list

There is still ample room for improvement – and opportunities for newcomers to contribute.

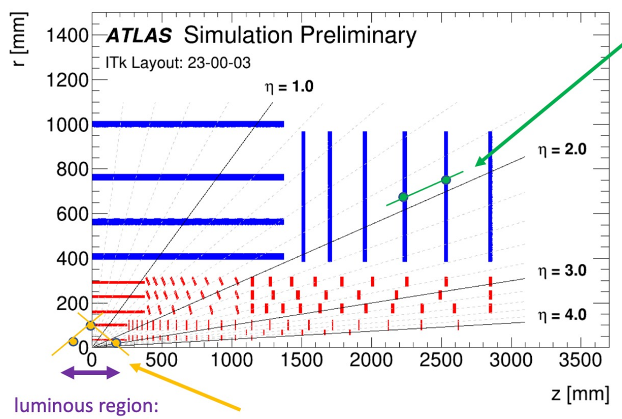
My personal wish list includes:

- tuning of algorithms for ultimate physics performance

- more efficient, light-weight GNNs, specifically designed to deal with some of the peculiarities of our graphs, including large variations of connectivity from one region of the graph to another

- computing performance gains for deployment on GPUs
    - simpler GNN models
    - dedicated CUDA kernels

- computing performance gain for deployment on CPUs and GPUs
    - pruning of the models ("kill neurons that have little impact")



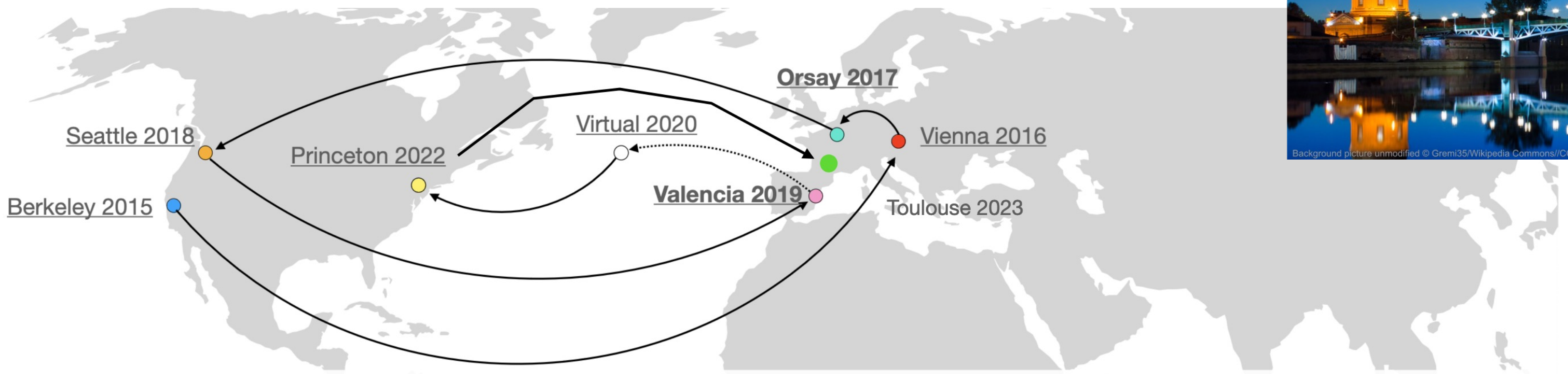
In this region, it is relatively clear in which direction to look for the next hit.

luminous region:
-200 < z < 200
at r = 0

In this region, the direction is less clear.
In addition, this is where the density of hits is largest.

One can easily have >10 edges on a given node.

Conferences



Connecting the Dots, 2023 edition: link

Concise summary: link

The *Connecting The Dots* workshop series brings together experts on track reconstruction and other problems involving pattern recognition in sparsely sampled data. While the main focus will be on High Energy Physics (HEP) detectors, the *Connecting The Dots* workshop is intended to be inclusive across other scientific disciplines wherever similar problems or solutions arise.

Advertisement

We are currently working on the organisation of a international workshop on *Heterogeneous Data and Large Representation Models in Science* in Toulouse in the early fall of 2024.
The workshop is sponsored by the CNRS AISSAI initiative.

Please do get in touch with me if you are interested.

AISSAI
AI for science, science for AI

Summary

Feasibility of GNN-based track reconstruction on realistic (fullsim, $<\mu>$ = 200) samples has been demonstrated.

Physics performance is getting close to that of classical algorithms (combinatorial Kalman filter).

Promising inference speed on GPU has been shown.

There is still ample room for improvements on our way towards deployment in production in a real experiment.

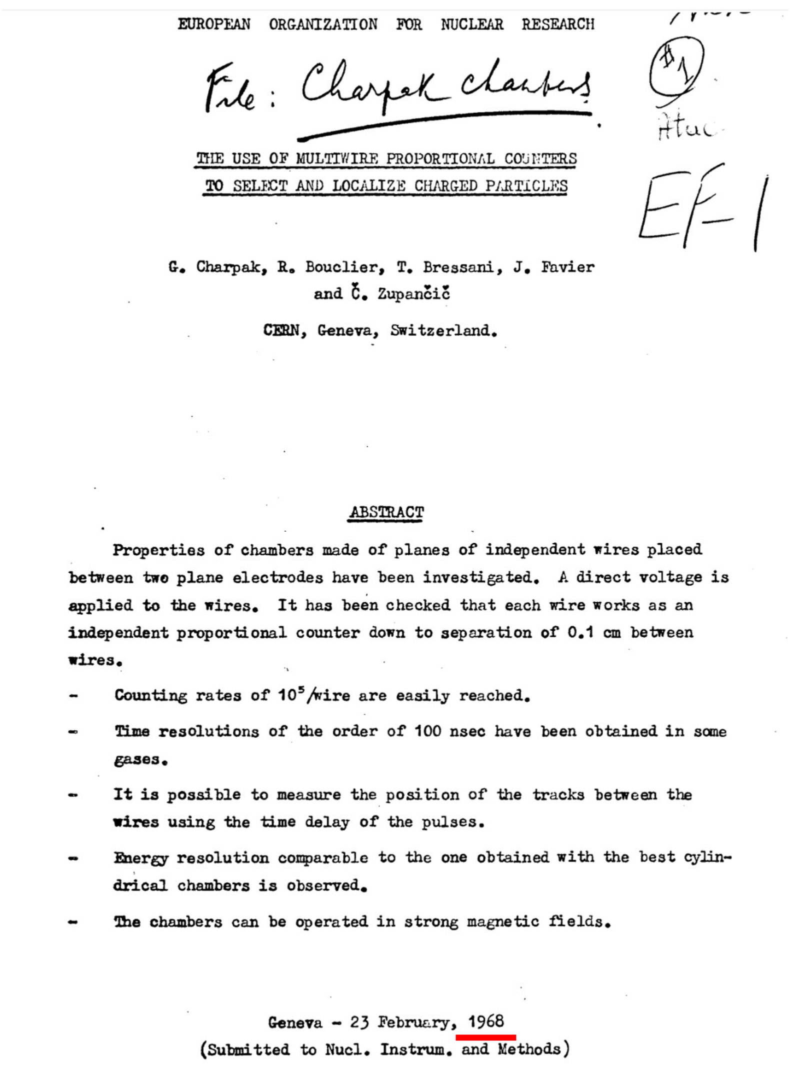
Would like to play around with GNN-based tracking yourself ?
Would like to contribute your own studies ?

There is publicly available software that can be helpful for getting started, including acorn.

(link)

Backup material

Electronic readout !


EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

File : Charpak chambers

EF-1

THE USE OF MULTIWIRE PROPORTIONAL COUNTERS
TO SELECT AND LOCALIZE CHARGED PARTICLES

G. Charpak, R. Bouclier, T. Bressani, J. Favier
and Č. Zupančič

CERN, Geneva, Switzerland.

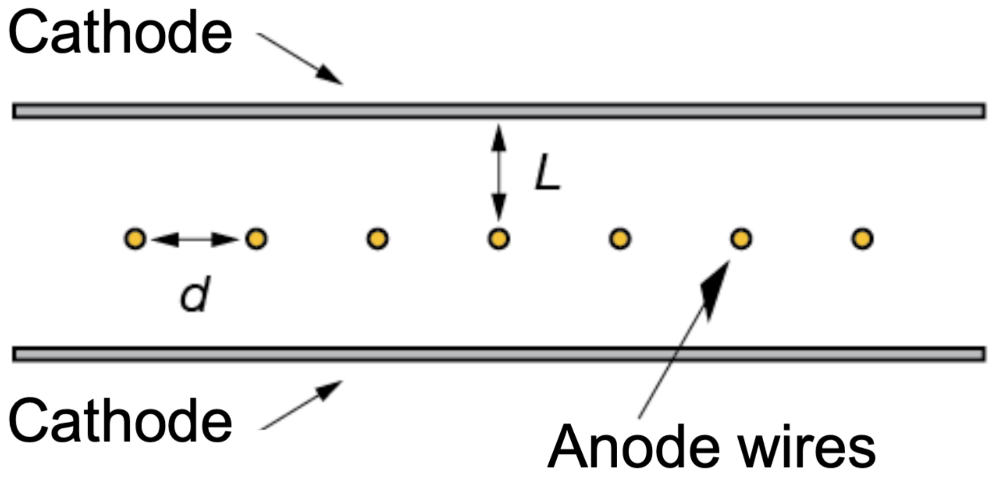
ABSTRACT

Properties of chambers made of planes of independent wires placed between two plane electrodes have been investigated. A direct voltage is applied to the wires. It has been checked that each wire works as an independent proportional counter down to separation of 0.1 cm between wires.

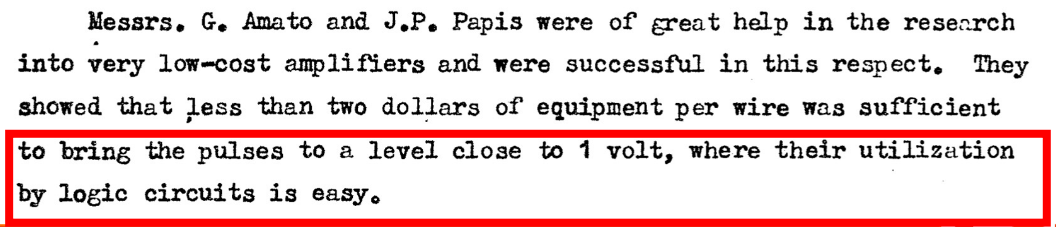
- Counting rates of 10^5/wire are easily reached.
- Time resolutions of the order of 100 nsec have been obtained in some gases.
- It is possible to measure the position of the tracks between the wires using the time delay of the pulses.
- Energy resolution comparable to the one obtained with the best cylindrical chambers is observed.
- The chambers can be operated in strong magnetic fields.

Geneva - 23 February, 1968
(Submitted to Nucl. Instrum. and Methods)

Multi-wire proportional chamber



Nobel prize
in physics (1992)


Messrs. G. Amato and J.P. Papis were of great help in the research into very low-cost amplifiers and were successful in this respect. They showed that less than two dollars of equipment per wire was sufficient to bring the pulses to a level close to 1 volt, where their utilization by logic circuits is easy.

Fifteen times more data

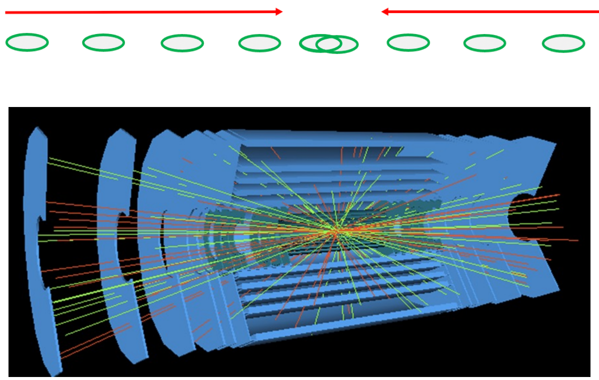
$25\ fb^{-1}$	$150\ fb^{-1}$	$350\ fb^{-1}$	$3000\ fb^{-1}$	Integrated luminosity
2011 -2012	2015 -2018	2022 -2025	2029 -2040	
Run 1	Run 2	Run 3	HL-LHC (Run 4, 5, …)	

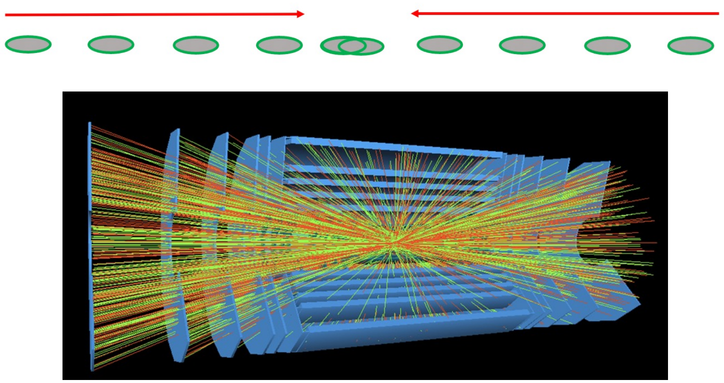
High luminosity phase:

- more data
- 5 times more protons/bunch
- more complex events
- highly granular detectors

40 million crossings of pairs of proton bunches per second !

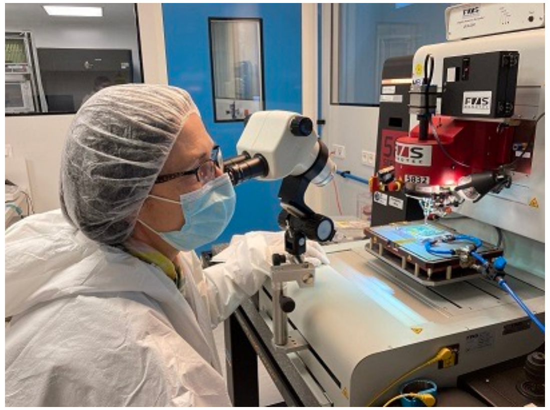
High luminosity: how ? Cannot reduce distance between bunches any further. More protons/bunch !





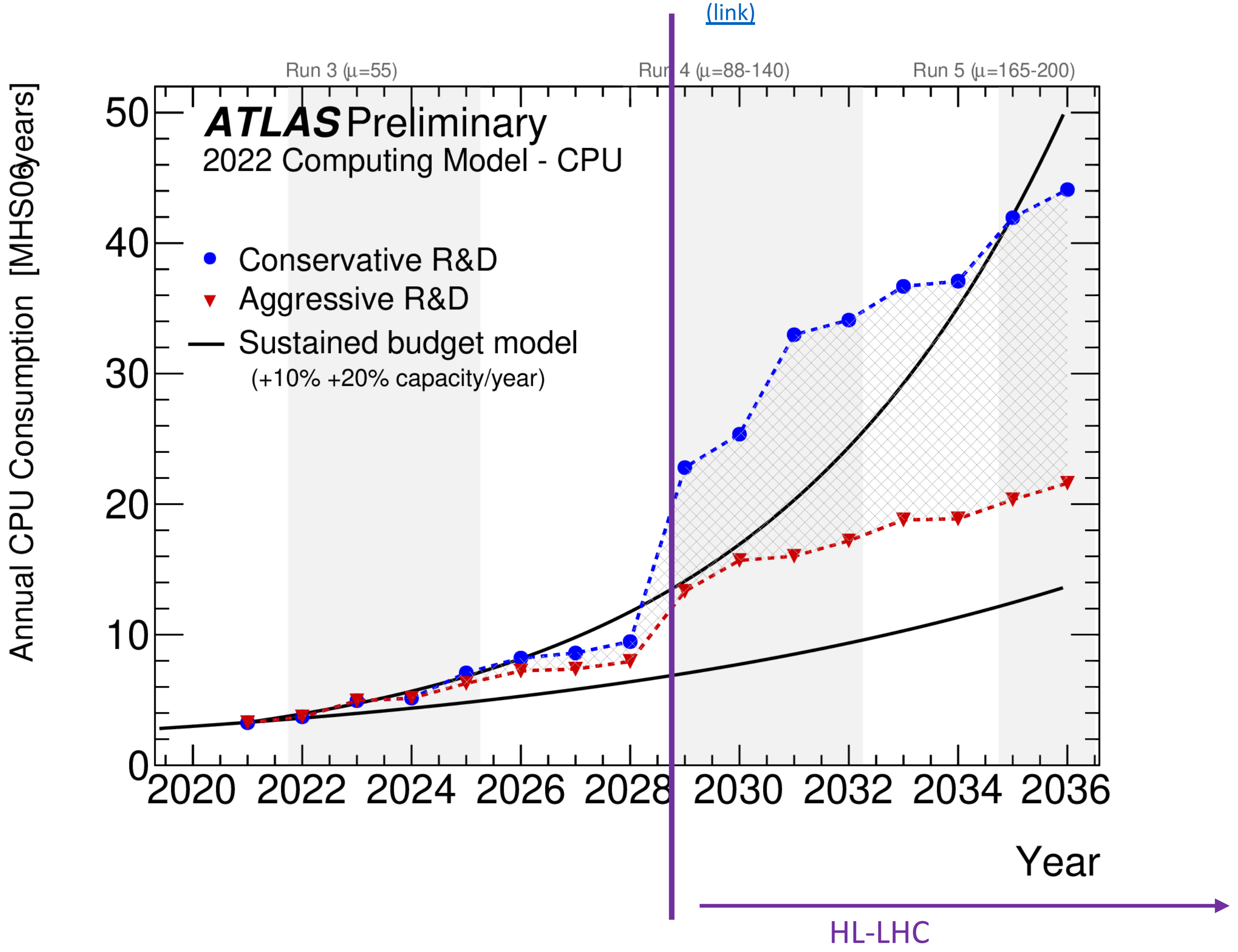
Work towards HL-LHC, today :

- beam injection chain
- construction of new detector components
- design of computing models



Computing resources

ATLAS Software and Computing HL-LHC Roadmap,
CERN-LHCC-2022-00 (2022)
(link)



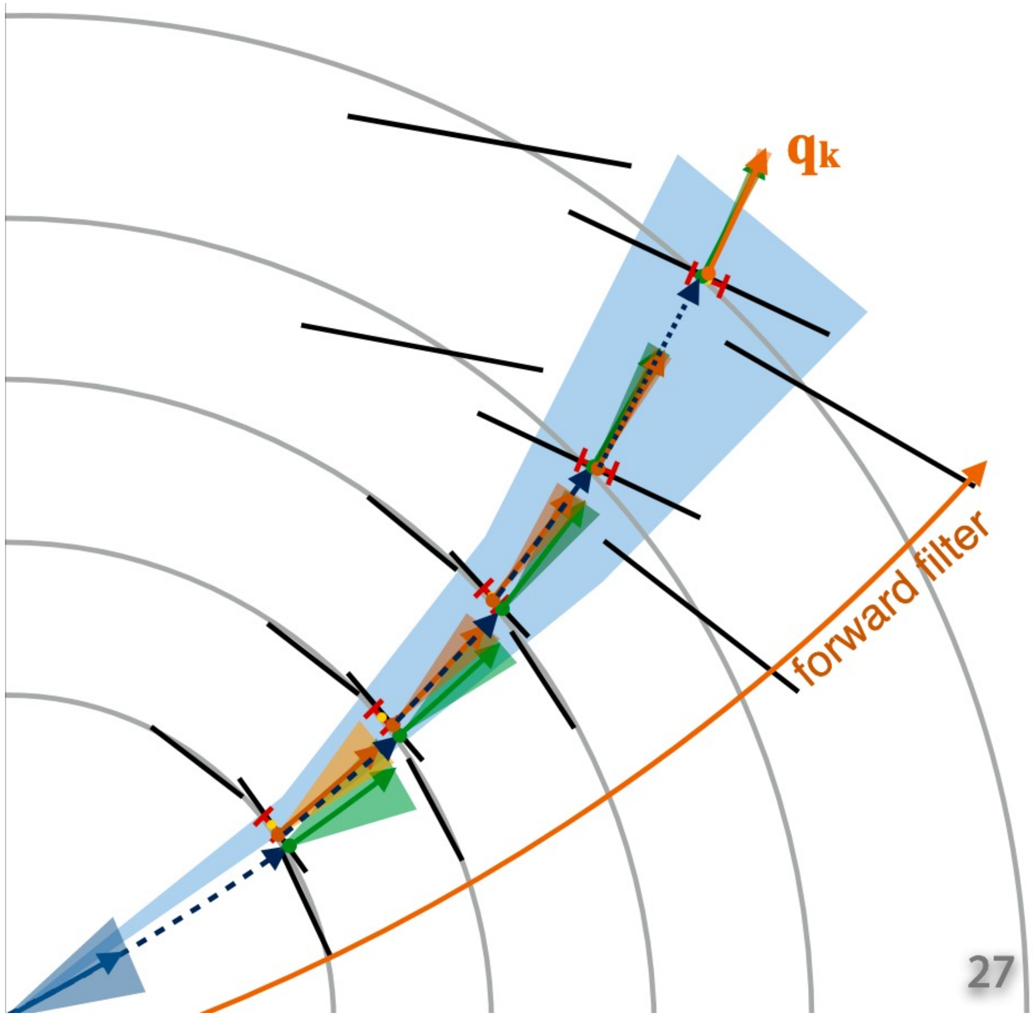
Resources used today:
- O(1) million de CPU cores running continuously
- O(1) exabyte of storage

At HL-LHC: with our current computing model, we would face a significant shortage of computing resources

→ **need to make important changes**

→ **… or live with cuts into our physics programme**

Classical algorithms for track reconstruction



1. propagate p_{k-1} and its covariance C_{k-1} :

$$\boldsymbol{q}_{k|k-1} = \boldsymbol{f}_{k|k-1}(\boldsymbol{q}_{k-1|k-1})$$

$$\boldsymbol{C}_{k|k-1} = \boldsymbol{F}_{k|k-1}\boldsymbol{C}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}}_{k|k-1} + \boldsymbol{Q}_k$$

with Q_k ~ noise term (M.S.)

2. update prediction to get $q_{k|k}$ and $C_{k|k}$:

$$\boldsymbol{q}_{k|k} = \boldsymbol{q}_{k|k-1} + \boldsymbol{K}_k[\boldsymbol{m}_k - \boldsymbol{h}_k(\boldsymbol{q}_{k|k-1})]$$

$$\boldsymbol{C}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{C}_{k|k-1}$$

with K_k ~ gain matrix :

$$\boldsymbol{K}_k = \boldsymbol{C}_{k|k-1}\boldsymbol{H}^{\mathrm{T}}_k(\boldsymbol{G}_k + \boldsymbol{H}_k\boldsymbol{C}_{k|k-1}\boldsymbol{H}^{\mathrm{T}}_k)^{-1}$$

Triggering (ATLAS)

ATLAS EXPERIMENT

(link)

CERN-LHCC-2022-004 / ATLAS-TDR-029-ADD-1
01/03/2022

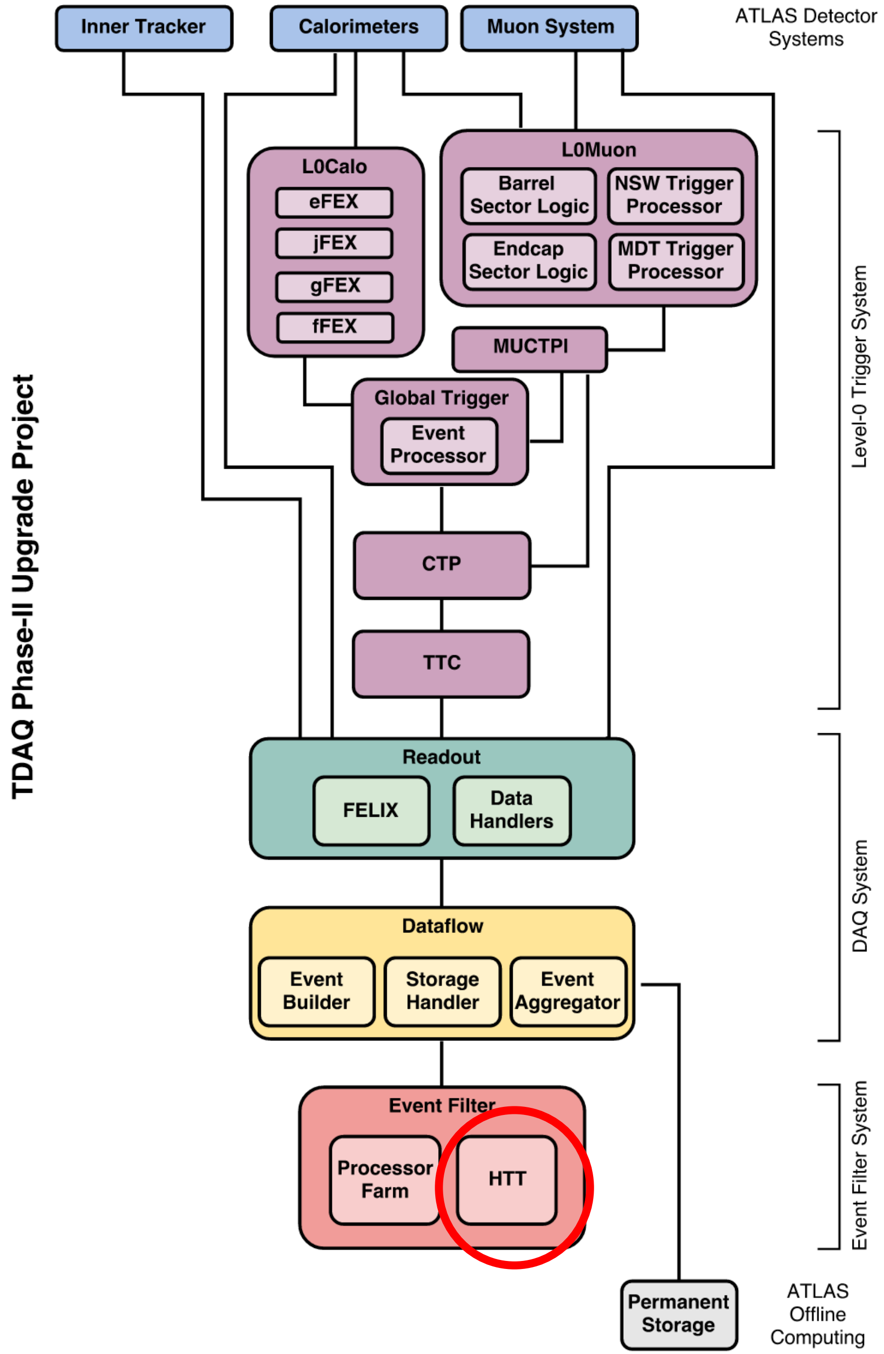
Technical Design Report for the Phase-II Upgrade of the ATLAS Trigger and Data Acquisition System - EF Tracking Amendment

The ATLAS Collaboration

Reference: v1.2

Created: March 1, 2022
Last modified: March 1, 2022
Prepared by: **The ATLAS Collaboration**

Abstract
This Technical Design Report Amendment describes revised plans for Event Filter Tracking in the upgrade of the ATLAS Trigger and Data Acquisition system for the High Luminosity LHC. The motivation to change the baseline for Event Filter Tracking is explained. Next, a description of the requirements for Event Filter Tracking and the definition of the proposed baseline to meet these requirements are presented. Several demonstrations using various hardware and software are reported in support of this proposal. Finally, the organization and resources needed to deliver the new baseline are set out.



Event rate: 40 MHz

after hardware-based L0 trigger: 1 MHz

detailed detector readout after L0 accept

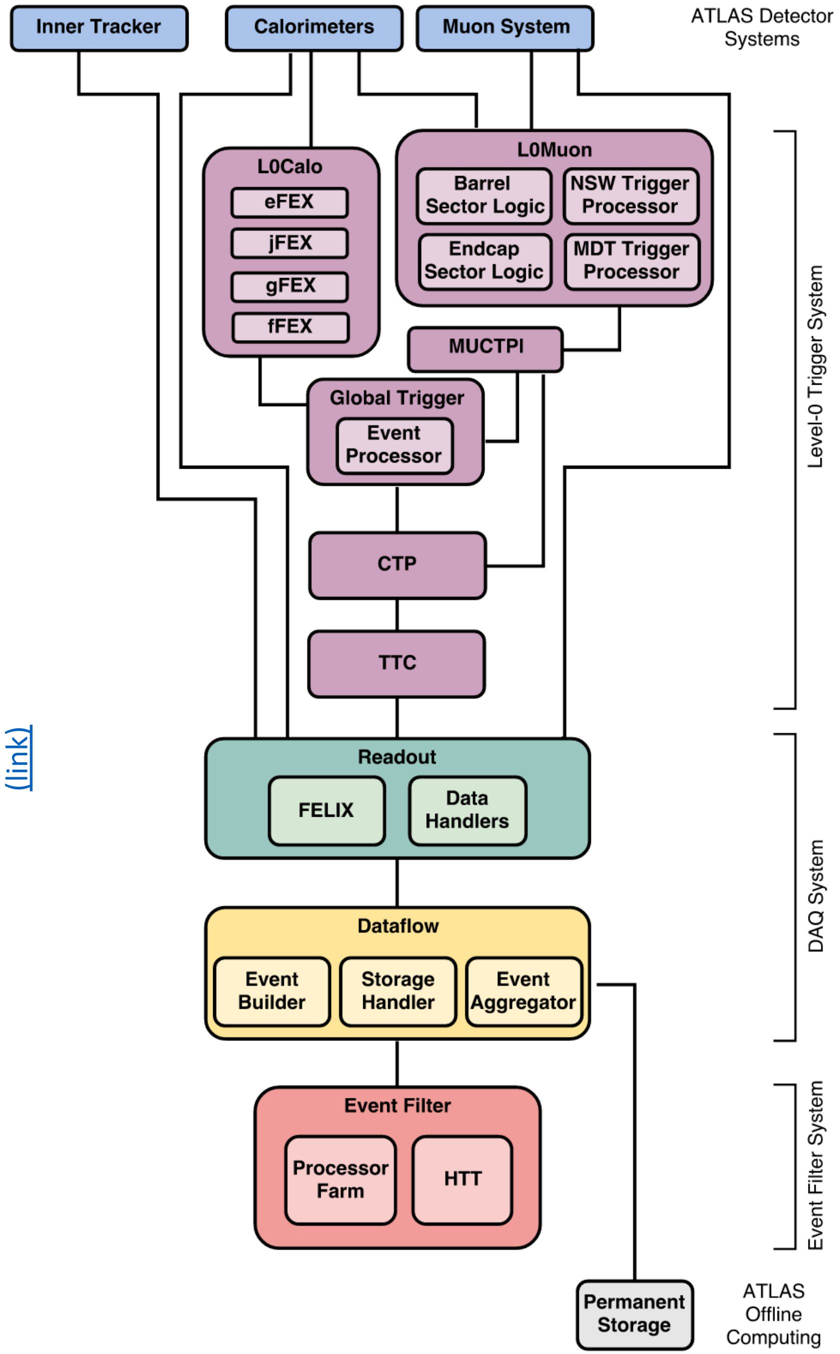
after event filter (to tape): 10 kHz

Triggering (ATLAS)

"Recording data at the LHC is like drinking from a fire hose"

TDAQ Phase-II Upgrade Project

(link)



Event rate: 40 MHz

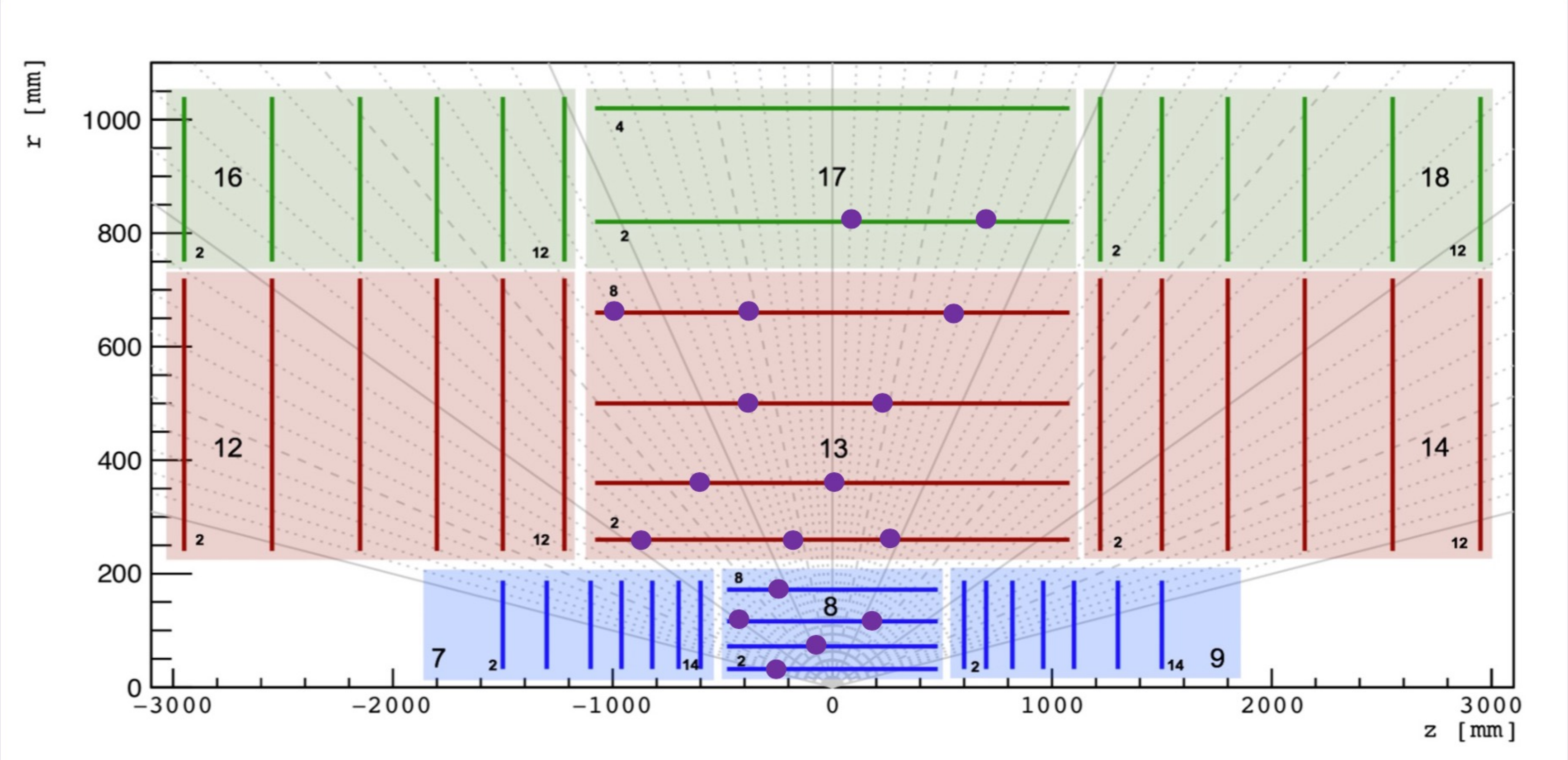
after hardware-based L0 trigger: 1 MHz

detailed detector readout after L0 accept

after event filter (to tape): 10 kHz

The TrackML dataset

- **Generation (Pythia8):** 1000 $t\bar{t}$ events from *pp* collisions
  - $\sqrt{s} = 14\,TeV, \mu = 200$ (HL-LHC conditions), pile-up modeling using A3 tune
- **Simulation:** *Generic detector* simulated with fast simulation of ACTS framework



18 728 silicon modules

hit

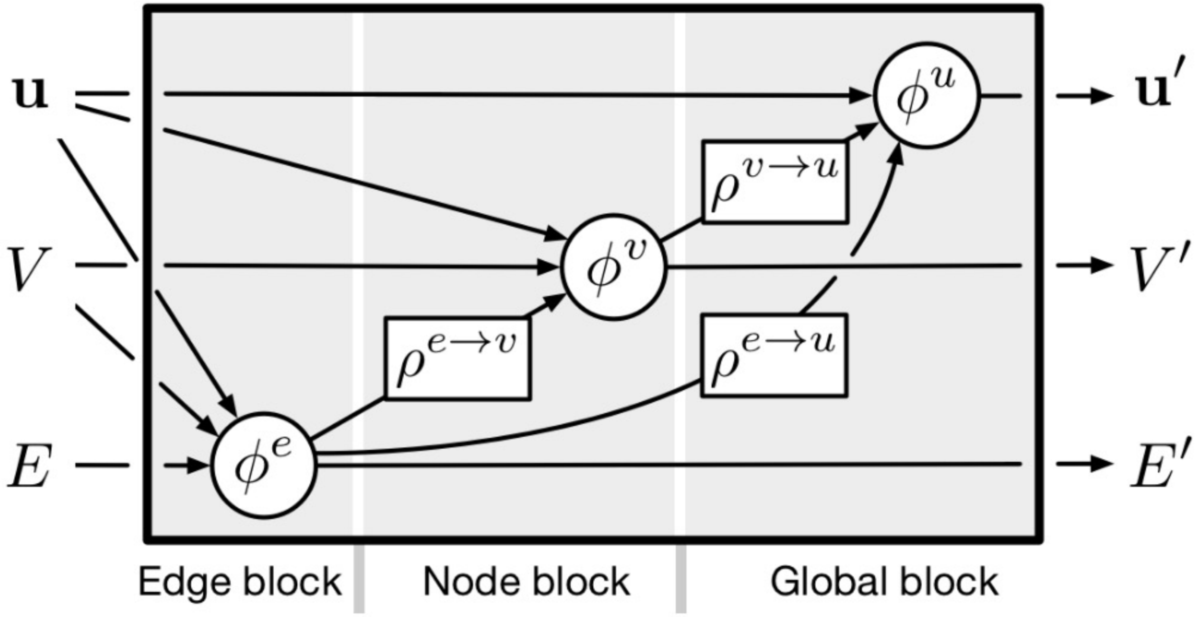
Graph (neural) networks

Oct 2018 (link)

Relational inductive biases, deep learning, and graph networks

Peter W. Battaglia[1,*], Jessica B. Hamrick[1], Victor Bapst[1], Alvaro Sanchez-Gonzalez[1], Vinicius Zambaldi[1], Mateusz Malinowski[1], Andrea Tacchetti[1], David Raposo[1], Adam Santoro[1], Ryan Faulkner[1], Caglar Gulcehre[1], Francis Song[1], Andrew Ballard[1], Justin Gilmer[2], George Dahl[2], Ashish Vaswani[2], Kelsey Allen[3], Charles Nash[4], Victoria Langston[1], Chris Dyer[1], Nicolas Heess[1], Daan Wierstra[1], Pushmeet Kohli[1], Matt Botvinick[1], Oriol Vinyals[1], Yujia Li[1], Razvan Pascanu[1]

[1]DeepMind; [2]Google Brain; [3]MIT; [4]University of Edinburgh



(a) Full GN block

A GN block contains three "update" functions, ϕ, and three "aggregation" functions, ρ,

$$\begin{aligned} \mathbf{e}'_k &= \phi^e\left(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u}\right) & \bar{\mathbf{e}}'_i &= \rho^{e\to v}\left(E'_i\right) \\ \mathbf{v}'_i &= \phi^v\left(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right) & \bar{\mathbf{e}}' &= \rho^{e\to u}\left(E'\right) \\ \mathbf{u}' &= \phi^u\left(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}\right) & \bar{\mathbf{v}}' &= \rho^{v\to u}\left(V'\right) \end{aligned}$$

where $E'_i = \{(\mathbf{e}'_k, r_k, s_k)\}_{r_k=i,\ k=1:N^e}$, $V' = \{\mathbf{v}'_i\}_{i=1:N^v}$, and $E' = \bigcup_i E'_i = \{(\mathbf{e}'_k, r_k, s_k)\}_{k=1:N^e}$.

Graph (neural) networks

Oct 2018 (link)

Relational inductive biases, deep learning, and graph networks

Peter W. Battaglia[1,*], Jessica B. Hamrick[1], Victor Bapst[1], Alvaro Sanchez-Gonzalez[1], Vinicius Zambaldi[1], Mateusz Malinowski[1], Andrea Tacchetti[1], David Raposo[1], Adam Santoro[1], Ryan Faulkner[1], Caglar Gulcehre[1], Francis Song[1], Andrew Ballard[1], Justin Gilmer[2], George Dahl[2], Ashish Vaswani[2], Kelsey Allen[3], Charles Nash[4], Victoria Langston[1], Chris Dyer[1], Nicolas Heess[1], Daan Wierstra[1], Pushmeet Kohli[1], Matt Botvinick[1], Oriol Vinyals[1], Yujia Li[1], Razvan Pascanu[1]

[1]DeepMind; [2]Google Brain; [3]MIT; [4]University of Edinburgh

Interaction networks

Interaction Networks (Battaglia et al., 2016; Watters et al., 2017) and the Neural Physics Engine Chang et al. (2017) use a full GN but for the absence of the global to update the edge properties:

$$\begin{aligned}
\phi^e\left(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u}\right) &:= f^e\left(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}\right) &&= \mathrm{NN}_e\left(\left[\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}\right]\right) \\
\phi^v\left(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right) &:= f^v\left(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right) &&= \mathrm{NN}_v\left(\left[\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right]\right) \\
\rho^{e\rightarrow v}\left(E'_i\right) &:= &&= \sum_{\{k:\, r_k=i\}} \mathbf{e}'_k
\end{aligned}$$

Simplify GNN -> inference on FPGA

Computing and Software for Big Science (2021) 5:26
https://doi.org/10.1007/s41781-021-00073-z

(link)

ORIGINAL ARTICLE

Charged Particle Tracking via Edge-Classifying Interaction Networks

Gage DeZoort[1] · Savannah Thais[1] · Javier Duarte[2] · Vesal Razavimaleki[2] · Markus Atkinson[3] · Isobel Ojalvo[1] · Mark Neubauer[3] · Peter Elmer[1]

Received: 12 July 2021 / Accepted: 13 October 2021 / Published online: 15 November 2021

Abstract

Recent work has demonstrated that geometric deep learning methods such as graph neural networks (GNNs) are well suited to address a variety of reconstruction problems in high-energy particle physics. In particular, particle tracking data are naturally represented as a graph by identifying silicon tracker hits as nodes and particle trajectories as edges, given a set of hypothesized edges, edge-classifying GNNs identify those corresponding to real particle trajectories. In this work, we adapt the physics-motivated interaction network (IN) GNN toward the problem of particle tracking in pileup conditions similar to those expected at the high-luminosity Large Hadron Collider. Assuming idealized hit filtering at various particle momenta thresholds, we demonstrate the IN's excellent edge-classification accuracy and tracking efficiency through a suite of measurements at each stage of GNN-based tracking: graph construction, edge classification, and track building. The proposed IN architecture is substantially smaller than previously studied GNN tracking architectures; this is particularly promising as a reduction in size is critical for enabling GNN-based tracking in constrained computing environments. Furthermore, the IN may be represented as either a set of explicit matrix operations or a message passing GNN. Efforts are underway to accelerate each representation via heterogeneous computing resources towards both high-level and low-latency triggering applications.

Work has also been done to accelerate the inference of deep neural networks with heterogeneous resources beyond GPUs, like field-programmable gate arrays (FPGAs) [49–57]. This work extends to GNN architectures [29, 58]. Specifically, in Ref. [29], a compact version of the IN was implemented for $p_T > 2\,\text{GeV}$ segmented geometric graphs with up to 28 nodes and 37 edges, and shown to have a latency less than 1 μs, an initiation interval of 5 ns, reproduce the floating-point precision model with a fixed-point precision of 16 bits or less, and fit on a Xilinx Kintex UltraScale FPGA.

While this preliminary FPGA acceleration work is promising, there are several limitations of the current FPGA implementation of the IN:

1. This fully pipelined design cannot easily scale to beyond $\mathcal{O}(100)$ nodes and $\mathcal{O}(1000)$ edges. However, if the initiation interval requirements are loosened, it can scale up to $O(10,000)$ nodes and edges.
2. The neural network itself is small, and while it is effective for $p_T > 2\,\text{GeV}$ graphs, it may not be sufficient for lower-p_T graphs.
3. The FPGA design makes no assumptions about the pos-

Acknowledgements

During the preparation of these slides, I have made extensive use of the following talks:

- Charline Rougier, “Towards a realistic track reconstruction algorithm based on graph neural networks for the HL-LHC”, vCHEP 2021, May 17-21 2021
  (link)

- Catherine Biscarat et Sylvain Caillou, « Comment la future phase de haute luminosité du collisionneur LHC du CERN, qui doit permettre l'étude des interactions du boson de Higgs, bouscule notre façon de calculer », Séminaire SFP de la section Midi-Pyrénées, 29 octobre 2021
  (link)

- Markus Elsing, “Pattern recognition in HEP (track reconstruction)”, Learning to Discover: advanced pattern recognition, Institut Pascal, October 14-25 2019
  (link)

- Material from Daniel Murnane (Berkeley LBL) and Charline Rougier (L2IT)

I thank the following colleagues for our frequent, fruitful discussions:

Alexis Vallier, Alina Lazar, Chen-Hsun Chan, Christophe Collard, Daniel Murnane, Heberth Torres, Jared Burleson, Mark Neubauer, Minh-Tuan Pham, Paolo Calafiura, Sebastian Dittmeier, Sylvain Caillou, Xiangyang Ju