k4Clue: Empowering Future Collider Experiments with CLUE

Erica Brondolin,
Felice Pantaleo, Marco Rovere
Introduction
CLUstering of Energy

- CLUE (CLUstering of Energy) is a fast density-based clustering algorithm for the next generation of sampling calorimeter with high granularity in HEP


- It uses energy density - rather than individual cell energy - to establish seeds, outliers, and followers in 2D planes.

- GPU-friendly, i.e. suitable for the upcoming era of heterogeneous computing in HEP

- Standalone repo:CLU - kalos
  gitlab.cern.ch
Step 1: Building Data Structure

- Querying neighborhood is a frequent operation in density-based clustering → **fast**!
- Build **Fixed-Grid Spatial Index** for hits on each layer:
  - Each tile in the grid hosts indices of hits inside it and has a fixed length of memory to store the hosted indices. It is independent by the detector granularity.
- To find the neighborhood hits $N_d(i)$ of $i$-hit, we only need to loop over hits in $\Omega_d(i)$

\[
d\text{-searchBox:} \\
\Omega_d(i) = \{ j : j \in \text{tiles touched by square window}(x_i \pm d, y_i \pm d) \} \\
\]

\[
d\text{-neighborhood:} \\
N_d(i) = \{ j : d_{ij} < d \} \subset \Omega_d(i) \\
\]
Step 2: Local energy density

- Calculate local energy density ($\rho_i$) in a distance ($d_c$)
  - Each hit $j$ weighted by the deposited energy ($E_j$)
  - For each hit, calculate $\rho_i$

$$\rho_i = \sum_{j \in N_d(i)} E_j \times f(d_{ij}); \quad f(d_{ij}) = \begin{cases} 1, & \text{if } i = j \\ k, & \text{if } 0 < d_{ij} \leq d_c \\ 0, & \text{if } d_{ij} > d_c \end{cases}$$

convolution kernel $k = 0.5$

build data structure
density
Step 3: Find “closest higher hit”

- Calculate “Nearest-Higher” hit within $N_{dm}(i)$
  - Define $d_m = o_f \cdot d_c$
  - Find the closest hit with higher local energy density, $n_{hi}$
    $$n_{hi} = \begin{cases} \arg\min_{j \in \mathcal{N}_{dm}(i)} d_{ij}, & \text{if } |\mathcal{N}_{dm}(i)| \neq 0, \mathcal{N}_{dm}(i) = \{j : j \in N_{dm}(i), \rho_j > \rho_i\} \\ -1, & \text{otherwise} \end{cases}$$
  - Calculate the separation distance $\delta_i = \text{dist}(i, n_{hi})$

[Diagrams showing data structure, density, and nearest higher hits]
Step 4: Classify hits

- Promote as **seed** if $\rho_i > \rho_c$, $\delta_i > d_c$
- Demote as **outlier** if $\rho_i < \rho_c$, $\delta_i > o_f \cdot d_c$
- Assign unique, progressive cluster ID to each cluster
  - ** Followers** are defined and associated to their closest seed

Example of decision plot

![Decision Plot](image)
Clustering procedure recap

**build data structure**
Query the neighborhood of a point by looping over the points in $N_d$ in the bins touched by the tiles intersected by $d_c$.

**density**
Hit position and energy used to calculate the hit's local energy density $\rho_i$ and its distance $\delta_i$ to the nearest hit with higher local density.

**nearest higher**
Define the nearest-higher of each hit as the hit with the local energy density higher then the hits itself and within a distance of $d_m = o_f \times d_c$.

**find seed**
Use following criteria:
- seed: $\rho_i \geq \rho_c$ and $\delta_i \geq d_c$;
- outlier: $\rho_i < \rho_c$ and $\delta_i \geq (o_f \times d_c)$.

**assign clusters**
Register each remaining point as a follower to its nearest-higher.

**Input parameters:**
- $d_c$: Critical Distance
- $o_f$: Outlier Delta Factor
- $\rho_c$: Minimum Local Density
CLUE in the HGCAL reconstruction
CMS High Granularity Calorimeter

- Phase-2 upgrade of CMS is needed to cope with HL-LHC phase
  - A significant increase in the instantaneous luminosity (5 - 7.5x)
- Imaging calorimeter with very fine lateral and longitudinal segmentation, and precision timing capabilities
  - Covering $1.5 < \eta < 3.0$

<table>
<thead>
<tr>
<th>Both endcaps</th>
<th>Silicon</th>
<th>Scintillators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>$\sim 620 \text{ m}^2$</td>
<td>$\sim 400 \text{ m}^2$</td>
</tr>
<tr>
<td>Channel size</td>
<td>0.5 - 1 cm$^2$</td>
<td>4 - 30 cm$^2$</td>
</tr>
<tr>
<td>#Modules</td>
<td>$\sim 30'000$</td>
<td>$\sim 4'000$</td>
</tr>
<tr>
<td>#Channels</td>
<td>$\sim 6 \text{ M}$</td>
<td>240 k</td>
</tr>
<tr>
<td>Op. temp.</td>
<td>$-30 , ^\circ\text{C}$</td>
<td>$-30 , ^\circ\text{C}$</td>
</tr>
</tbody>
</table>

Ref.

CE-E : Electromagnetic Endcap Calorimeter
CE-H : Hadronic Endcap Calorimeter

~0.28 m
~2 m

Erica Brondolin (erica.brondolin@cern.ch)
HGCAL Software Reconstruction

- The HGCAL reconstruction framework is **TICL (The Iterative Clustering)**
- It starts by calibrating deposited energy in individual cells, also called RecHits → an order of $10^5$ RecHits in the HGCAL detector for events @ 200 pileup
- CLUE clusters the RecHits in the same layer to produce **Layer Clusters (LCs)**
The package
Key4hep in a nutshell

The common software vision: key4hep

**key4hep** is a huge ecosystem of software packages adopted by all future collider projects, complete workflow from generator to analysis

- Event data model: EDM4HEP for exchange among framework components
  - Podio as underlying tool, for different collision environments
  - Including truth information
- Data processing framework: Gaudi
- Geometry description: DD4hep, ability to include CAD files
- Package manager: **Spack**: source /cvmfs/sw.hsf.org/key4hep/setup.sh
Integrating CLUE in Key4hep

- **k4Clue v01-00** (doi: [10.5281/zenodo.7851995](https://doi.org/10.5281/zenodo.7851995))
  - It’s adapted to the common event data model, **EDM4hep**
  - It includes a wrapper class to run in the **Gaudi** software framework
  - It’s included in the new **Key4hep** releases managed by Spack
Additional features w.r.t. **kalos/Clue**

- **Cluster hits in the entire 4π detector region**
  - Definition of the tessellated space (LayerTile) in the standalone version defines coordinates and searches only in the transverse plane
  - Modified basic structure of the LayerTile and the search algorithm to allow for the definition of a cylindrical surface
    \[ x \rightarrow r\Phi \quad y \rightarrow z \]

- **Template CLUE algorithm classes**
  - To allow the possibility of defining several different calorimeter layouts
  - A dedicated documentation page in the package (**include/readme.md**) allows the user to follow a simple but detailed step-by-step procedure to introduce and test the preferred layout.

- **GitHub CI & EDM4hep Validation**
  - **edm4hep:CLUECalorimeterHit**: CalorimeterHit class with specific methods related to the CLUE algorithm
Performance evaluation
- CLD & CLICdet -
ECAL of CLICdet & CLD

- 40 layers of 5x5 mm$^2$ Silicon cells & W
- The main difference between the two calorimeters lies in the layout parameters → To compensate for a lower detector solenoid field, the CLD design starts from a larger radius both in the barrel and in the endcap region w.r.t. CLICdet.
  - Further details in backup
ECAL of CLICdet & CLD

- 40 layers of 5x5 mm² Silicon cells & W
- The main difference between the two calorimeters lies in the layout parameters → To compensate for a lower detector solenoid field, the CLD design starts from a larger radius both in the barrel and in the endcap region w.r.t. CLICdet.
  - Further details in backup
- 500 events of single gamma at 10 GeV generated perpendicular to the surface with Geant4 General Particle Source
  - Main reason: no conversion in the tracker volume

Gun from vertex

→ many secondaries simulated

Gun from ECAL boundaries

→ mostly primaries
Parameters tuning

- Input parameters tuned for CLD
- Same ones tested also for CLICdet (similar geometry, same granularity)
- Critical Distance ($d_c$) is established by geometry granularity to contain (minimum) the close neighbors cells:
  - $d_c = 15$ mm
Parameters tuning

Varying the outlier factor ($o_f$)

Varying the minimum local density ($\rho_c$)

- CLD, 10 GeV gamma
  - $o_f = 1, \rho = 0.02$
  - $o_f = 2, \rho = 0.02$
  - $o_f = 3, \rho = 0.02$

- CLD, 10 GeV gamma
  - $\rho_c = 0.1, o_f = 0.01$
  - $\rho_c = 0.2, o_f = 0.02$
  - $\rho_c = 0.3, o_f = 0.02$

Erica Brondolin (erica.brondolin@cern.ch)
Higher energies

- 500 events with single gamma (from ECAL surface) at 100 GeV
- $d_c = 15.00$, $\rho_c = 0.02$, $\alpha_f = 3.0$
Multiple gamma event

- Produced with normal gun, i.e. particles generated from vertex
- 1 event with 500 single gammas each produced with 10 GeV

Only simulated calorimeter hits are shown
CLICdet results

- Using same input parameters selected for CLD

Comparison with Pandora Clusters not completely equitable comparison (it includes a dedicated calibration procedure), but comparable results in terms of energy linearity and resolution.
Performance evaluation
- Noble Liquid Calo -
Noble Liquid ECAL for FCC-ee

- 12 layers, only barrel considered
  - cell size in $\Phi$: 17.9 mm - 20.7 mm
  - cell size in $\eta$: ~ 20 mm
- Sample (if not stated otherwise):
  - 500 single gamma at 10 GeV
  - $\theta_{[\text{min}, \text{max}]} = [50, 130]$
Parameters tuning

- 500 events with single gamma (from vertex) at 10 GeV
- \( d_c = 40.00, \ \rho_c = 0.03, \ \phi_f = 3.0 \)

Input parameters:
- \( d_c \): Critical Distance
- \( \phi_f \): Outlier Delta Factor
- \( \rho_c \): Minimum Local Density
Comparison with other cluster algorithms

- **Sliding window**: It considers the calorimeter as a two-dimensional grid in $\eta$-$\phi$ space, neglecting the longitudinal segmentation of the calorimeter.

- **Topological clustering**: It starts with a seed cell and then adds topologically connected calorimeter cells.

CLUE creates about $\sim$10 clusters per event (up to few GeV per layer)
Noise in Liquid Argon Calorimeter

- High level of noise in the detector
- In the topoclustering, there is no filter directly at the beginning for the noise, but this is done using cuts in the algorithm itself
- The main observable is the cell significance $\xi_{\text{cell}}$ which is defined as the absolute value of the ratio of the cell signal to the expected noise in this cell
  \[ \xi_{\text{cell}} = \frac{|E_{\text{cell}}|}{\sigma_{\text{cell}}^{\text{noise}}} \]
- CLUE hits w/noise selected with filter of $> 2\sigma_{\text{noise}}$
Comparison with other cluster algorithms

No noise

Including noise

LCWS 2023

Erica Brondolin (erica.brondolin@cern.ch)
Low(er) energy

- 500 events with single gamma (from vertex) at 2 GeV
  Motivated by flavor physics searches at Z peak
Conclusions & Outlook
Conclusions

● k4Clue package (v01-00) has improved upon the standalone CLUE
  ○ Run on the full detector (barrel & endcap)
  ○ Adapted for different types of calorimeters

● Analysis on three different future calorimeters has demonstrated the good performance for single gamma events
  ○ Good performance even in the presence of noise
  ○ Compared favorably to other baseline algorithms

This work highlights the adaptability and versatility of the CLUE algorithm for a wide range of experiments and detectors, as well as its potential for future high-energy physics experiments beyond CMS
  ○ Improvements from k4clue also under discussion to use the developments also in CMS (Phase-2 barrel region)
Conclusions

● Final article summarizing k4clue and its performance for future collider detectors almost ready
  ○ Computing time under study

● This research was supported by the CERN Strategic R&D Programme on Technologies for Future Experiments

● Special thanks go to the Key4hep team and the FCC-ee liquid calorimeter software experts for the support
Backup
Integrating CLUE in Key4hep

- **GitHub CI** to ensure that the modifications or additions to the software do not break the clusterization process
  - In the latest release was modified to focus on C++ code and EDM4hep data
  - Test on both current key4hep release and nightlies
- **Validation**
  - edm4hep:CLUECalorimeterHit
    - CalorimeterHit class with specific methods related to the CLUE algorithm
  - CLUEHistograms class to produce ntuples
CLICdet results

- Using same input parameters selected for CLD

Comparison with Pandora Clusters not completely equitable comparison (it includes a dedicated calibration procedure), but comparable results in terms of energy linearity and resolution
Noble Liquid Calo

Pre-filtering

- CLUE hits w/noise selected with filter of $> 2\sigma_{\text{noise}}$
Noble Liquid Calo
Comparison with other cluster algorithms

No significant effect on CLUE clusters - about ~10 per event (one per layer)

More than one SW and Topo cluster per event, but most of them with low energy
Noble Liquid Calo

Summary for 10 GeV gammas

**Sliding Window**
- Cluster energy $> 1$ GeV

**Topological**