Applications multi-GPUs au CC-IN2P3 pour les analyses et simulations : apports, expériences

Workshop GPU@CC-IN2P3

Gilles Grasseau, LLR
3-5 avril 2019, CC-IN2P3

Intro.

Outlines

Reprises DecaLog ComputeOps

Context

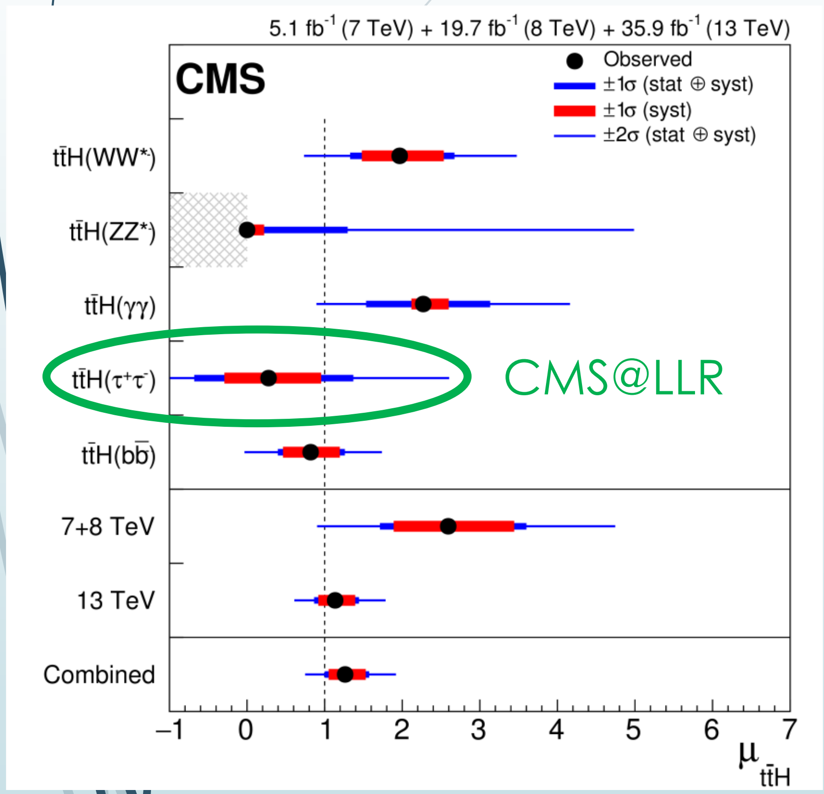
Master-projet IN2P3 DAS « Calcul & Données »
DecaLog (G.Grasseau)

- ComputeOps (C.Cavet)
- **Reprises** (D.Chamont)

Two parts

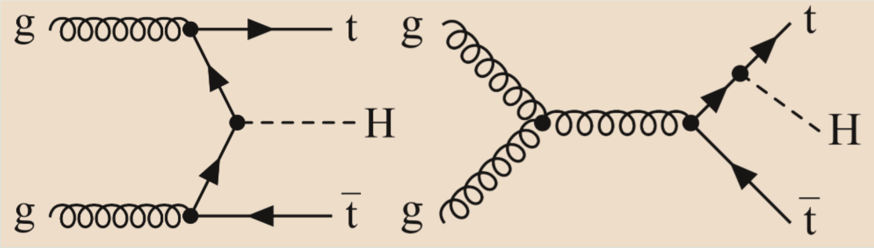
- MEM-GPU production code multi-GPUs for ttH analysis
- Applications potentielles pour la production au CC-IN2P3

Recent discovery of H boson in *ttH* channel



We (CMS@LLR) contributed to the $t\bar{t}H \to \tau\tau$ sub-channel

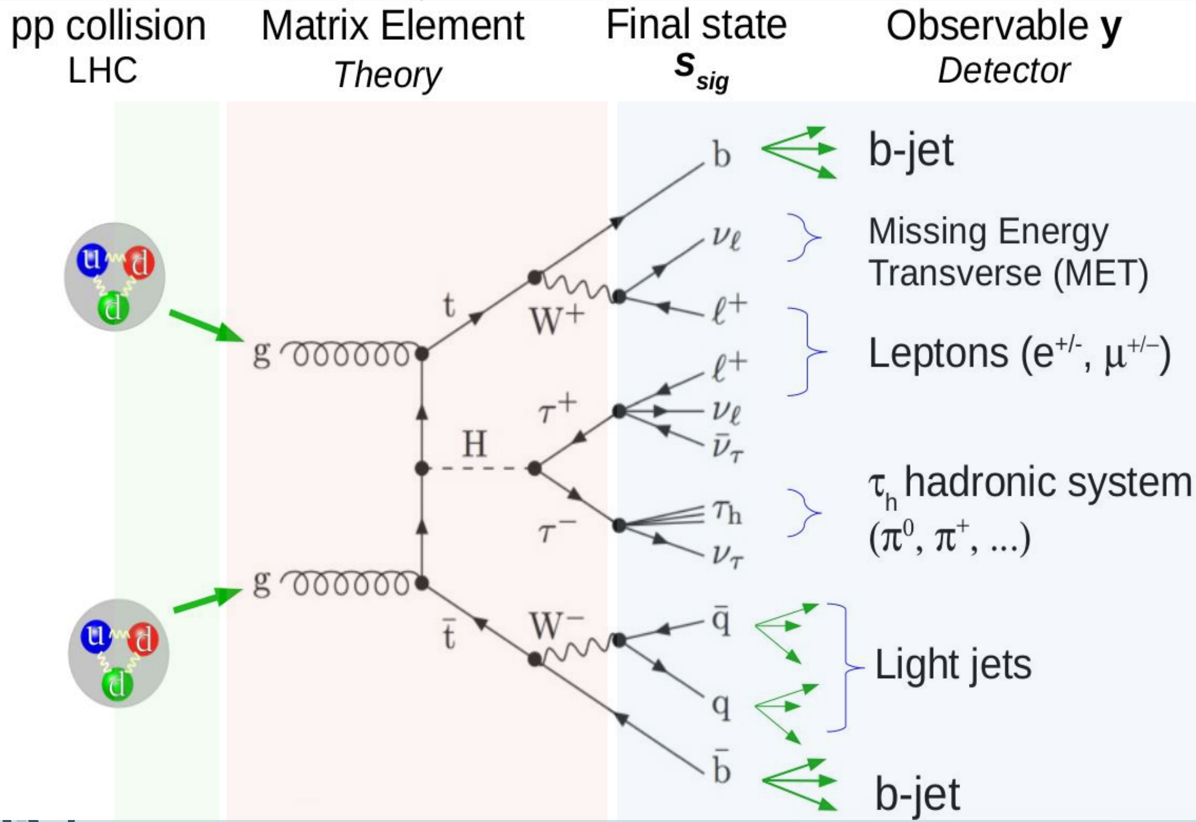
- Higgs decays into $\gamma\gamma, ZZ, WW\ \tau\tau$ and $b\bar{b}$ (2018) final states have been observed (discovery 2012)
- In the SM, the Higgs boson couples to fermions with a strength proportional to the fermion mass (Yukawa coupling)
- Probing the coupling of the Higgs boson to the t quark, the *heaviest* known fermion, was a high priority
- The decay to the $t\bar{t}$ final state is not kinematically possible
- The Higgs boson in association with $t\bar{t}$ final state can result from the *fusion* of a $t\bar{t}$ pair or through a *radiation* of t quark



- First observation* of the simultaneous production of a Higgs boson with a $t\bar{t}$ pair (channel) April 2018

*A. M. Sirunyan et al. (CMS Collaboration), "Observation of $t\bar{t}H$ Production", Phys. Rev. Lett. 120, 231801 (2018)

Matrix Element Method (MEM)



MEM is an *unsupervised* method (theory-driven) which is important to have among the supervised ones (Machine Learning, …)

Principle:

- select a Signal final state S_{sig} :

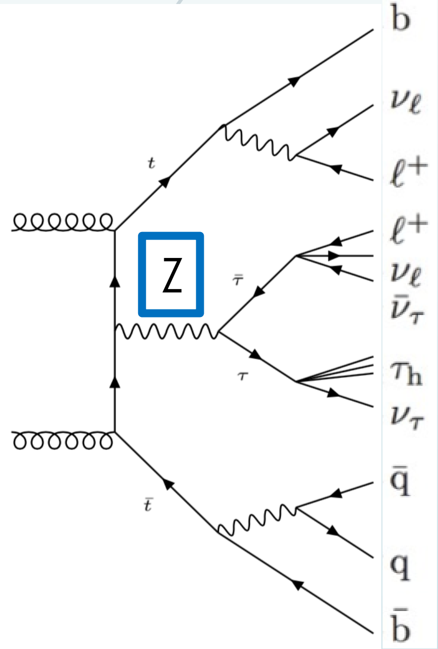
  $b\bar{b}, q\bar{q},\ \tau_{had}$, *2 leptons same sign*
- compute a weight quantifying the probability that an observed event **y** matches a theoretical model **x**
- vary the theoretical model (Signal, background(s))
- deduce a likelihood ratio

p processes — Parton Density Function (PDF) — Kinematics constrains — Matrix Element — Transfer Function *Response of the detector*

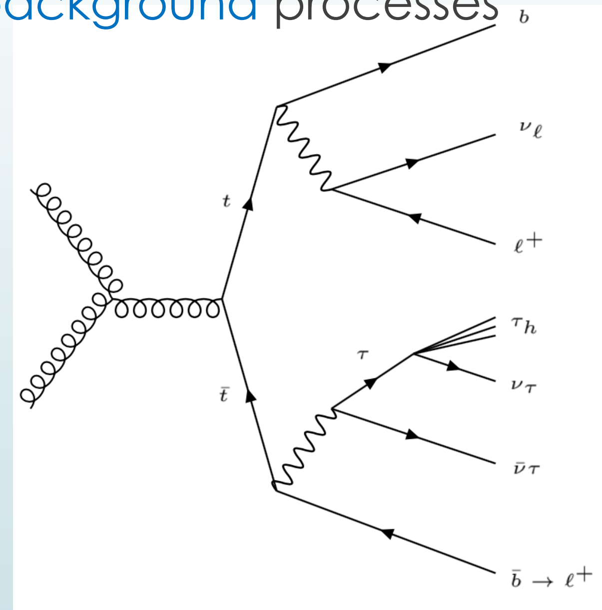
$$w_i(\mathbf{y}) = \frac{1}{\sigma_i}\sum_p \int d\mathbf{x}\, dx_a\, dx_b \frac{f(x_a, Q) f(x_b, Q)}{x_a x_b s}\delta^2\left(x_a P_a + x_b P_b - \sum p_k\right)|\mathcal{M}_i(\mathbf{x})|^2 W(\mathbf{y}||\mathbf{x})$$

MEM: time-consuming computations

- Multiple scenarios to consider (compute one integral for each) : the signal process and the background processes

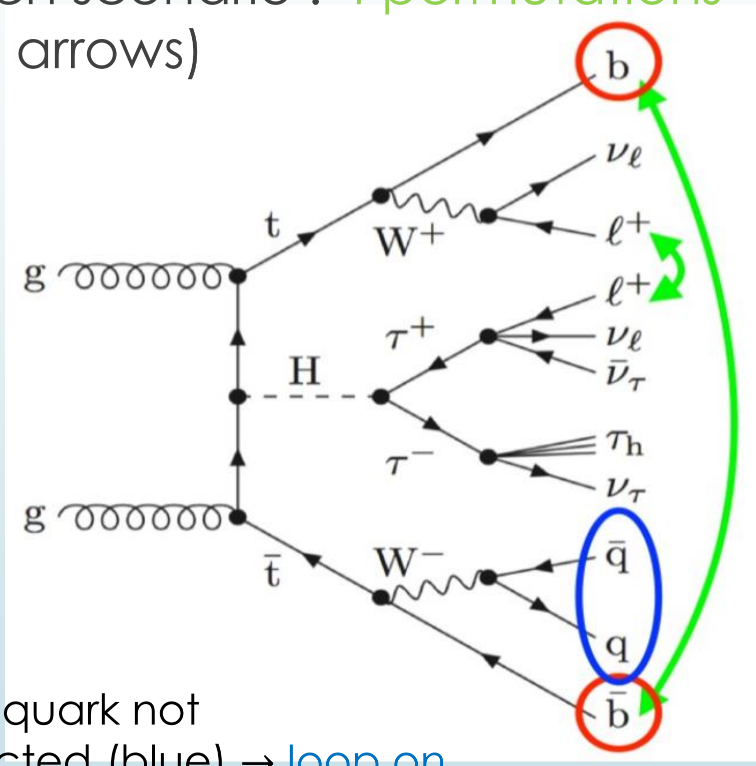


Irreducible background



One background: one non-prompt lepton produced in a b decay

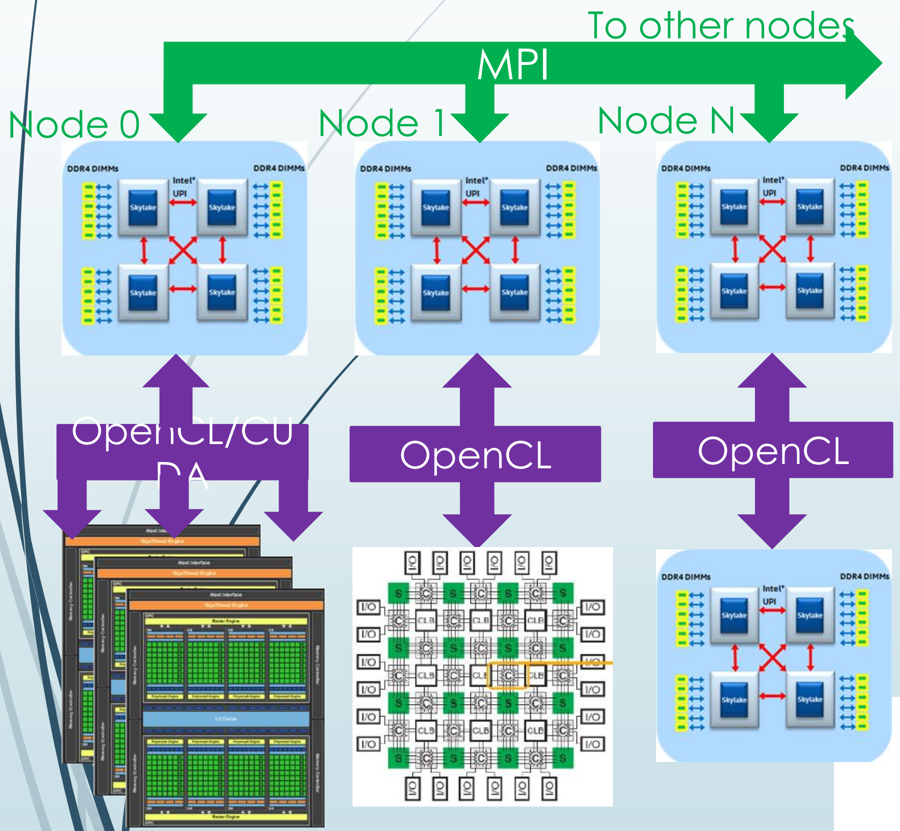
- For each scenario : 4 permutations (green arrows)



Only one quark not reconstructed (blue) → loop on all "light-jets"

*(1+3) * 4 [* #Ligth-jets]* Integrals with a dimension from 3 to 7.
They are computed if they are kinematically possible

The MEM implementation



Multi-GPUs on one node

Other hardware *OpenCL compliant*

- The processing time for a typical data set (2395 evts) 55 days (14 hours / 96 cores)
- MEM code features: MPI/OpenCL/Cuda to aggregate numerous computing resources (HPC)
- Main kernel (one Vegas iteration)
  - developed a MadGraph extension to generate the OCL/Cuda kernel codes
  - LHAPDF lib.: Fortran to C-kernel translation
  - ROOT tools: Lorentz/geometric arithmetic's
  - $\rightarrow$big kernels (10-20 x 10^3 lines)
- OpenCL / Cuda bridge (IBM+NVidia)
- Lot of asynchronous to deal with (MPI, communication/computing, several GPUs per nodes)

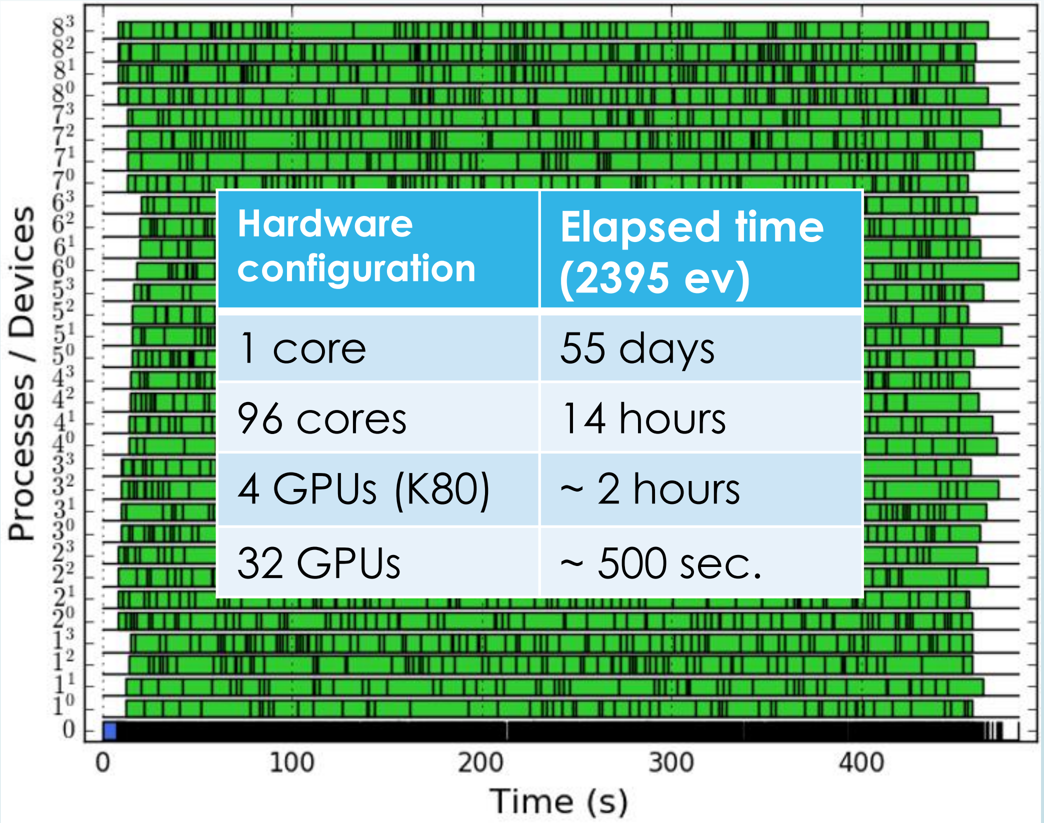
MEM-GPU

MEM-GPU Performance

Performance

- Exploitation : 3rd PhD thesis (C. Martin-Perez) - *ttH*
- Enabling these computations (MPI, GPU) help our physicists (LLR) to have an impact in this CMS collaboration analysis
- Improvements to do:
  - Performances
  - Code generation extension
- Ready for HL-LHC analysis

CC-IN2P3 – NVidia K80 platform



Hardware configuration	Elapsed time (2395 ev)
1 core	55 days
96 cores	14 hours
4 GPUs (K80)	~ 2 hours
32 GPUs	~ 500 sec.

G. Grasseau et al, *Deployment of a MEM code for the ttH channel analysis on GPU's platform - CHEP'18 Proceeding*

R&D Projects (1)

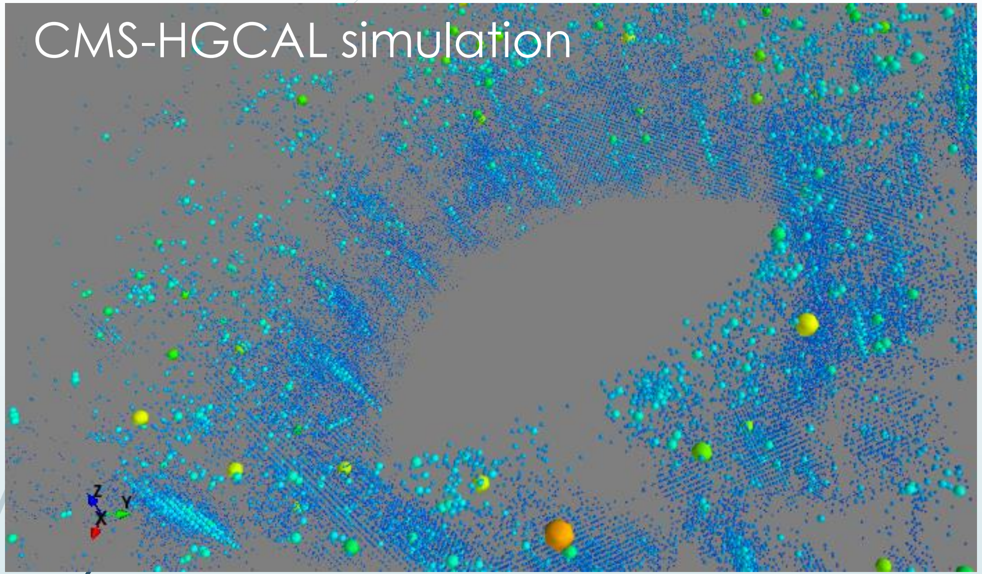
R&D Projets

- « Patmos » : Simulation MC simulation de transport - neutron transport (Tripoli, ~Geant) : heterogeneous HPC, evaluate different programing paradigms: CUDA, OpenMP, OpenACC, Kokkos SEMA/CEA
- « MC simulations for medical or nuclear physics applications (accurate) and image reconstruction for Tomography TEP/TEMP/TMD image data » - Multi-GPUs - IPHC
- "Nuclear Astrophysics simulations », supernova simulation - OpenACC - IPNO
- « Tracking with Kokkos » : CPU et GPU approach - LAL
- « FPGA/OpenCL » : bridge FPGA/GPU, ML & FPGA interactions -
- « HGCAL-3D » DL for sub-detector HGCAL (ObjectDetection) - LLR

Other close activities

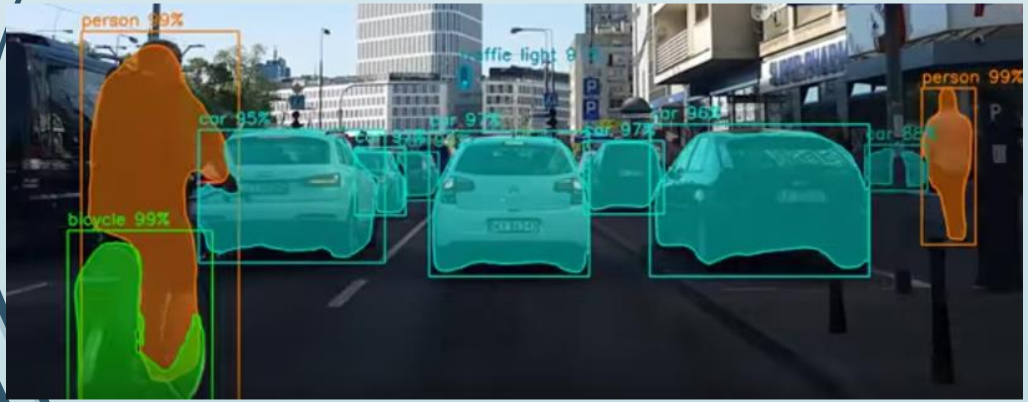
- Simulations MC Corsika (vectorisation)- LUPM,
- CTA - *ASTERICS-OBELICS International School* (*Advanced software programming for astrophysics and astroparticle physics*) 8-12 avril 2019, LAPP https://indico.in2p3.fr/event/18333/page/2035-scientific-programme
  - Machine learning class
  - GPU class (2018)

DL for new CMS sub-detector

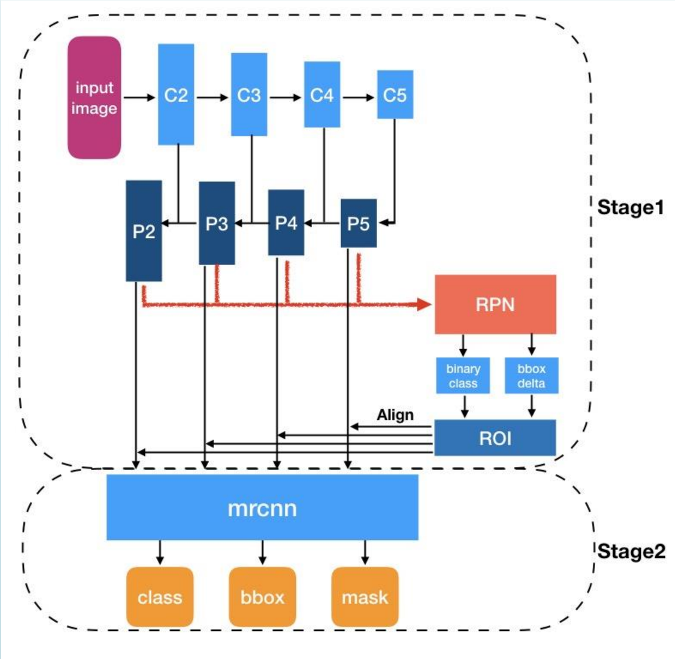


GSOC'19 proposal, HAhRD (LLR)

Mask R-CNN for **3D**



- R&D project motivated motived by par les physicists & collaboration for HGCAL
- Mask-RCNN : reduced problem to 2D HGCAL "images" works
- 3D Training … need a high computing power (CC-IN2P3)



Conclusion

- Great benefit to use GPU platforms for our analysis
- Rewriting for GPUs improve CPU performance
- "Reprises" gathers expertise, experiments in GPU (programing models), … Next Meeting F2F April 9th
- Importance of Production GPU-Platforms (CC-IN2P3)
- Importance of Development GPU-Platforms (more local) : to build the application, optimize it, no tools limitations, feeds the CC-IN2P3 GPU platforms
- Platform ACP (Labex P2IO) – 2 x NVidia V100