ACAT 2010
Summary Track 2
Data analysis - Algorithms and Tools

Liliana Teodorescu

Brunel UNIVERSITY WEST LONDON

Topics

Statistics
Multivariate Analysis
Event Reconstruction Algorithms
Software Packages for Data Analysis
Paralellisation of Algorithms

Statistics

Absorbing systematic effects to obtain a better background model in a search for new physics

Sascha Caron[1], Glen Cowan[2], Eilam Gross[3],
Stephan Horner[1] & Jan Erik Sundermann[1]

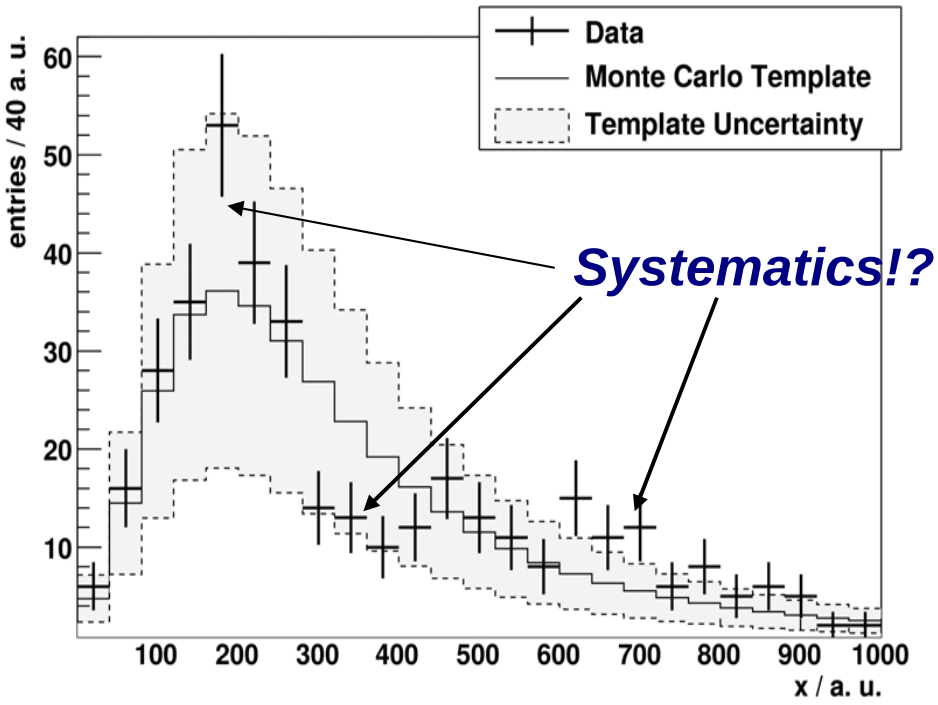
[1]Physikalisches Institut, University of Freiburg
[2]Physics Department, Royal Holloway, University of London
[3]Dep. of Particle Physics, Weizmann Institute of Science, Rehovot

Method

New approaches to obtain a background prediction for the Signal Region:

To verify Monte Carlo find region in phase space, Control Region, satisfying:
- ideally contain only known physics
- observable of interest x - similar physical meaning and dependence on systematic effects in Control and Signal Region

Measurement on Control Region



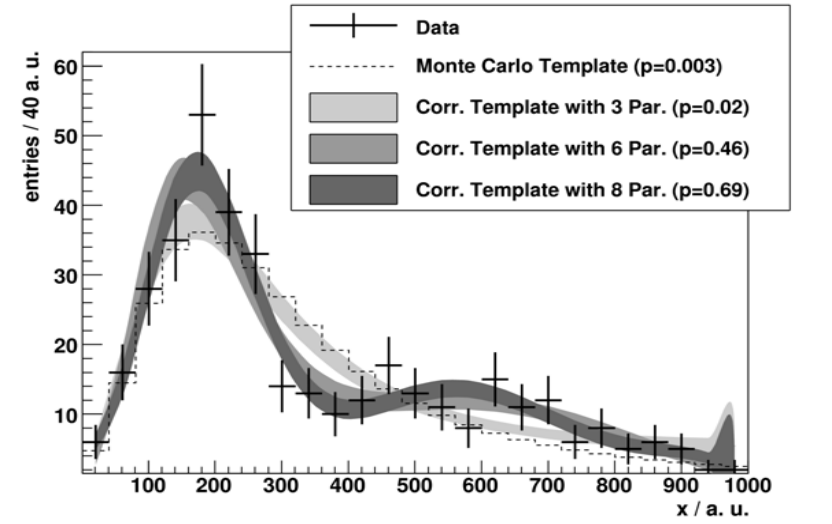
1. Multiply the MC template with a correction function

*Model_x = Template * Polynomial with x parameters*

2. Fit the modified template to the data to determine parameters

3. Use successively more complex correction functions until satisfactory goodness-of-fit is reached (*p*-Value)

Model selection and uncertainty in the starting model



Absolute goodness-of-fit:

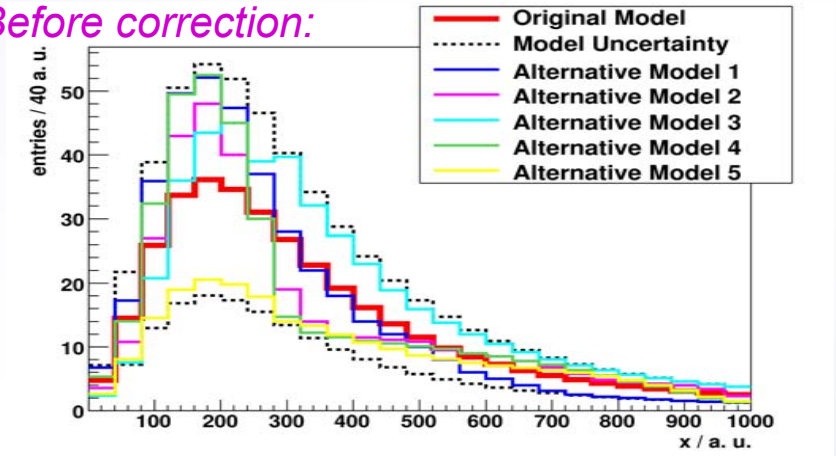
p(Model_0) = 0.0027
p(Model_1) = 0.0033
p(Model_5) = 0.33
p(Model_7) = 0.46
p(Model_8) = 0.69
p(Model_9) = 0.63

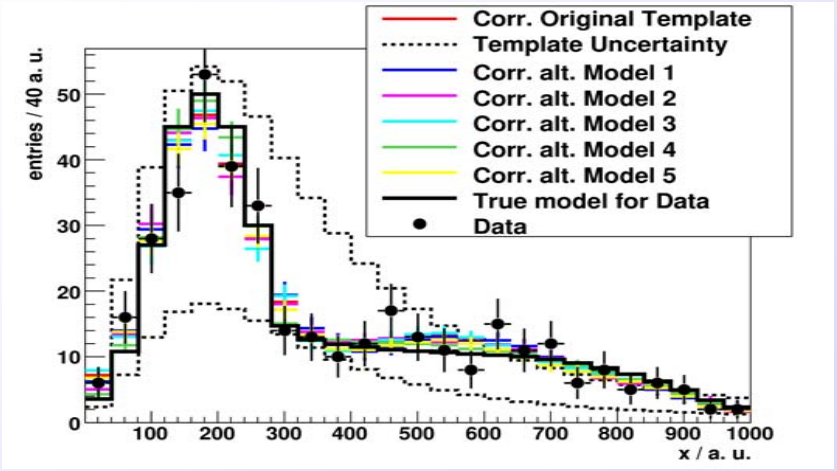
Relative goodness-of-fit:

p(Model_0 | Model_1) = 0.15
p(Model_7 | Model_8) = 0.04
p(Model_8| Model_9) = 0.80

Alternative starting templates - vary Monte Carlo prediction according to known systematic effects - has little effect

Before correction:

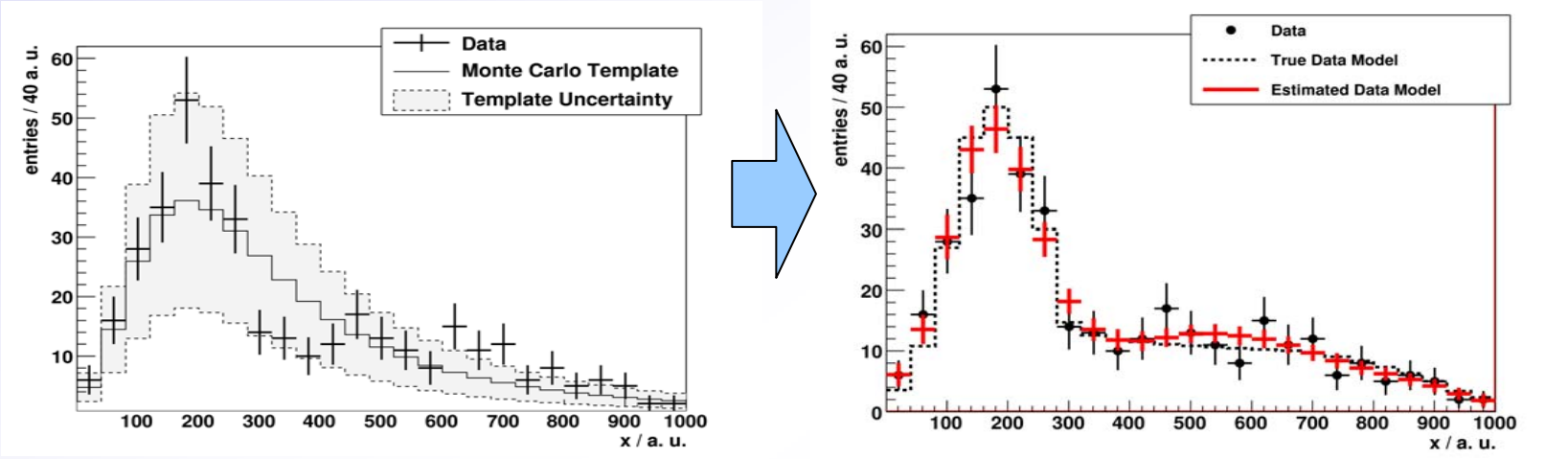




Bkg. estimation

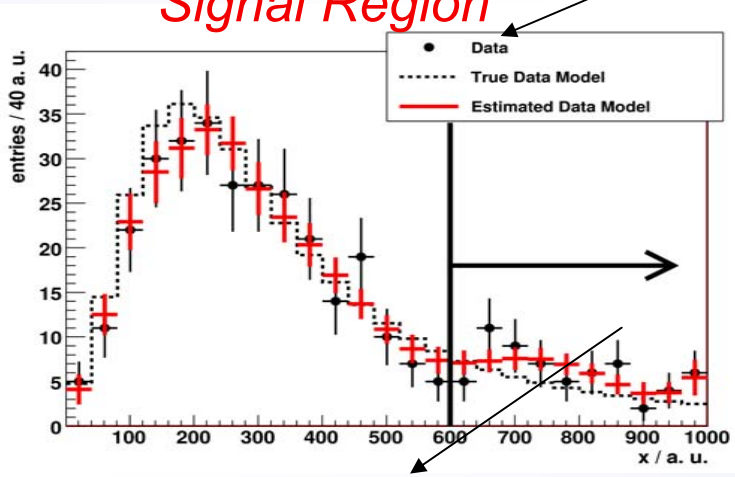
Control Region

Large systematics absorbed and uncertainty reduced!



Signal Region

from Control Region!



Region of interest to look for new physics

Expected bkg events

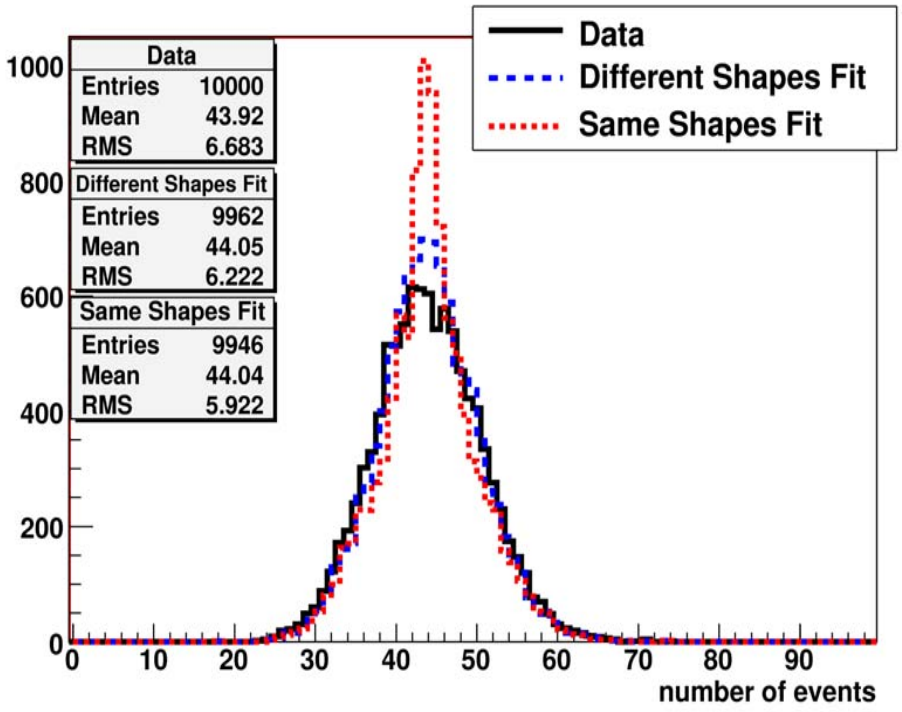
Sum up bins taking into account the correlation
Compare with the data from Control Region

Model	Number of expected events	Relative error
Original prediction (MC template)	43.9 ± 21.9	50%
Corrected model	59.9 ± 7.6	12.7%
Data as model	62.0 ± 7.9	12.7%

Significance

Many experiments - 10.000 toy data sets from true model and apply method

Same starting templates as before



Discovery Significance

	x > 600 a. u.: 99 events counted Bgrd predicted: (true value 43.89)	Significance:
Data	43.92 ± 6.683	5.01
Different Shapes	44.05 ± 6.222	5.15
Same Shapes	44.04 ± 5.922	5.25

Equivalent to 4% luminosity increase

Method has smaller uncertainty than using the data as a model

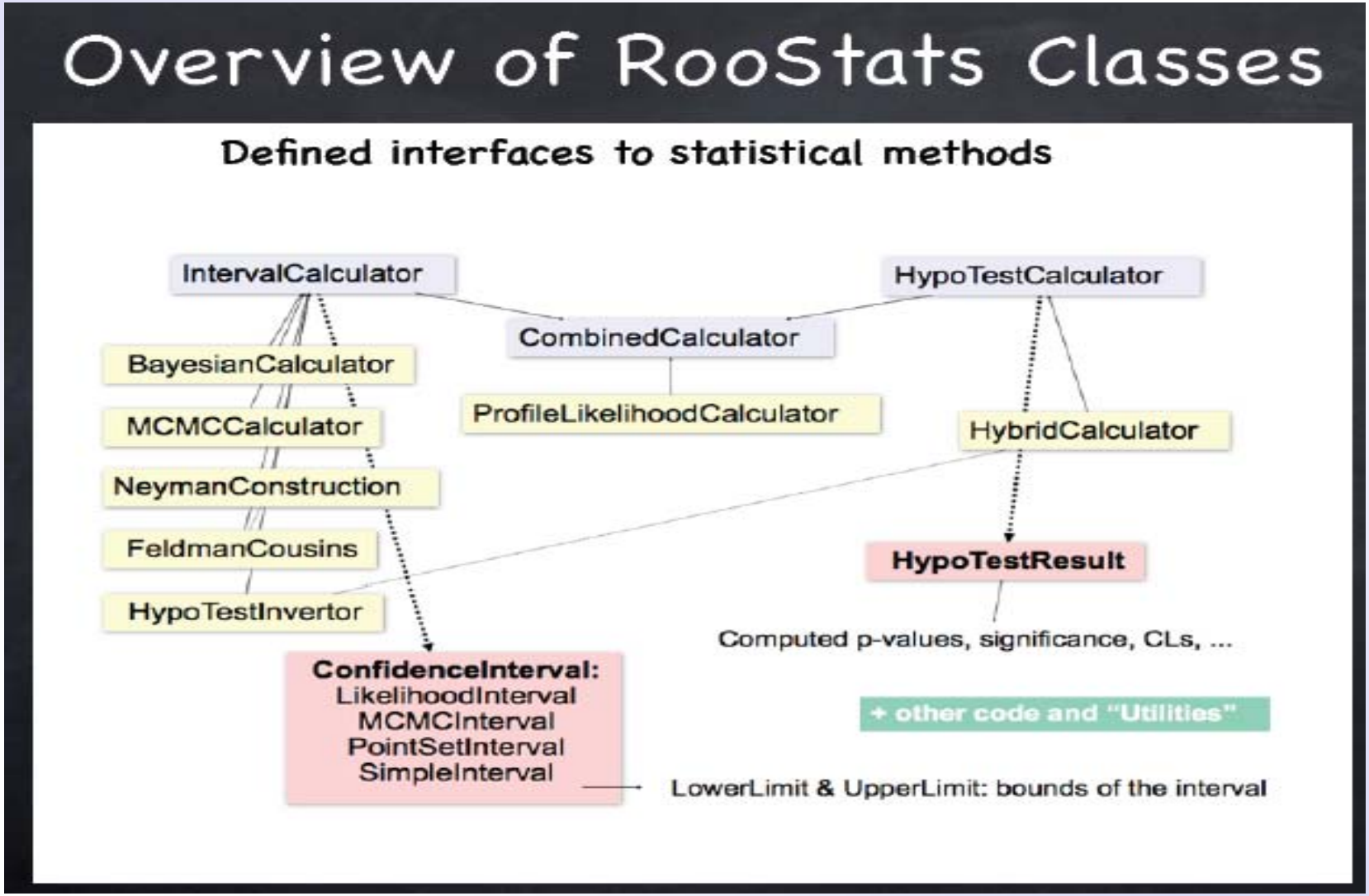
The RooStats Project

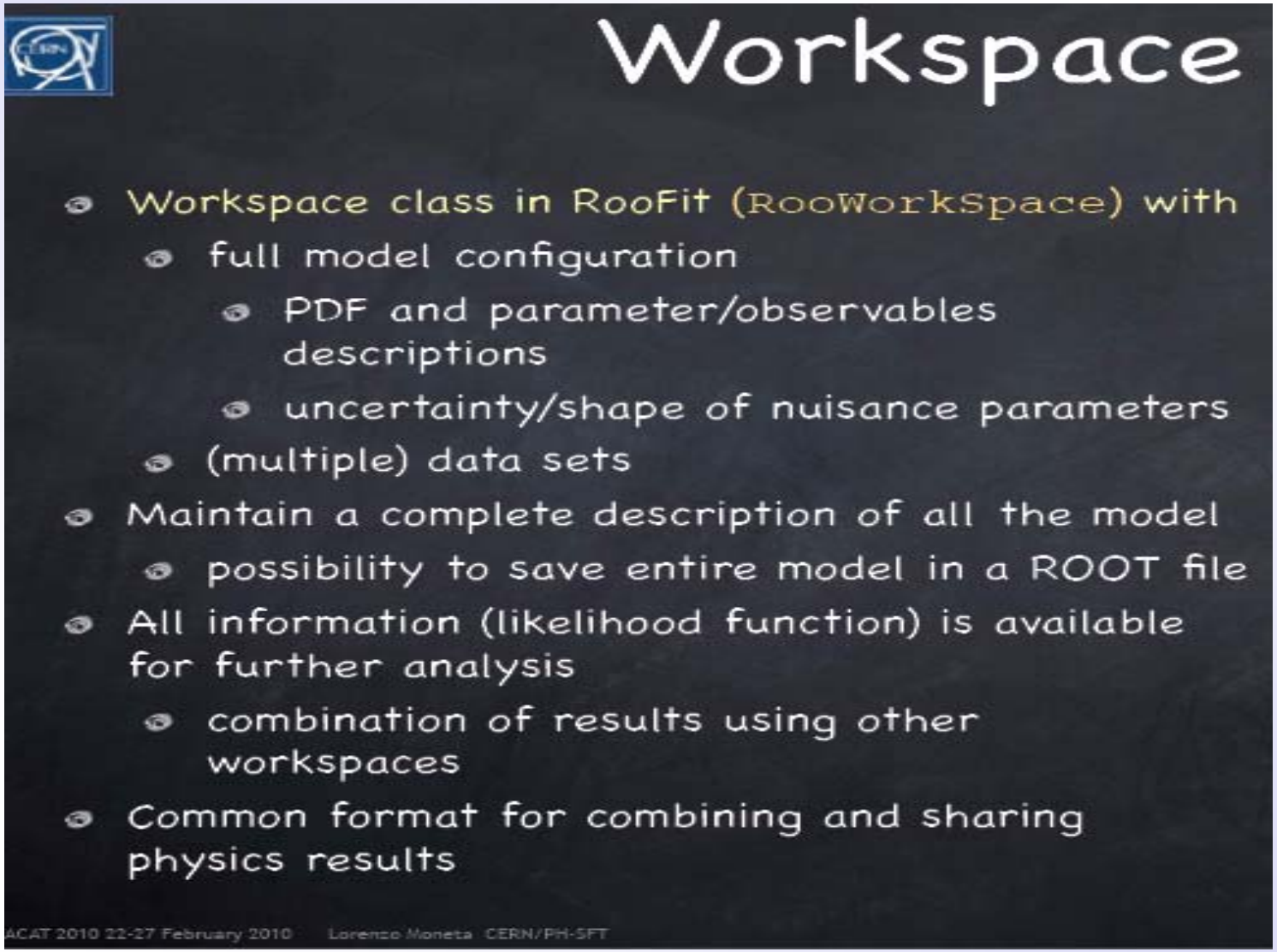
K.Cranmer (New York University, ATLAS)
L. Moneta (CERN, ROOT)
G. Schott (Karlsruhe Institute of Technology, CMS)
W. Verkerke (Nikhef, ATLAS and RooFit)
and also contributions from
D. Piparo (CMS), M. Pelliccioni (CMS), A. Lazzaro (ATLAS)
N. Ruthmann (CMS), K. Belasco (ATLAS), M. Wolf (CMS)

presented by A. Lazzaro

- Common framework for statistical calculations
  - work on arbitrary models and datasets
  - implement most accepted techniques (frequentists, Bayesian and likelihood based methods)
  - provide utility for combinations
- Built on top of RooFit
  - provides generic and convenient description of models (probability density function or likelihood functions)

Overview of RooStats Classes




Workspace
Workspace class in RooFit (RooWorkSpace) with
full model configuration
PDF and parameter/observables descriptions
uncertainty/shape of nuisance parameters
(multiple) data sets
Maintain a complete description of all the model
possibility to save entire model in a ROOT file
All information (likelihood function) is available for further analysis
combination of results using other workspaces
Common format for combining and sharing physics results
ACAT 2010 22-27 February 2010 Lorenzo Moneta CERN/PH-SFT

Multivariate Analysis

Data Analysis Techniques with *T*MVA

Jörg Stelzer* (DESY, Hamburg)

Presented by Dr. Attila Krasznahorkay (New York University)

13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Jaipur, India
February 22-27, 2010

*For the TMVA developer team: A. Höcker, P.Speckmayer, J.Stelzer, J. Therhaag, E. v. Törne, H.Voss and many contributors

New developments

TMVA – tool for multivariate analysis

- implements multivariate classification and regression – adapted for HEP

New developments

TMVA 4 first released with ROOT 5.24. New Features:

Multivariate multi-target regression

Generic boosting

Category classifier

New methods: PDEFoam and Bayesian Neural Net

Framework changes: chained data preprocessing, weight-files in xml format (text format can still be read), internal reorganization to prepare for composite classifiers

TMVA development moved to ROOT SVN repository

TMVA releases more tightly coupled to ROOT releases,

Future plans

Finish multiclass classification

Framework implemented. Some classifiers already extended to work for multiple classes (MLP, FDA, BDTGradBoost)

Cross-Validation

Overtraining protection

Automated classifier tuning

Combination of classifiers

E.g., simple classifiers can first combine uncorrelated variables, their output can be fed into other classifiers → performance improvement

Classifying extremely imbalanced data sets

Markward Britsch[1], Nikolai Gagunashvili[1,2], Michael Schmelling[1]

[1]Max-Planck-Institut für Kernphysik, [2]University of Akureyri, Iceland

Classifier

Imbalance data sets – data sets with much more background than signal
In this study background to signal ration ~ 3000

- *Used RIPPER classifier – rule based*

```
(V1 >= 1.039316) and (V2 <= 0.307358)
and (V3 <= 0.270767) and (V4 >= 0.800645)
=> class=Lambda
(V1 >= 0.637403) and (V2 <= 0.159043)
and (V3 <= 0.12081) and (V5 >= 149.2332)
and (V3 >= 0.003371)
=> class=Lambda
=> class=BG
```

- *introduced a cost factor as a weight for the events incorrectly classifies -> new rules developed*

	pred. BG	pred. signal
tr. BG	0	$C(\mathrm{BG}, \mathrm{s})$
tr. signal	$C(\mathrm{s}, \mathrm{BG})$	0

- *apply bagging*

Data and preselection

- $D^0 \to \pi^+ + K^-$
- LHCb minimum bias Monte Carlo, $3.6 \cdot 10^7$ events from 2006, $\sqrt{s} = 14$ TeV
- candidates: pairs of differently charged tracks passing through full spectrometer
- distance of closest approach < 10 mm
- use 14 geometric and kinematic variables

- for preselection: extra classification step:
  1. preclassification incl. bagging – high cost for loosing D^0 $\to$ keep almost all D^0s, reduce background (BG)
  2. classify including bagging with high cost for wrongly accepted BG
  3. to produce ROC curve: scan cost x (one classifier model per point in ROC curve)

	pr. BG	pr. D^0
tr. BG	0	1
tr. D^0	200	0

preselection cost matrix

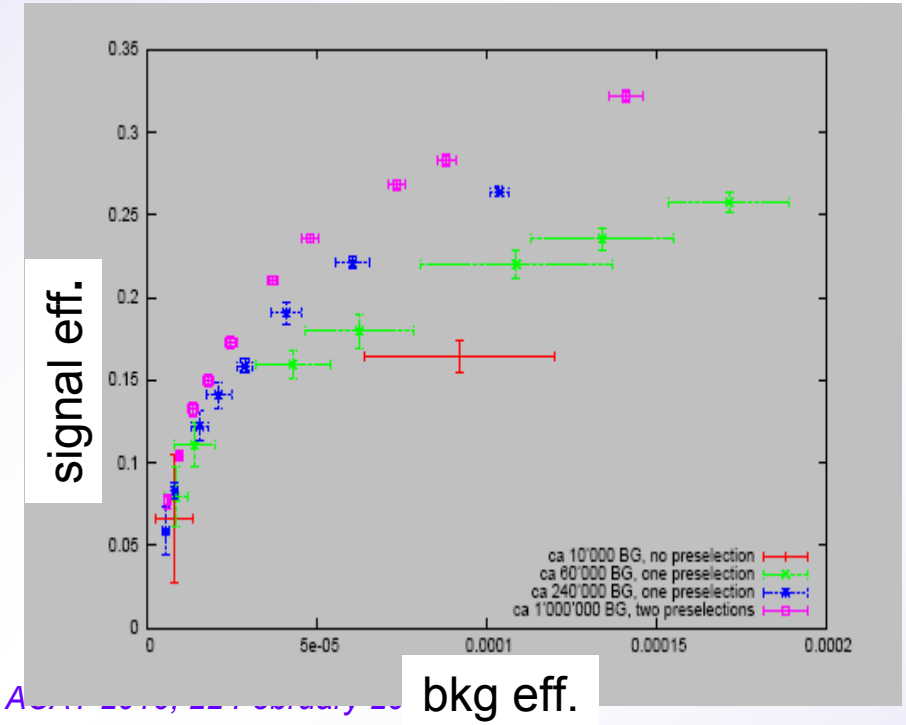
	pr. BG	pr. D^0
tr. BG	0	x
tr. D^0	1	0

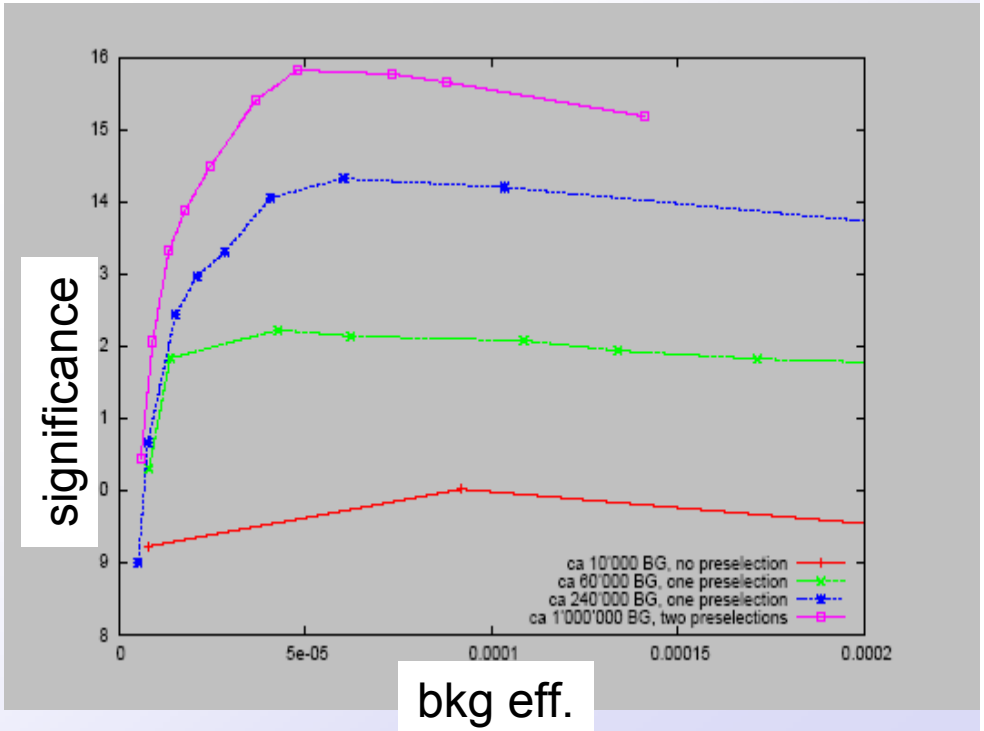
main cost matrix

Results

- training data sets: same number of signal
  increasing number of background

data set	# BG	# sig.	# presel.
test	$6.5 \cdot 10^6$	1827	–
training small	ca 10'000	1851	0
training mid	ca 60'000	1851	1
training larger	ca 240'000	1851	1
training largest	ca 1'000'000	1851	2

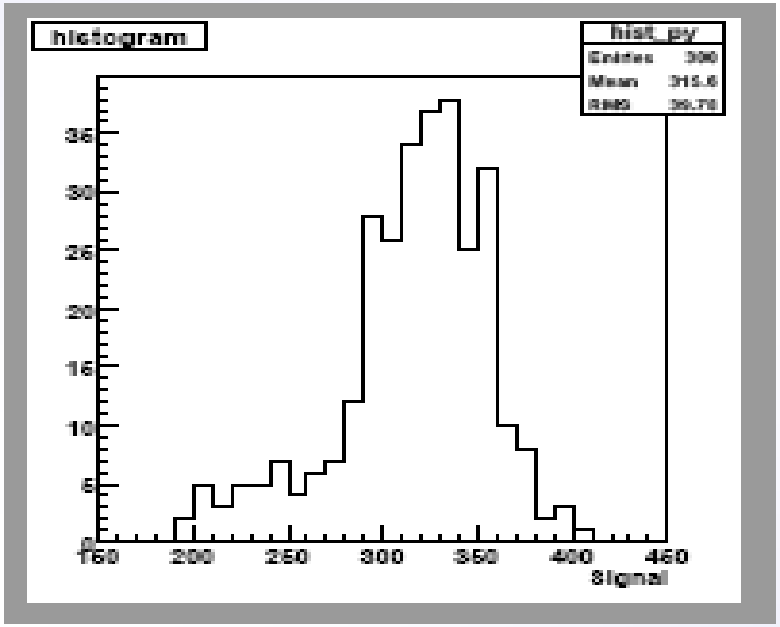


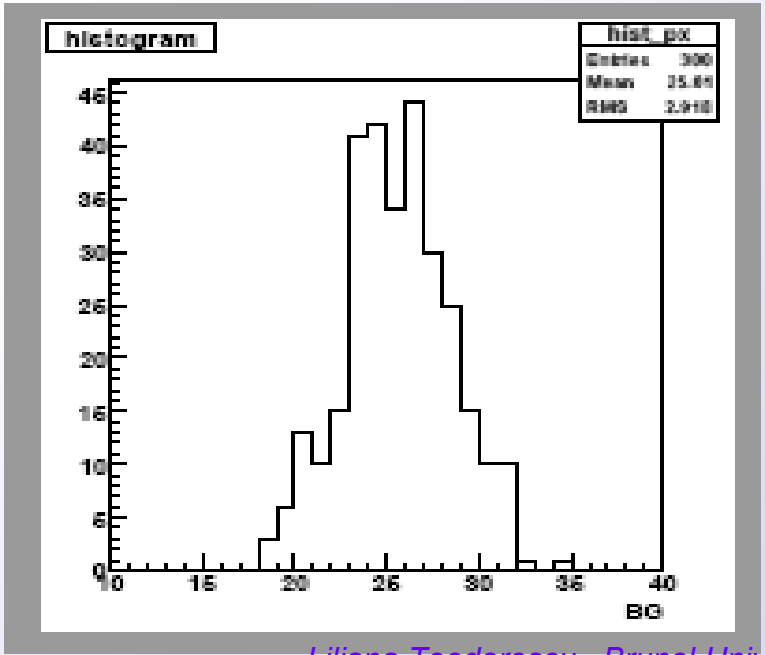


Errors on ROC curve

This is our (ad hoc) method:

- do each main selection 10 times with different random seeds
- take the mean FPR and TPF as the point in ROC space
- similar to using 10 cross-validation samples
- take the standard deviations (SD) as errors in *x* and *y*
- the result is what you have seen in the plots





Analysis of Photoluminescence measurement data from interdiffused Quantum Wells by Real coded Quantum inspired Evolutionary Algorithm

Ashish Mani and C. Patvardhan

Dayalbagh Educational Institute, Agra, India

Quantum inspired Evolutionary Algorithms (QiAE)

EA

Initialize
Evaluate
Do {
Select Parents
Recombine
Mutate
Evaluate
Select for next Generation
} While (!Termination_Criteria)

QiEA

- *EA combined with concepts of Quantum Computing*
- *Incorporates Quantum Mechanics principles such as Superposition, Entanglement, Interference and Measurement.*
- *Principles mostly utilized are superposition and measurement for improving diversity.*

Qubit and QiEA

Qubit

- *the smallest information element in quantum computing (quantum analog of a classical bit)*
- *represented by a the vector $|\psi\rangle$*

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle.$$

$|\alpha|^2$ and $|\beta|^2$ - the probability amplitudes of the qubit to be in state $|0\rangle$ and $|1\rangle$

$$|\alpha|^2 + |\beta|^2 = 1.$$

QiEA

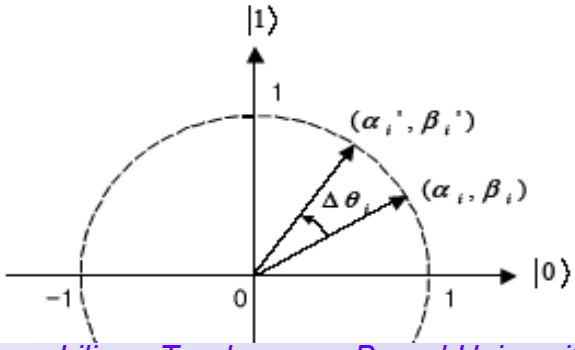
Uses a q-bit representation of the chromosome (candidate solution)

$$Q = \{q_1, q_2, \ldots, q_n\}$$

Each q-bit is defined as a pair of numbers (α,β) – representation of the solution

Evolution – apply rotation (quantum gate) on each

$$\begin{bmatrix} \alpha_i' \\ \beta_i' \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}$$



QiEA - version with two qubits

$|\psi_1(t)\rangle$ *stores the variables of the solution (which will be optimised)*

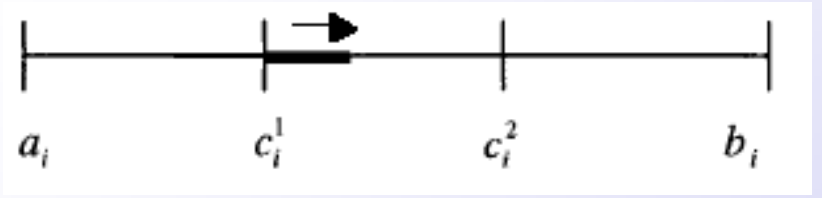
$|\psi_2(t)\rangle$ *stores scaled and ranked objective function value of the solution*

define the rotation crossover operator used for evolving the first qubit

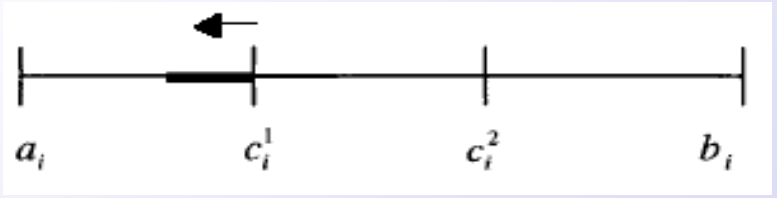
$$\psi_{1i}(t+1) = \psi_{1i}(t) + f(\psi_{2i}(t), \psi_{2j}(t))*(\psi_{1j}(t) - \psi_{1i}(t))$$

$f(\psi_{2i}(t), \psi_{2j}(t))$ *generates a random number either between* $(0, |\alpha_{2i} - \alpha_{2j}|)$ *or* $(0, ||\alpha_{2j}|^2 - |\alpha_{2i}|^2|)$.

Rotation towards Best or Better



Rotation away from Worst



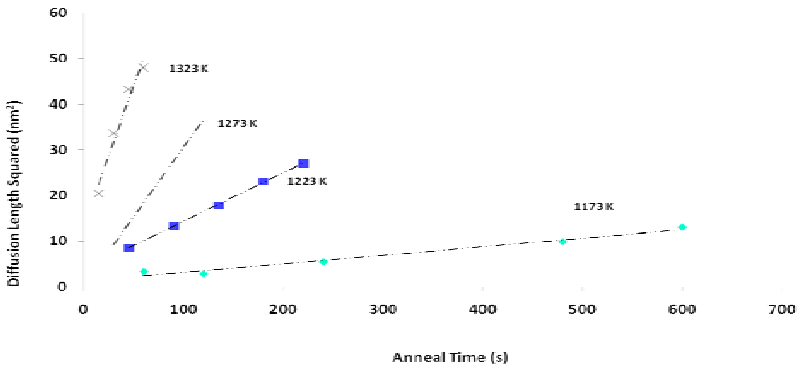
QiEA for photoluminescence

Interdiffusion coefficient $L_D^2 = 4*D(T)*t$ *where* $D(T) = Do * exp(-Ea/(kB*T))$

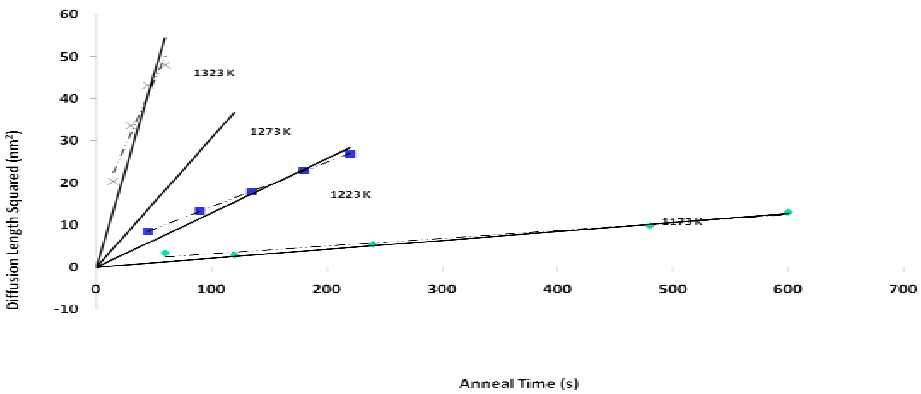
where kB is Boltzmann constant ; T is annealing temperature

Parameters to be determined: activation energy-Ea; interdiffusion prefactor -Do,

Experimental data



QiEA fit



Fitness Function: $e = \sum_i \sum_j [L_D^2 (t_j, T_i) - 4 * t_j * Do * exp(-Ea/(kB*T_i))]^2$

Temperature (K)	RQiEA			GA			LSA
	D_o X 10^{-3} (cm^2/s)	E_a (eV)	D (T) X 10^{-16} (cm^2/s)	D_o X 10^{-3} (cm^2/s)	E_a (eV)	D (T) X 10^{-16} (cm^2/s)	D (T) X 10^{-16} (cm^2/s)
1173	1.00	3.090	0.744	1.06	3.099	0.536	0.46
1223	0.99	3.031	2.560	1.01	3.037	3.208	2.63

Analysis algorithms

Electron/Jet Neural Discrimination based on Nonlinear Independent Components for ATLAS Second-Level Trigger

Eduardo Simas, José Seixas and Luiz Calôba

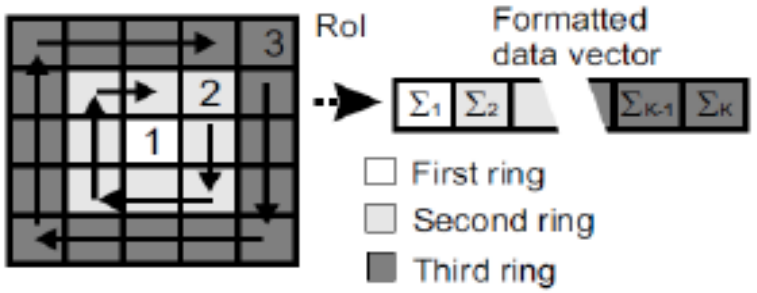
Signal Processing Laboratory

Federal University of Rio de Janeiro – Brazil

Signal Pre-processing

ACAT 201

- Here the calorimeter signals are pre-processed using a ring-like structure:

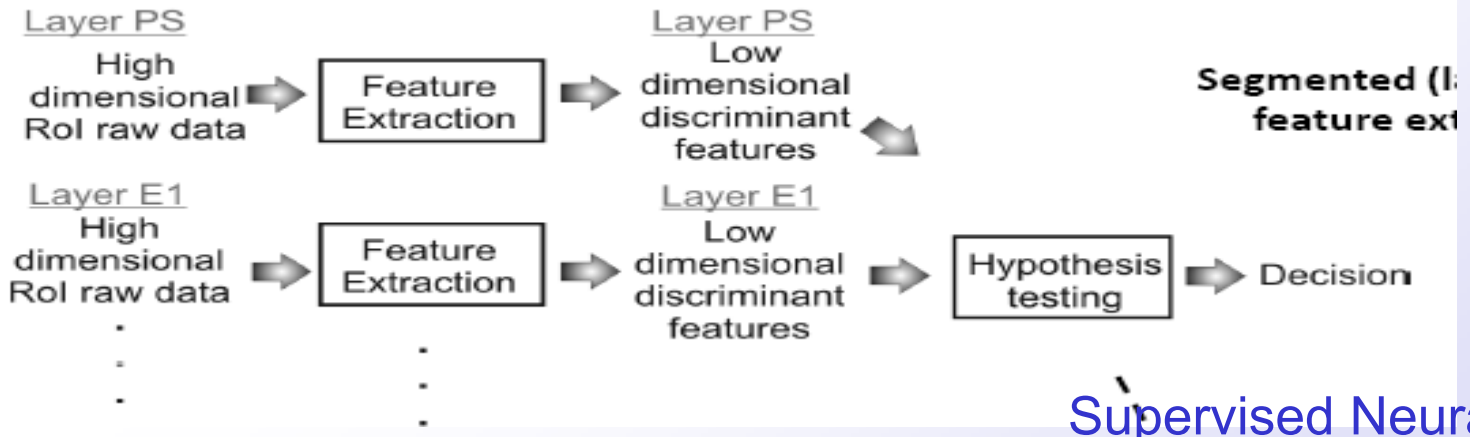


a: select a calorimeter layer;

b: find the hottest cell → Ring 1;

c: select the cells around the first ring → Ring 2;

d: repeat this procedure over the RoI .

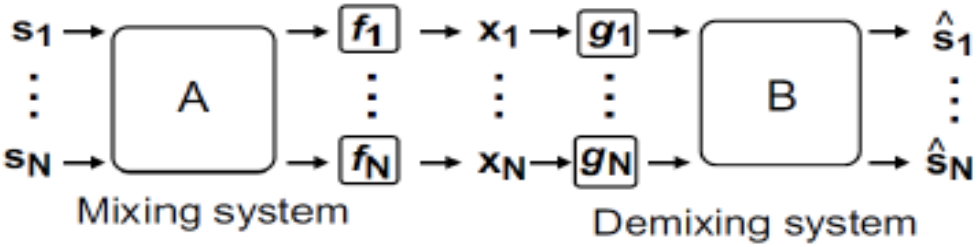


Nonlinear Independent Component Analysis

Supervised Neural Classifier

Post-Nonlinear ICA

- The post-nonlinear (PNL) ICA model is described through:



- For *i=1, ... , N*, the measured signals are defined as:

$$x_i = f_i\left(\sum_{j=1}^{N} a_{ij} s_j\right)$$

- An estimation of the independent signals:

$$\hat{s}_i = \sum_{j=1}^{N} b_{ij} g_j(x_j)$$

- Each nonlinear function g_k is modeled by a two-layer MLP neural network:

$$g_k(x_k) = \sum_{h=1}^{N_H} \sigma_h \tanh(\omega_h x_k + \eta_h)$$

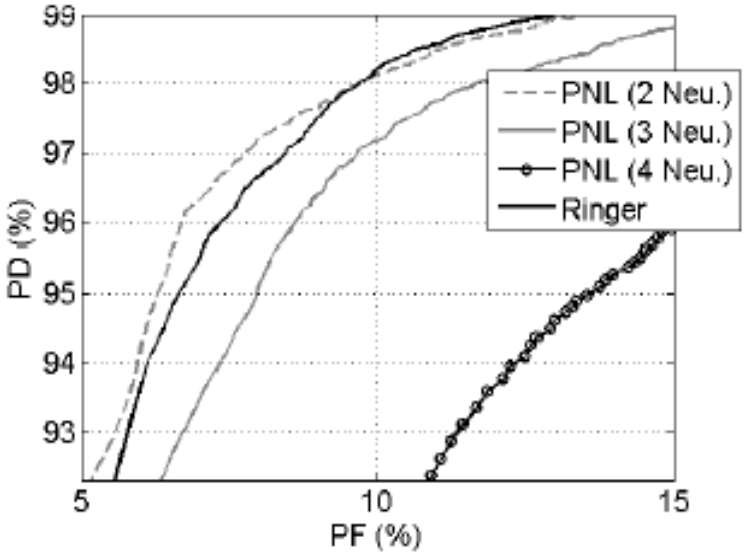
Where N_H is the number of hidden neurons.

Results

- Number of neurons used to estimated each nonlinear function in the PNL model:
  - The nonlinearities are expected to be smooth (the calorimeter is approximately linear) ;
  - The same number of hidden neurons was used to estimate each nonlinearity;
  - By increasing the number of neurons the discrimination efficiency decreases.

Discriminator	Best SP x 100	PF for PD=97%
Ringer	94.35	(8.67 ± 0.20) %
PNL (2 Neur.)	94.70	(7.69 ± 0.35)%
PNL (3 Neur.)	93.70	(9.67 ± 0.38)%
PNL (4 Neur.)	90.83	(17.39 ± 0.40)%

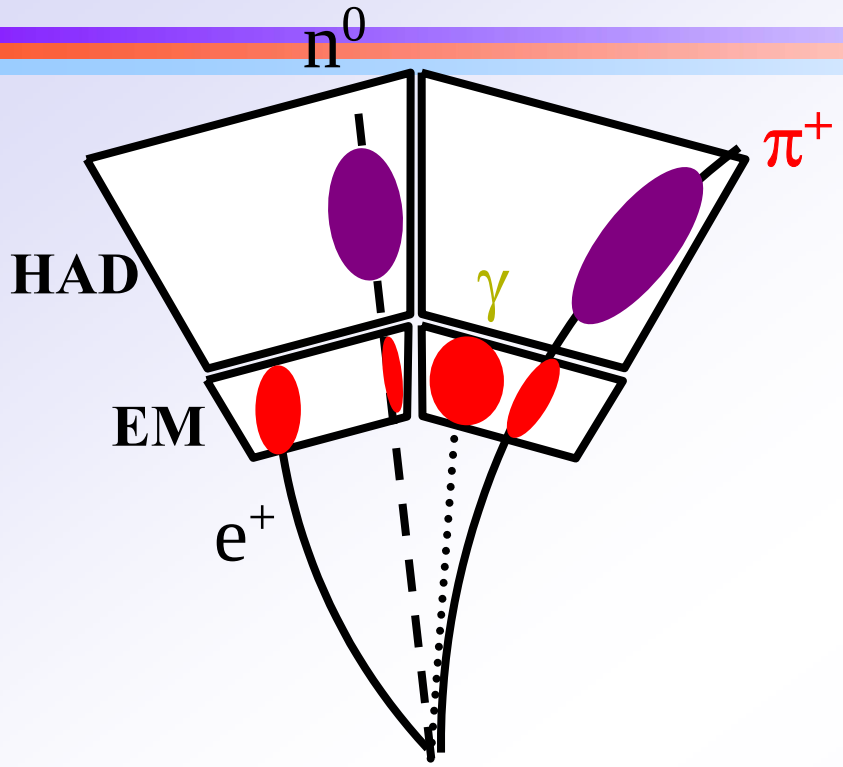
P_D – Probability of Detection (Electron Efficiency)
P_F – Prob. of False Alarm (Jet Acceptance)



Likelihood-based Particle Flow Algorithm at CDF for Accurate Energy Measurement and Identification of Hadronically Decaying Tau Leptons

Andrey Elagin, Alexei Safonov
Texas A&M University
on behalf of the CDF collaboration

Particle Flow Algorithm



High calorimeter segmentation is required

Challenge for Particle-Flow at CDF: large calorimeter segmentation

Calorimeter approach:

$E_{jet} = E^{EM} + E^{HAD}$

$\delta E/E^{EM} \sim 0.2/\sqrt{E}$; $\delta E/E^{HAD} \sim 0.5/\sqrt{E}$

70% in the energy measurement depends on poor resolution of hadron calorimeter

Particle Flow Algorithm:

$E_{jet} = E_{tracks} + E_n + E_\gamma$

$\delta p_T/p_T^2 \sim 0.001$

only 10% utilize hadron calorimeter

$$E_\gamma = E^{EM} - E^{EM}_{tracks}$$

$$E_n = E^{HAD} - E^{HAD}_{tracks}$$

Hadronically decaying taus (important for Higgs analysis) are similar to jets
⇒PFA should improve their identification and energy resolution

Likelihood PFA

Construct a likelihood based on particle type-specific signatures in the detector.

$$L = f(\textit{detector responses} \mid \textit{particles and energies})$$

Use MC to determine response functions - particle's PDF.

EM, HAD Calorimeter

$$\text{p.d.f.} = f_{\pi}^{EM,HAD}(E^{EM}, E^{HAD} \mid E_{\pi})$$

$$\text{p.d.f.} = f_{\gamma}^{EM,HAD}(E^{EM}, E^{HAD} \mid E_{\gamma})$$

$$\text{p.d.f.} = f_{\pi\gamma}^{EM,HAD}(E^{EM}, E^{HAD} \mid E_{\gamma}, E_{\pi})$$

Central Electromagnetic Shower (CES)

$$\text{p.d.f.} = f_{\gamma}^{CES}(E^{CES(1)}, E^{CES(2)} \mid E^{1}_{\gamma}, E^{2}_{\gamma}) = f_{\gamma}^{CES}(E^{CES(1)} \mid E^{1}_{\gamma}) \times f_{\gamma}^{CES}(E^{CES(2)} \mid E^{2}_{\gamma})$$

Likelihood function

$$L = f_{\pi\gamma}^{EM,HAD}(E^{EM}, E^{HAD} \mid E_{\gamma}, E_{\pi}) \times f_{\gamma}^{CES}(E^{CES} \mid E_{\gamma})$$

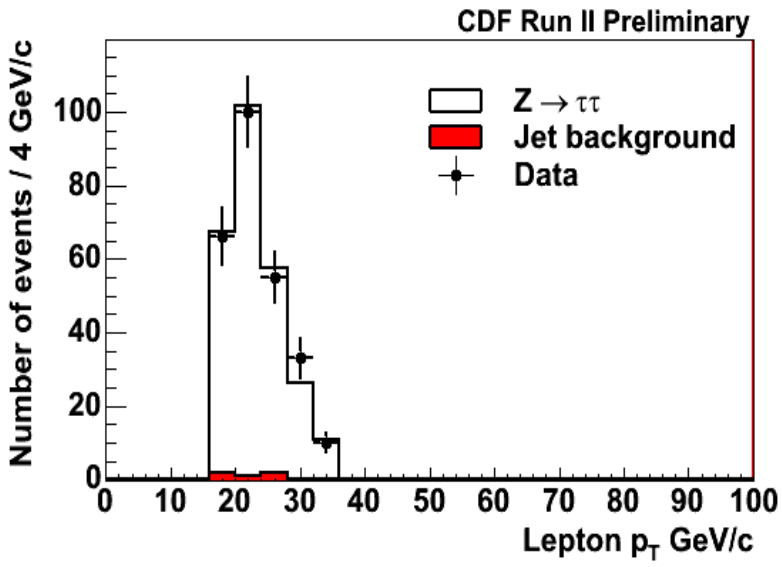
- E_{π} – measured in the Tracker, <u>fixed parameter</u>
- E_{γ} – likelihood maximization
- visible tau energy $E^{\tau}_{rec} = E_{\pi} + E_{\gamma}$
- use width of $L(E_{\gamma})$ to estimate uncertainties

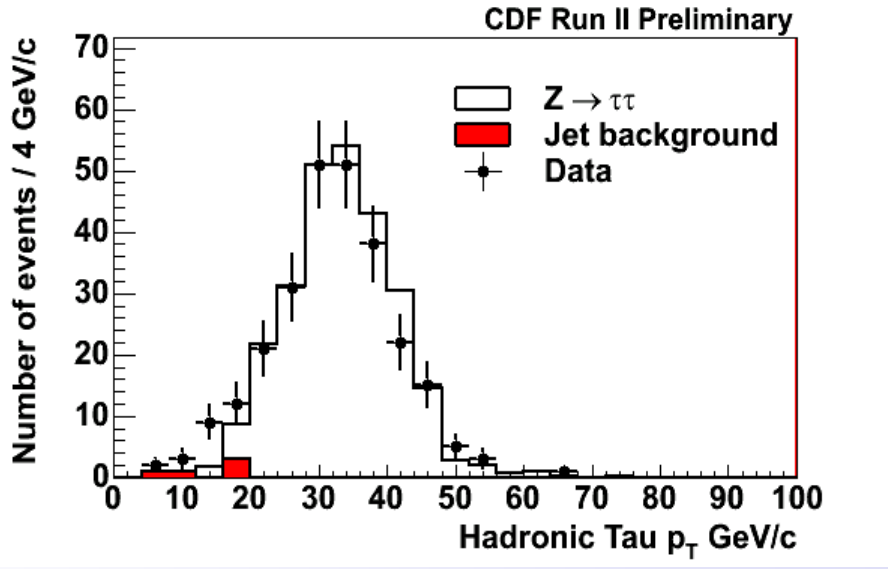
Tests with data

Tau selection in data ($Z\rightarrow\tau_h\tau_e$ and $Z\rightarrow\tau_h\tau_\mu$)

p_T distributions of leptons (e/μ) and hadronic taus agree well with MC.

Events with 1 prong hadronic taus



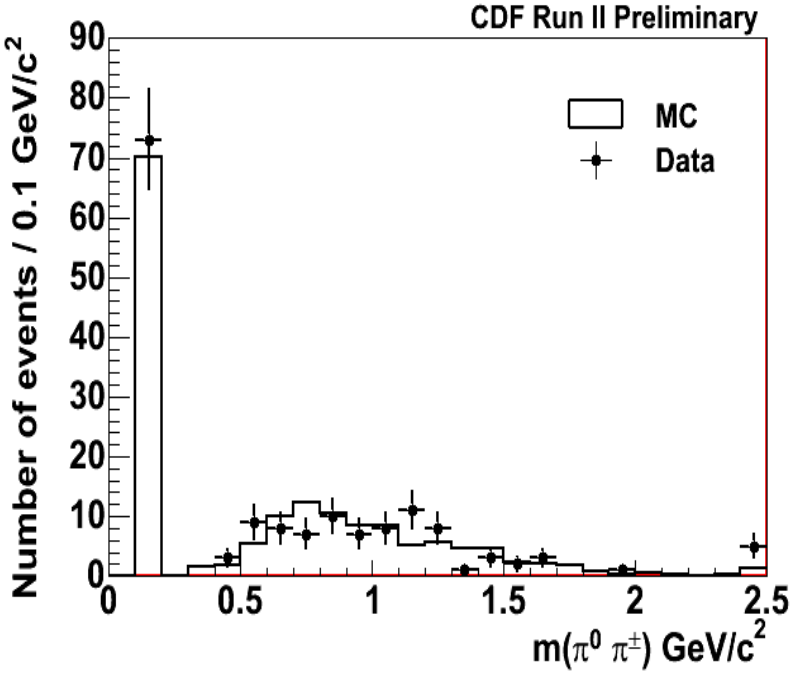


These events with low jet background are used as a reference tau sample to test algorithm performance with data.

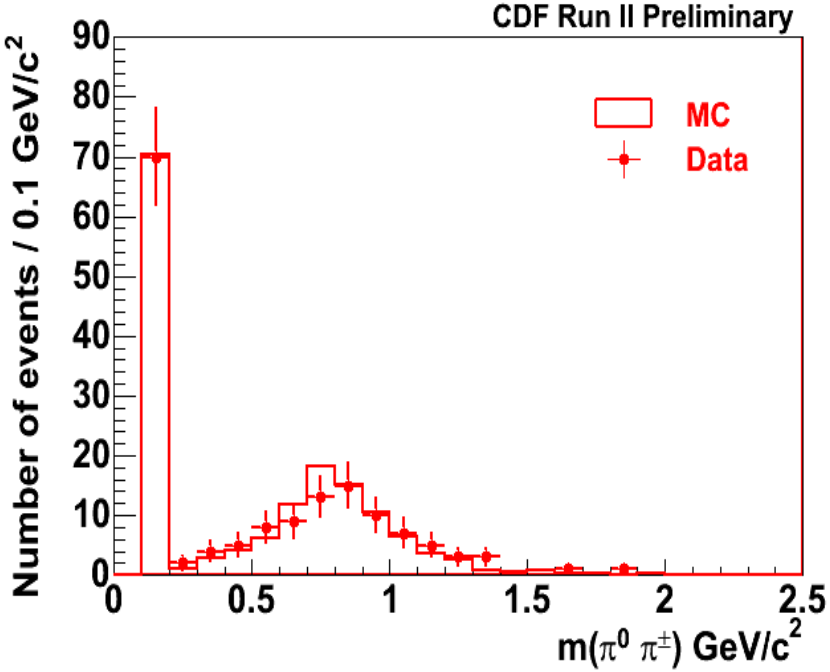
Algorithm performance

Events with 1 prong hadronic taus

Standard CDF algorithm



Likelihood based method



Selection of taus with smaller reconstructed invariant mass using likelihood based algorithm will provide higher efficiency and better suppression of the backgrounds

Fourier Transforms as a Tool for Analysis of Hadron-Hadron collisions

James Monk & Mario Campanelli
University College London

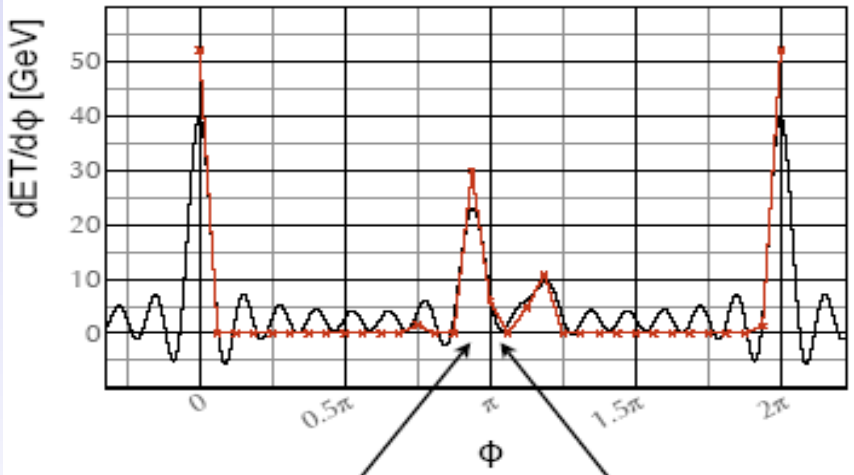
Analysis of gaps between the jets – experimental signature for diffractive pomeron excahnge

Fourier transform – used to separate features of different scale (objects of different sizes) in the event

Fourier decomposition of a single event

Plot curve obtained from summation of Fourier terms (black) over input grid of ET bins (red)

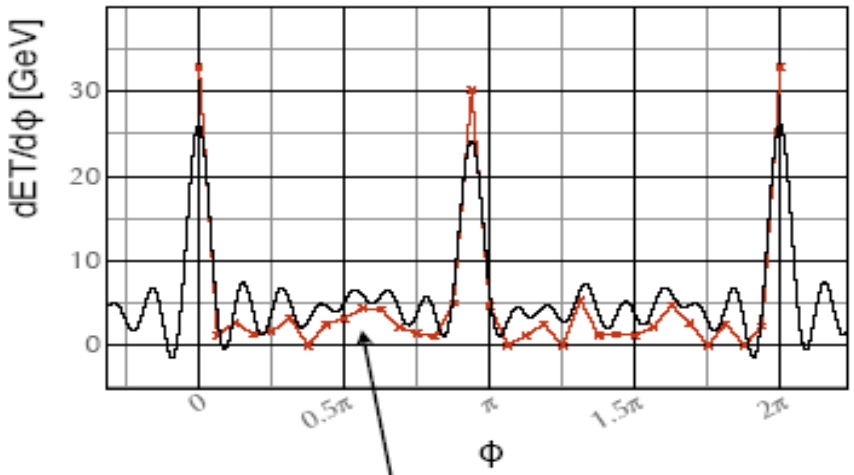
Colour singlet exchange, no underlying event



Note second jet is always in region $\phi < \pi$

Second jet split in two

QCD 2->2 with underlying event



FT output matches radiation between jets

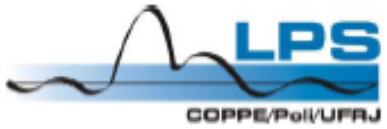
Shows Fourier Transform works for events with little inter-jet radiation or with radiation populating the inter-jet region

- The effect of different features such as underlying event, hadronisation, jets and mini-jets are confined to separate regions of the coefficient space.
- "Gaps" present in diffractive events should appear as a depletion in certain coefficients, which are not necessarily the ones affected by pile-up or underlying event.

AC

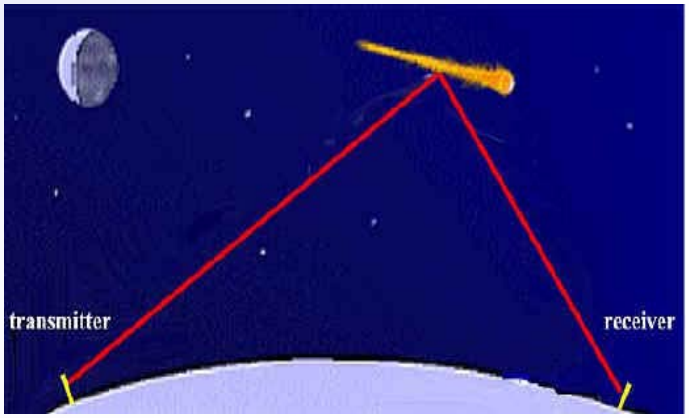
Online Filtering for Radar Detection of Meteors

- LEITE, E. C. V.
- ALVES, G. O.
- SEIXAS, J. M. de
- ALMEIDA JUNIOR, F. M. L.
- TAKAI, H.
- VIANNA, C. S.



The radio meteor scatter technique

- *Very High Frequencies (VHF) radio waves – 30 to 300 MHz – sent by a transmitter, scattered on the meteor's trails and detected by a receiver*
- *other events can be detected, e.g. cosmic rays, atmospheric phenomena (lightening, planes etc)*
- *Motivation of meteor study - determine signal parameters to study the ozone layer at 80 to 100 km above sea level;*



Signal Detection:

- *In frequency-domain: Cumulative power spectrum analysis and narrowband demodulation.*
- *Online filtering: storage requirement reduction (~10 gigabytes per day of raw data) ; efficient classification.*

Data

Acquired at Custer Institute and Observatory, by the Brookhaven National Laboratory team from June 1st to 7th, 2008;

Meteor trails

The meteor trail is a plasma with a characteristic frequency:

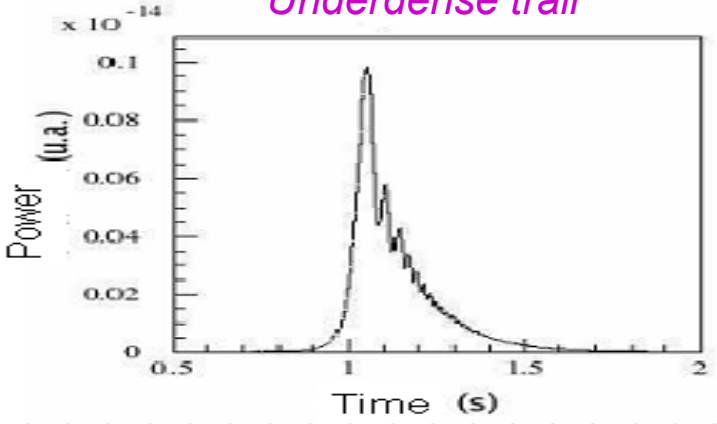
$$\omega_p = \sqrt{\frac{N_e e^2}{\epsilon_0 m_e}}$$

Two types:

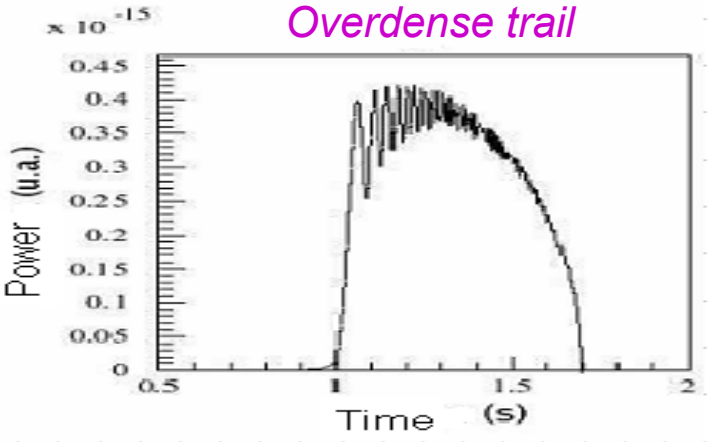
Underdense trails: Low N_e. The scattering is done by individual electrons. Fraction of events: 90%. Duration: tenths of seconds.

Overdense trails: High N_e . Fully reflect the incident wave and the trail is treated as a cylinder reflector. Fraction of events 10%. Duration: few seconds.

Underdense trail

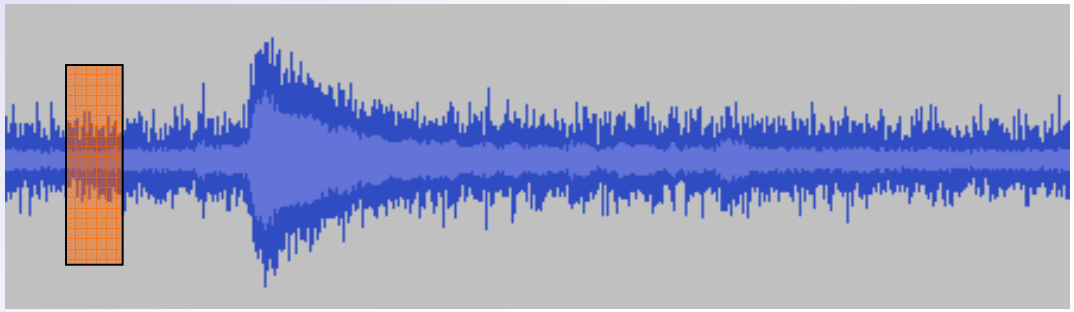


Overdense trail

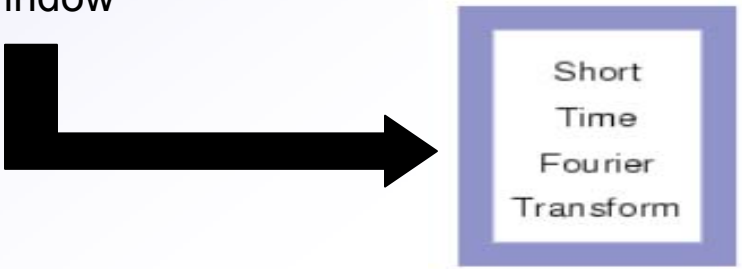


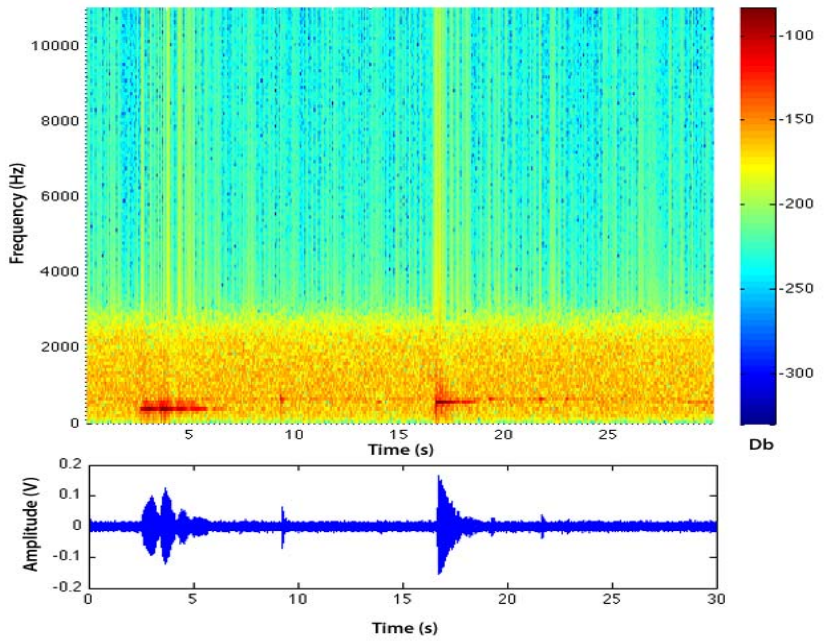
Power Spectrum Analysis (PSD)

Short-time Fourier Transform on time windows (30s): spectrogram



Window



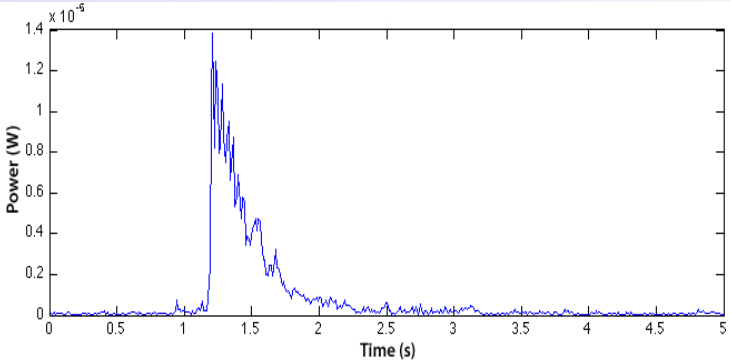


The PSD determines the frequency bin in which information is maximun;

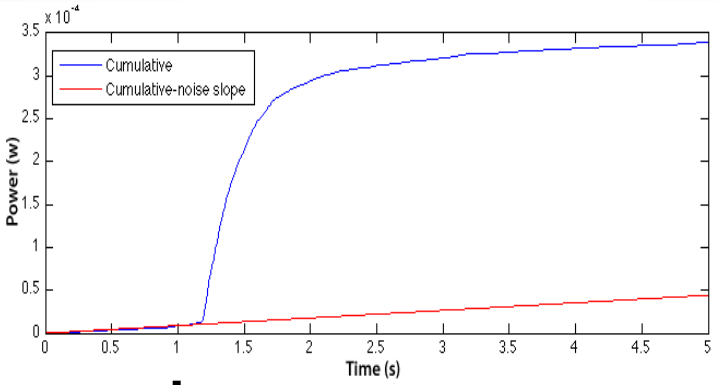
The power peak value for each data window is stored and analysed

PSD and Online Decetion

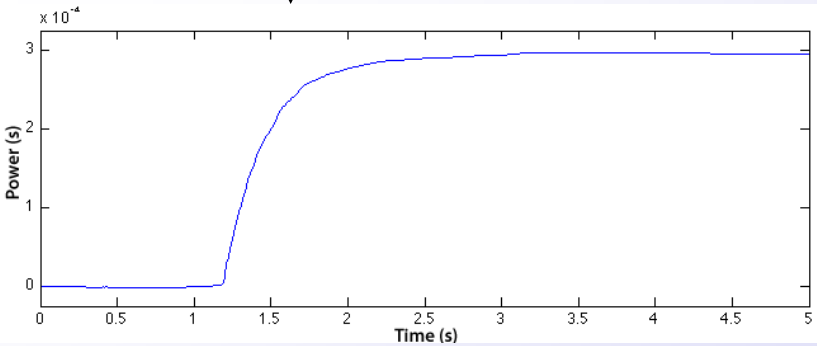
- *Power peak received over time for an underdense trail:*



Accumulating process



Bkg. substraction



- *Fluctuations in the slope:*

  Threshold values for defining the start and end samples of the triggered event.

Online Signal Detection:

- *Requires a buffer to store a block of data (30 seconds);*
- *Processes the current data block;*
- *Time required for processing each block of data: ~ 0.7 seconds;*

Results and Perspective

Detection Results:

- *50 blocks of data, which contain 261 meteor events (visual inspection)*
- *246 events correctly detected (94.2% efficiency);*
- *False Alarm Rate: ~1 fake event per 100 seconds of data;*
- *Avoids 220 fake events to be recorded (144 MB less per hour);*
- *The Filter avoided to record 3.45 GB of noise each 24h.*

Perspectives:

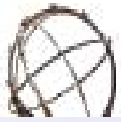
- *Accumulate received power at full width at half maximum, instead of using peak values;*
- *Data processing under white noise conditions;*
- *Optimal stochastic detection: Matched Filter;*
- *Extract signal envelope: Narrowband Demodulation*;

Software for analysis

FATRAS —

A Novel Fast Track Simulation Engine for the ATLAS Experiment

Sebastian Fleischmann
on behalf of the ATLAS Collaboration



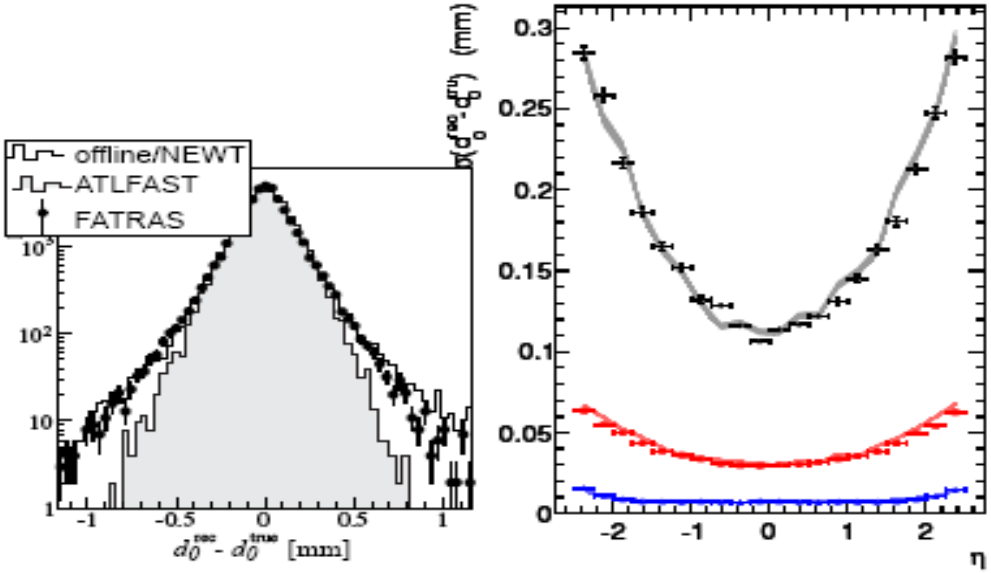
Features

- ATLAS is a new track simulation concept between full Geant4 simulation and conventional fast detector simulations
  - The full reconstruction chain can be run on ATLAS output
  - Speed improvement mostly due to simplified Tracking Geometry and extrapolation
  - (Nearly) no parametrisations needed
  - All important physics effects included, like multiple scattering, brem, conversions, particle decays, hadronic interactions
  - Allows studies to be performed that cannot easily be done either with full simulation or conventional fast simulations

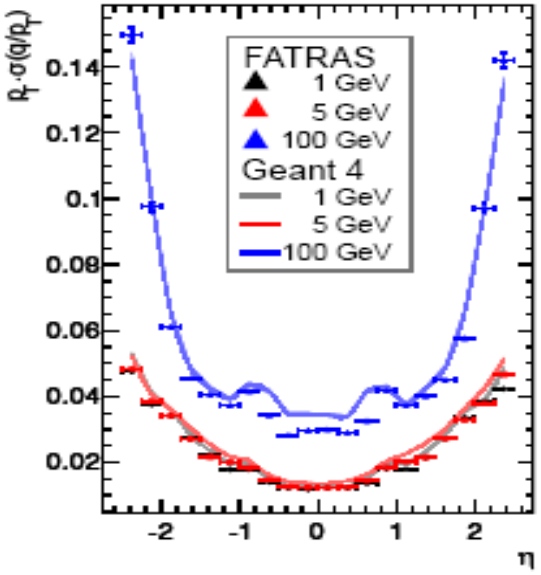
Current tests- example

Track parameter resolutions

- Single muon events with p_T =1 GeV, 5 GeV, 100 GeV
- In general good agreement, but still some parameters to tune in the digitisation
- In particular tails better described than in ultra-fast sim



Transverse impact parameter d_0



Inverse track momentum q/p_T

Visual Physics Analysis (VISPA)

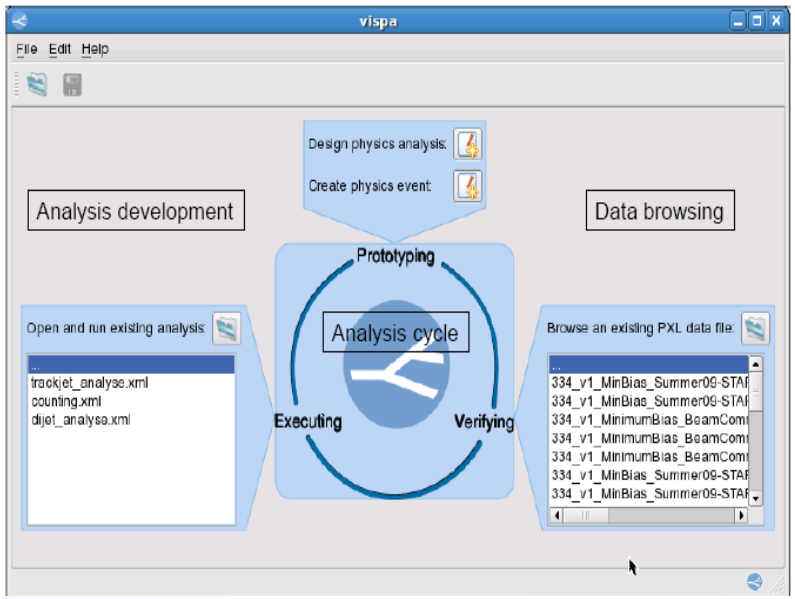
A graphical development environment for physics data analysis

M.Brodski, M.Erdmann, R.Fischer, **Andreas Hinzmann**,
T.Klimkovich, D.Klingebiel, M. Komm, G.Müller,
T.Münzer, J.Steggemann, T.Winchen

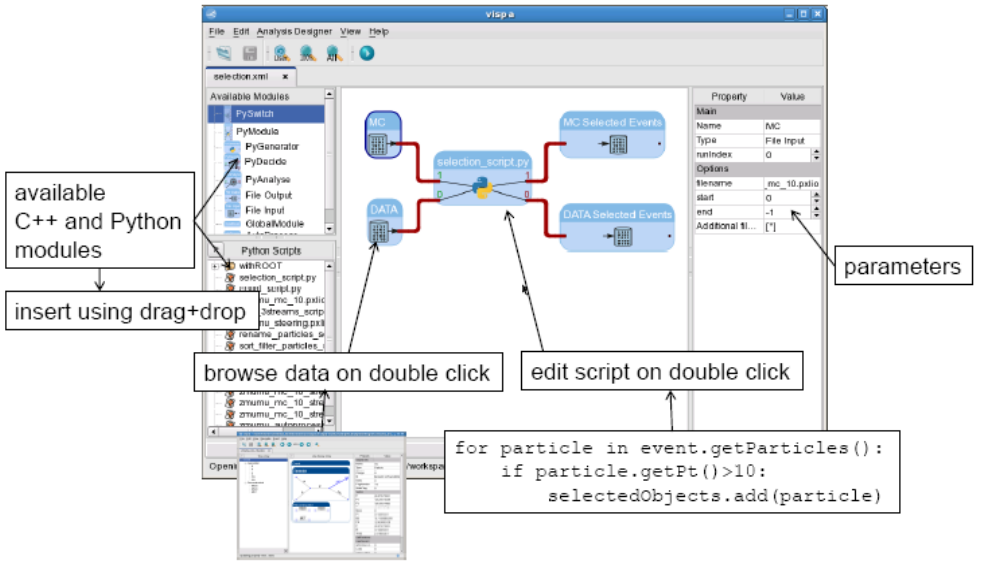
ACAT 2010, 22 February 2010

Current status

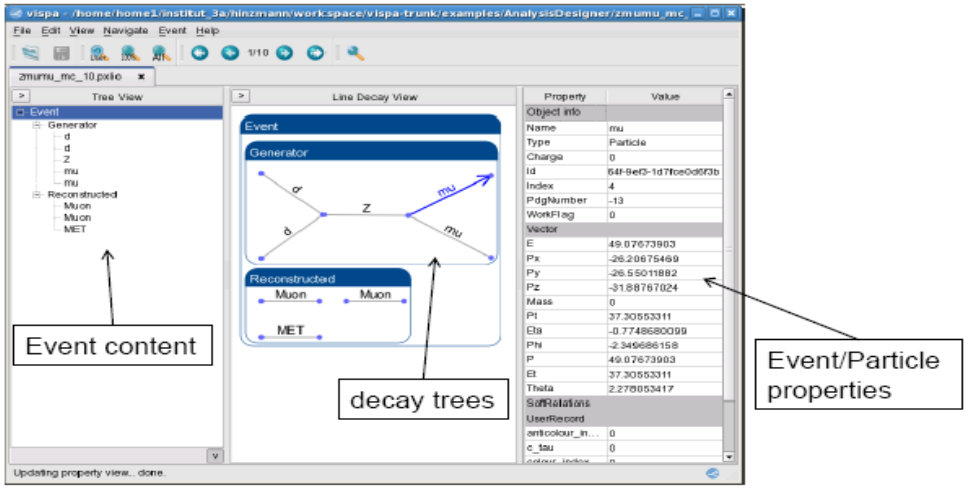
IDE for Physics analysis



Visual development of analyses



Visual representation of data

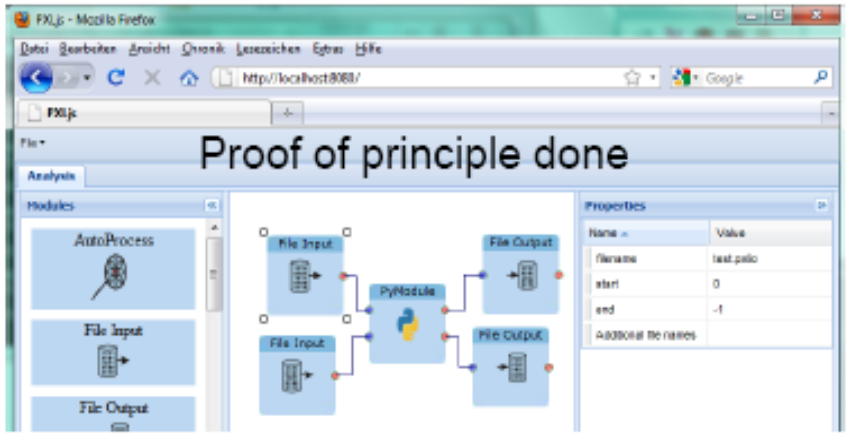


- Exchange of modules:
  - Well defined module interface allows reuse
  - Share common modules within group
  - Outlook: Central database of modules on th
- Exchange of analyses:
  - VISPA allows automatic tar-ball creation
- Exchange between platforms: Linux, Windows

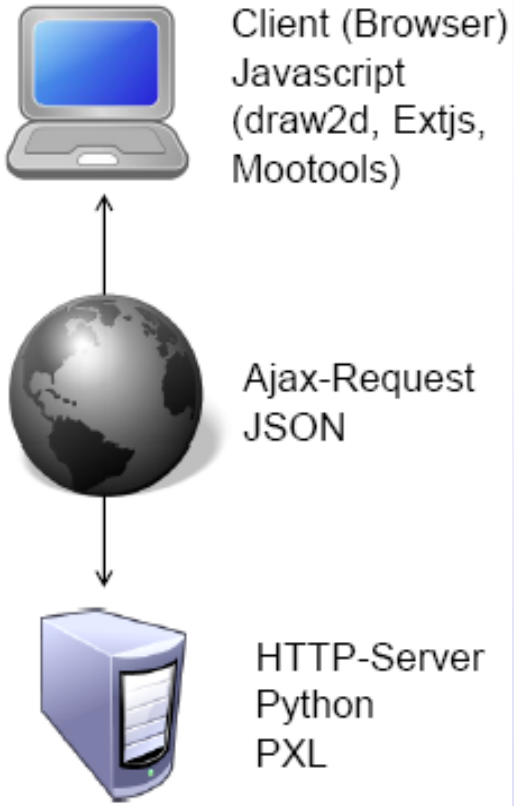
Future plans

Outlook: VISPA@WEB

- VISPA analyses using web browser
  - No installation needed
  - Modules and data centrally maintained: on the web or within institute
  - Analysis performed on server
  - Good solution for teaching



- Development of security concept for user data and modules



WatchMan Project

Computer Aided Software Engineering applied to HEP Analysis Code Building for LHC

Riccardo-Maria BIANCHI, Renaud BRUNELIERE (Freiburg University)

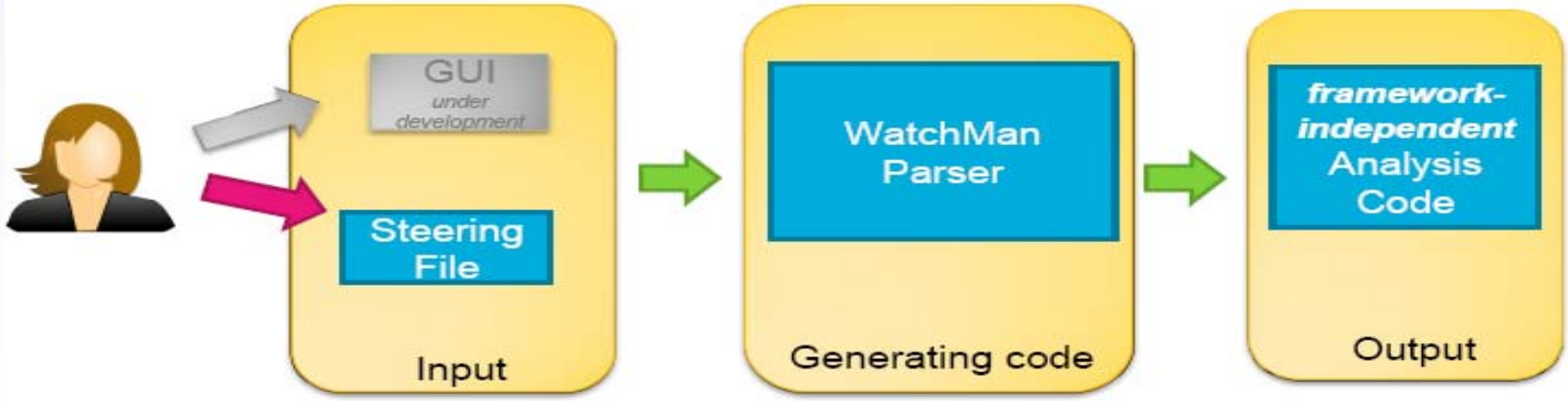
ACAT 2010, 22 February 2010

WatchMan: Analysis Code Generator

- WatchMan is an Analysis Code Generator:
  ...it automatically builds Analysis Code, from user settings.

Automated Building of the Analysis Code

WatchMan is a Code Generator, a **Software Factory**, aimed to build AnalysisCode easily



The user defines her/his analyses (as many as wanted) in an easy way, and the actual analysis code is **dynamically generated**

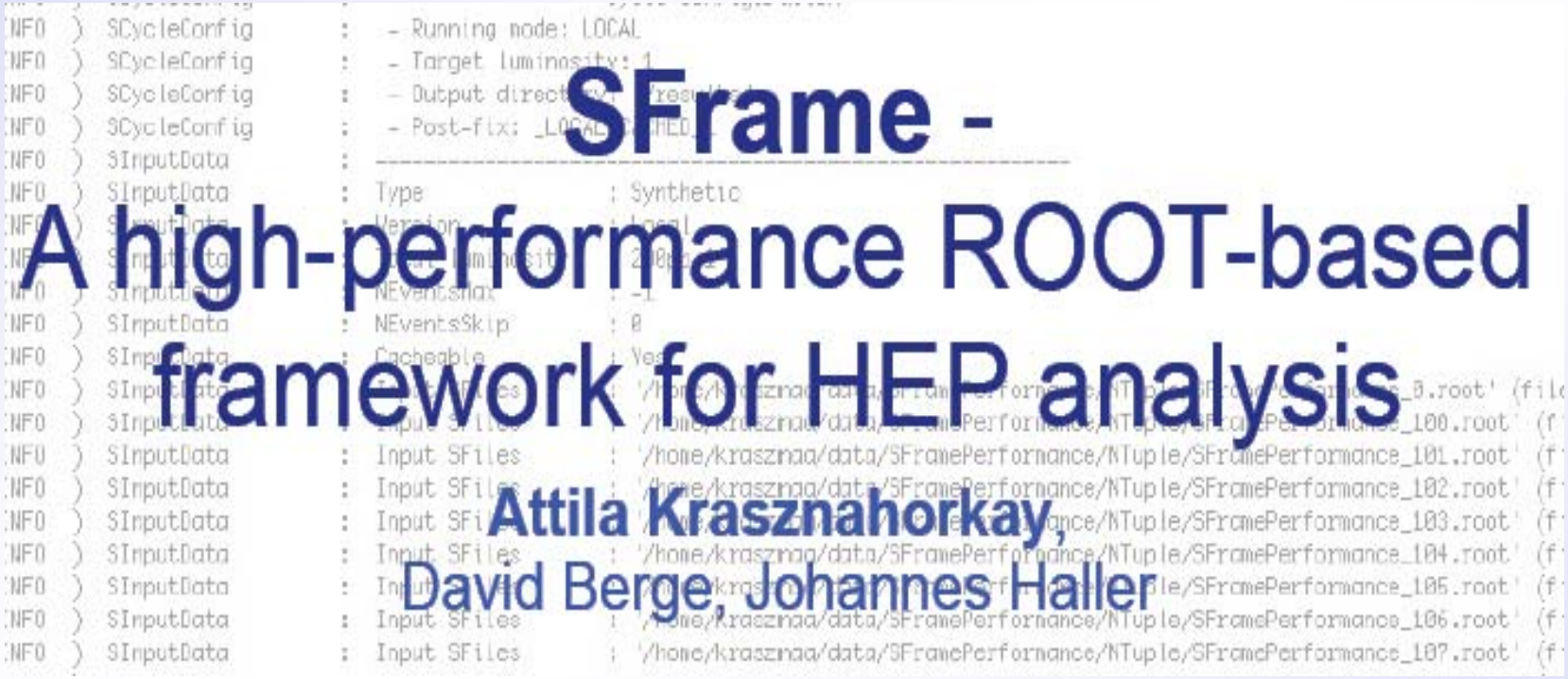
Features

- WatchMan only depends on ROOT and Python. And it's a stand-alone self-contained package.
  Download and try it!

- The user can:
  - define as many physics analyses / channels as wanted
  - Define many object selection definitions
  - add UserData to output Ntuple
  - plug in user-defined custom code

- Combining all this in one steering file in order to get a Ntuple with all that info and objects.

- **Trigger**:
  the user can ask for a list of triggers to be checked, to flag or skim events according to them
- **Skimming**:
  events can be skimmed or not according to the event selection (skimmed if it does not pass any channel)
- **Overlap Removal flags:**
  particles can be removed or only flagged if overlapping
- **Modularity:**
  Object Selection, Overlap Removal and Event Selection can be switched on/off independently, for each channel or for group of channels.
- **Cut-Flow Table:**
  For every analysis/channel the info on event selection cuts passed by each event is stored. Easy to extract then a cut-flow table



SFrame - A high-performance ROOT-based framework for HEP analysis

Attila Krasznahorkay,
David Berge, Johannes Haller

SFrame features

SFrame – analysis framework, initially done for ATLAS analysis outside the ATLAS offline software, now generally available on SourceForge

Features

- *Makes code sharing within a group very simple -> Can share "analysis cycles" for common tasks (dataset cleaning, etc.)*
- *Arranges input files into InputData blocks (blocks that should be handled homogeneously)*
- *Keeps track of the integrated luminosity of the InputData blocks*
- *Provides a way of scaling the Monte Carlo to the data integrated luminosity*
- *Easy-to-use functions for reading/writing TTrees*
- *Simple interface for writing various TObjects into the output file.*
- *etc.*

Processing speed

From 5 consecutive runnings.

Input location		Local	XRootD
Athena		1.77±0.02 kHz	N/A
ACLiC		3.85±0.03 kHz	3.77±0.04 kHz
CINT		259.0±2.2 Hz	N/A
PyROOT		127.2±2.2 Hz	123.1±1.6 Hz
SFrame	LOCAL	4.04±0.02 kHz	4.02±0.03 kHz
	PROOF-Lite	15.92±0.15 kHz	15.81±0.13 kHz
	PROOF	N/A	29.53±0.17 kHz

MC4QCD

web based analysis and workflow tool for Lattice QCD

Massimo Di Pierro

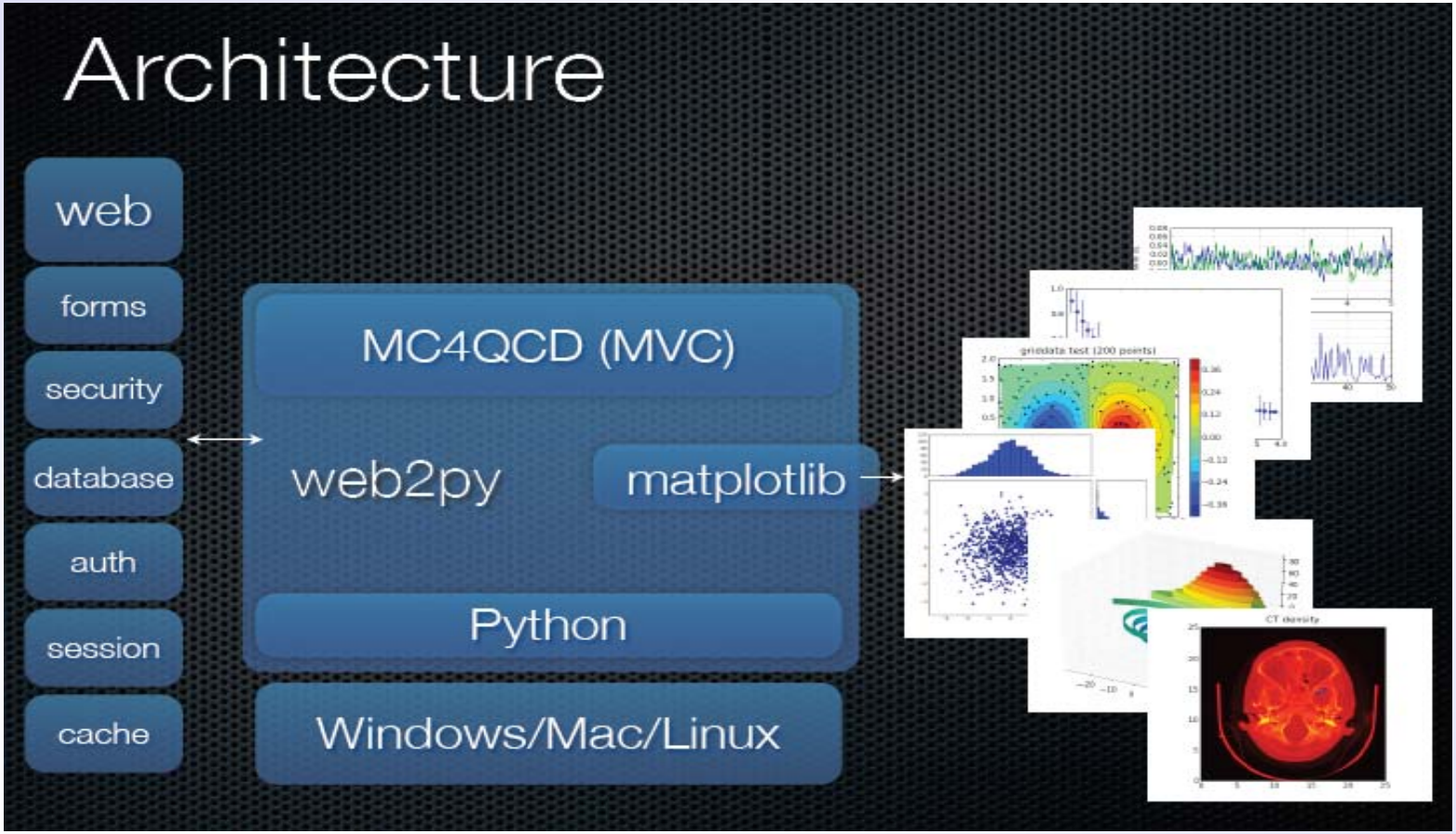
School of Computing

DePaul University

Chicago

Advantages

- Very general and easy to use
- Users can share they data and they results (optional)
- Users easily find their data
- Users repeat old analysis on new data
- Users can post comments on each other data/results including latex expressions.
- Improves collaboration and reduces development time
- All functions can be scripted for batch processing



web2py – framework for building web applications

- can be used for other applications than this specialised MC4QCD

Paralellisation

Parallelization of the SIMD Kalman Filter for Track Fitting

Rama Malladi

R. Gabriel Esteves
Ashok Thirumurthi
Xin Zhou
Michael D. McCool
Anwar Ghuloum

ACAT 2010, 22 February 2010

Track fitting - problem, optimisation

Problem

- *Given a track (sequence of observations z_k) → determine the state of the particle at each observation point.*
- *State is usually given by tuple of values:* $r_k = (x, y, t_x, t_y, q/p)$

Kalman Filter - optimisation

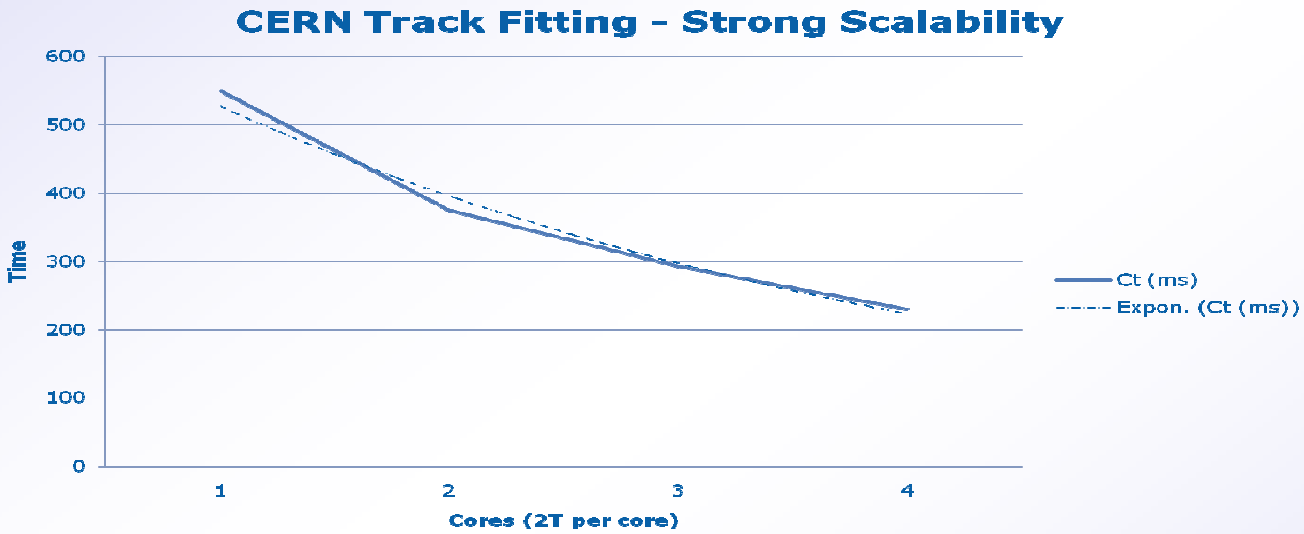
- *Implementation*
  - *Take advantage of modern hardware features: multicore + vectors*
  - *Obvious parallelization opportunity: over tracks*

Ct: C for Throughput

- *Dynamic run-time environment*
  - *Managed data/code*
  - *Dynamic code generation*
  - *Autovectorization*
- *Domain-specific language embedded in C++*

Performance, Future work

Scalability with hyperthreading



Future work

- ***Parallelism - use of more powerful algorithms:***
- ***Explore other parallelisation strategies -*** *Pipeline strategy if applicable*
- ***Integration with C++ frameworks:***
  - *Ct allows construction of reusable frameworks*
  - *Ct allows development of application-specific languages*
  - *Explore parallelization of entire frameworks like genfit instead of experiment-specific kernels*

Fast parallel tracking algorithm for the muon detector of the CBM experiment at FAIR

Andrey Lebedev[1,2] Claudia Höhne[1] Ivan Kisel[1] Gennady Ososkov[2]

Fast Parallel Ring Recognition Algorithm in the RICH Detector of the CBM Experiment at FAIR

Semen Lebedev
GSI, Darmstadt, Germany and LIT JINR, Dubna, Russia
Claudia Höhne
GSI, Darmstadt, Germany
Ivan Kisel
GSI, Darmstadt, Germany
Gennady Ososkov
LIT JINR, Dubna, Russia

Parallel Approach to Online Event Reconstruction in the CBM Experiment

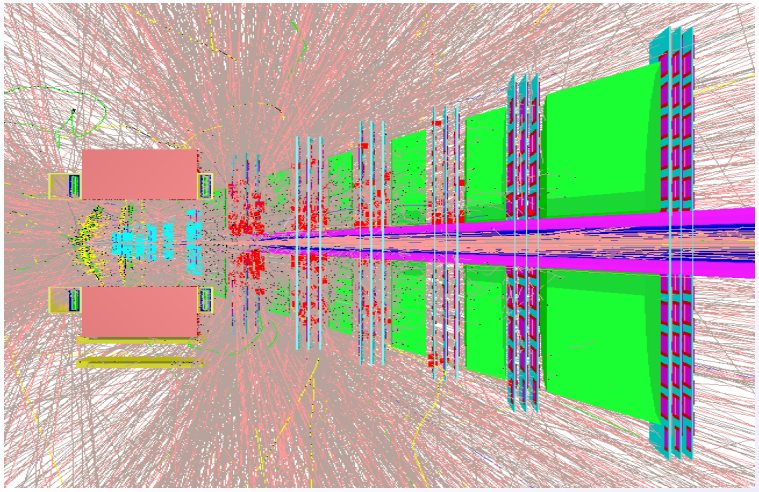
Ivan Kisel
GSI, Darmstadt, Germany
(for CBM Collaboration)

ACAT 2010, 22 February 2010

Experiment and tracking

CBM experiment

- *fixed target experiment*
- study *pp, pA and AA collisions at 8-45 AGeV beam energy*
- *track multiplicities of 800 charge part./reaction*
- *reaction rate up to 100 MHz*



Fast tracking algorithms required

Parallel methods for tracking investigated

- *SIMD – Single Instruction Multiple Data*
- *Multithreading*

Algorithm optimisation – needed for SIMD

- *Minimize access to global memory*
- *Simplification of the detector geometry*
- *Computational optimization of the Kalman Filter*

SIMDization of the track fitter

- *tracks packed into one vector and fitted in parallel*

→ *For muon detector*

Performance of the parallel tracking

Simulation:

- 1000 UrQMD events at 25 AGeV Au-Au collisions + 5 μ + and 5 μ - embedded in each event

	Initial version	Parallel version
Efficiency [%]	94.7	94.0

Speedup of the track finder

	Time [ms/event]	Speedup
Initial	730	-
Optimization	7.2	101
SIMDization	4.8	1.5
Multithreading	1.5	3.3
Final	1.5	487

Computer with 2xCPUs Intel Core i7 (8 cores in total) at 2.67 GHz

Online event reconstruction

Track reconstruction in STS/MVD
and displaced vertex search
are required in the first trigger level.

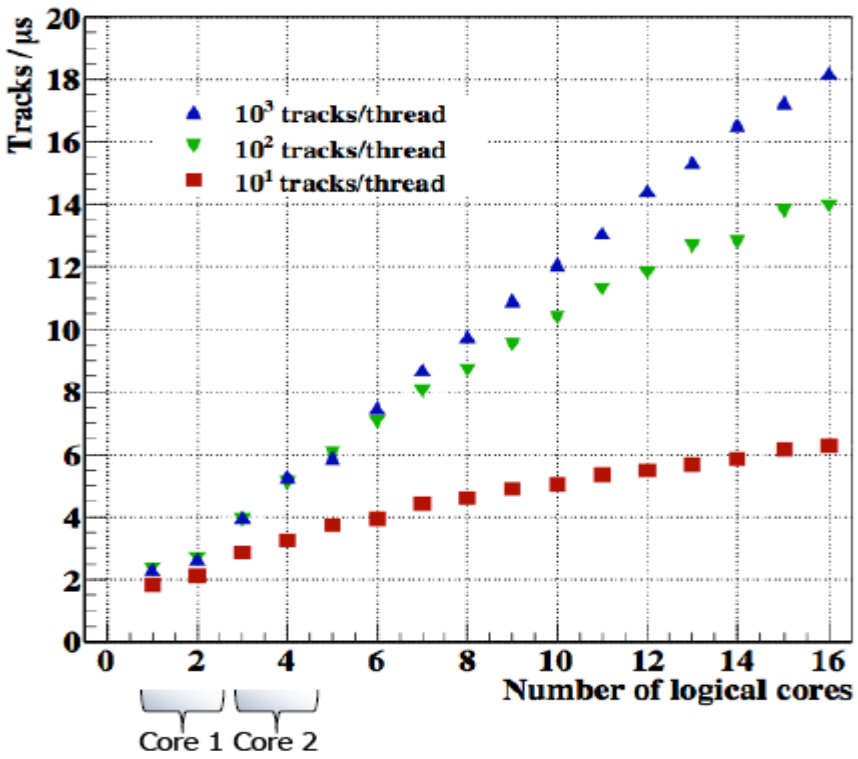
Reconstruction packages:

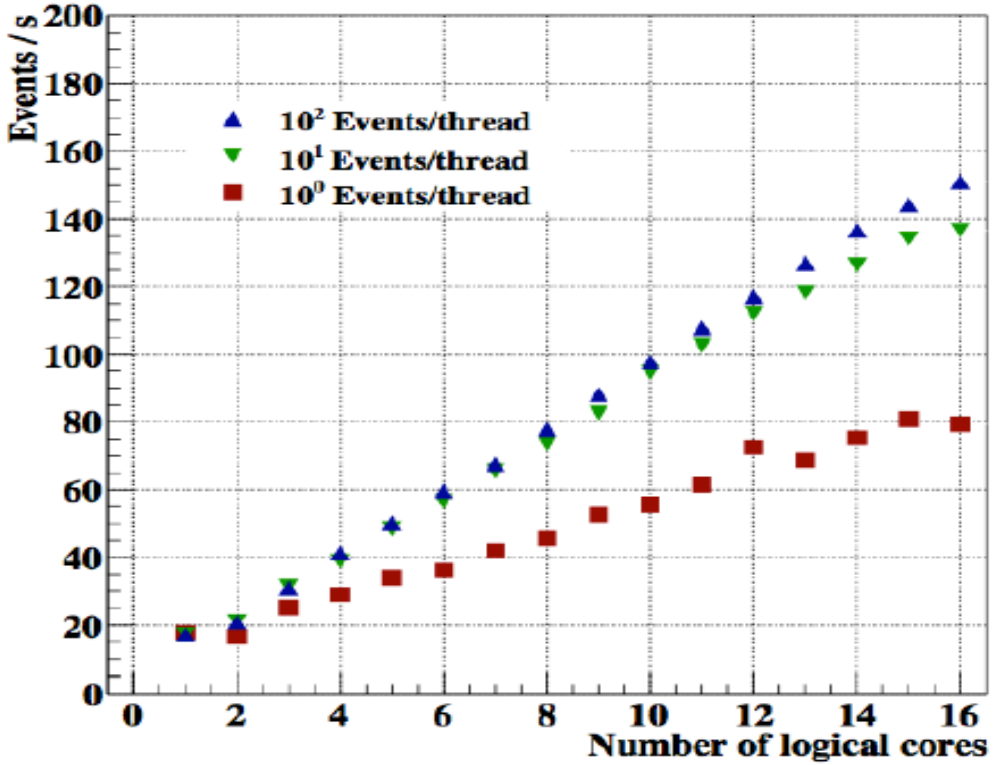
track finding	Cellular Automaton (CA)
track fitting	Kalman Filter (KF)
vertexing	KF Particle

Scalability of the KF Track Fit

Scalability of the CA Track Finder

2xNehalem = 8 Cores





RICH - Ring Recognition

Ring candidates found with

Hough Transform: large combinatorics => slow
Localized Hough Transform: much less combinatorics => fast

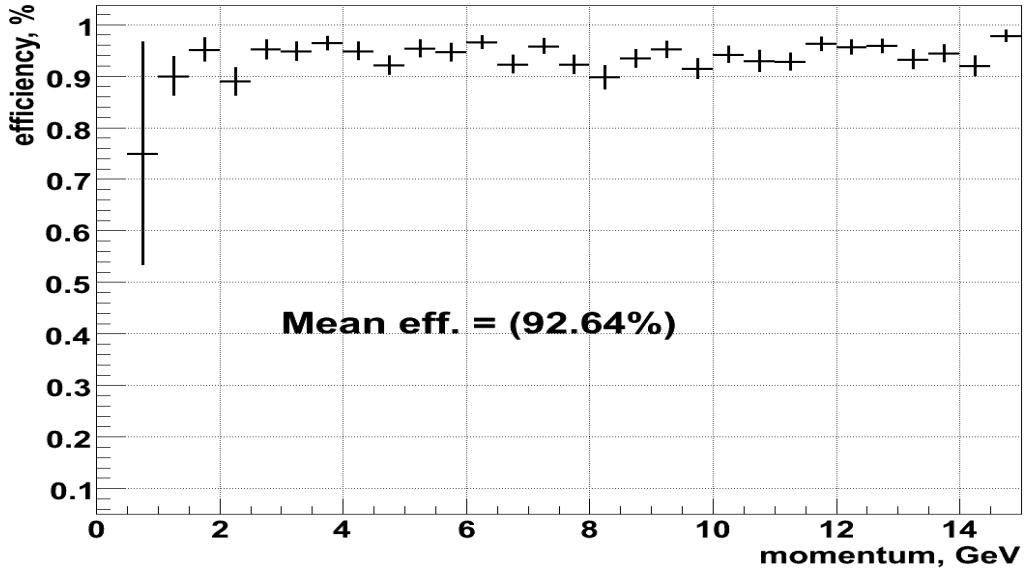
Optimization of the Hough Transform combinatorics

- *Divide hits into several groups*
- *Run Hough Transform of each group independently*

Parallel methods for tracking investigated

- *SIMD – Single Instruction Multiple Data*
- *Multithreading*

Simulation: central Au+Au collisions at 25 AGeV beam energy (



	Time, ms	Speedup
Initial	357	-
Optimization	5.8	62
Parallelization	3	2
Final	-	119

Future plans

Parallelization is now a Standard in the CBM Reconstruction

CBM Tracking Workshop (15-17 June 2009, GSI):

- Tracking Presentations
- Tracking Discussion
- "Future Intel CPU Architecture"
- "OpenCL Parallel Language"
- Training Day on SIMD/MT

Next Tracking Workshop in May 2010

Algorithm	Vector SIMD	Multi-Threading	NVIDIA CUDA	OpenCL	Speedup	Speed/PC
STS	+	+	+	+	10000	6.5 ms
MuCh	+	+			500*	1.5 ms
TRD	+	+			500*	1.5 ms
RICH	+	+			100**	3.0 ms
Vertexing	+	Future			1.5***	20µs
Open Charm Analysis	+				1.5***	20µs
User Reco/Digi		Future				
User Analysis						

+March 2009
+October 2009

*Single hit access should be avoided
**Reformulation of the algorithm is probably necessary
***Avoid accessing the main memory –> approximation of the magnetic field map

Highlights

- *parllelization of reconstruction algorithms – obvious trend*
- *new algorithms for data analysis tried*
- *QiEA – new algorithm we learned – to be tried*
- *automatic code generation for Data Analysis ?*