

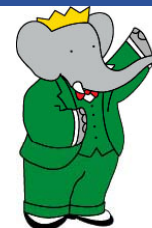
The logo for EGEE (Enabling Grids for E-science) features the letters 'e', 'G', 'e', and 'e' in a stylized font. The 'G' is yellow, while the other letters are blue.

Enabling Grids for E-science



**GridPP**

UK Computing for Particle Physics



**MANCHESTER**  
1824

The University of Manchester

# EasyGrid: a job submission system for distributed analysis using grid

*James Cunha Werner*

[jamwer2000@hotmail.com](mailto:jamwer2000@hotmail.com)

<http://www.geocities.com/jamwer2002/>

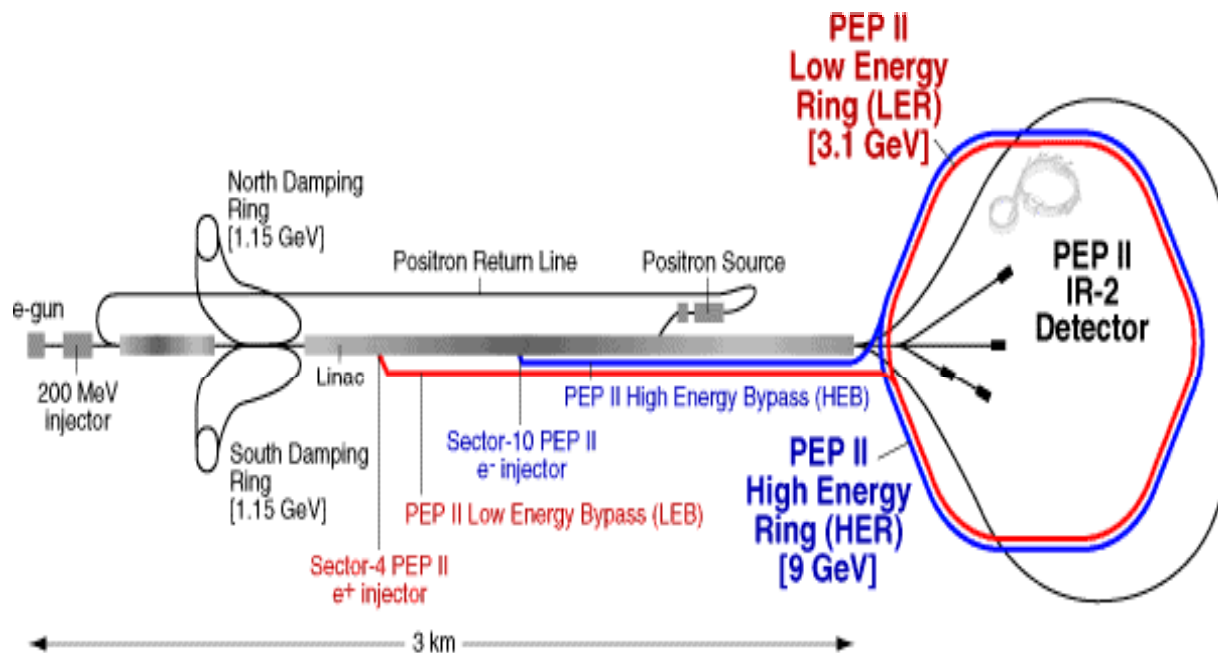
[www.eu-egee.org](http://www.eu-egee.org)



Information Society  
and Media



- BaBar is a high-energy physics experiment running since 1999 at Stanford University/SLAC to throw light on how the matter-antimatter symmetric Big Bang can have given rise to today's matter-dominated universe.
- BaBar analysis was a conventional centralized software (850 packages).
- The project goal was to study grid performance and develop gridification algorithms – 5 papers published and 20 international talks.



## **TauUser data:**

- 18,000 files → each user has thousands of different results
- 500,000,000 events raw data
- 800,000,000 simulated Monte Carlo events

## **Raw data:**

- 1,000,000 files / 20,000 categories
- 4,000,000,000 events raw data
- 4,000,000,000 simulated Monte Carlo events

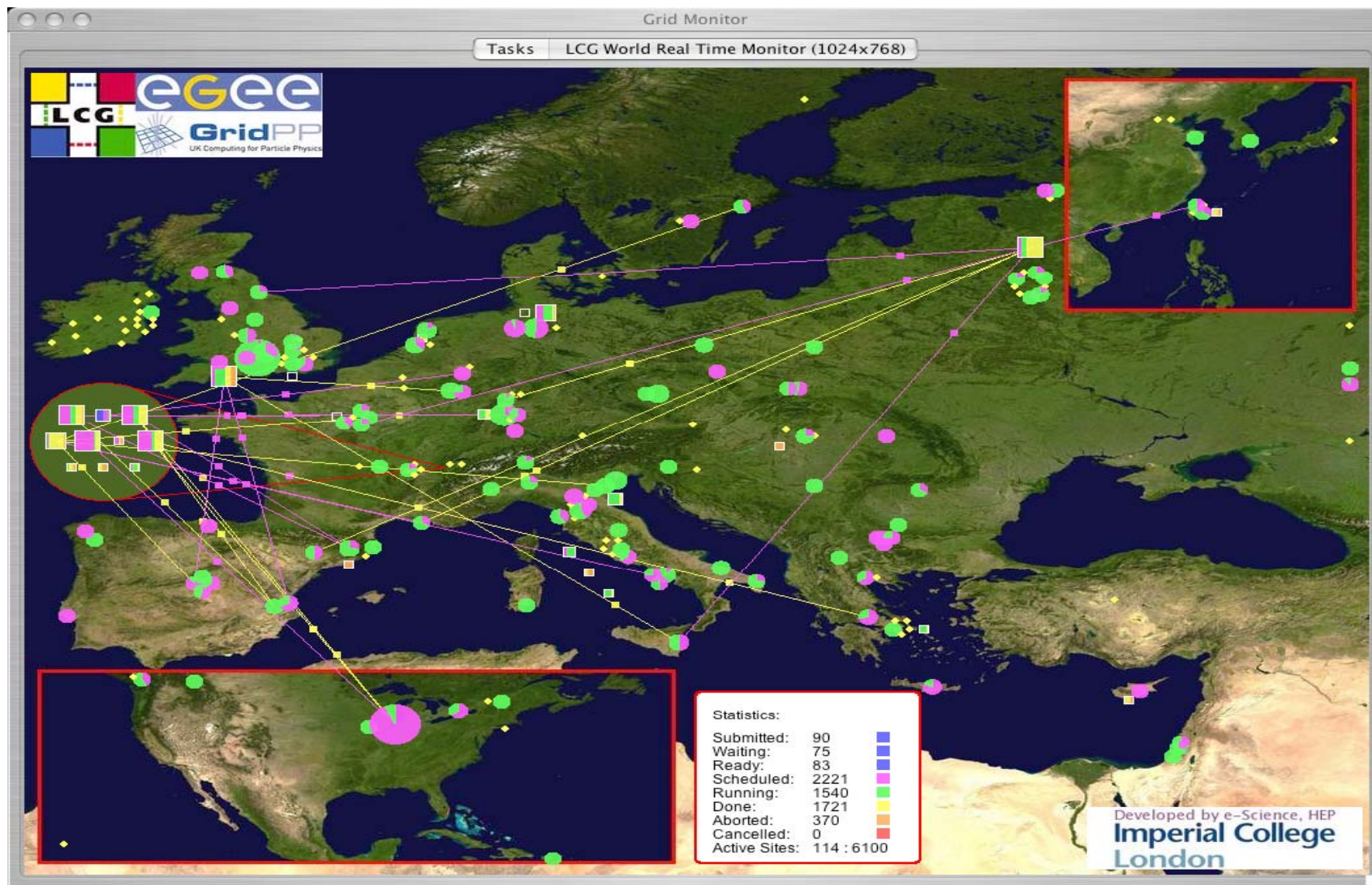
**Massive computational resources are required.**

**Grid computing is a strong candidate to provide them!**

- Complex data management: Distributed datasets around the world and several other support databases (conditions, configuration, bookkeeping metadata, and parameters).
- Distributed and heterogeneous hardware platform around the world (standards).
- Users do not have grid skills. Their interests were high energy physics, not grid.
- Reliability/performance should be at least the same as SLAC. Users have a fixed time to do their research, they will use the more efficient resource.

- **Grid middleware developed by CERN / Switzerland and GridPP/UK.**
- **Homogeneous common ground in a heterogeneous platform.**
  - User interface
  - Information system
  - Resource broker
  - Computer elements
  - Worker node
  - Storage Element

**Integration can be difficult for outsider users!**

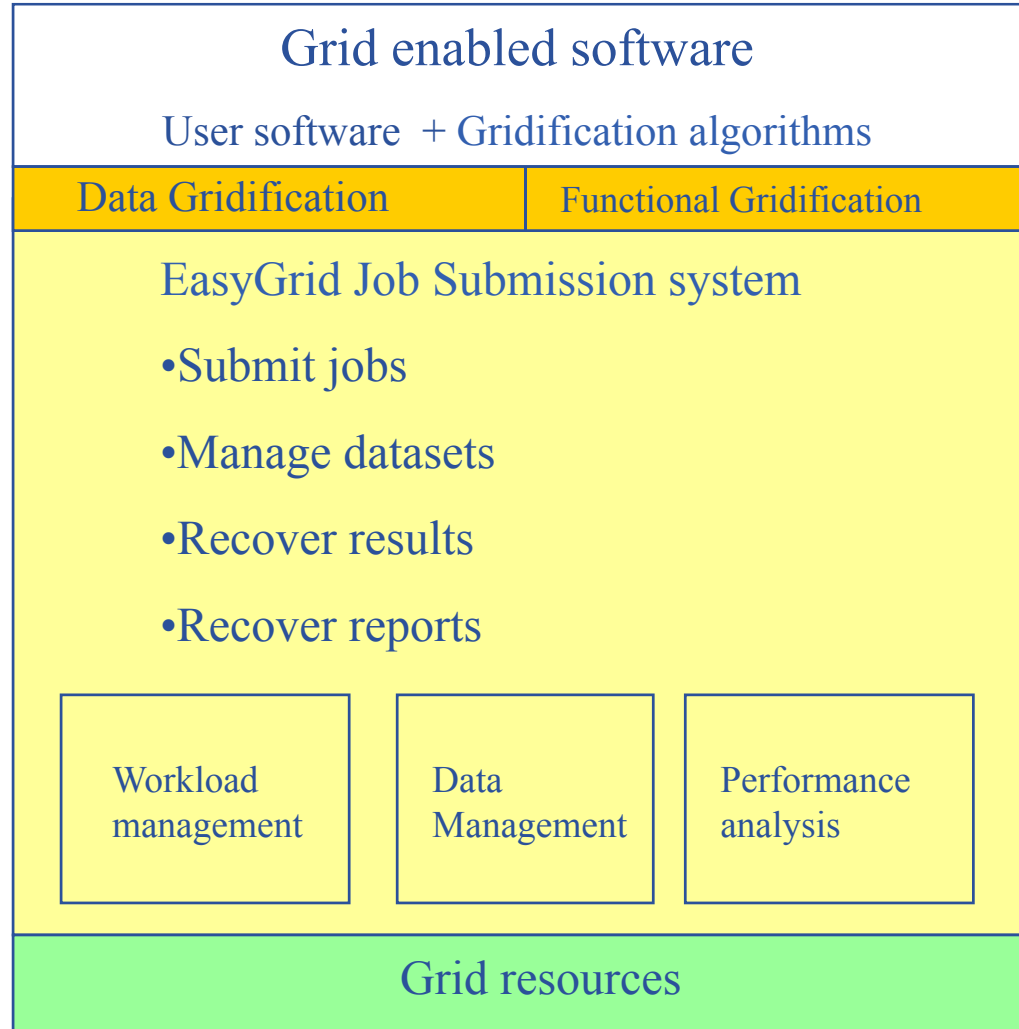
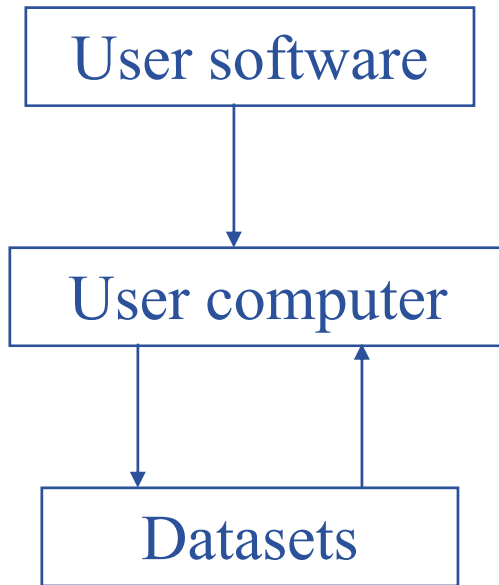


- It is an intermediate layer between Grid middleware and user's software. It integrates data, parameters, software, and grid middleware doing all submission and management of several users' software copies to grid.
- Performs DATA and TASK parallelism in grid.
- Web page: <http://www.hep.man.ac.uk/u/jamwer/>
- Paper: <http://www.geocities.com/jamwer2002/gridgeral.pdf>

> Easygrid BetaMiniApp Tau11-Run3

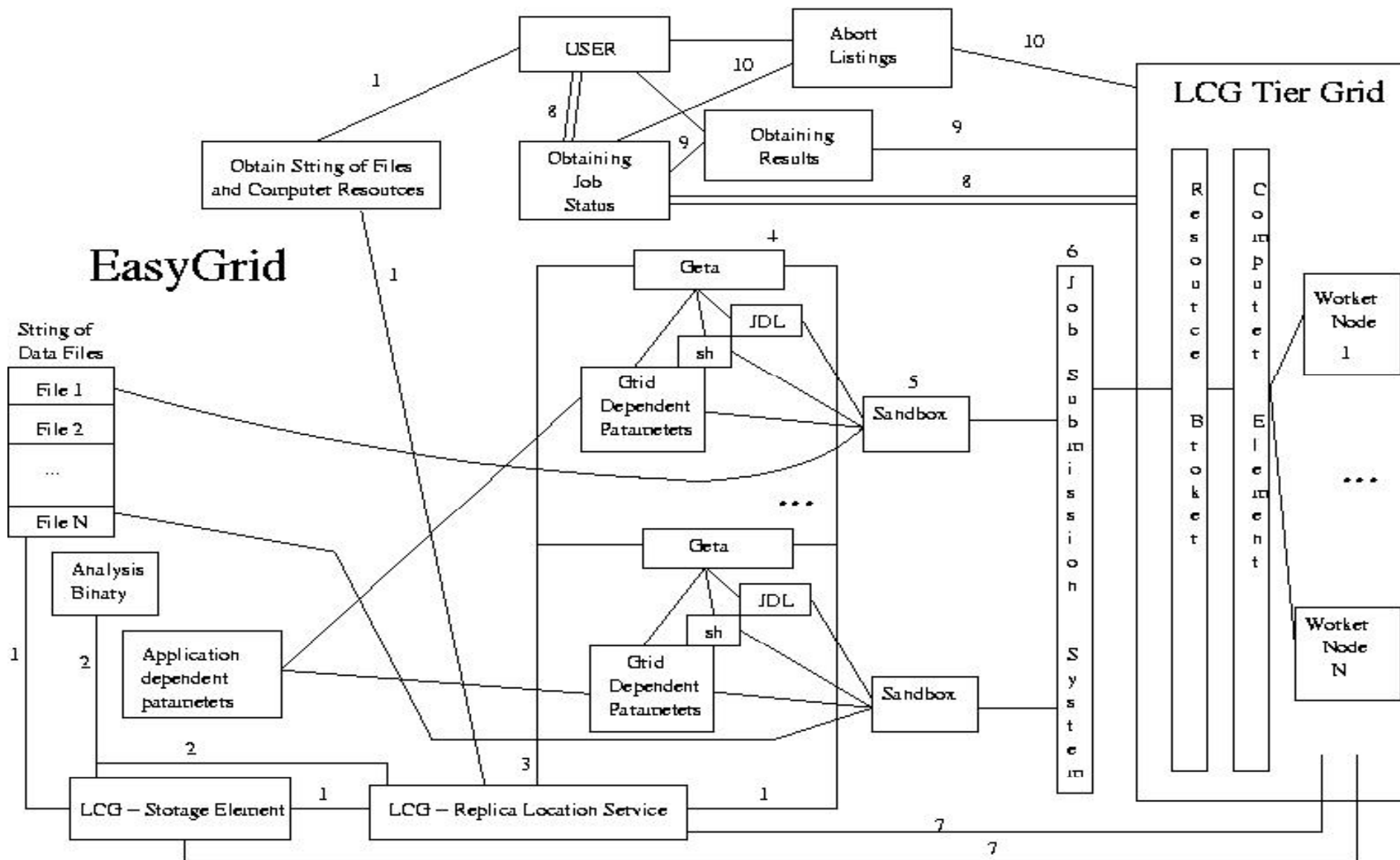
File name

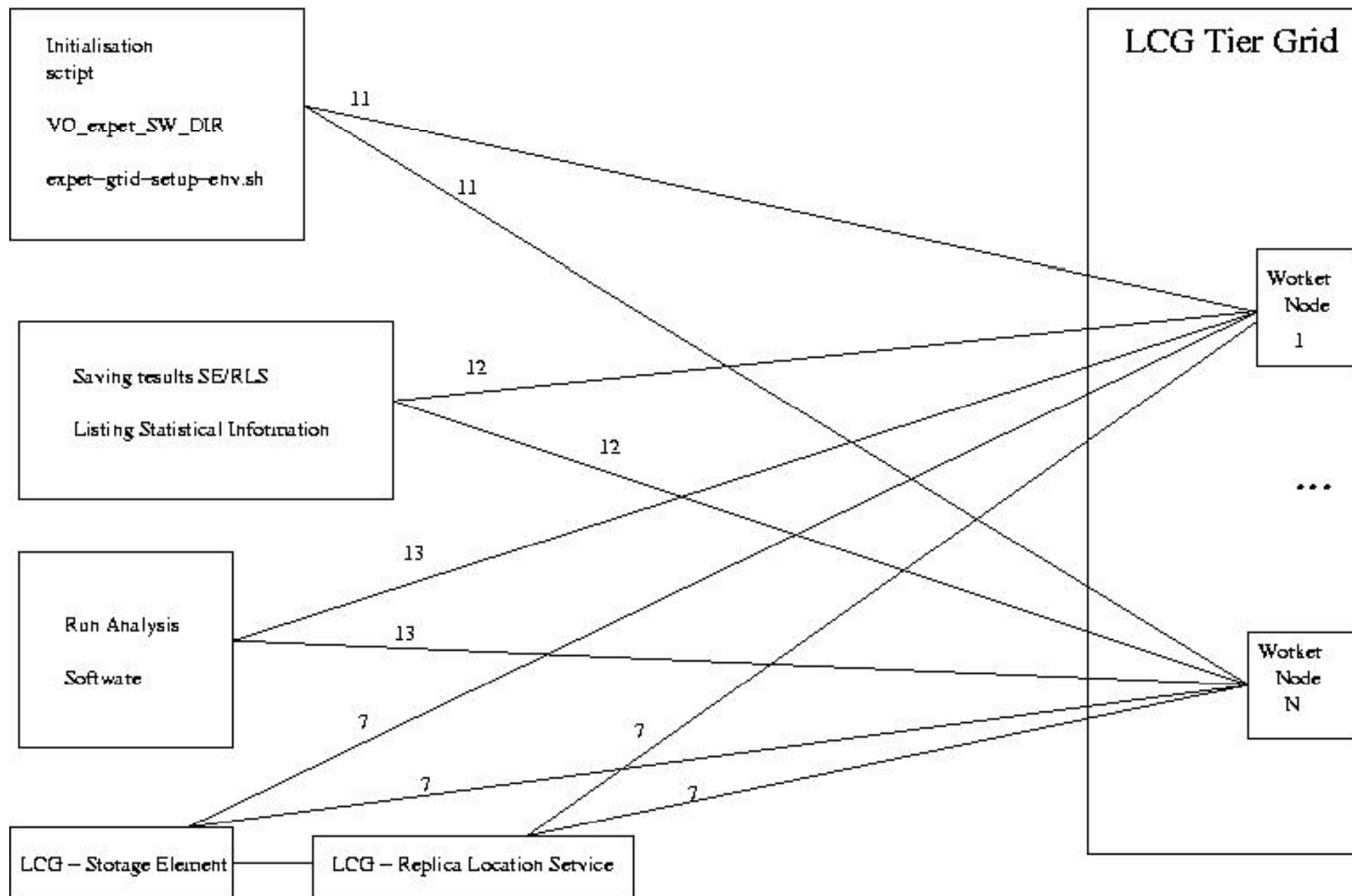
> BetaMiniApp Tau11-Run3.tcl



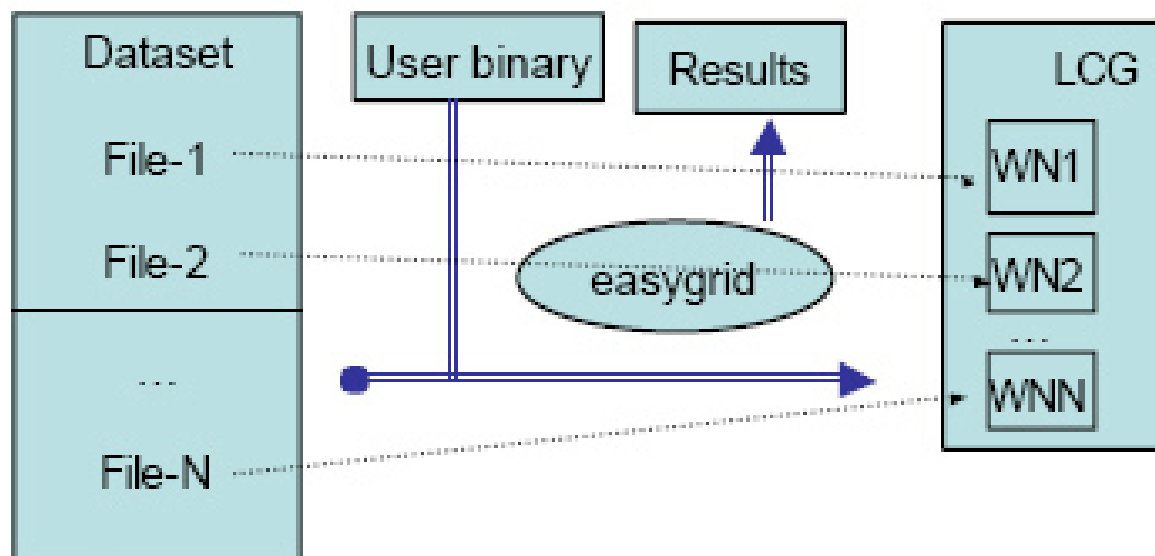
See <http://www.hep.man.ac.uk/u/jamwer/Grid2006.pdf> for more information







- **Each data file will be read by each copy of the binary code in parallel.**
- **EasyGrid Tasks:**
  - Copy binary code at closest storage elements.
  - Set environment in each worker node.
  - Start the binary code.
  - Recover results in user's directory.
  - Provide information in case software fails.
  - Tools for data management and replication.





Enabling Grids for E-science

# Data gridification benchmarks

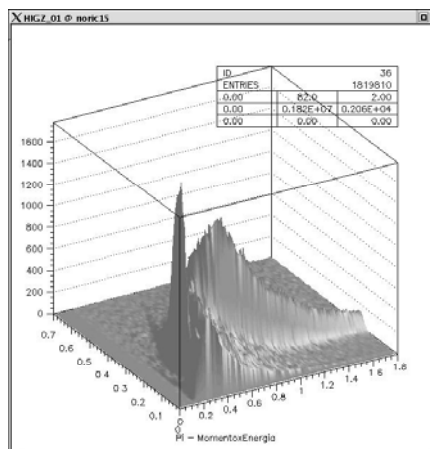
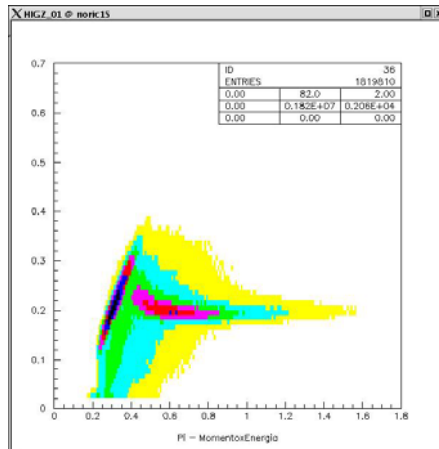
[www.eu-egee.org](http://www.eu-egee.org)



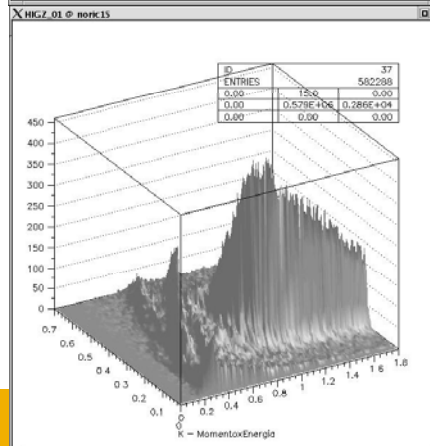
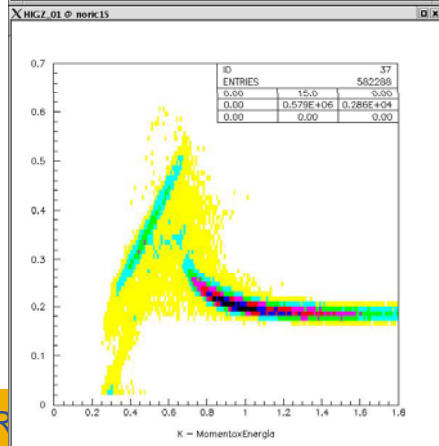
- Energy x Momentum for Tau 1N dataset.
- It contains 18,700,000 events.
- See <http://www.hep.man.ac.uk/u/jamwer/index.html#06>

## Monte Carlo Simulation

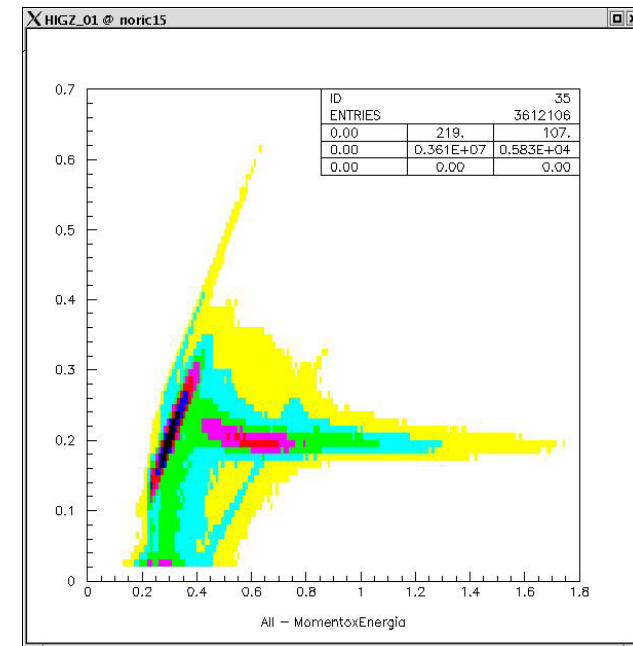
Pions



Kaons



Real data



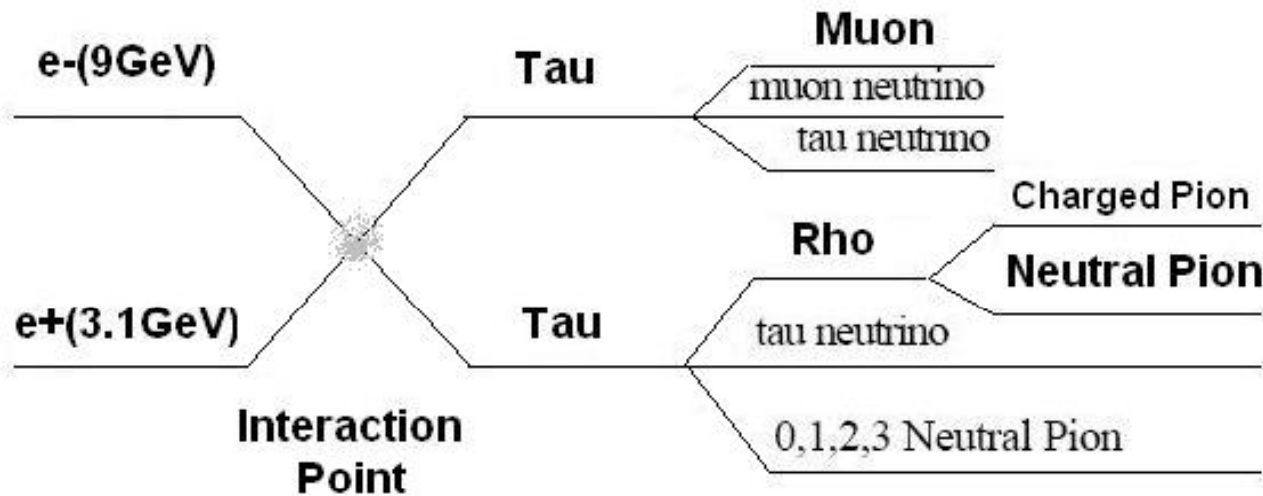
BbkDatasetTcl selected 482,303,947 events in dataset Tau11-Run[1,2,3,4]-OnPeak-R14.

Using easymoncar 4,890,000 events were simulated using Monte Carlo. Grid platform was used to run in parallel every data file selected by BbkDatasetTcl.

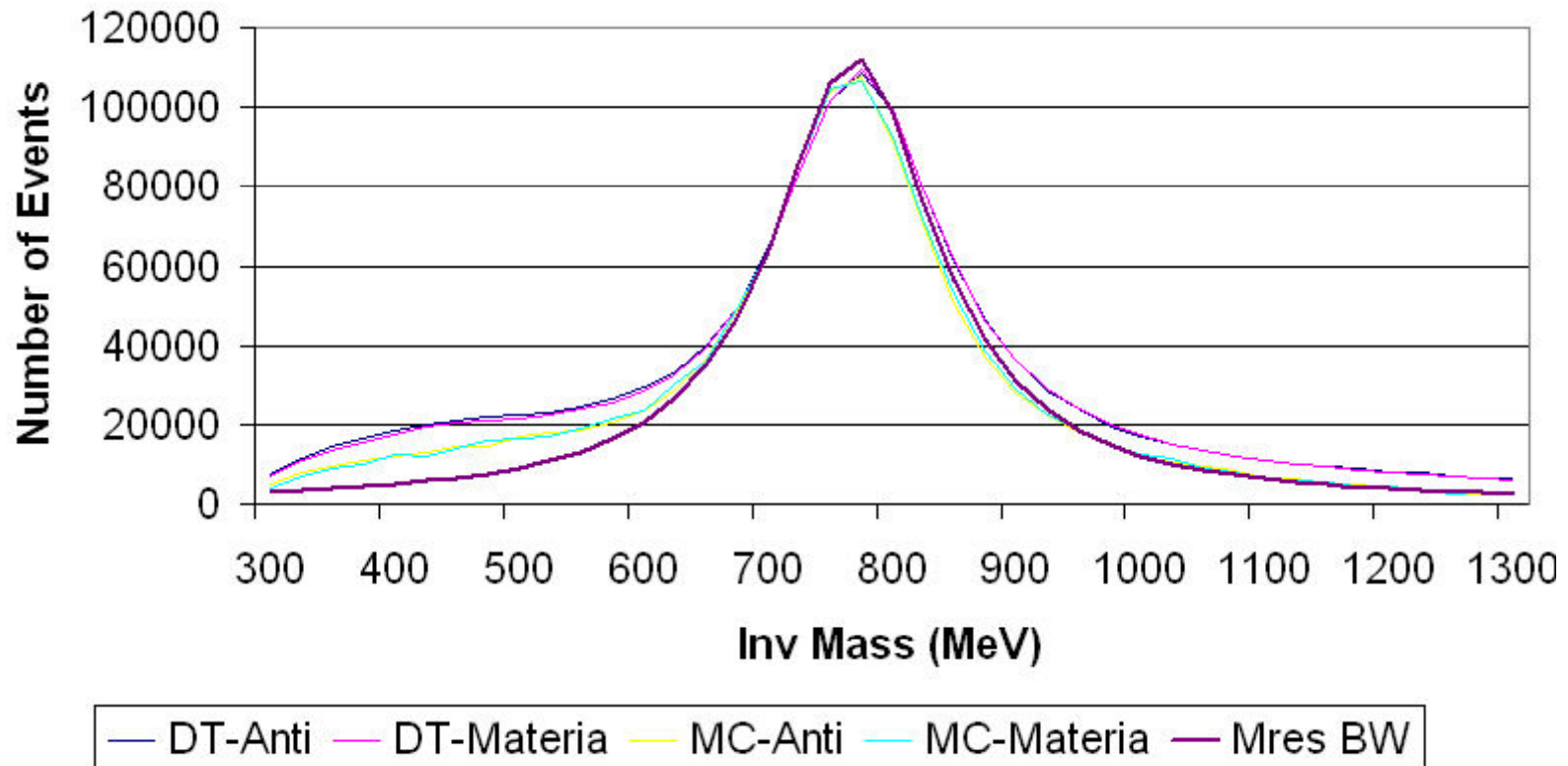
Run3 run at Manchester and Run1,2,4 at RAL.

Processing performance was 70,000 events per hour.

See <http://www.hep.man.ac.uk/u/jamwer/index.html#07>



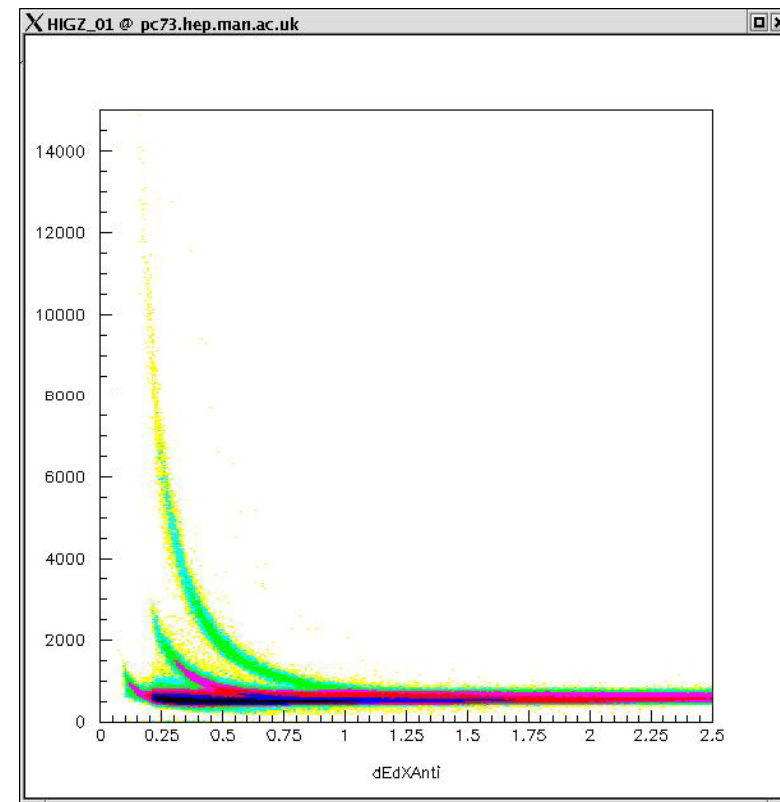
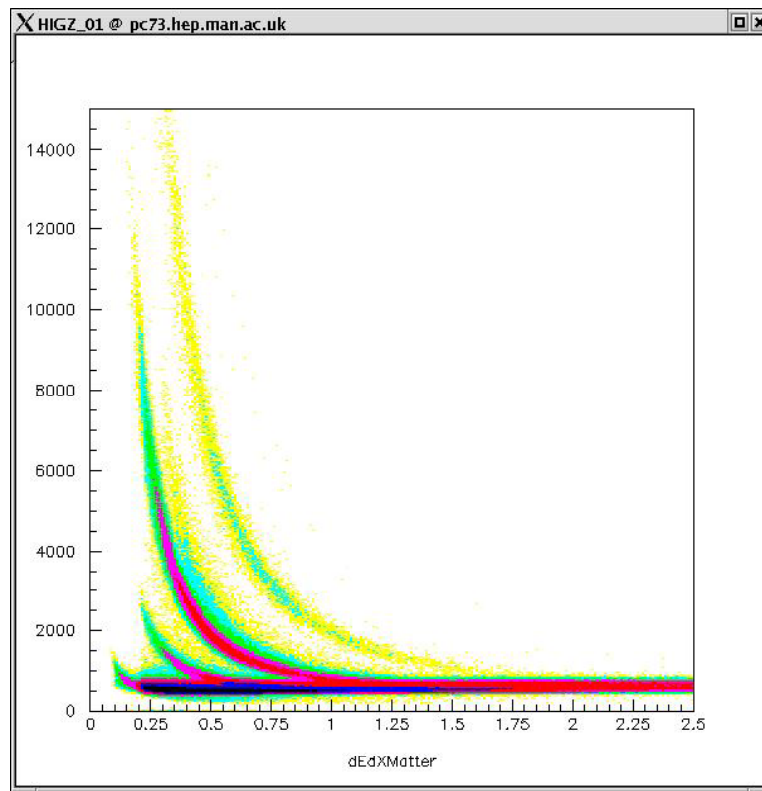
## Inv Mass Resonance 1 Pi0



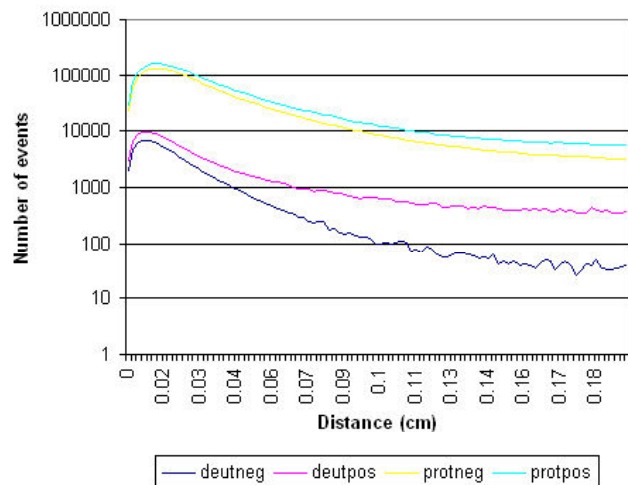
Parameters from Breit-Wigner mass distribution are: resonant mass 770 MeV, width 160 MeV and normalisation 4,500,000.



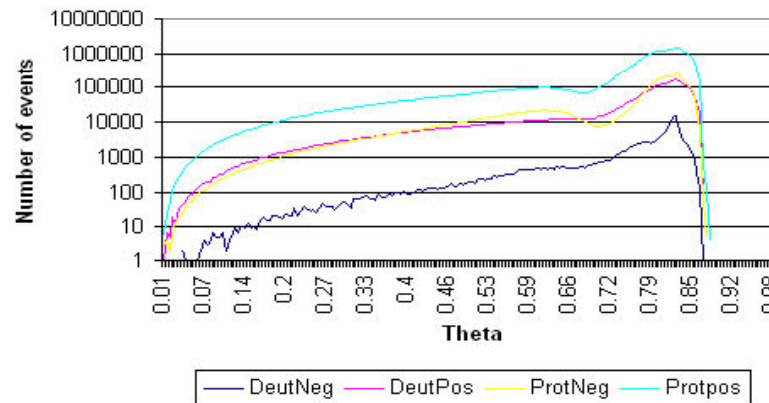
- The first task is to find where deuterons (and anti-deuterons) straps will be in  $dE/dx$  by momentum biparametric plots. The straps correspond to Pions, kaons, protons and deuterons respectively. The anti-matter plot almost does not have anti-deuteron events.
- There were 800 jobs searching in 2 million events each.
- See <http://www.hep.man.ac.uk/u/jamwer/index.html#08>



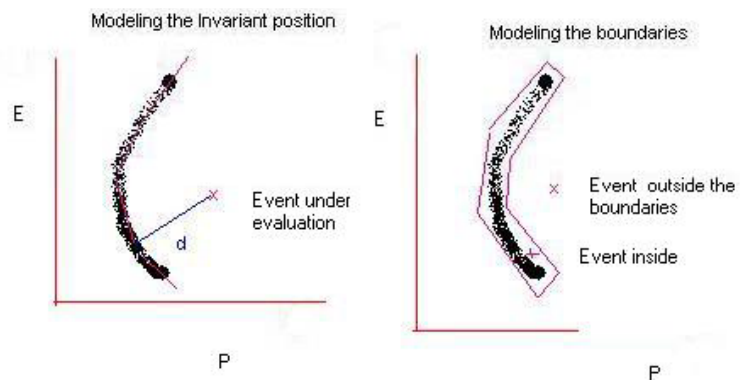
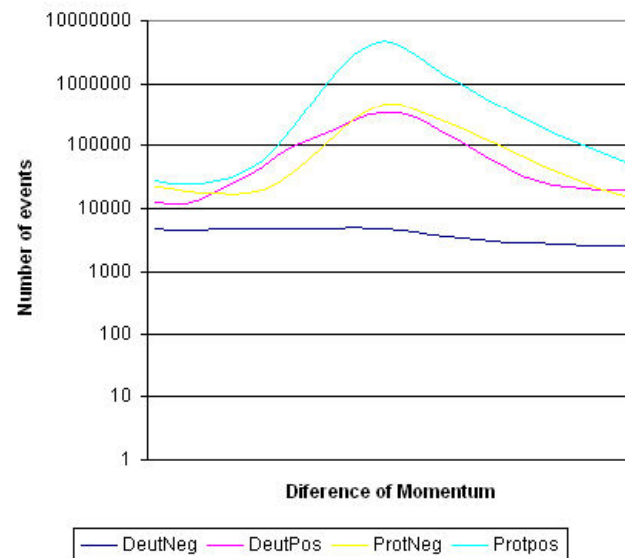
DOCA - detail



Theta Chewrenkov

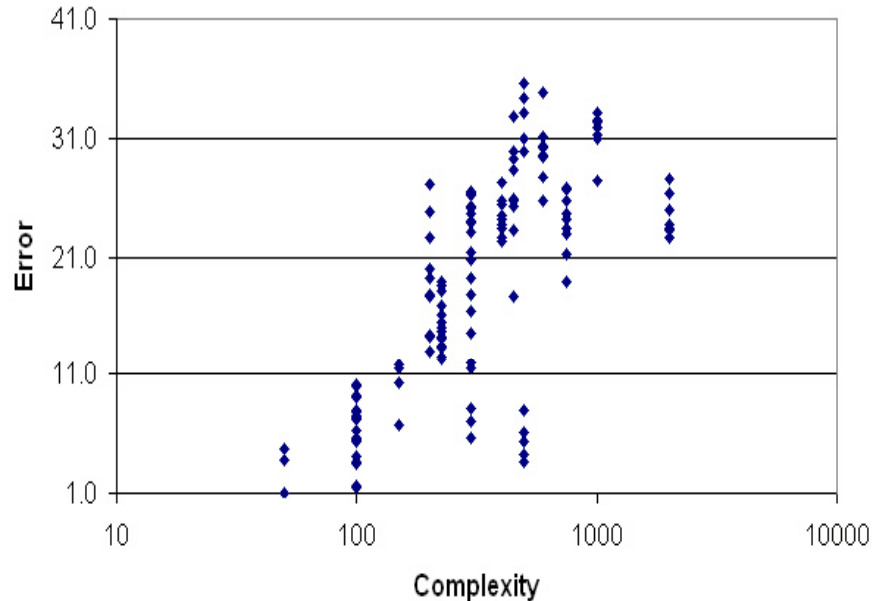


dE/dx Projection

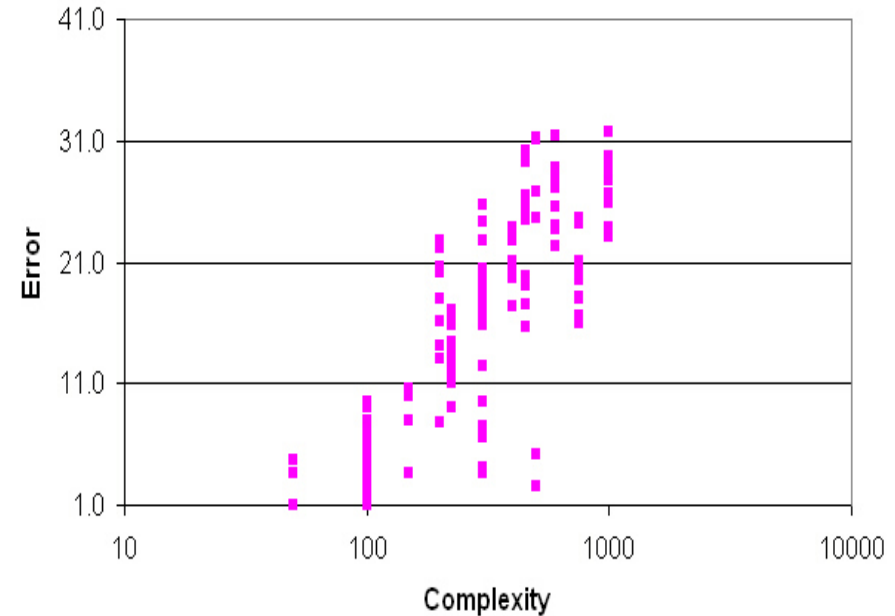


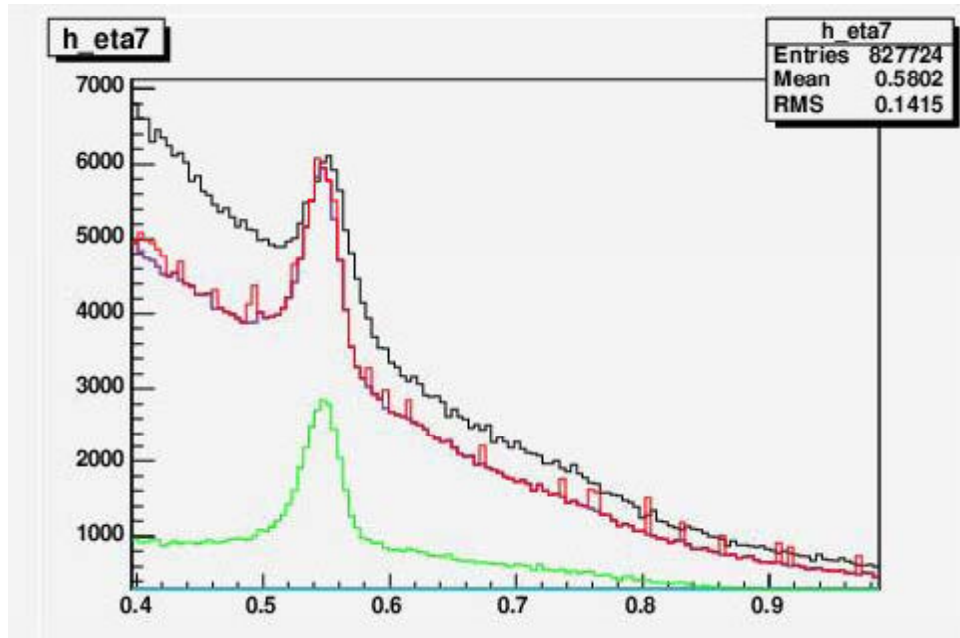
- **Job Shop Scheduling optimization using an always feasible map with genetic algorithm.**
- **161 data tests running in GA and MC.**

MC (1M/minute)

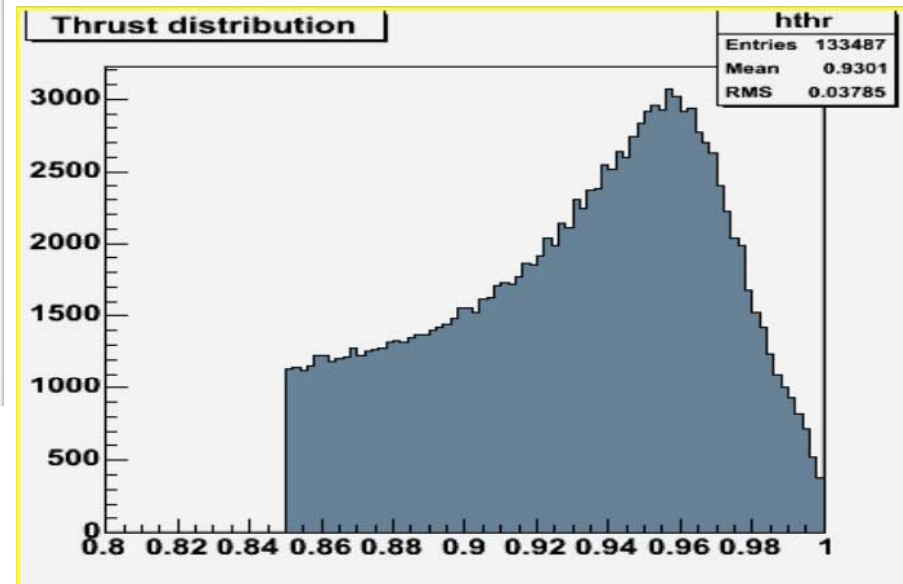


GA: 1000ind/ger, 50000 ger, 1seg/ger - gasched1



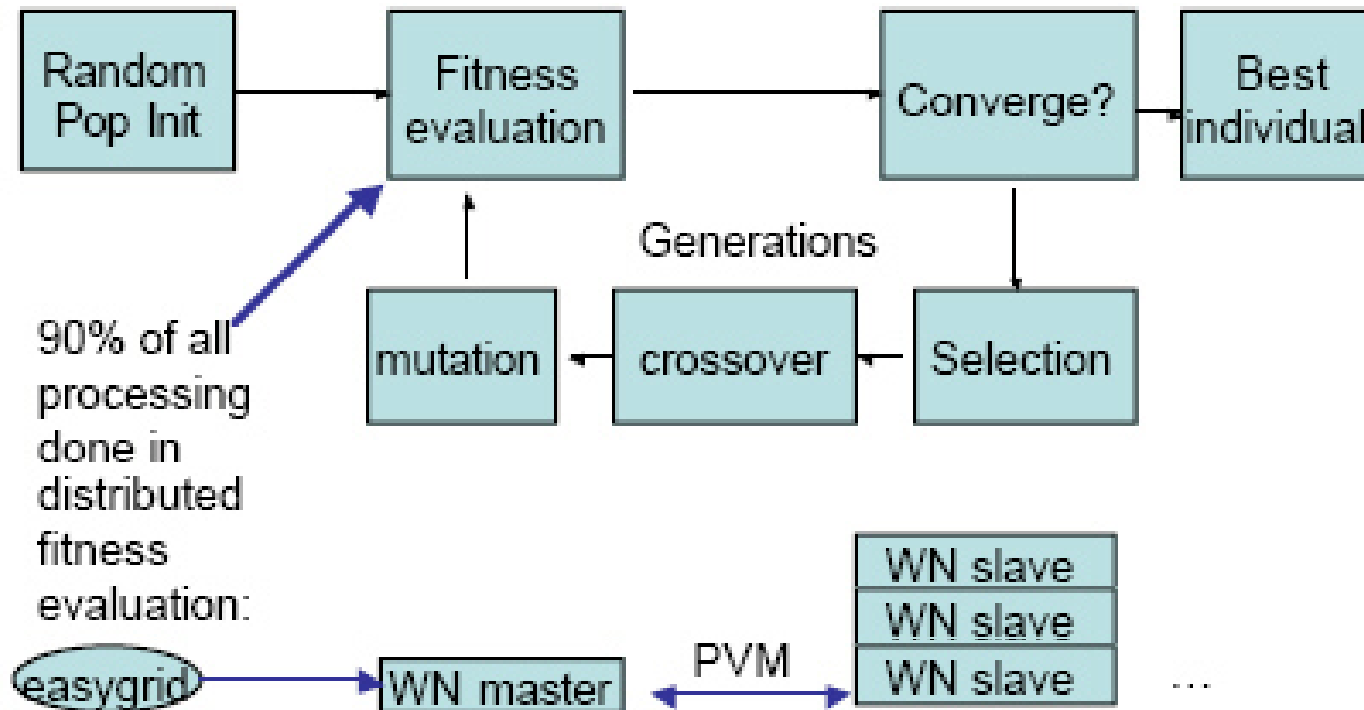


Source Dr Marta Tavera



Source Dr Mitchell Naisbit

- **One master binary code (or client) requesting services and managing load flow.**
- **EasyGrid Tasks:**
  - Set a task queue.
  - Search information system for services published in grid.
  - Establish sections in each worker node.
  - Start services and initialize software.
  - Send data for processing in each server.
  - Manages processing and re-submit in case of fail.
  - Manages notification and recover results in master.



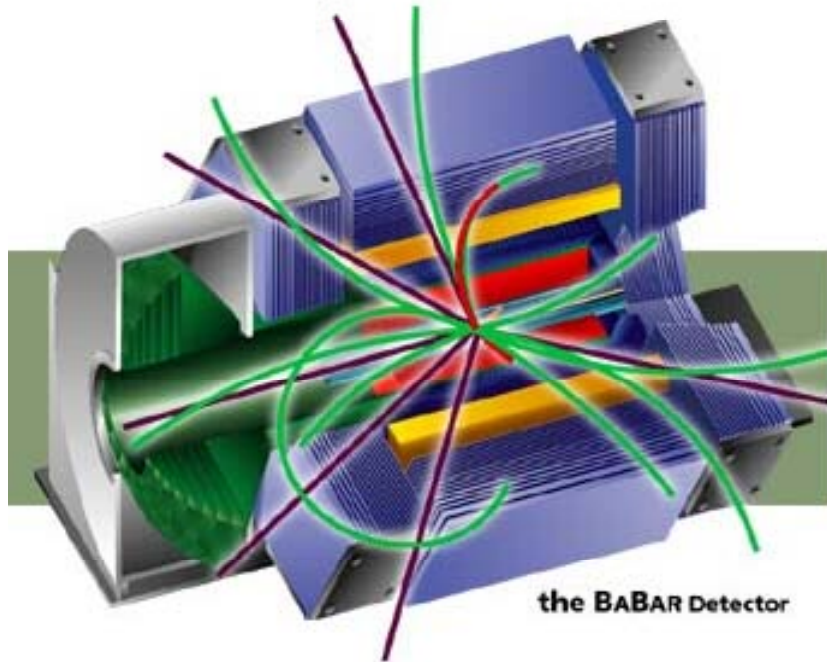


Enabling Grids for E-science

# Task gridification benchmark

[www.eu-egee.org](http://www.eu-egee.org)





Neutral Pions decays into 2  
Gammas, detected by

BaBar's Electromagnetic  
Calorimeter.

$$M^2 = \left(\sum E_i\right)^2 - \left(\sum P_i\right)^2$$

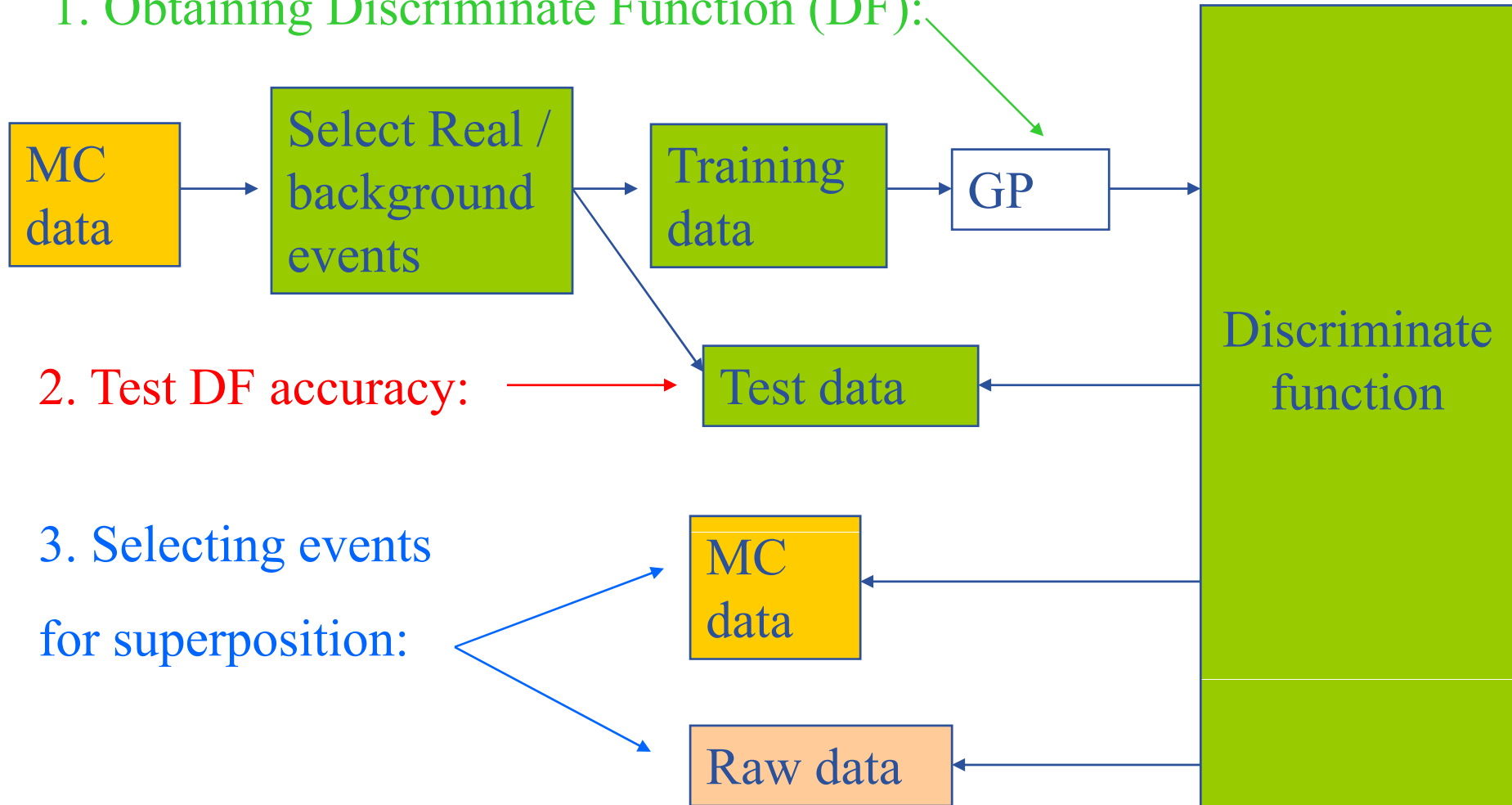
Two background gammas could have neutral pion invariant  
mass just by chance.

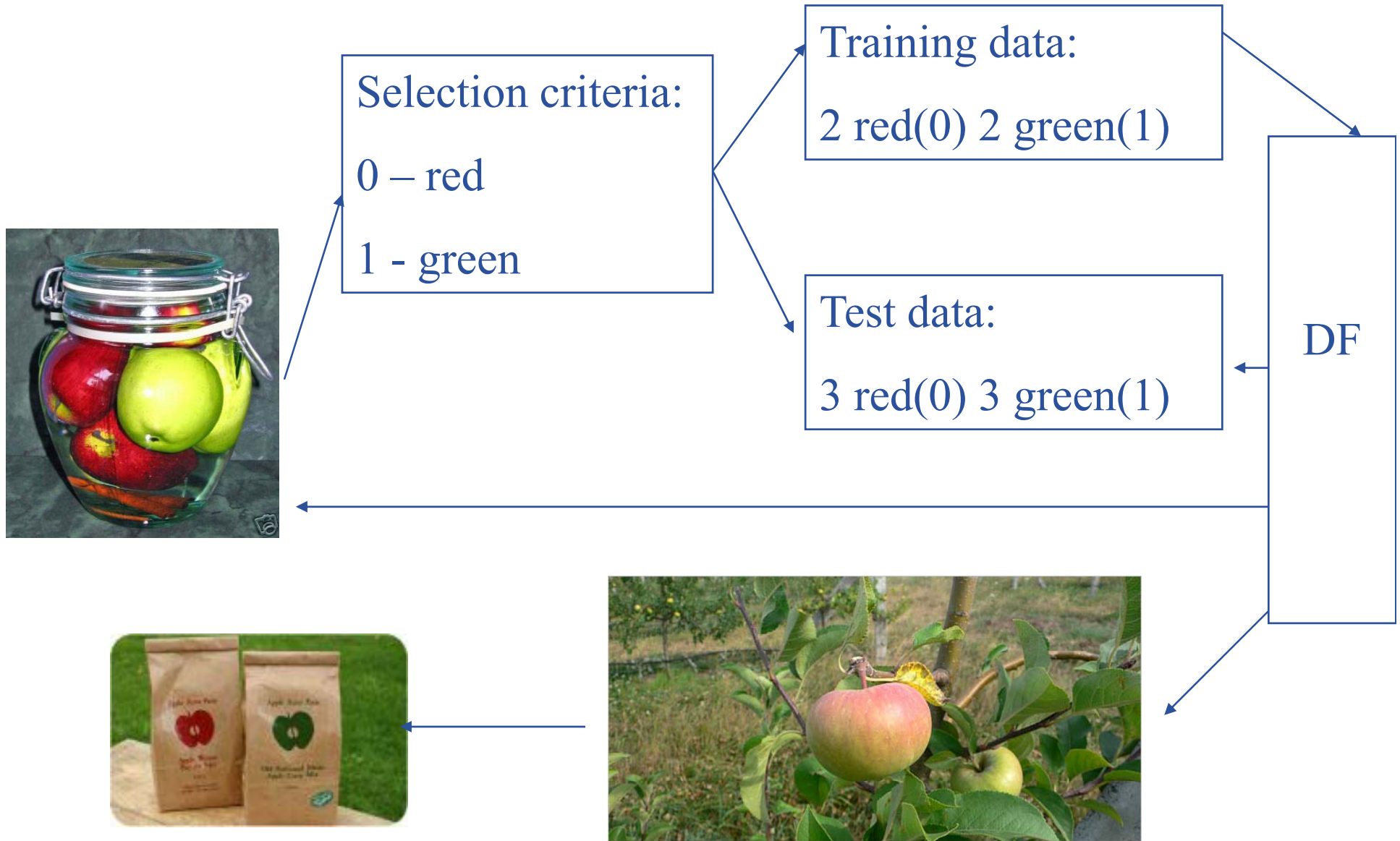
How to discriminate them using artificial intelligence ???



- **Mathematical model obtained with GP maps the variables hyperspace to a real value through the discriminator function, an algebraic function of kinematics variables.**
- **Applying the discriminator to a given pair of gammas:**
  - if the discriminate value is bigger than zero, the pair of gammas is deemed to come from pion decay.
  - Otherwise, the pair is deemed to come from another (background) source.
- **Paper:**  
<http://www.hep.man.ac.uk/u/jamwer/gphep.pdf>
- **Poster:**  
<http://www.hep.man.ac.uk/u/jamwer/loP2007.ppt>

## 1. Obtaining Discriminate Function (DF):





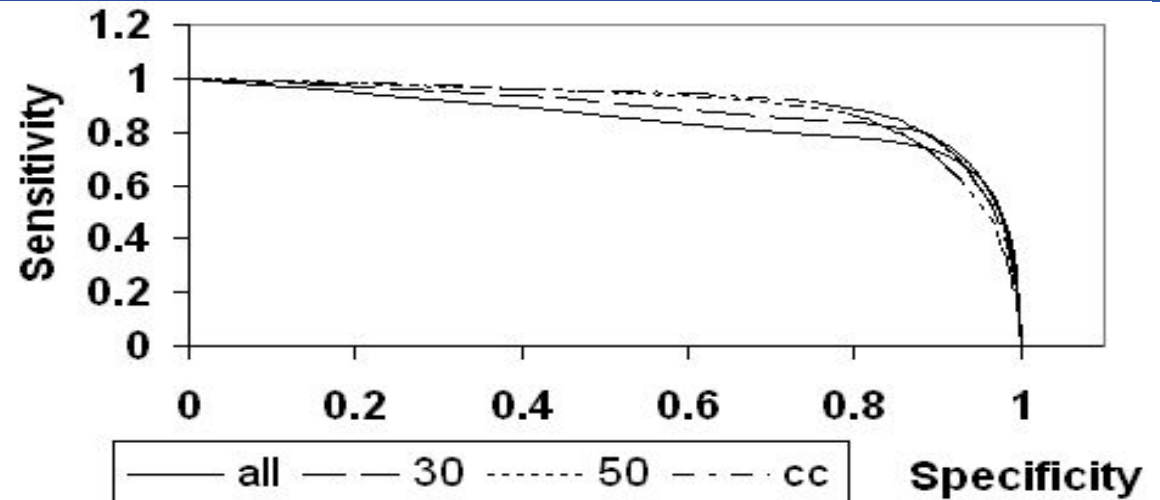
- **Reverse Polish Notation**
- **The population size is 500 individuals;**
- **Crossover and mutation probabilities are 60% and 20% respectively.**
- **Every generation, 20 best individuals are copied as they are (without crossover and mutation) and half population is generated randomly and replace the worse individuals.**
- **Algebraic operators have been used with kinematics data.**
- **The service we have distributed in grid was fitness evaluation, in parallel by many WN .**
- **482,303,947 BaBar's detector events and 20,489,668 MC events**

	<b>Standalone</b>	<b>1 node / 2 slaves</b>	<b>5 nodes / 10 slaves</b>
<b>Time(1,000s)</b>	80	47	19
<b>Improvement</b>		58%	24%

- **Monte Carlo (MC) generators integrates particle decays models with detector's system transfer function.**
- **MC events contain all information from each track particle and gamma radiation, which allows select high purity training dataset (96%+).**
- **Events with real neutral pion were selected and marked as “1”.**
- **Events without real pions into MC truth and invariant mass reconstruction in the same region of real neutral pions where also selected and marked as “0”.**

- **all gammas without energy cut (60,000 real and background records for training, and 60,000 real and 44527 background for test),**
- **more energetic than 30 MeV electronics' noise threshold (32,000 real and background records for training and test),**
- **more energetic than 50 MeV (15,000 real and background records for training and test),**
- **more energetic than 30MeV, lateral moment between 0.0 and 0.8, and have hit more than one crystal in the electromagnetic calorimeter - the conventional cut for neutral pion(16,000 real and background records for training and test).**

- $\alpha$ : Sensitivity or efficiency.
- $\beta$ : specificity or purity.
- $\gamma$ : accuracy.

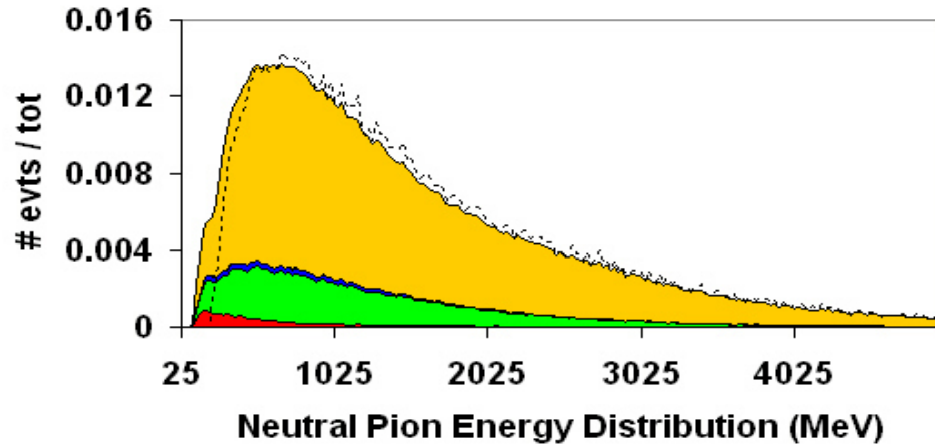


Steps	Real	a. all gammas		b. E>30MeV		c. E>50 MeV		d. Usual Cuts	
		D<0	D>0	D<0	D>0	D<0	D>0	D<0	D>0
Train	0	53559	6441	28143	3857	12147	2853	13620	2380
	1	15834	44166	6316	25684	2180	12820	2400	13600
	$\alpha$	74		80		85		85	
	$\beta$	89		88		81		85	
	$\gamma$	81		84		83		85	
Test	Real	D<0	D>0	D<0	D>0	D<0	D>0	D<0	D>0
	0	39531	4996	27510	4490	11799	3201	13319	2681
	1	16480	43520	6726	25274	2222	12778	2566	13434
	$\alpha$	72		79		85		84	
	$\beta$	89		86		79		83	
$\gamma$	79		82		82		84		

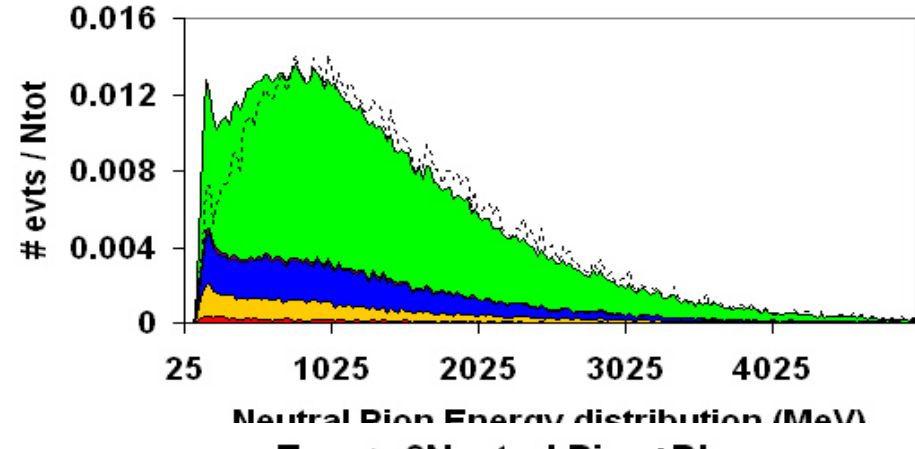
- **Cumulative plot of energy distribution for 1, 2, 3 and 4 neutral pion decays using all gammas NPDF.**
- **Contamination effect can be seen from MC energy distribution.**
- **The agreement between Monte Carlo and experimental data is conclusive about method's convergence and accuracy.**



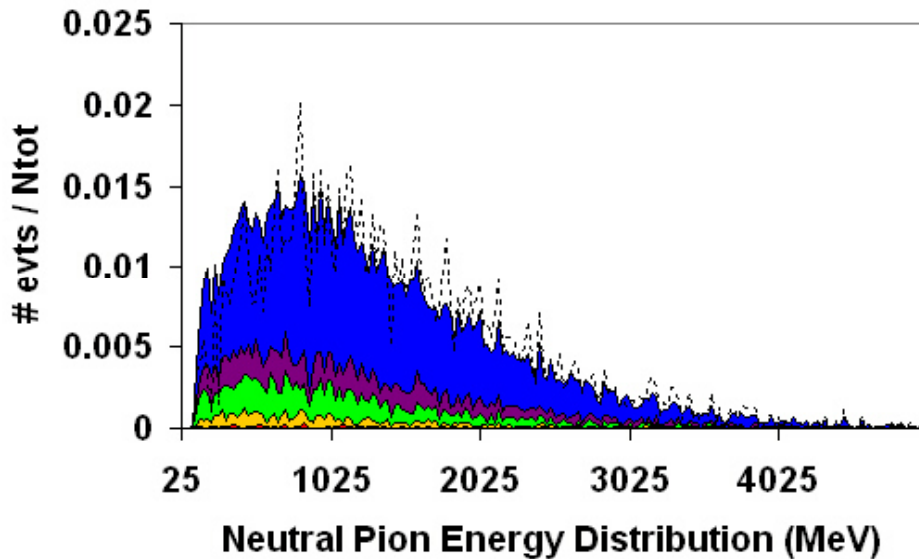
**Tau -> Rho(770) -> Pion+Neutral Pion**



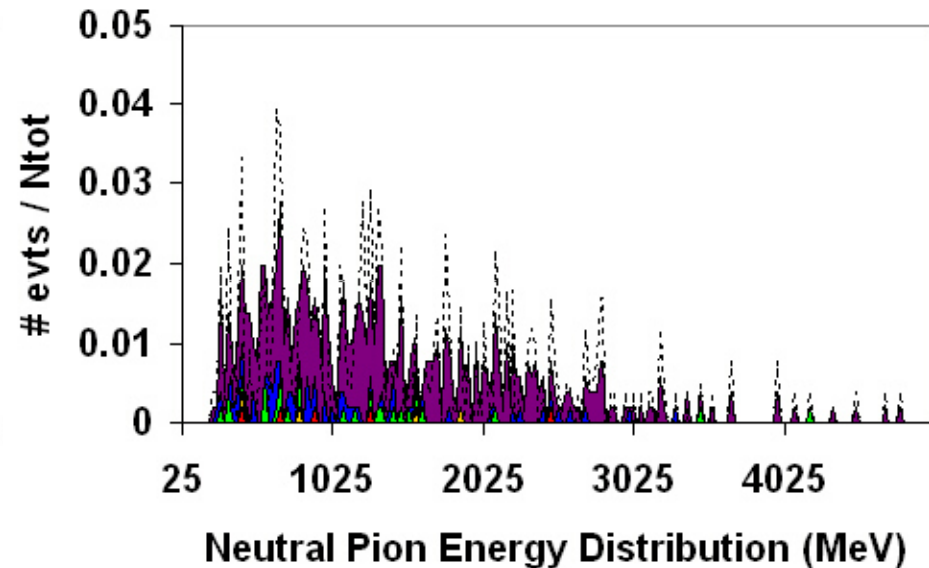
**Tau -> Neutral Pion+Rho(770)**



**Tau -> 2 Neutral Pion+Rho**



**Tau -> 3Neutral Pion+Rho**



■ 0Pi0 
 ■ 1Pi0 
 ■ 2Pi0 
 ■ 4Pi0 
 ■ 3Pi0 
 - - - data

■ 0Pi0 
 ■ 1Pi0 
 ■ 2Pi0 
 ■ 3Pi0 
 ■ 4Pi0 
 - - - data

- Available since GridPP11 - September/2004:  
[http://www.gridpp.ac.uk/gridpp11/babar\\_main.ppt](http://www.gridpp.ac.uk/gridpp11/babar_main.ppt)
- Several benchmarks with BaBar experiment data:
  - Data Gridification:
    - Particle identification: <http://www.hep.man.ac.uk/u/jamwer/index.html#06>
    - Neutral pion decays: <http://www.hep.man.ac.uk/u/jamwer/index.html#07>
    - Search for anti deuteron: <http://www.hep.man.ac.uk/u/jamwer/index.html#08>
  - Functional gridification:
    - Evolutionary neutral pion discriminate function:  
<http://www.hep.man.ac.uk/u/jamwer/index.html#13>
- Documentation (main web page):  
<http://www.hep.man.ac.uk/u/jamwer/>

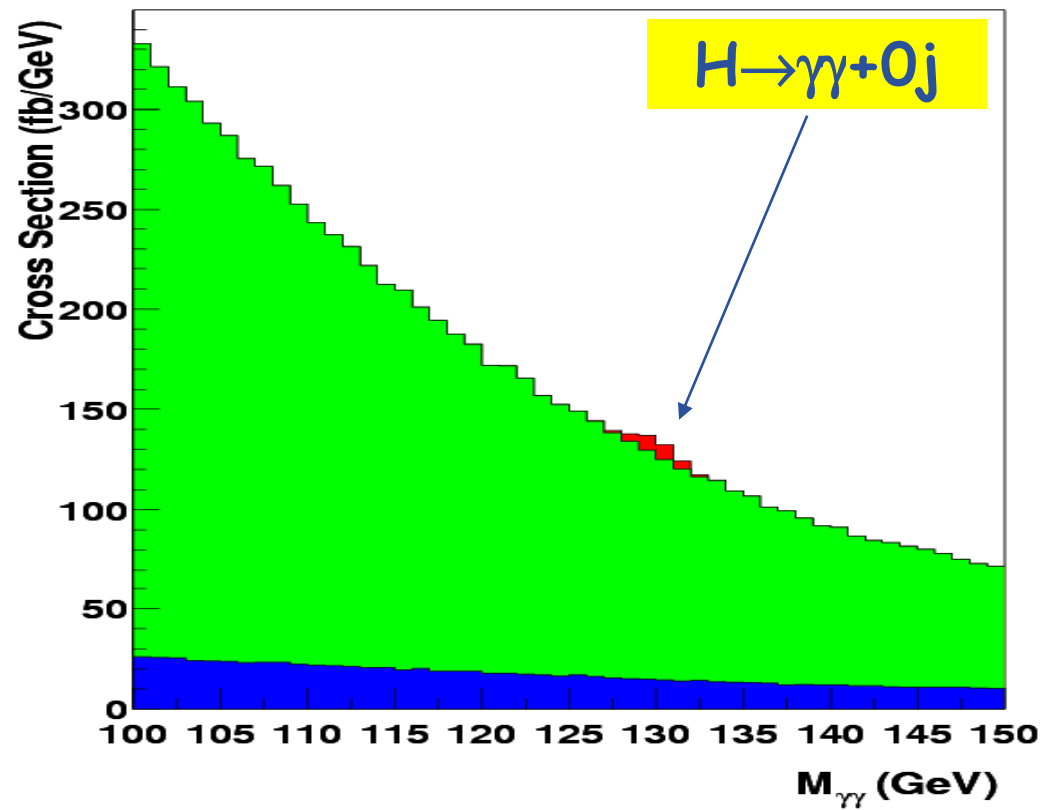
109 html files and 327 complementary files
- 60 CPUs production and 10 CPUs development farms running independently without any problem between November/2005 and September /2006.

- 20 international events:

<http://www.hep.man.ac.uk/u/jamwer/index.html#10>

- 5 refereed papers Int. Conferences.
- GridPP stand at IoP2006 and IoP2007.
- Contributions at GridPP web pages.

<http://www.gridpp.ac.uk/posters/>



- **EasyGrid is a framework for distributed analysis that works very well providing task and functional gridification capabilities.**
- **Genetic programming approach obtains neutral pion discriminate function to discern between background and real neutral pion particles. Background can produce a critical influence in systematic errors and constrain qualitative analysis.**
- **Results from hadronic tau decays analyzed in this paper showed genetic programming discriminate function has an important role in background reduction, improving analysis quality.**
- **The use of NPDF will allow the study of observable and check with values obtained from theoretical Standard Model, from a sample of events with high purity.**