Large High Altitude Air Shower Observatory(LHAASO) 数据处理

查敏

zham@ihep.ac.cn

2024高能物理计算暑期学校 @ 21st - 24th August, 2024

outline

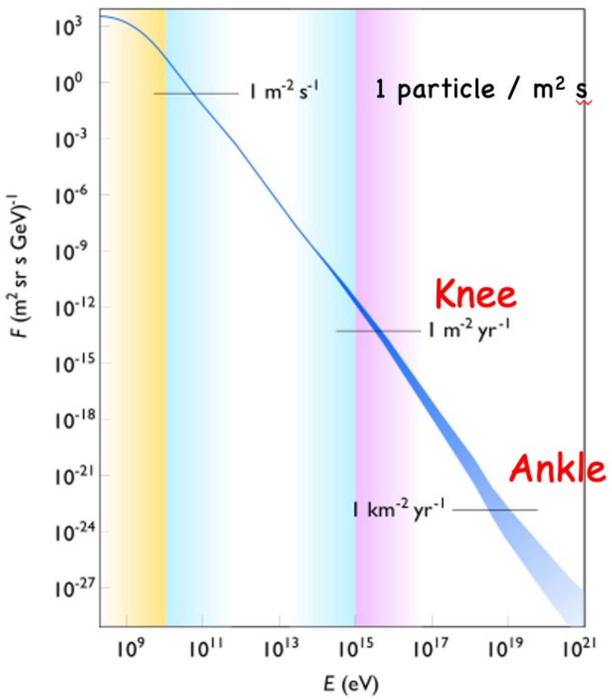
- **LHAASO and its detector**
  - LHAASO science
  - Detector calibration
- **Data production**
  - Data reconstruction
  - Data quality check
  - MC data production
- **Scientific data analysis**
  - Gamma ray astronomy related
  - Cosmic Ray related

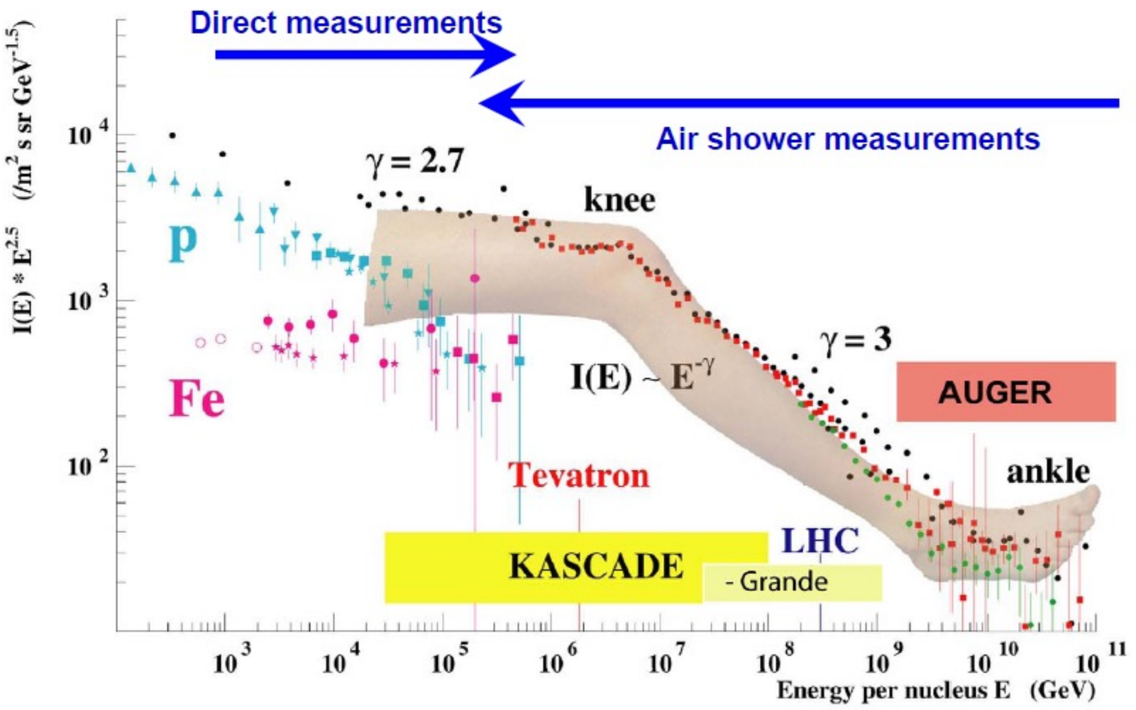
LHAASO “拉索”, Haizi Mountain 4410 m a.s.l. Daocheng, Sichuan Province, China

Location: 29°21' 27.6" N , 100°08'19.6" E

2021-07 completed built and in full array operation

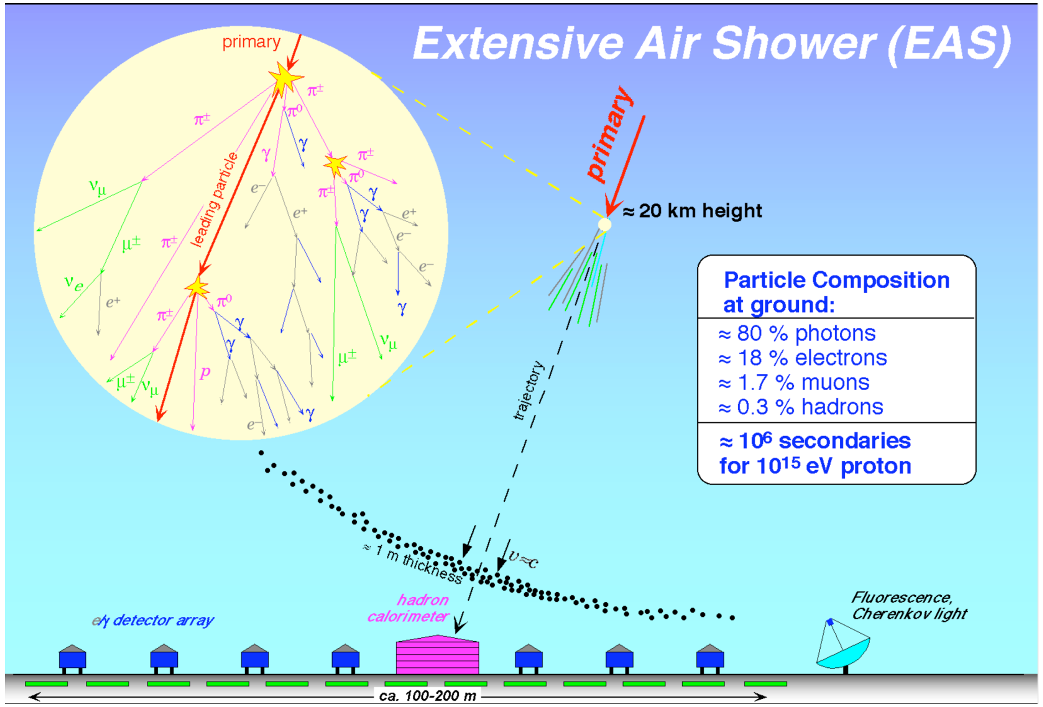
The Energy Spectrum for Cosmic Ray

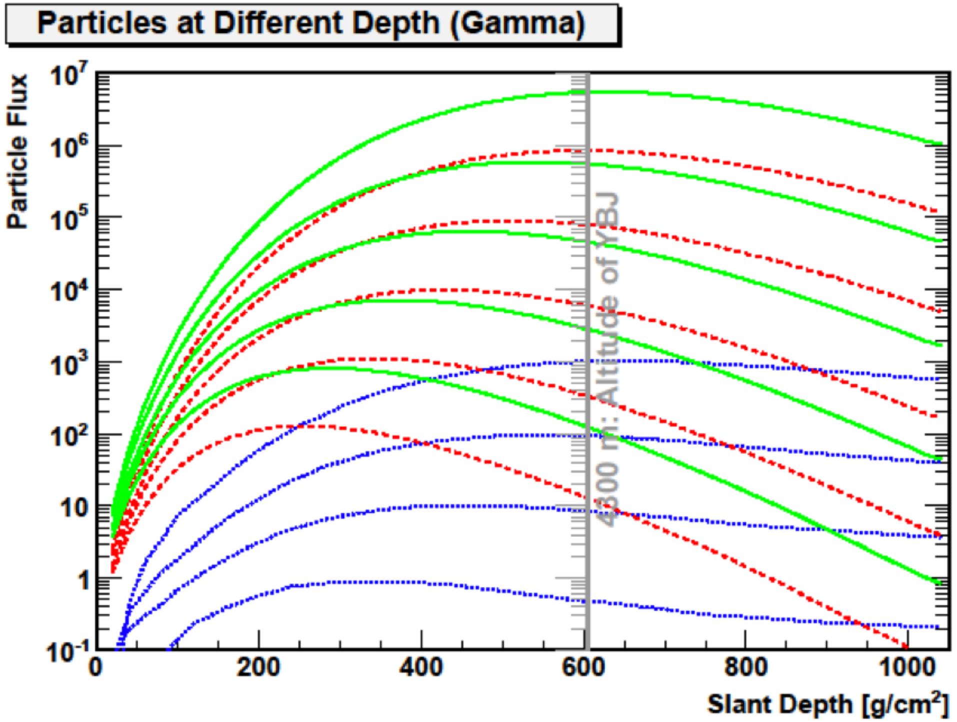




- 跨越12个量级；
- 最高能可达10^{20} eV！
- 近似幂指数 (power law)分布；
- 能量越高，流强越小。
- 加速机制？

高海拔: 降低阈能 + 膝区宇宙线的极大发展深度

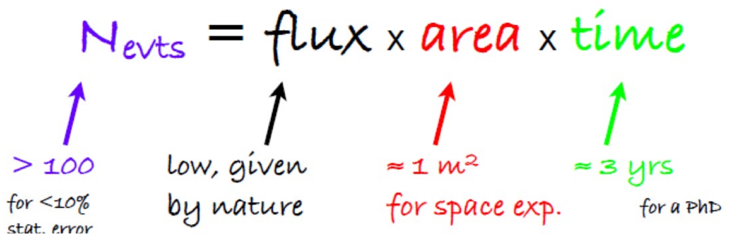


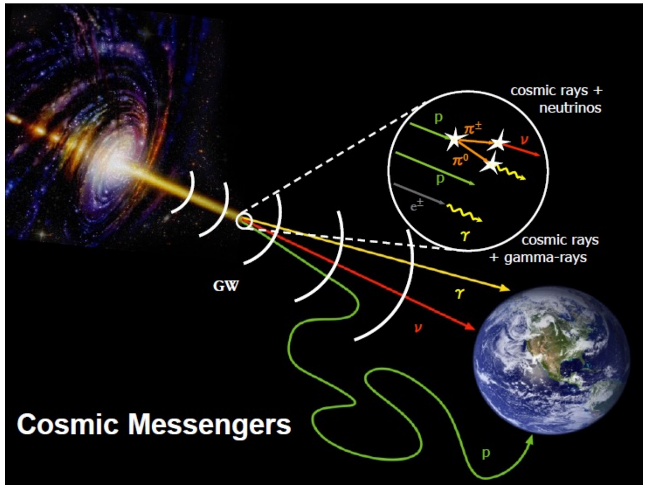


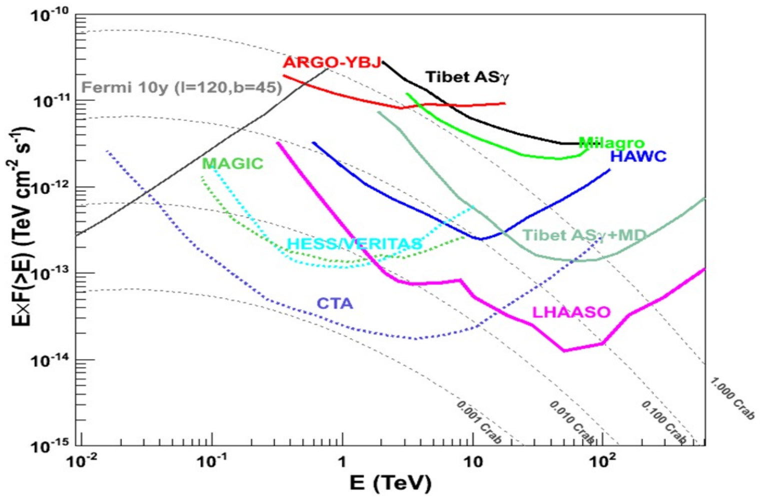
- Green: gamma;
- Red: electron;
- Blue: muon.
- From top to bottom:
  - 1 PeV;
  - 100 TeV;
  - 10 TeV;
  - 1 TeV;
  - 100 GeV.

通过测量簇射中的次级粒子来获取原初宇宙线的信息（方向、能量、成分）

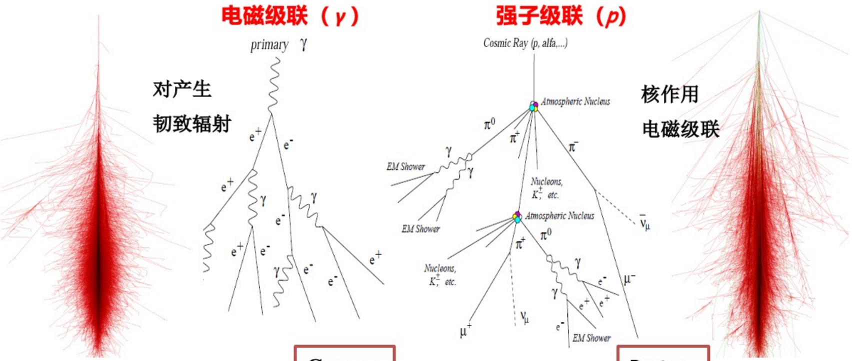
Ground-based air shower detection



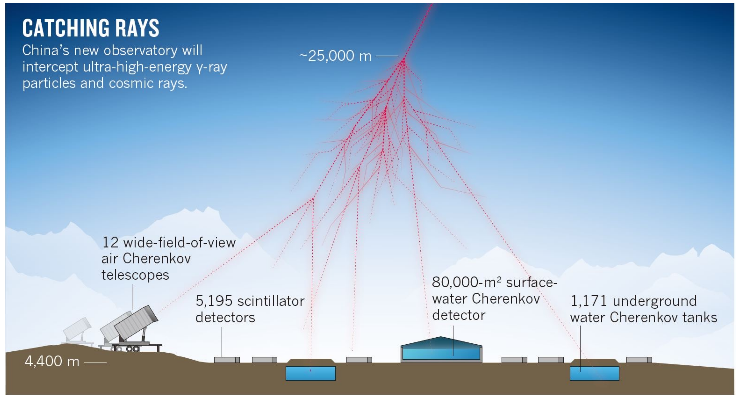




- High sensitivity: ~2% Crab @3TeV@100TeV
- Wide energy range: sub-TeV to 10 PeV
- Large FOV:~1.8 sr
- Detect air shower secondary particles: Gammas, electrons/positrons, muons, photons, hadrons, ...
- Measure the numbers / (or energy eqv.), arrival time, as well as lateral / longitudinal distribution.
- Reconstruct the direction, energy, type of the primary particle.



	Gamma	Proton
次级粒子分布：	单芯对称；集中分布；	多芯或单芯不对称；杂乱分布；
主要成分：	e±；γ；	e±；γ；核子；**μ子；**
首次作用点：	离地面较低；	离地面较高；



伽马射线是重要的宇宙信使之一

LHAASO

KM2A:
5216 ED/1m^2 + 1188 MD/36m^2
Area: 1.3 km^2

UHE gamma ray astronomy

WCDA:
3 pools， 3120 cells/25m^2
area: 78,000 m^2

VHE gamma ray astronomy

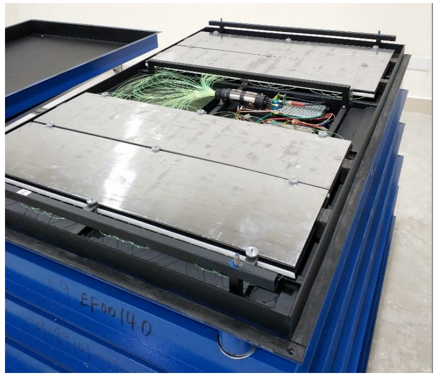
WFCTA:
18 telescopes

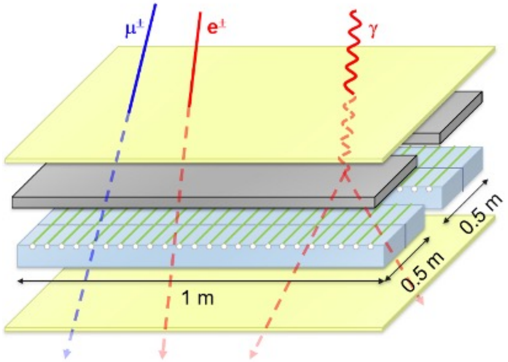
CR individual spectrum…

Planed neutron detectors + IACT
…

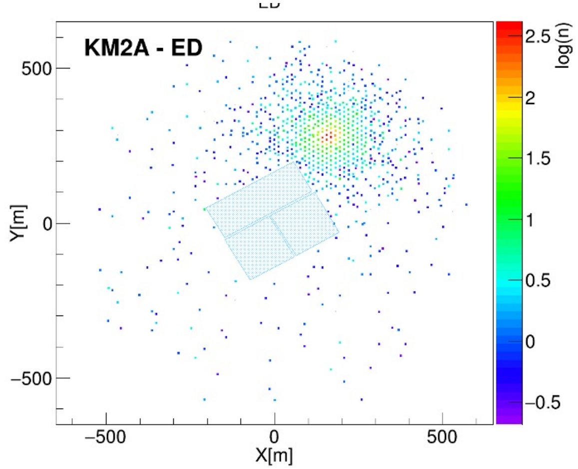
……

Electromagnetic Detector (ED)

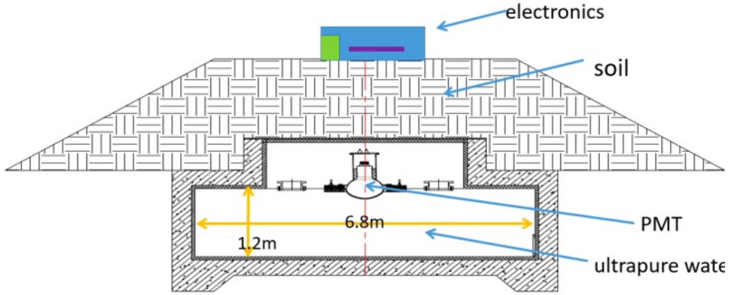




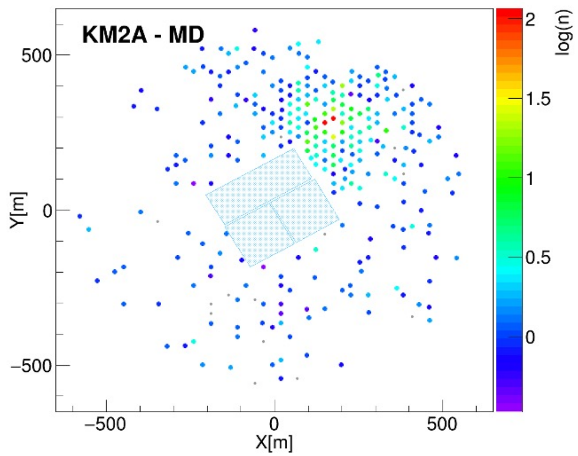
Ne=5381, N_{ED}=773

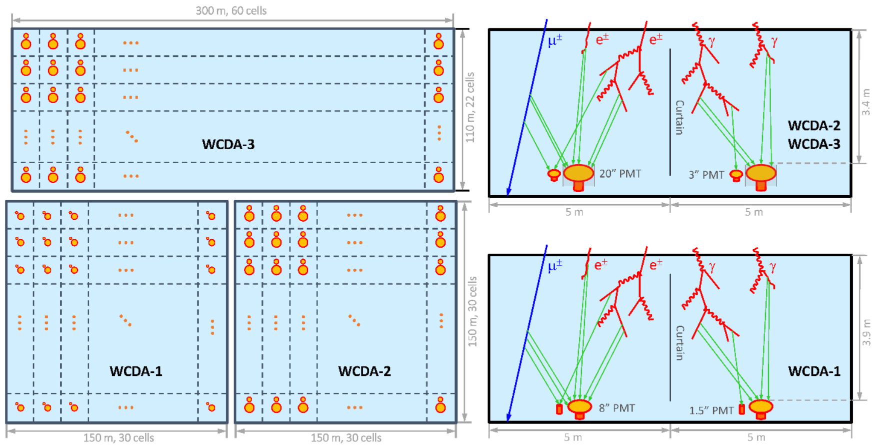


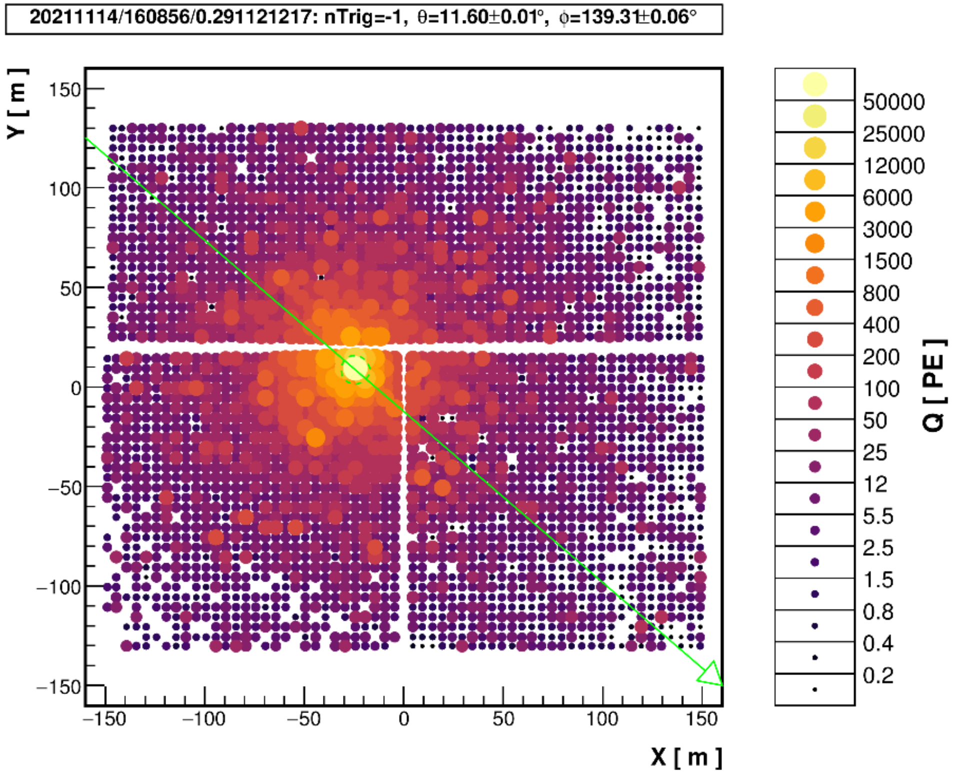
Muon Detector (MD)



Nμ=853, N_{MD}=279







- **Area:**
  78,000 m^2
- **Detector units:**
  3120
- **Energy Range:**
  0.1-30 TeV

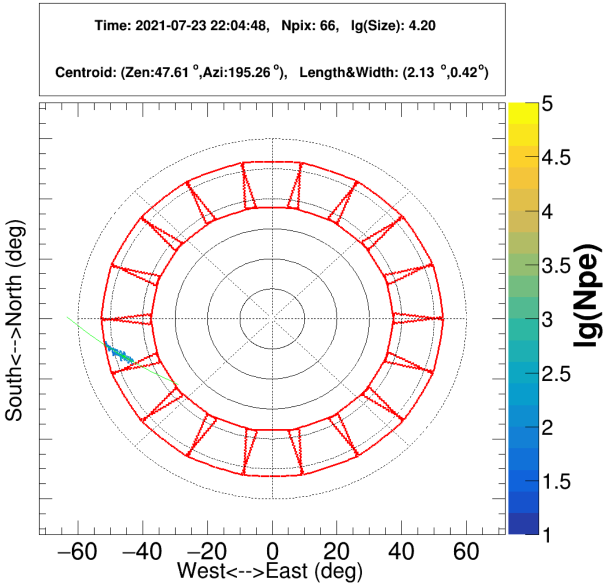
Wide Field of View Cherekov Telescope Array (WFCTA)

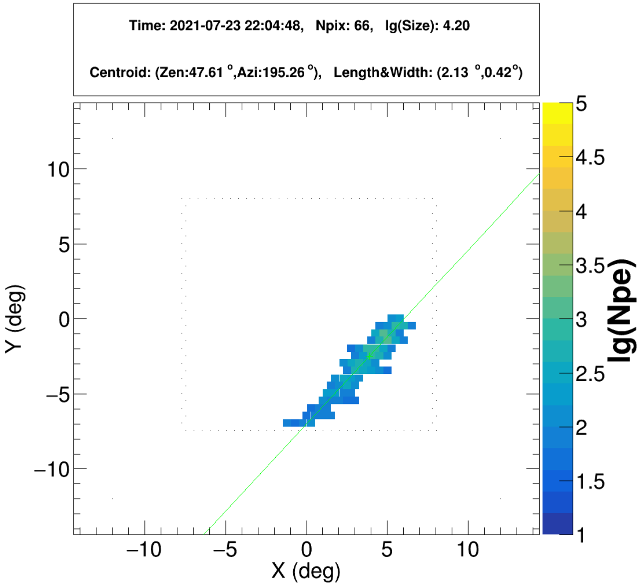
Mirror: 5 m^2 spherical mirror

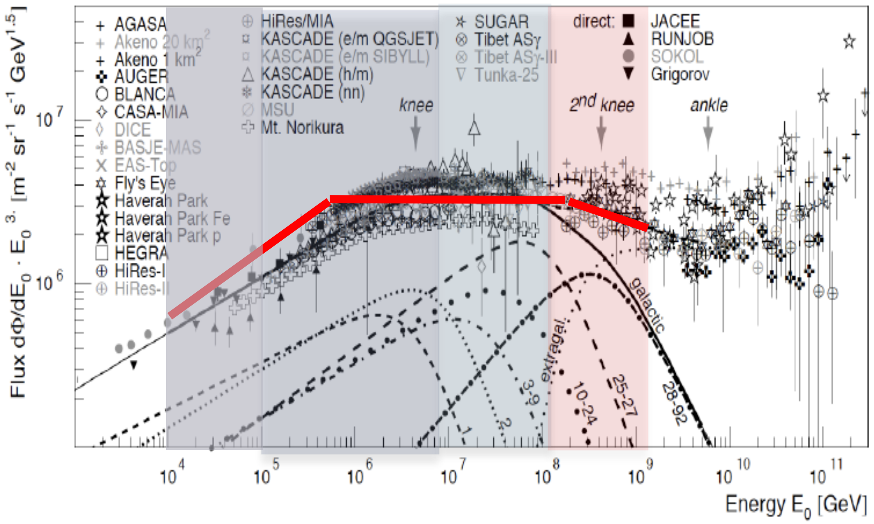
FOV: $16° \times 16°$ / telescope

Camera: 32x32 =1024 pixels /telescope

Pixel: $0.5°$ each

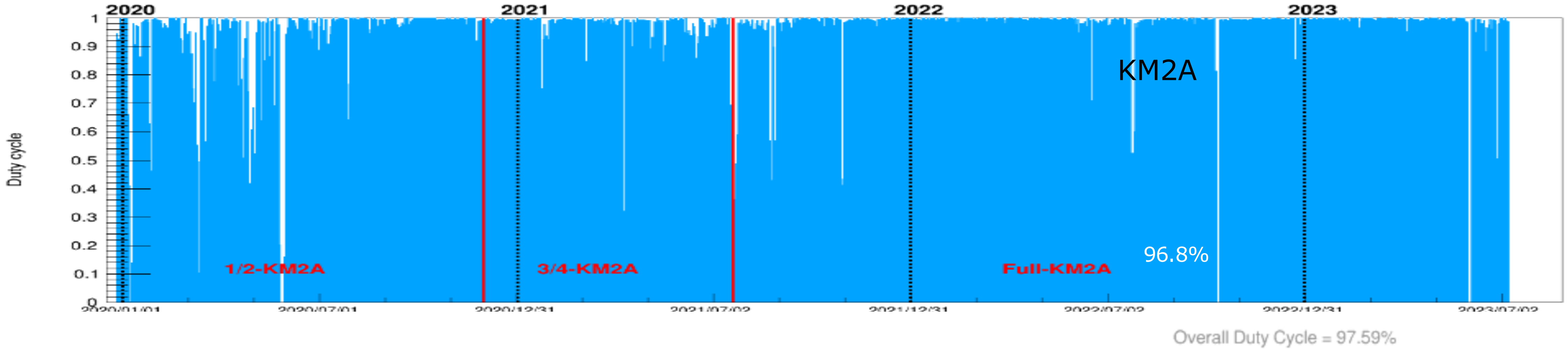


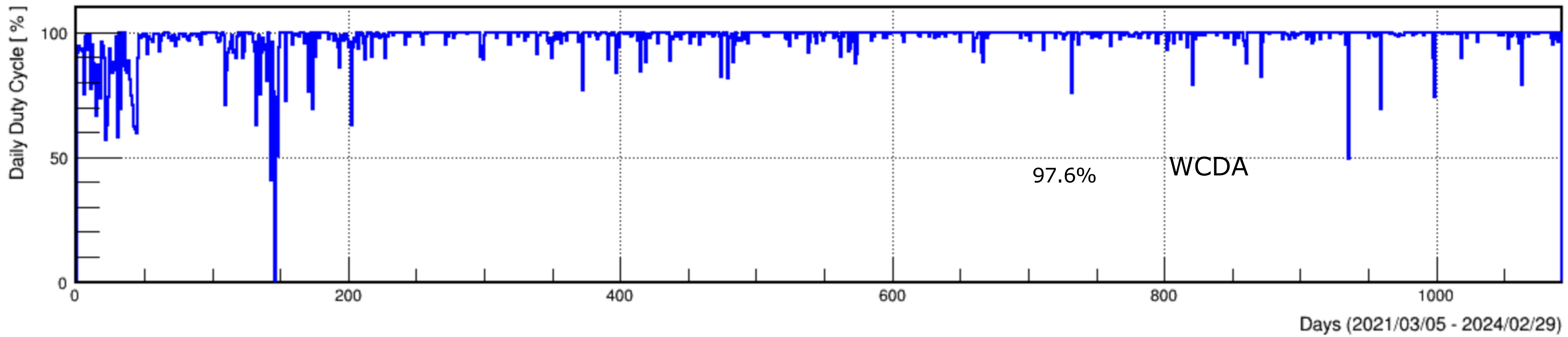




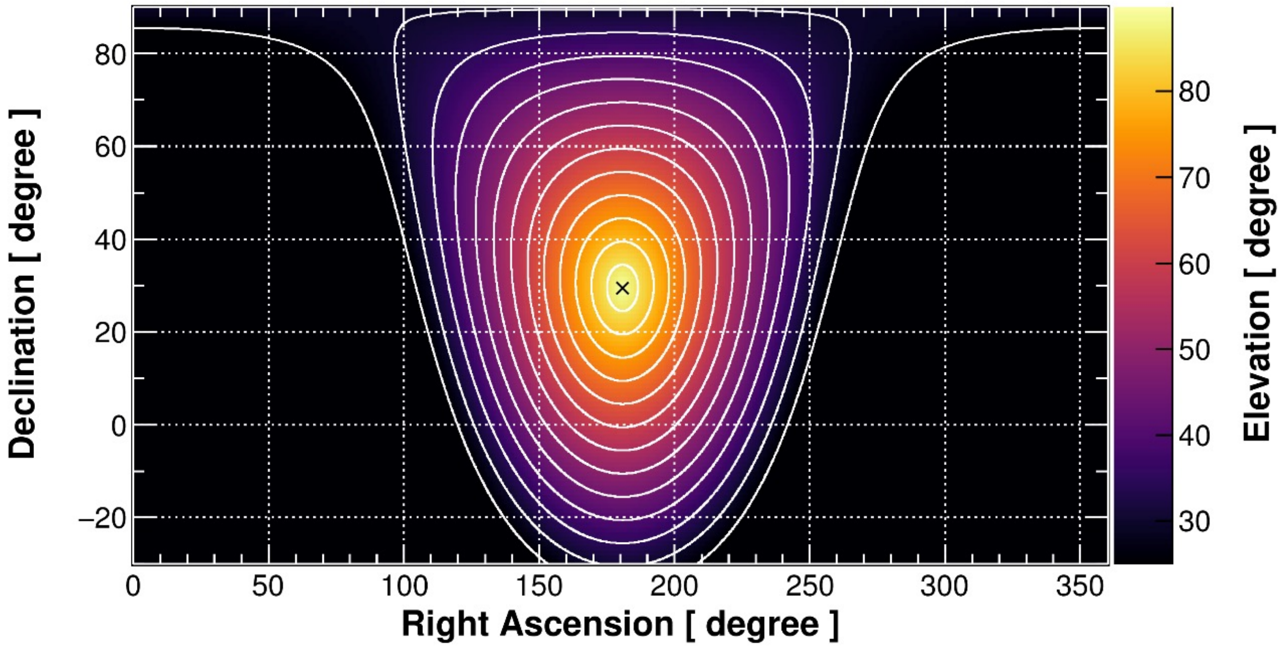
➢ 10TeV-200TeV/ 100TeV-10PeV / 10PeV-100PeV/ 100PeV-2EeV

Features: full duty cycle



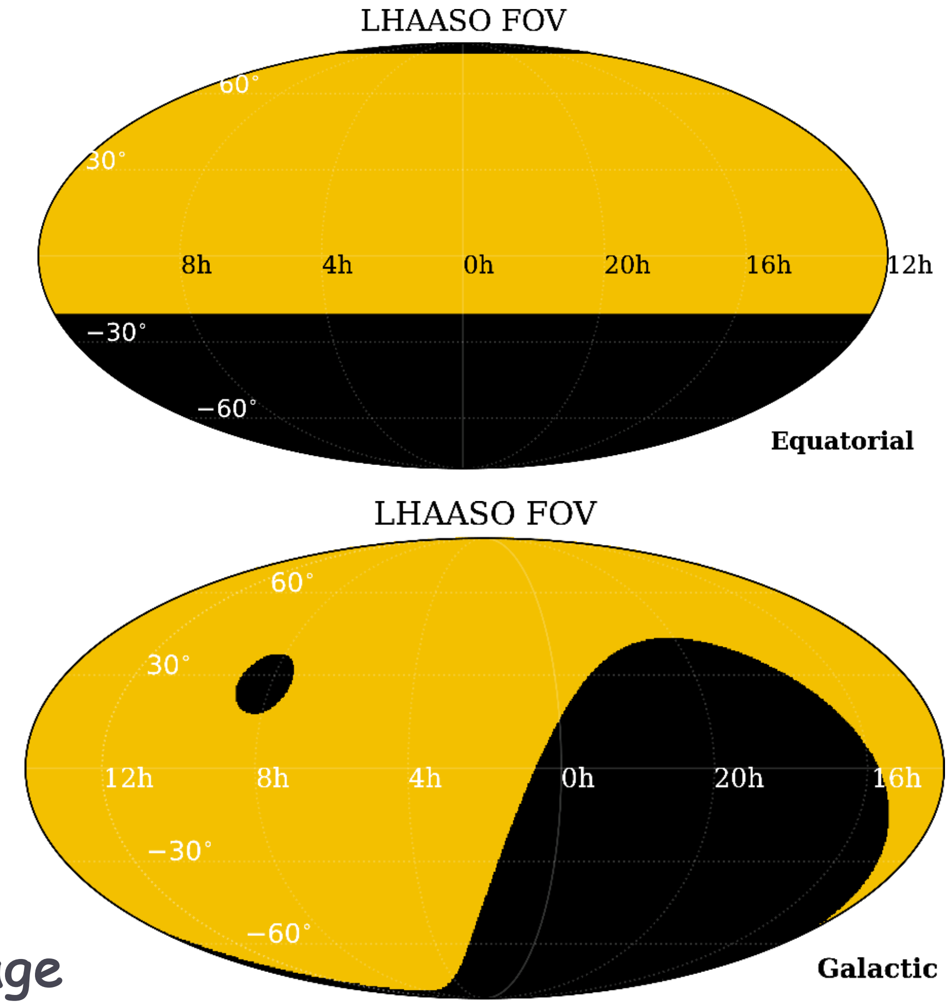


Features: wide field of view



Instant FOV

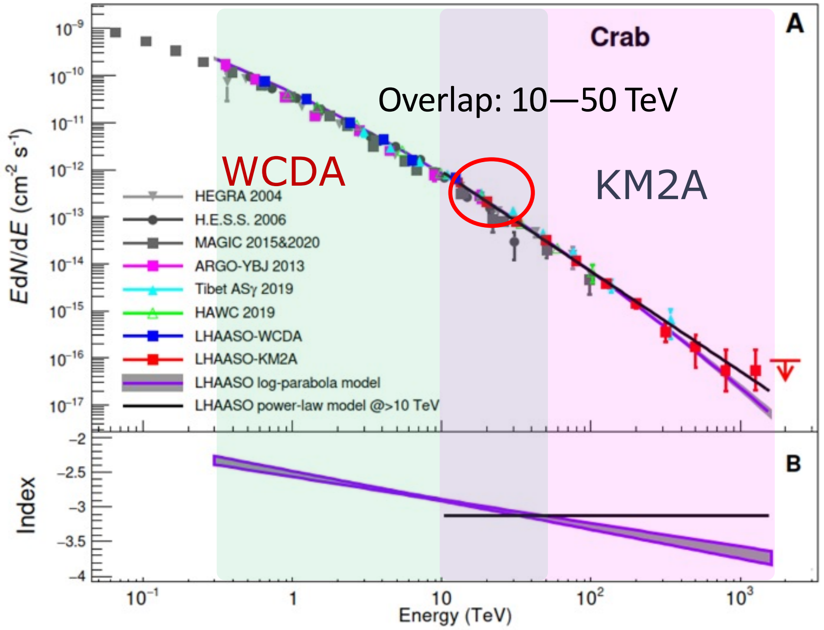
Daily/yearly FOV



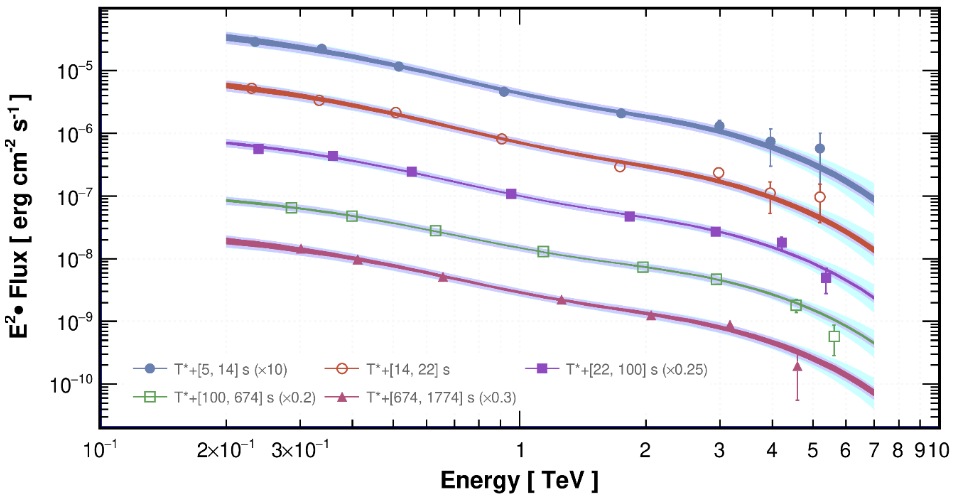
1/6 of the entire sky at any given moment.

The Earth's rotation further enables a 3/4 sky coverage

Features: wide energy range coverage



GRB 221009A

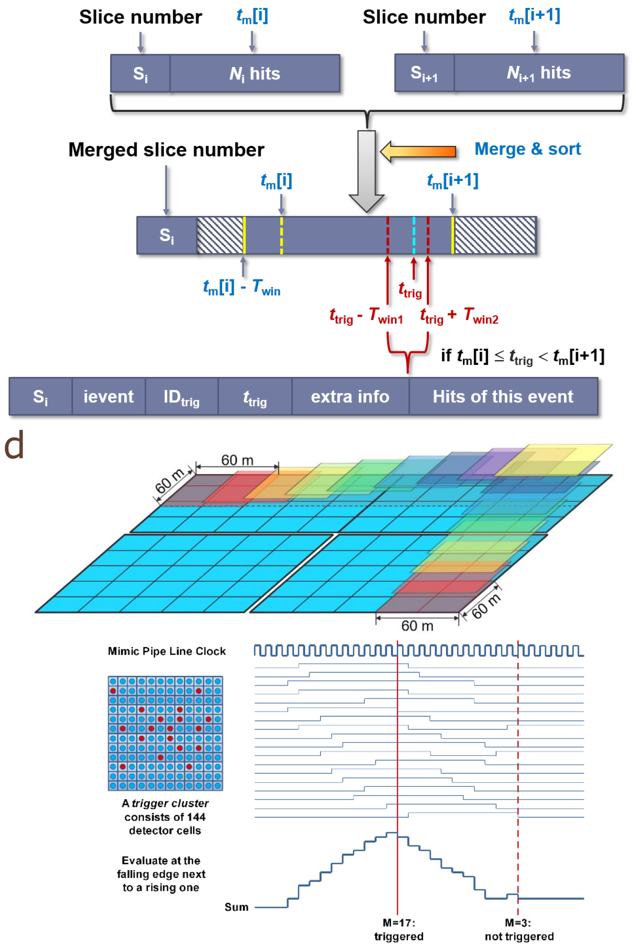


The lowest can reach < ~100 GeV?

- Covering 3.5 ~ 4 decdades of energy (200 GeV – 2 PeV)
  - Consistent with others < 100 TeV
  - Self cross-check between WCDA and KM2A; KM2A and WFCTA

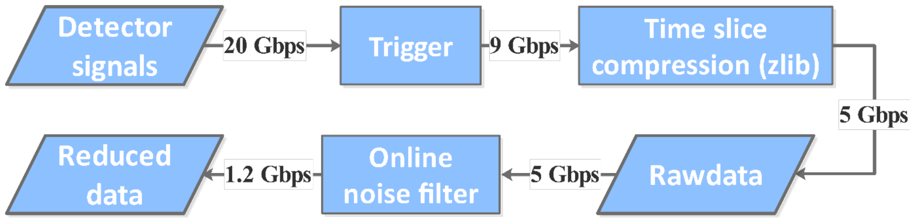
LHAASO Trigger

- Implemented on a computing cluster:
  - Soft trigger.
- Basic triggers:
  - KM2A (EDA + MDA), WCDA and WFCTA, independently;
  - 400 ns + 20 ED -> km2a
  - 250 ns + 30 DU -> WCDA
  - 3 parallel data streams;
  - for every stream, other detector hits in a time window are collected and stored.
- Special triggers:
  - Calibration;
  - For some special physics goals.
- Triggerless data:
  - Compact single counting signals (with precision lost) are cached;
  - Stored for up to 2 weeks;
  - For follow-up observations at very low energy threshold, on GRBs, Blazers, FRBs, neutrino counterparts, GW counterparts, etc.



Trigger logic of WCDA

LHAASO data volume：~12 PB/yr



KM2A原始数据：

- 触发率：2.6 kHz
- 数据量：0.20 Gbps = 2.2 TB/day = 760 TB/yr

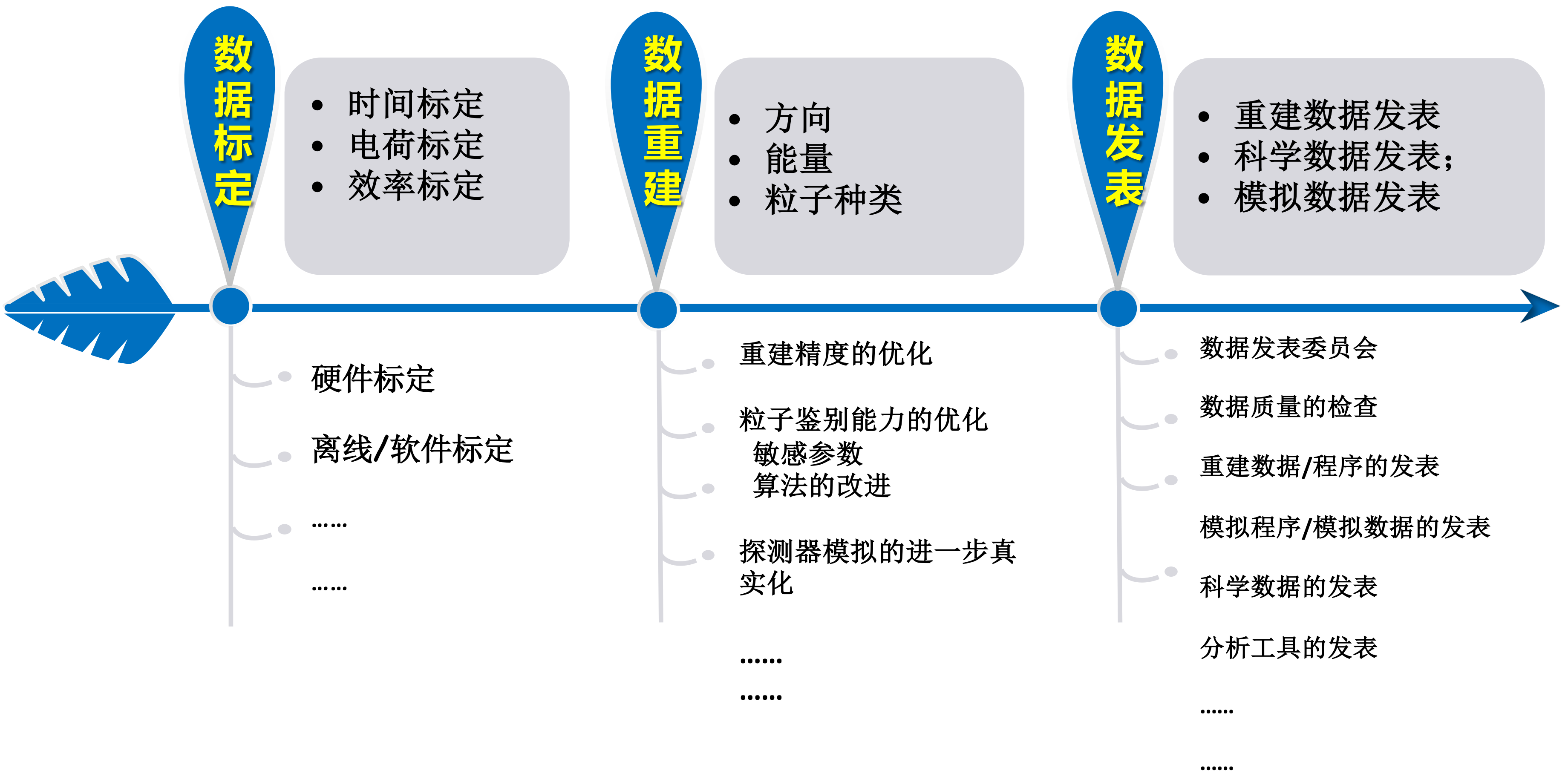
WFCTA原始数据：

- 触发率：1.1 Hz/telescope * 18 = 20 Hz
- 数据量：100 TB/yr （注意：1400 hour/yr）

WCDA原始数据：

- 触发率：34 kHz ➔ 160 kHz（降低单道阈值及触发多重度阈值）
- 数据量（噪声过滤前）：1.1 Gbps = 12 TB/day = 4.4 PB/yr ➔ 3.9 Gbps = 42 TB/day = 15 PB/yr
- 数据量（噪声过滤后）：0.42 Gbps = 4.5 TB/day = 1.6 PB/yr ➔ 1.2 Gbps = 12 TB/day = 4.3 PB/yr
- GRB数据（~3 triggers/week，LAT GCN only）：8.7 TB/burst = 1.3 PB/yr ➔ 30 TB/burst = 4.6 PB/yr

Pipeline of data production



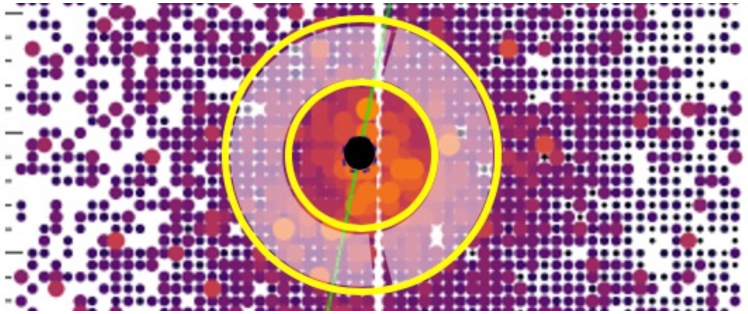
Calibration @ WCDA

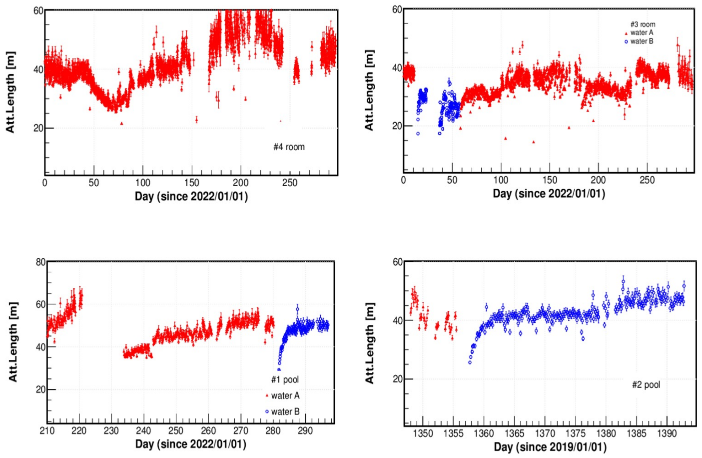
特色和难点:

- 电荷标定:
  - 4种不同类型PMT，每个PMT又分阳极（高增益）和打拿极（低增益）需要把8种信号归为一种
- 时间标定:
  - 探测器存在明显的Q-T（电荷与时间）关系，而且还包含R-T（芯距与时间）甚至与簇射方向关联，修正极其复杂
- 还有3个水池间的时间与电荷标定
- 水质及污染物等原因造成的探测效率的变化
- 探测器的复杂多变，需要定期或实时标定

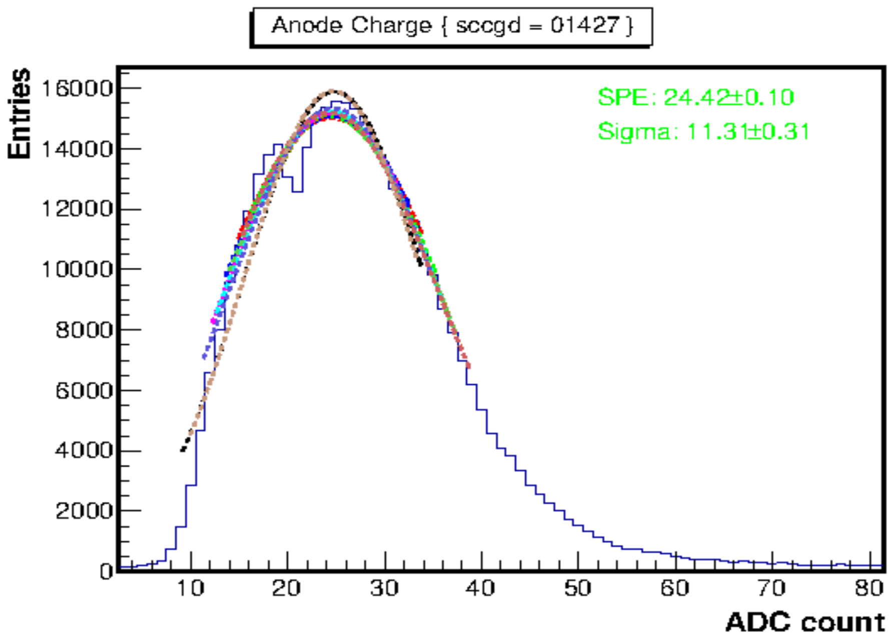
解决方案:

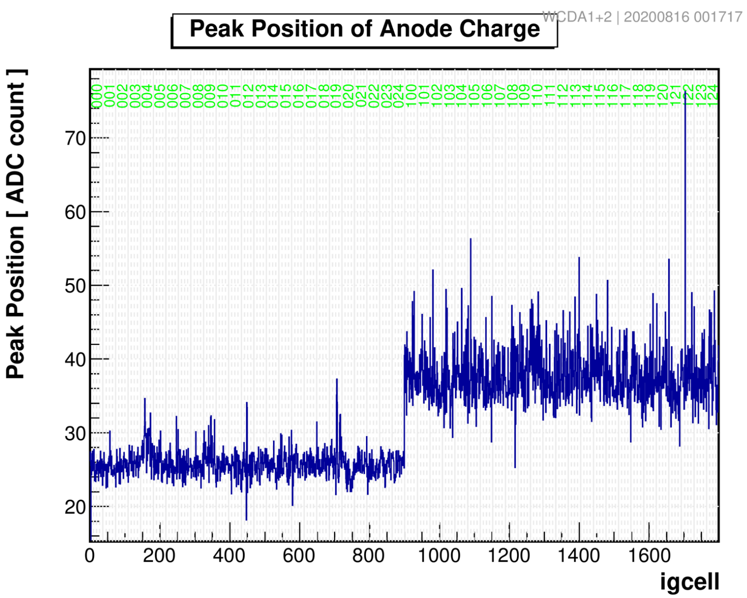
- 电荷标定：采用簇射信号，采用迭代拟合的方式；每次标定需要采用4天以上的数据；已经实现自动数据处理，得到标定结果。
- 时间标定：基于硬件标定，采用天量级的簇射事例完成时间偏差、Q-T、R-T的修正参数的计算
- 水池间标定：采用复杂算法，采用簇射事例的对称性，实现了每天一次的水池间的标定
- 根据簇射信号的多峰结构进行标定，并提出了CRS的方法，实现了不同单元（共3120）间的效率的实时（天量级）标定

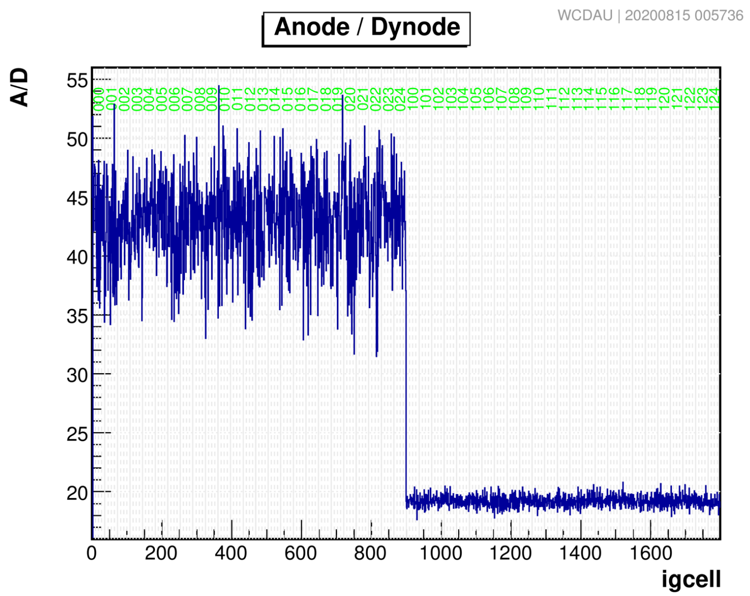


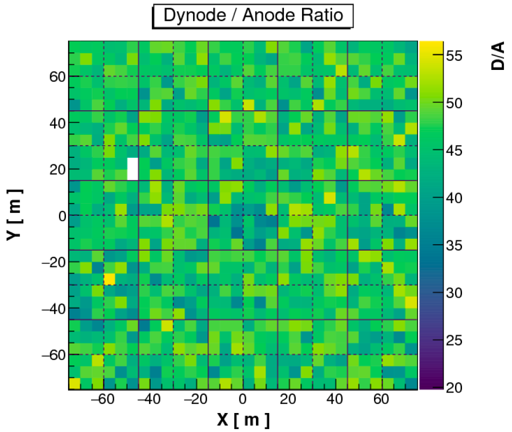


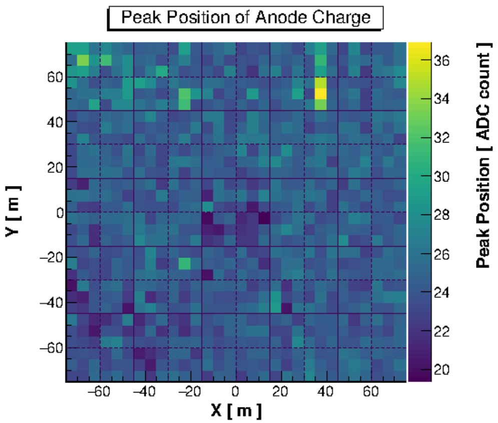
Charge calibration: SPE + AD ratio

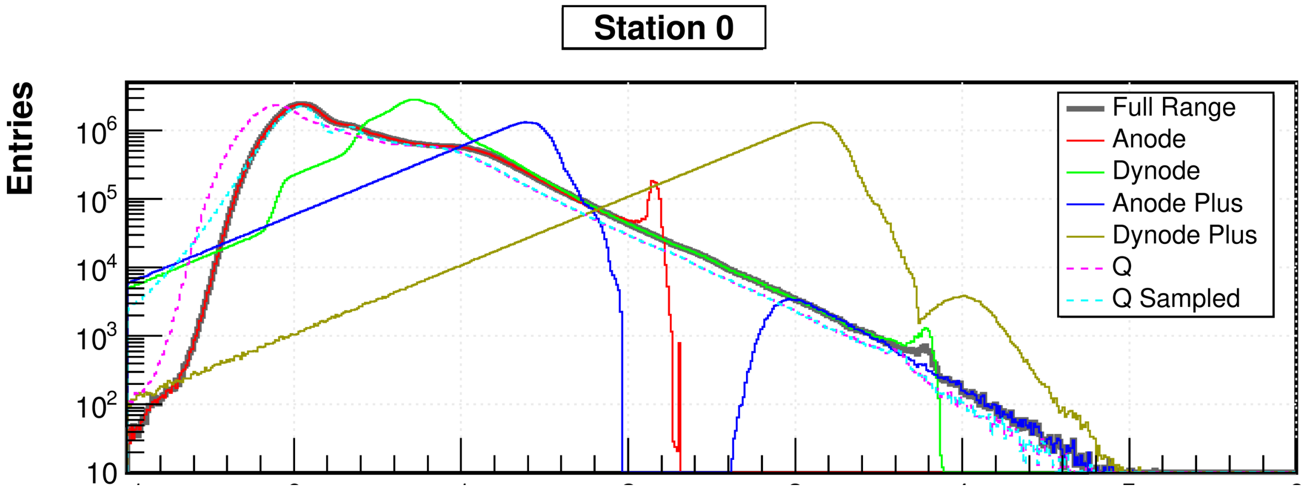




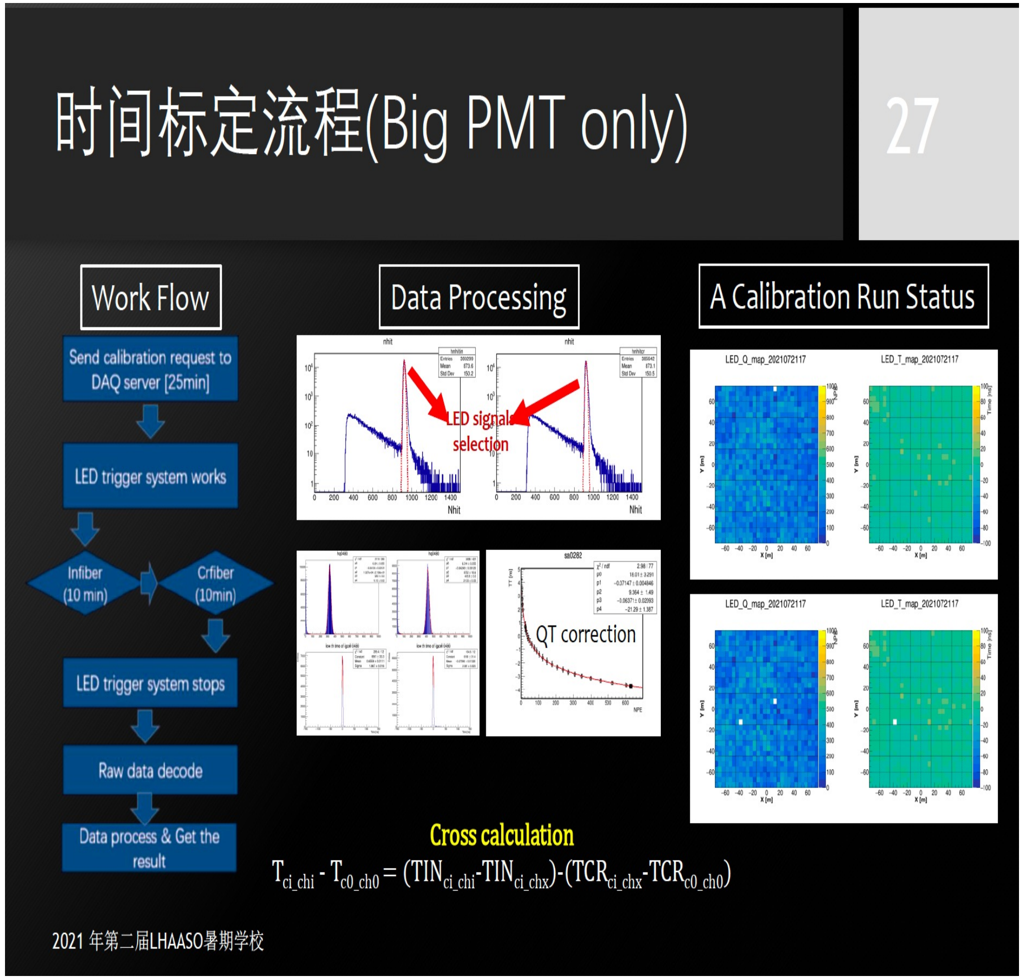


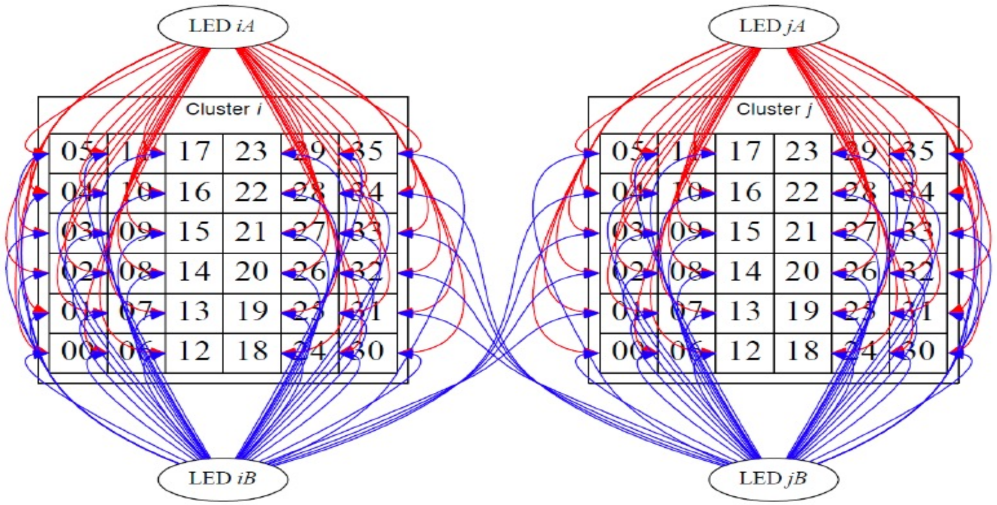


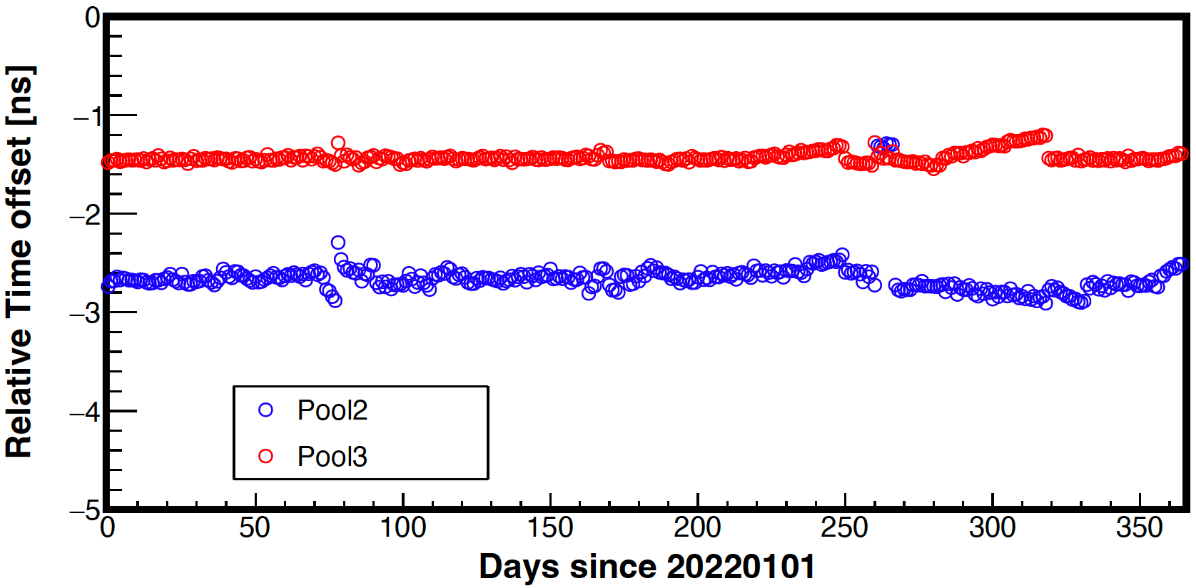




Time calibration @ WCDA



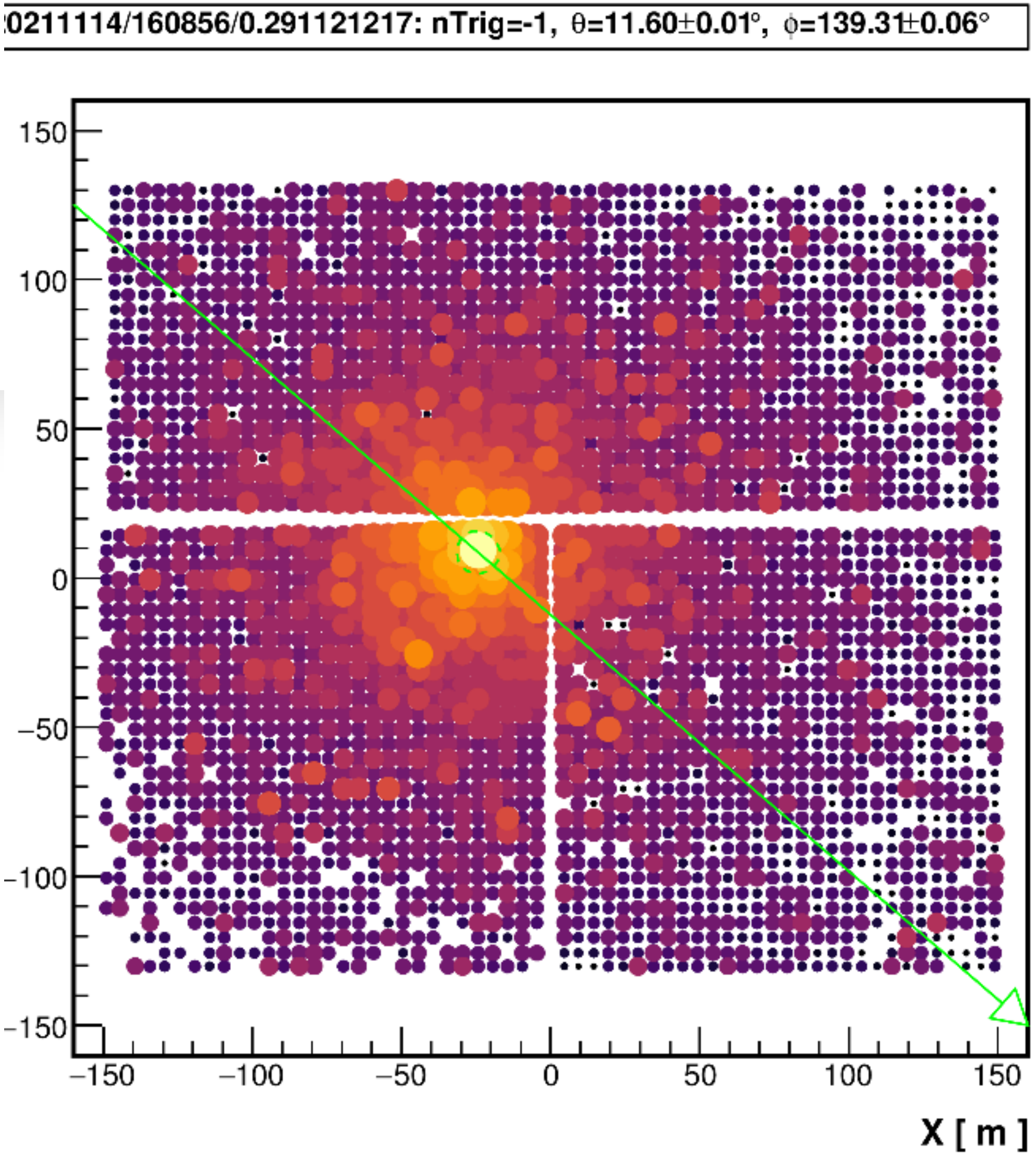




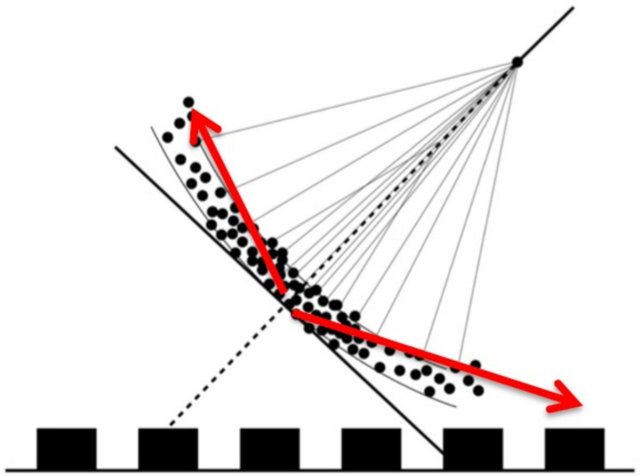
PMT相应差异，TDC测量精度的差异, 信号大小的差异 ....

Shower reconstruction

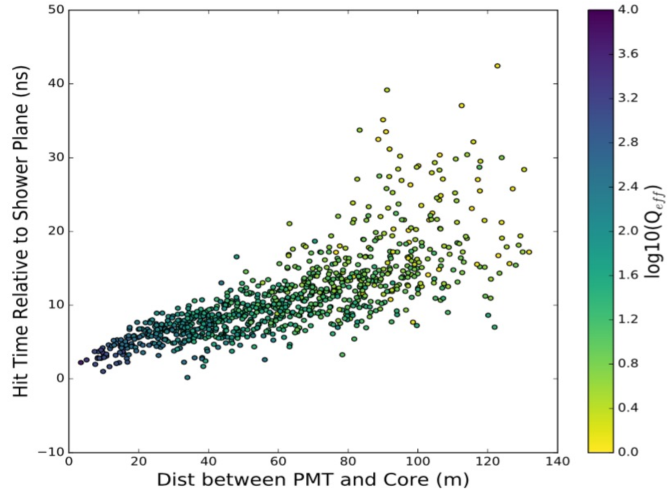
- Shower geometry reconstruction
  - direction + shower core
  - Npe, Np, Ti @ each detector unit
- Shower energy reconstruction
  - Lateral or longitudinal distribution of Shower
- Primary particle identification
  - Mass sensitive parameters → Nmuon



Classic way to reconstruct the direction



- 簇射前锋面到达阵列时，
- 第i个fired PMT 的坐标为(x_i, y_i, t_i)
- 未知参量(L, M, T_0) $L=\sin\theta\cos\varphi$, $M=\sin\theta\sin\varphi$,

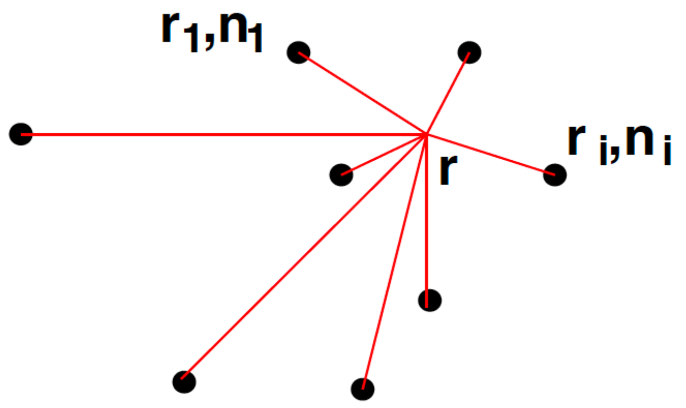


Direction reconstruction: 前锋面拟合

Plana fitting: $$\chi^2 = \sum_i w_i(c\cdot(t_i - T_0) - x_i \cdot L - y_i \cdot M)^2$$

Conical correction: $$\chi^2 = \sum_i w_i(c\cdot(t_i - T_0) - x_i \cdot L - y_i \cdot M - c\cdot(\alpha R_i))^2$$

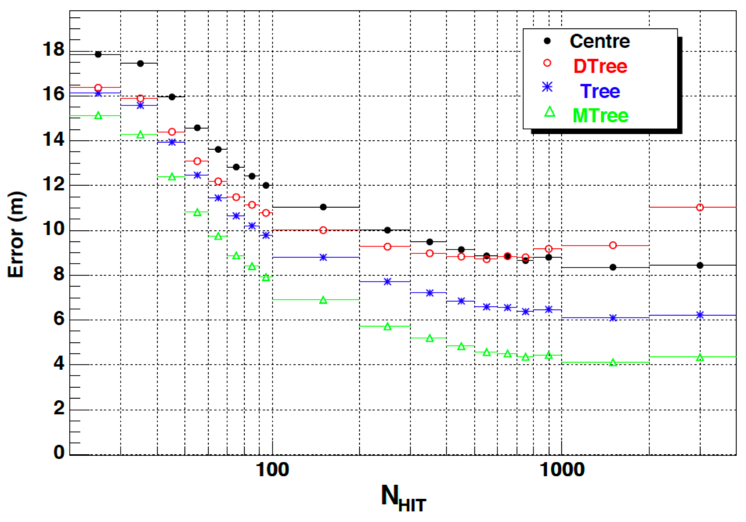
Classic way to reconstruct the core position

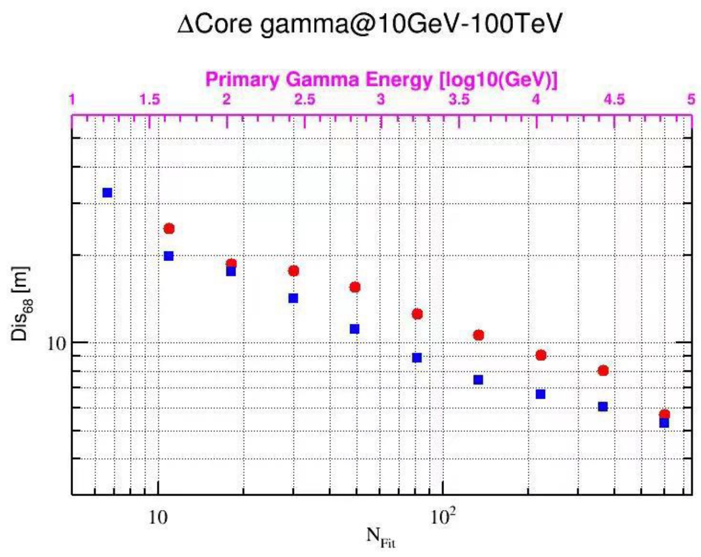


- Center of Gravity(COG)

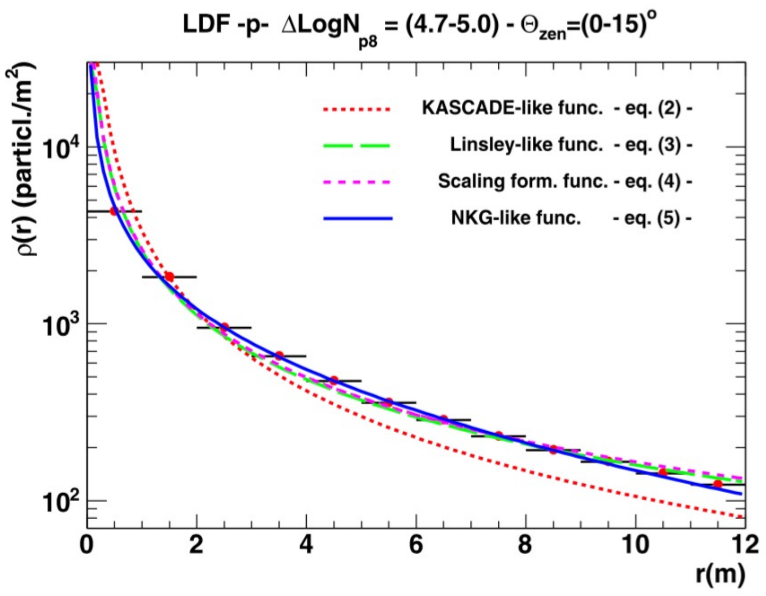
$$(X_c, Y_c) = \frac{\sum_{i=1}^{N} (x_i, y_i)\, n_i}{\sum_{i=1}^{N} n_i}$$

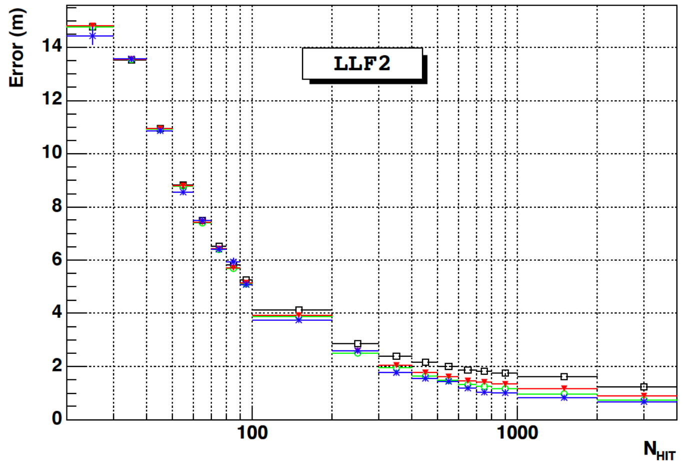
- Tree length algarithm





Shower Core reconstruction





- COG is initial seed;
- NKG function is analytical function, in principle it is closely related with direction.(x_c, y_c, θ, φ)

$$\rho_2(r) = N_e C(s)\left(\frac{r}{r_0}\right)^{s-\alpha}\left(1+\frac{r}{r_0}\right)^{s-\beta}$$

- different experiments use different NKG-like or nkg-modified functions;
- AGASA

$$\rho_4(r) = \frac{N_e}{r_0^2} C\left(\frac{r}{r_0}\right)^{-\alpha}\left(1+\frac{r}{r_0}\right)^{-(\beta-\alpha)}\left[1+\left(\frac{r}{10r_0}\right)^2\right]^{-\delta}$$

- AGRO-YBJ BigPad data

$$\rho(r) = A\left(\frac{r}{r_0}\right)^{s'-2}\left(1+\frac{r}{r_0}\right)^{s'-4.5}$$

- Likelihood algorithm

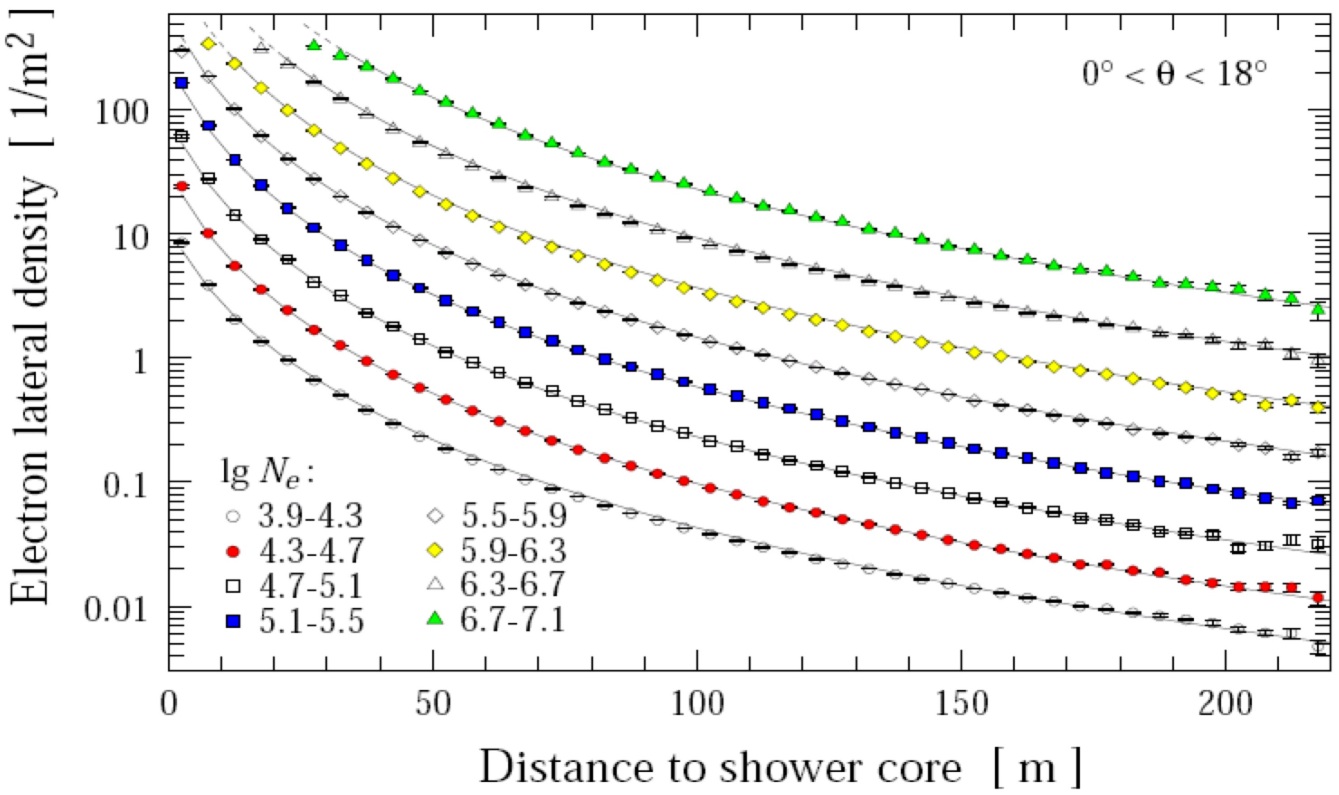
$$LF2 = \prod_{k=1}^{N_S} p_k(m_k)$$

Classic way to reconstruct the shower @ global fitting

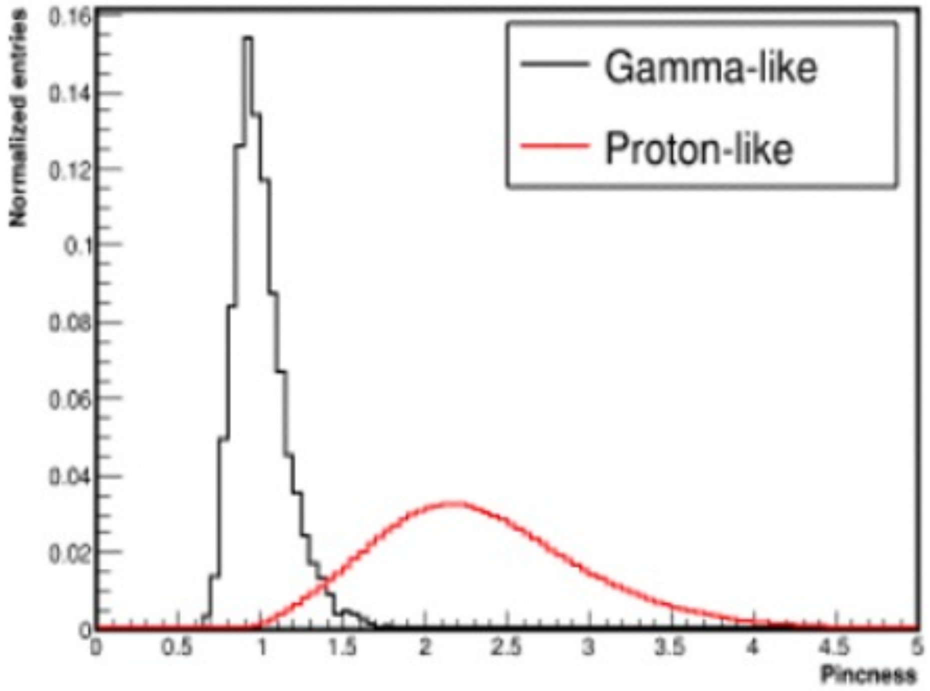
Lateral distribution(global fitting)

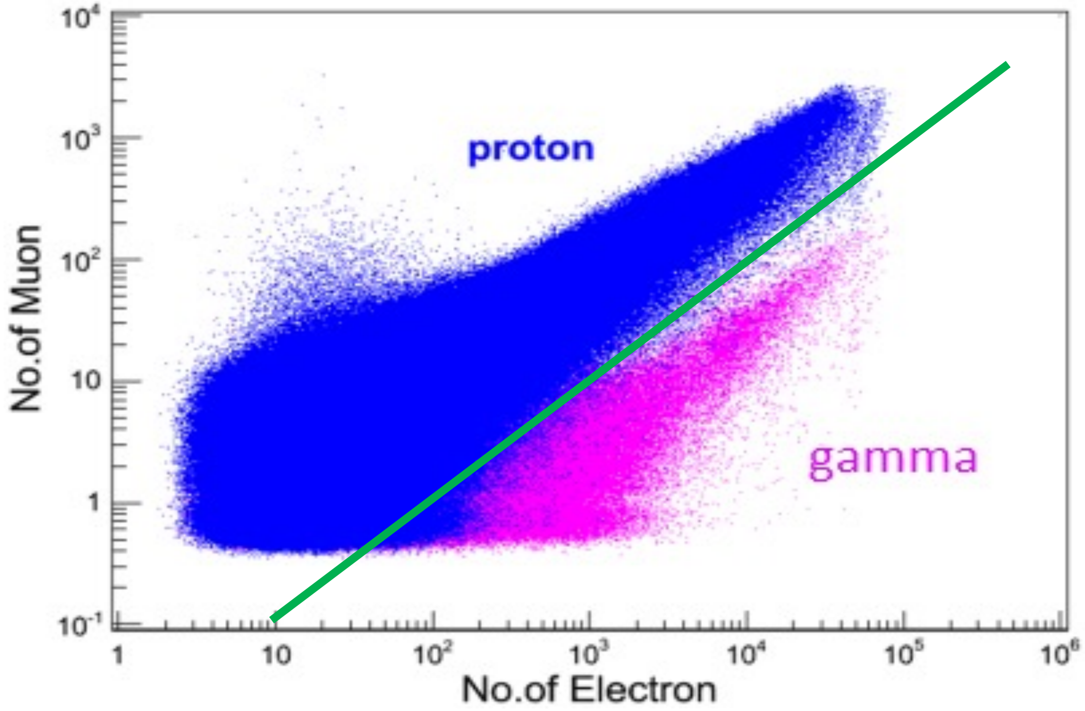
* To fit (xc, yc, theta, phi, Ne, rm, s)

$$\rho(r)=Ne.A(\frac{r}{rm})^{s-2}(1+\frac{r}{rm})^{s-4.5}$$

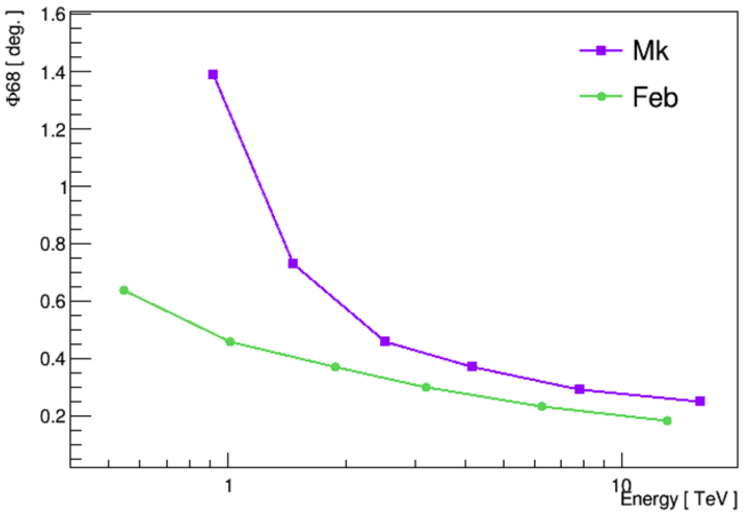


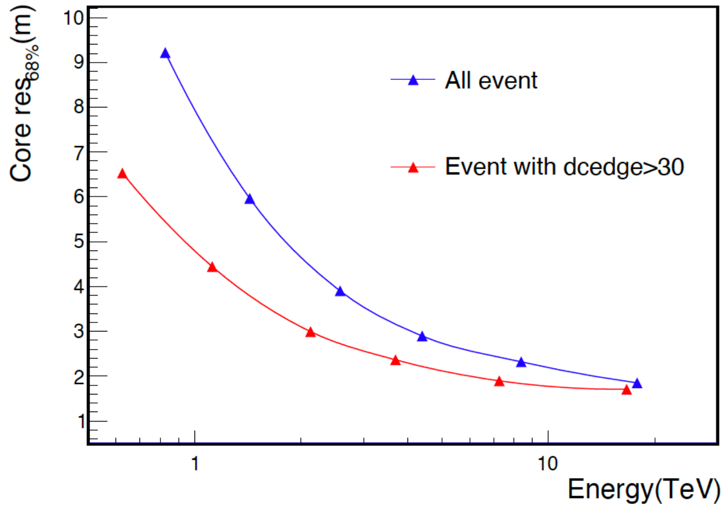
G/P separation

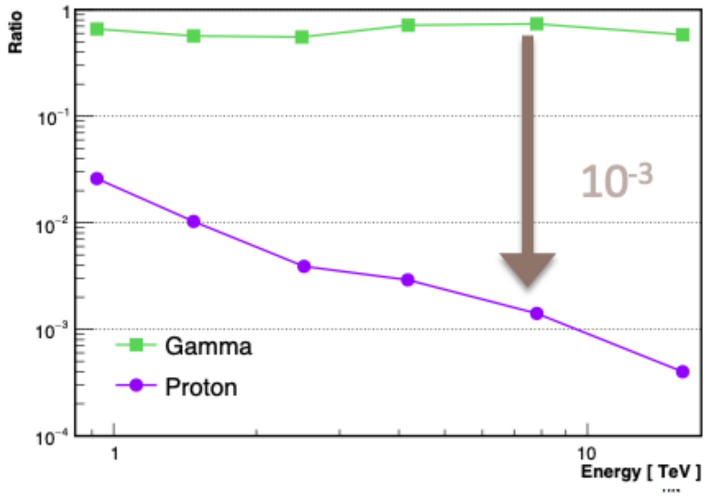


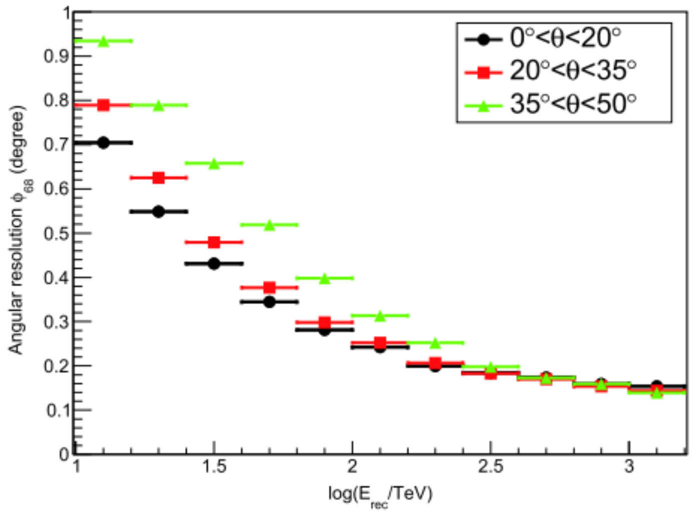


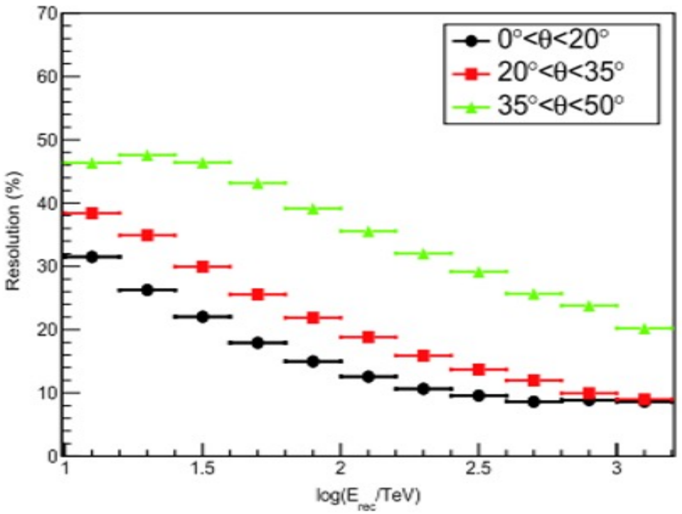
Shower reconstruction resolution

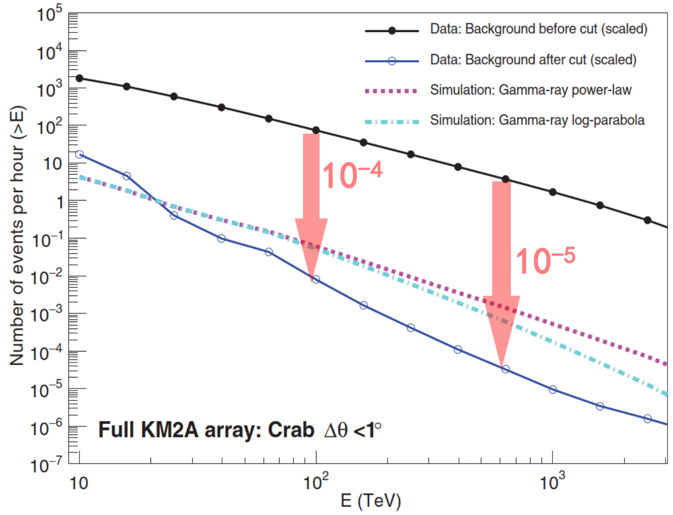






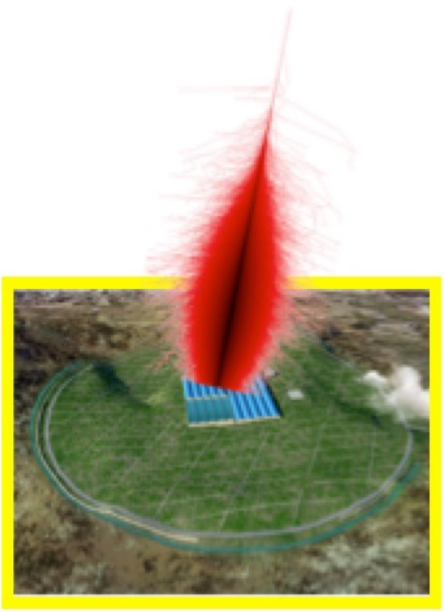






Ground-based Air Shower Array

Data production Board



CORSIKA / AIRES

QGSJET / EPOS / SIBYLL/…….

Detector response simulation

Data

MC samples

Reconstruction / Analysis Chain

Comparison MC .vs Data

Air shower simulation: CORSIKA
Output:
KM2A :shower secondary particles
WCDA : shower secondary particles
WFCTA: shower secondary particles
Cherenkov light

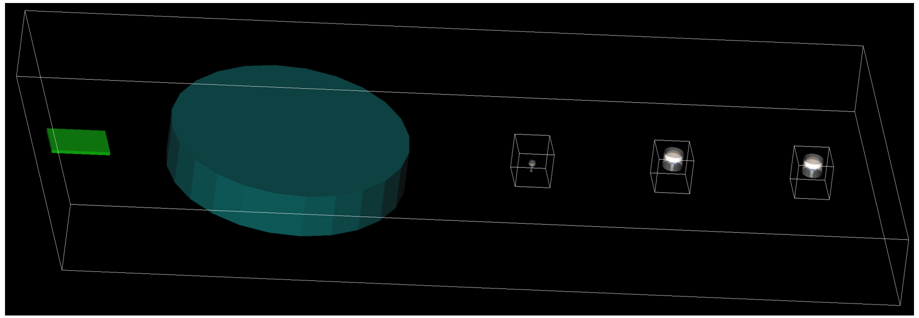
Geant4 based detectors simulation package
G4KM2A:
G4WCDA: two steps processing
WFCTA: c++ based ray tracing

Physics Data production
Level 1:
code + algorithm developer
Level 2:
user I/O to physics result

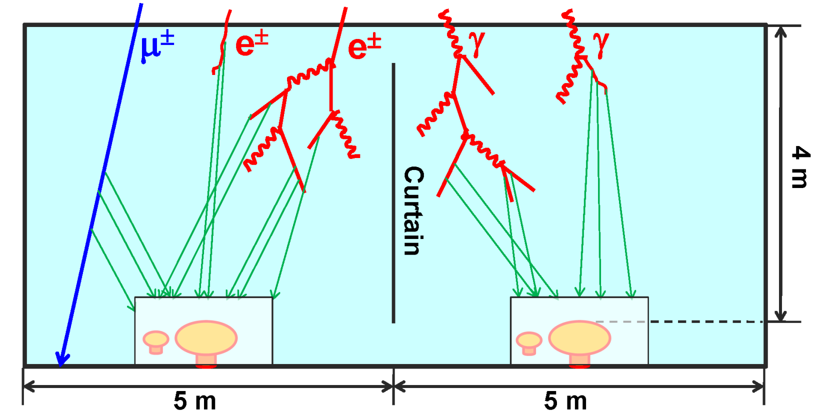
LHAASO MC simulation

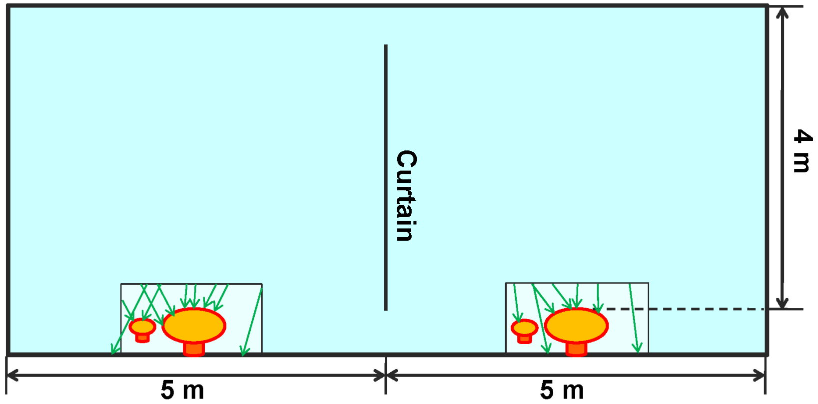
特色和难点

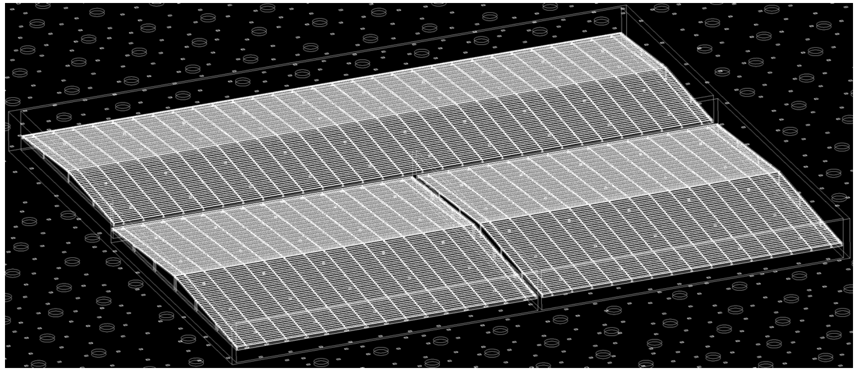
- 空气簇射的模拟：CORSIKA/COSMOS/AIRES
  - 模拟样本多
  - 伽马、质子到铁核5组或56种元素
    - Multiple samples: Crab orbit and isotropic samples.
  - 多种强相互作用模型的结合)
    - QGSJET， EPOS-LHC， SIBYLL， GHEISHA， FLUKA
  - 能量范围宽广（10 GeV - 10 PeV)
- 探测器模拟（GEANT4 为基础)
  - 切伦科夫光子数巨大，内存消耗量大、模拟缓慢
  - WCDA实验大厅结构复杂，并存在结合KM2A（包括ED和MD）探测器模拟的必要
  - 探测器存在若干不确定的参数（多变的水质、国际首次使用的20-cin PMT等)
  - IO @ ED, MD detector unit

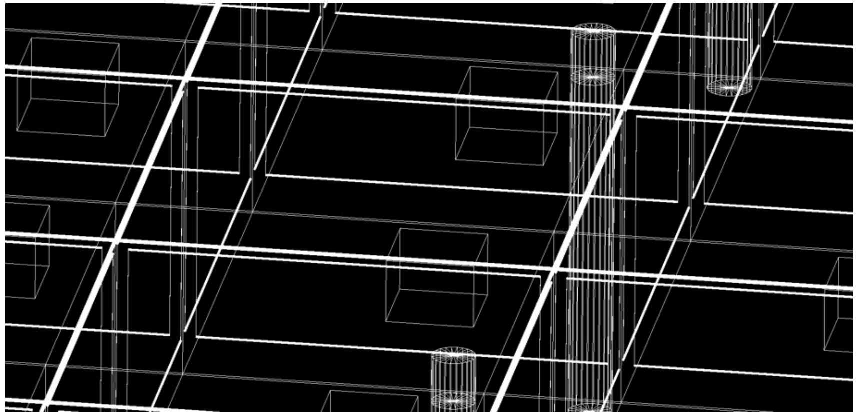


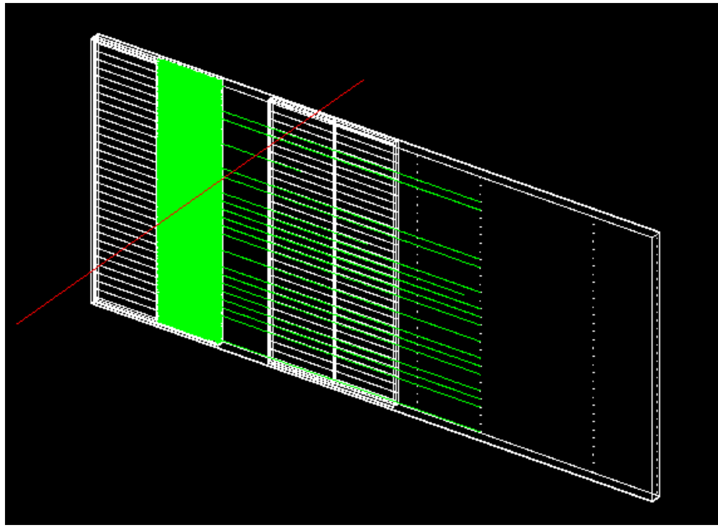
LHAASO MC status

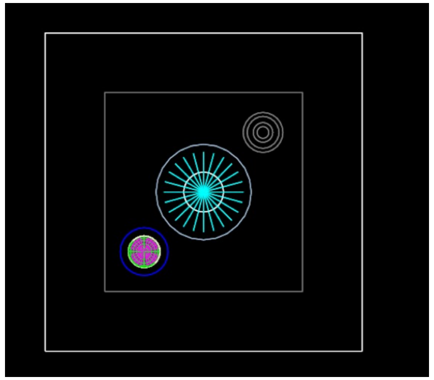


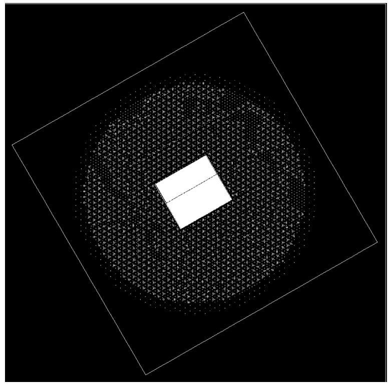


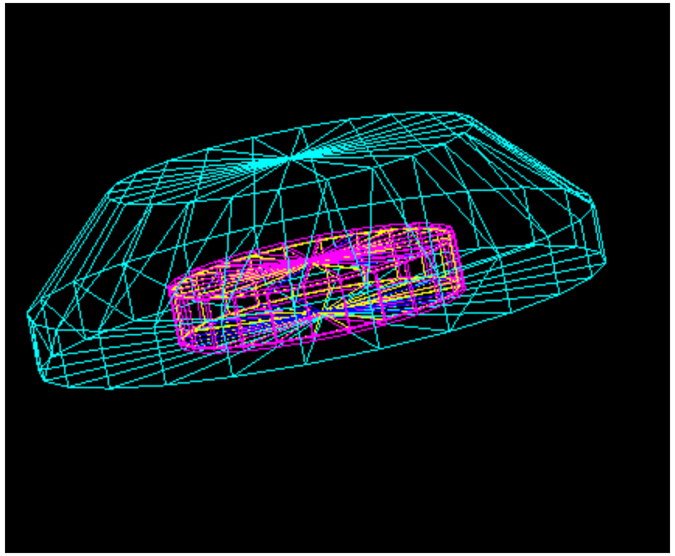






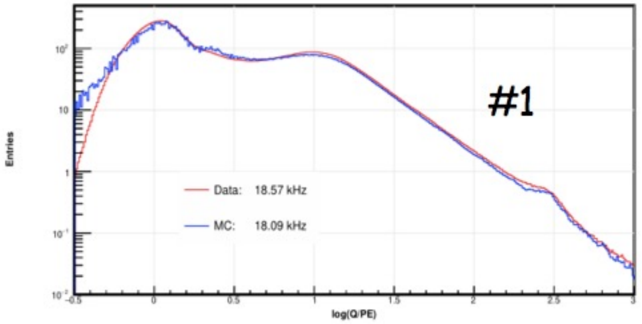


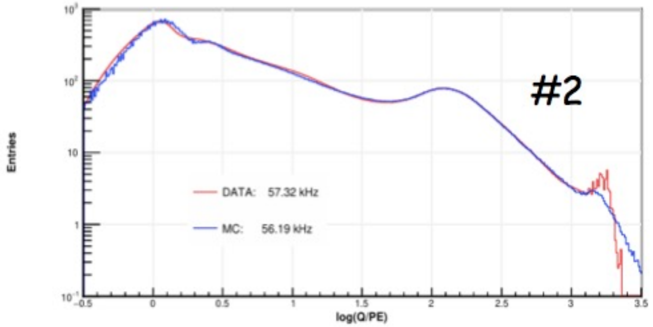


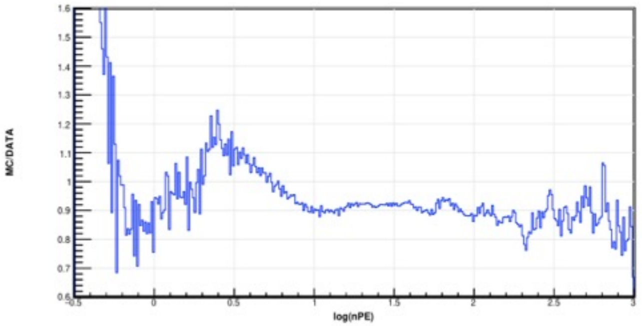


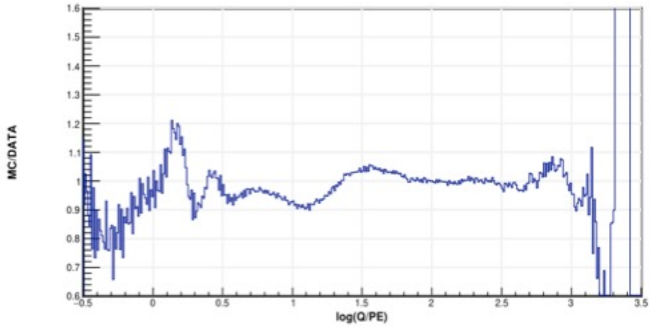
- 解决了内存耗尽
- 优化中间结果的存储
- 易于探测器真实化
- 简化各类探测器的统一模拟。
- Example: version >2.0

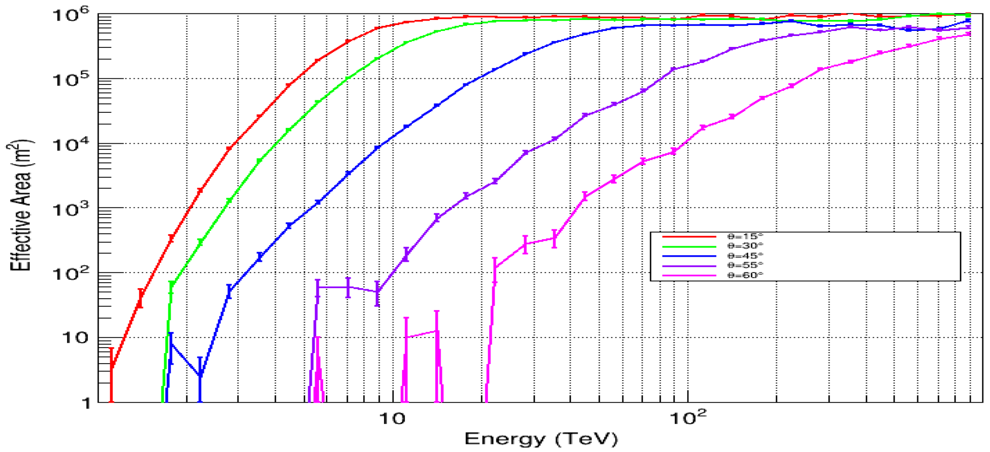
LHAASO MC status

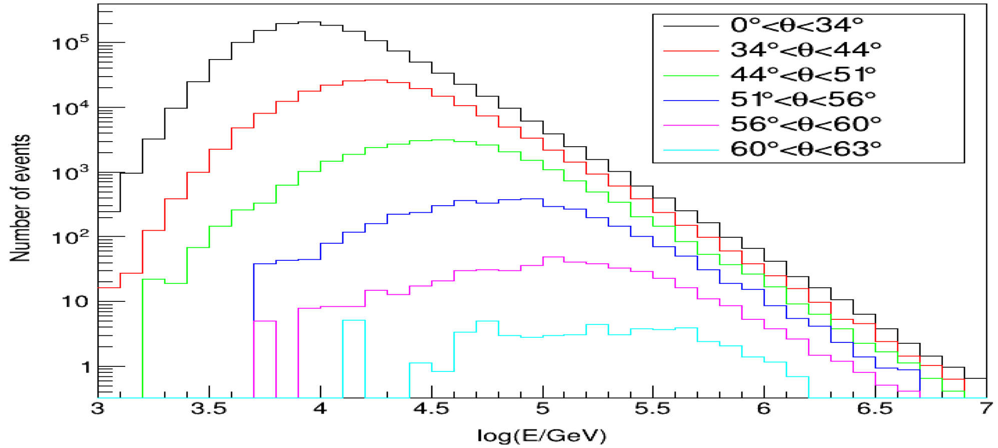












WCDA Data Production

Releasing working version: Mk

Releasing directory @ /publish/

1:progs/ 2:Mk/ 3:goodlist/ 4:Simulation/ 5:Skymap/

Reconstruction and Simulation programme @ progs/

- Reconstruction: Mk/ + test/test.sh
- Simulation: g4wcda/8.02run + test1.sh && test2.sh

Three physics data products in root format @ Mk/

- yyyy/mmdd → 2023/0101
- Readme.wcda → details about root elements
- Recdata/ -→ Standard reconstruction data 450 G/day
- Recgdata/ → Gamma-like reconstruction data 1.6 G/day
- Sampdata/ → specific sample data around the sources(crab) 100 G/day

File-list about Data quality Check @ goodlist/

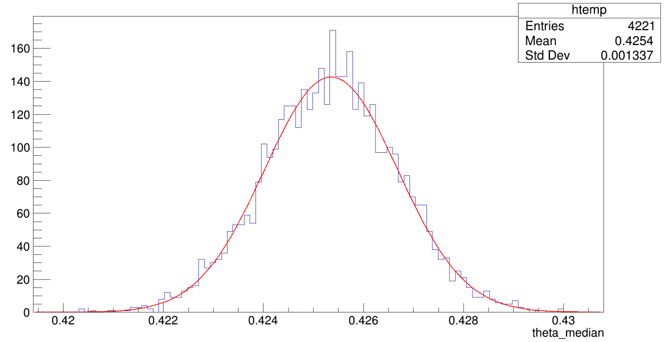
- Txt format: yyyy/mmdd.dat → 2023/0101.dat

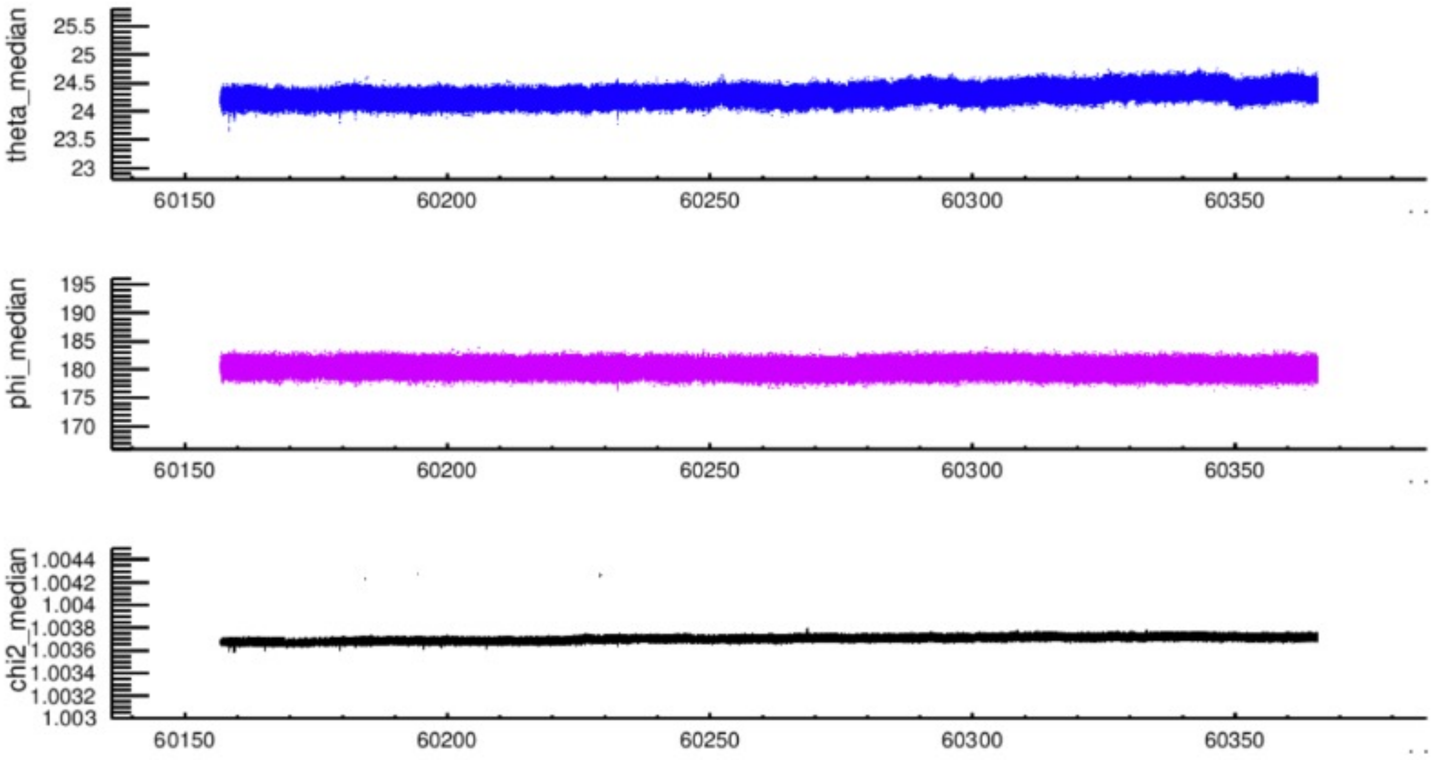
Two scientific data products in root format @

- One skymap data in root format @ skymap
- One simulation samples in root format @ simulation/
  - MC1 is for 20210305-20220930
  - MC2 is for 20210305 – 20240131

Data Quality Monitoring

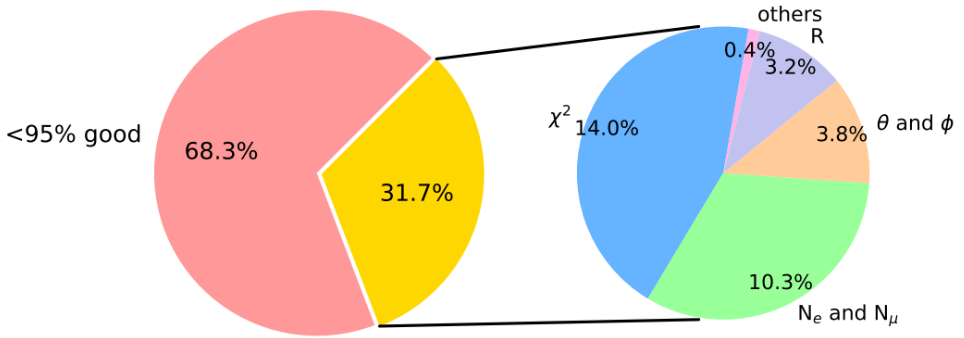
- Done by monitoring some parameters related with the daily stability of detector running and reconstruction;
- t_{live}, n_{hit}, θ, φ, x_c, y_c, χ^2 @ Nq05t30>150
- Over 5 sigma file is marked as bad file;
- On average around 3% file is marked as bad file.

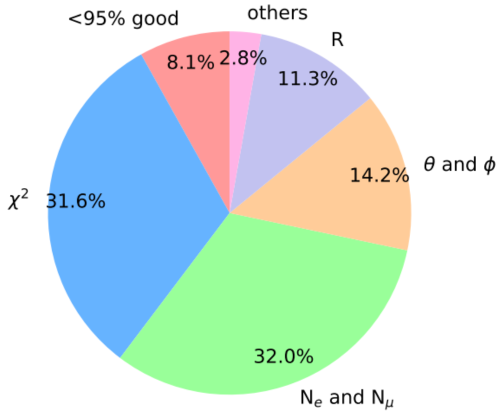


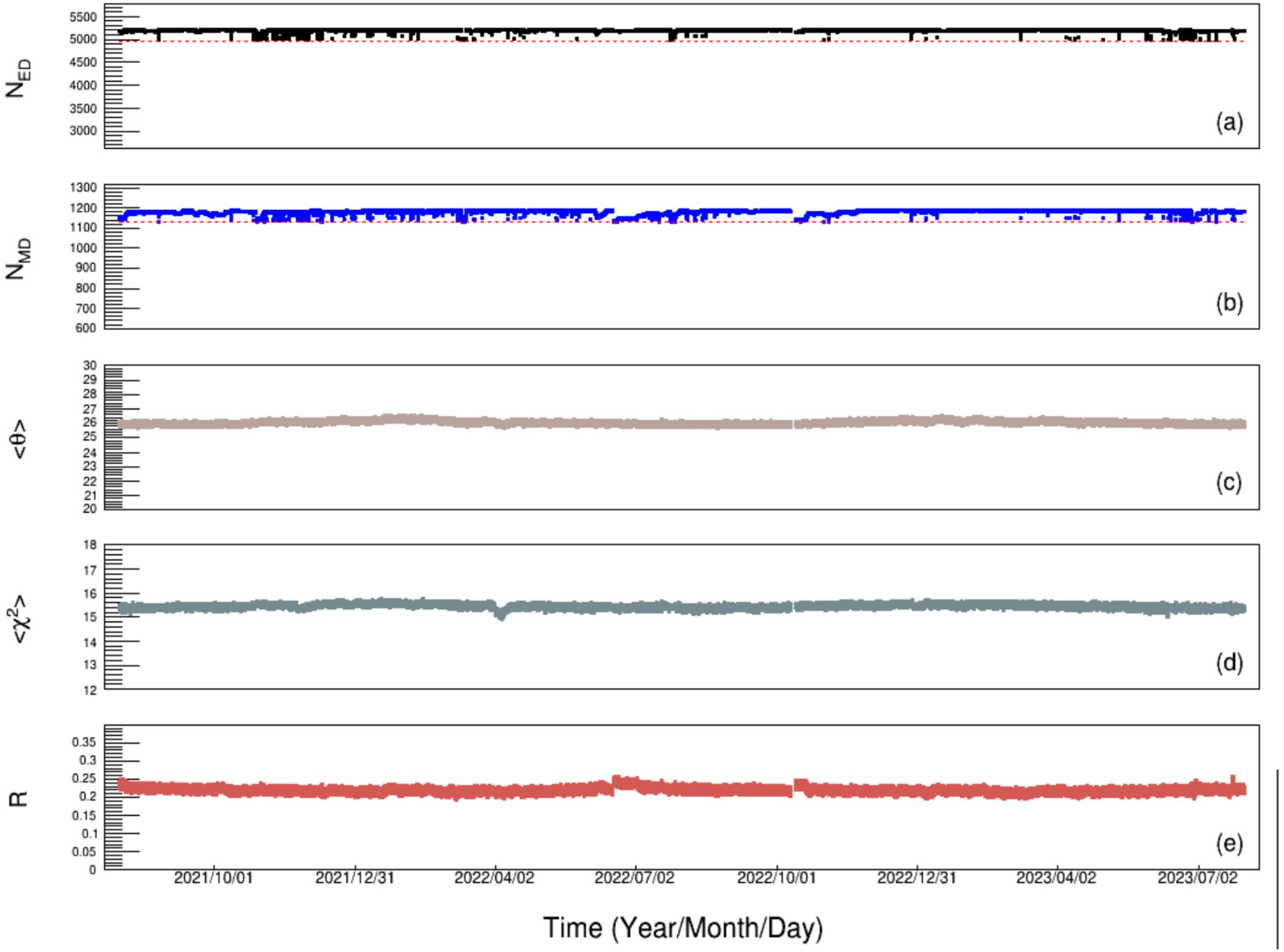


事例重建后：重建数据质量筛选

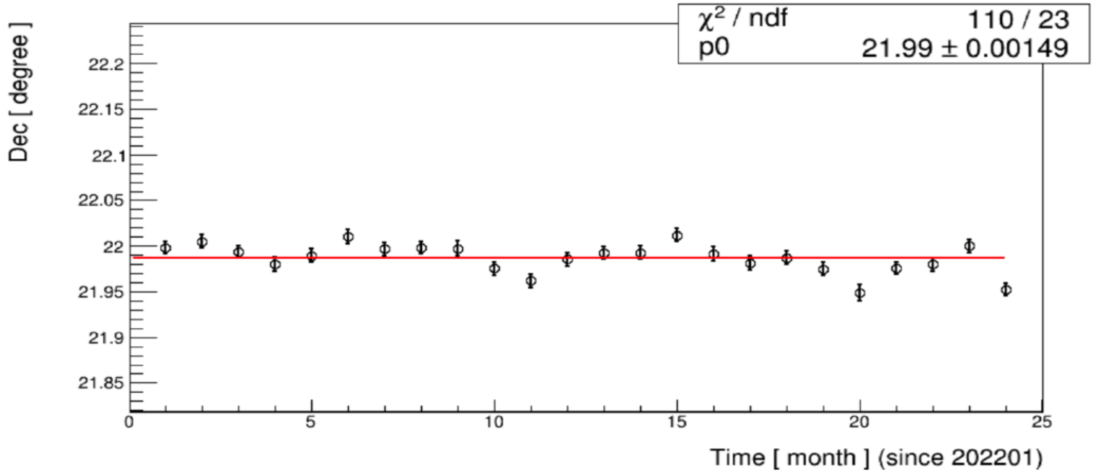
剔除1.77%数据

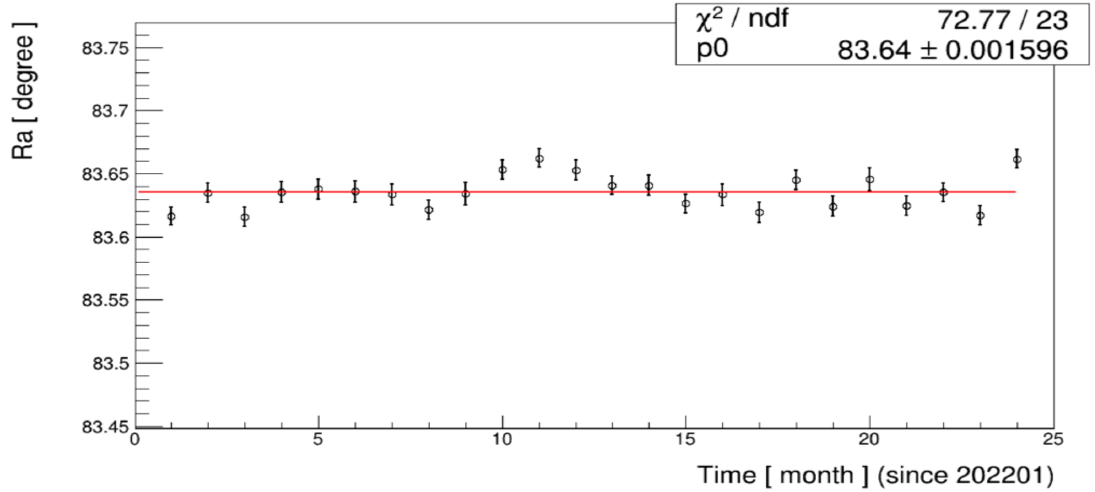


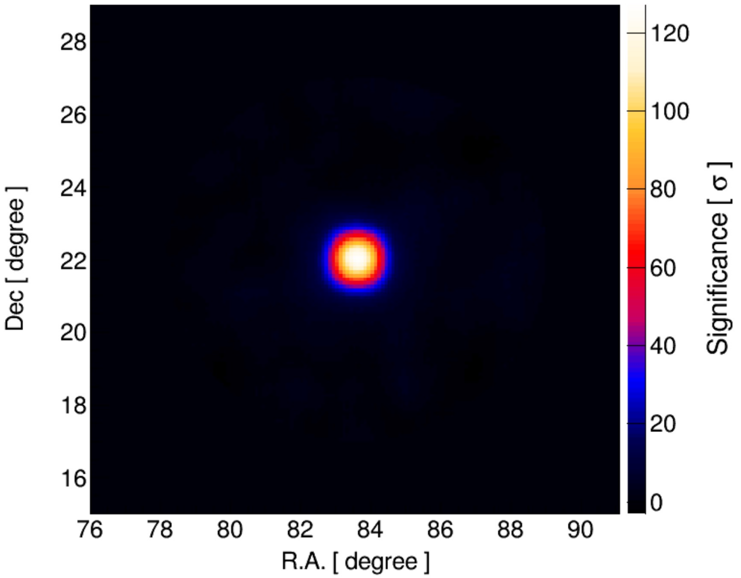


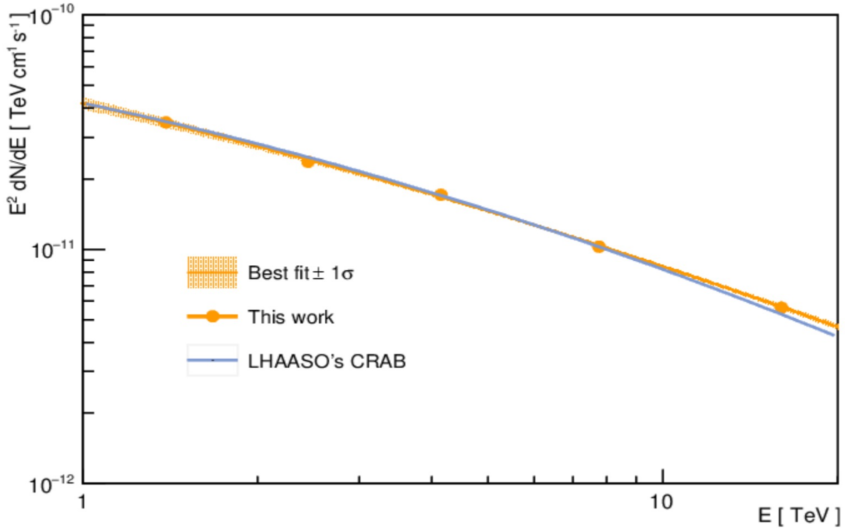


Crab Nebula monitoring @ N_{hit}>100



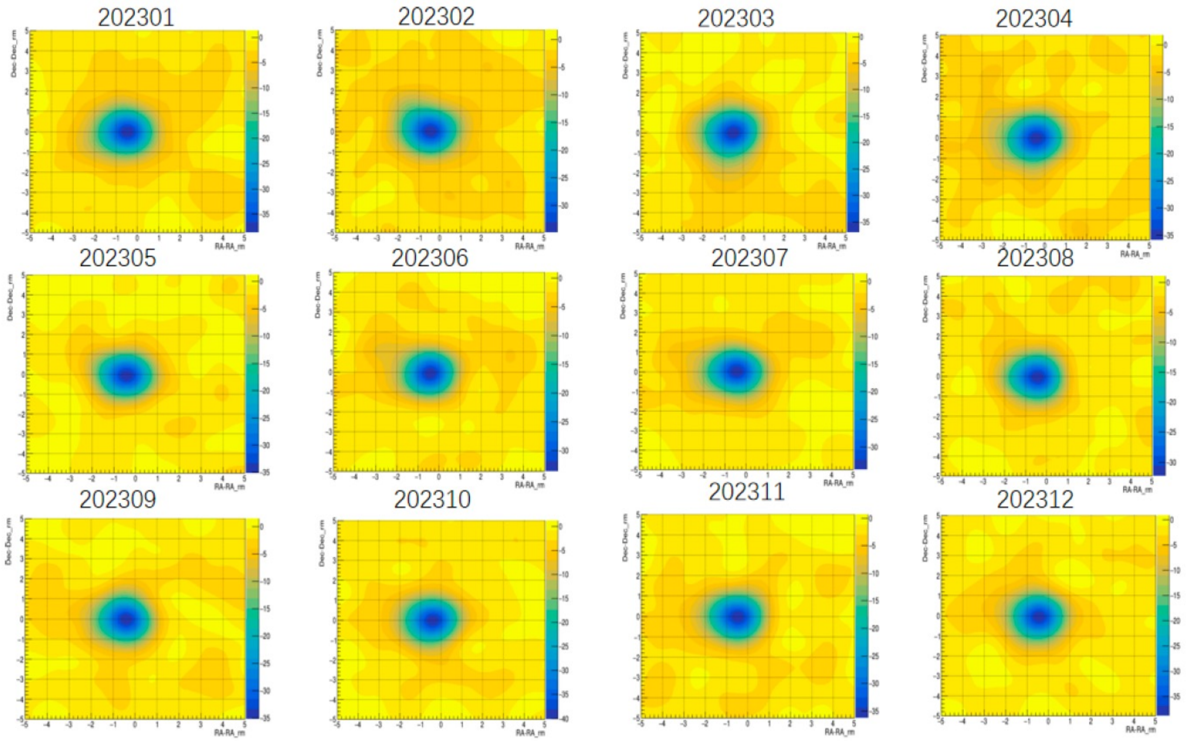


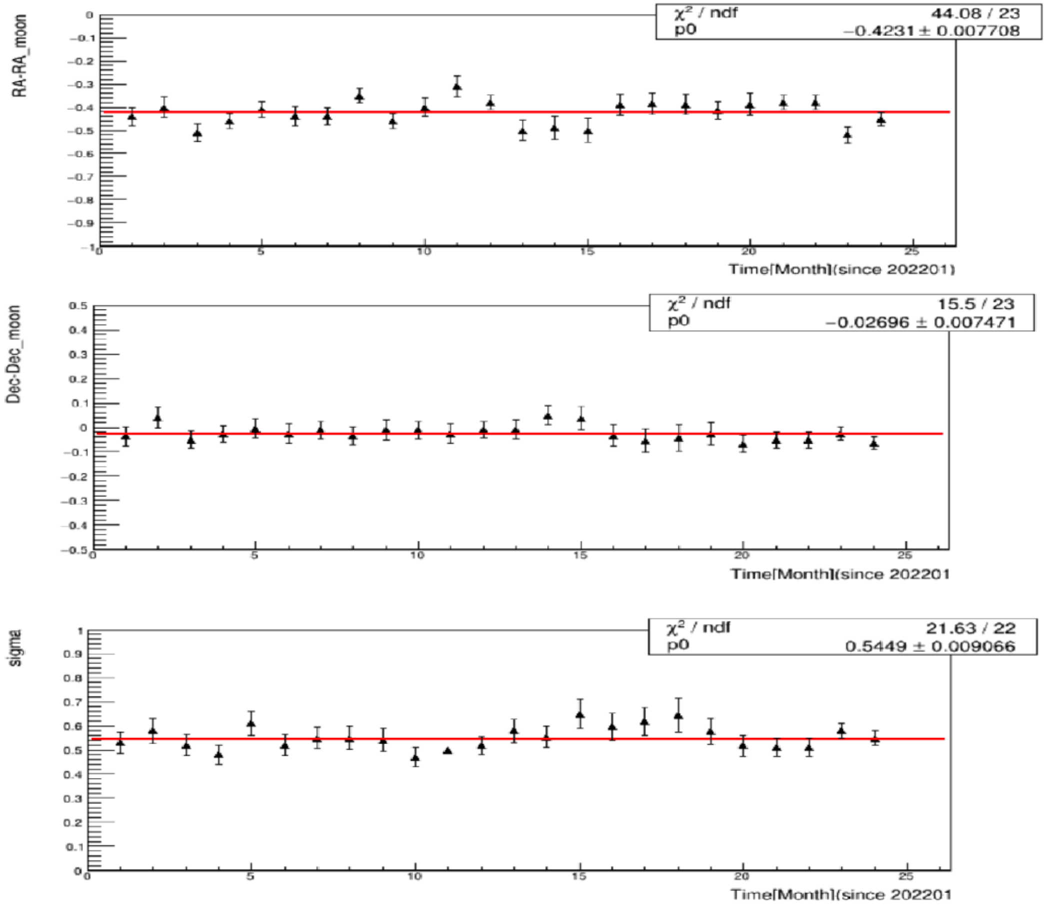




- N_{hit}>100 pointing error <0.1 deg

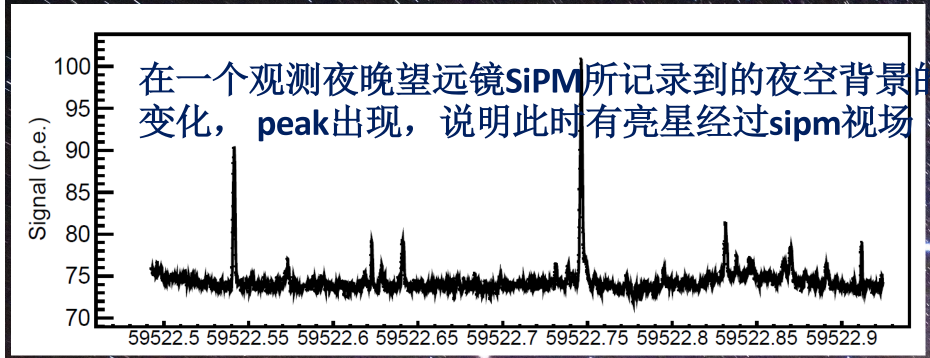
Moon shadow monitoring @ N_{hit}>100

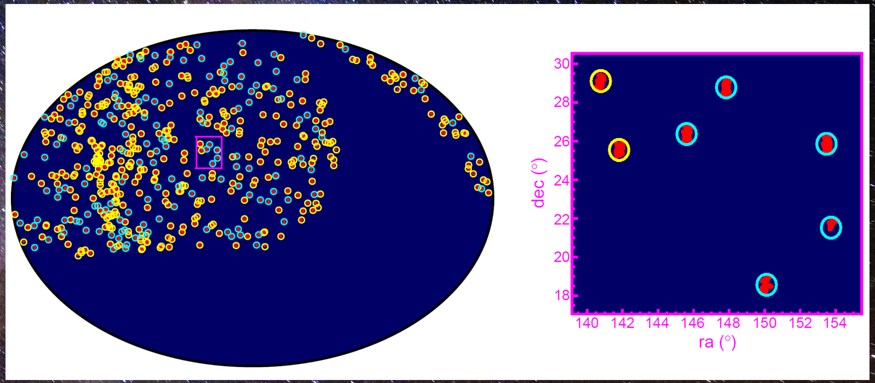


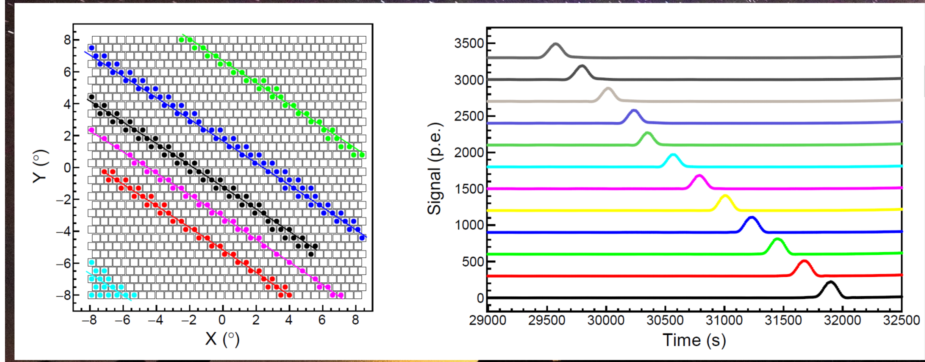


- N_{hit}>100 pointing error <0.1 deg

望远镜指向标定 夜空中明亮的恒星作为向导







WFCTA记录下记录下的6颗星的径迹（点）
与恒星的在某一望远镜指向下的径迹

方法特点：

用望远镜自己的观测数据，一个观测夜晚的数据可以完成标定，用时约10分钟；

有5颗星时，指向精度约0.02度

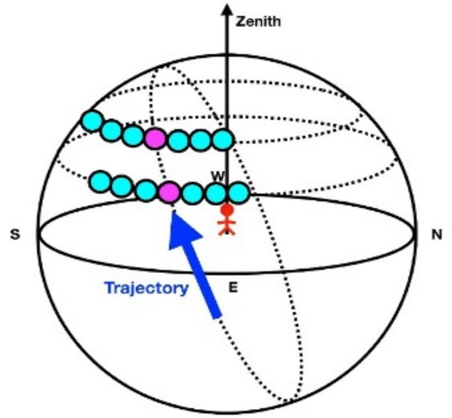
有15颗星时, 指向精度可以达到0.01度

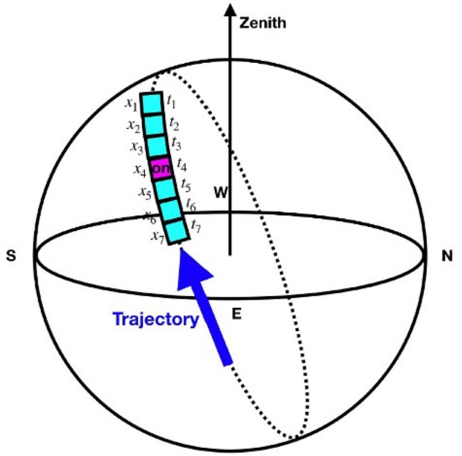
- 望远镜观测到的星
- 用于指向标定的孤立亮星
- 星表中星等小于5的亮星

天体源数据分析

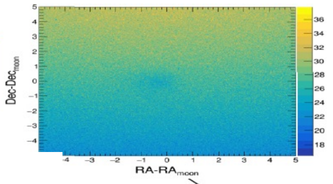
- **背景估计**
  - 等天顶角，等赤纬，
  - 时间交换法，直接积分
  - 环绕窗口……
- **天图分析**
- **显著性计算**
- **流强估计和能谱拟合**
  - Forward folding
  - 单源/多源分析
  - 复杂背景物理图像的考虑
- ……

$$S = \frac{N_S}{\hat{\sigma}(N_S)} = \frac{N_{on} - \alpha N_{off}}{\sqrt{N_{on} + \alpha^2 N_{off}}}.$$

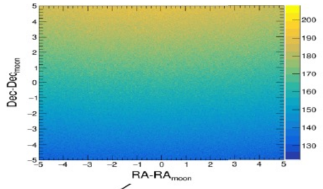




向源天图

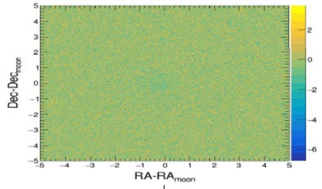


+



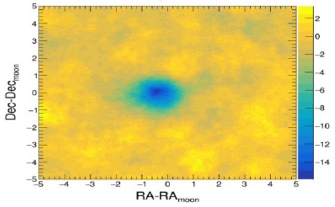
背景天图

$$n_\sigma = \frac{n_{obs} - n_b}{\sqrt{n_b}}$$



初步显著性天图

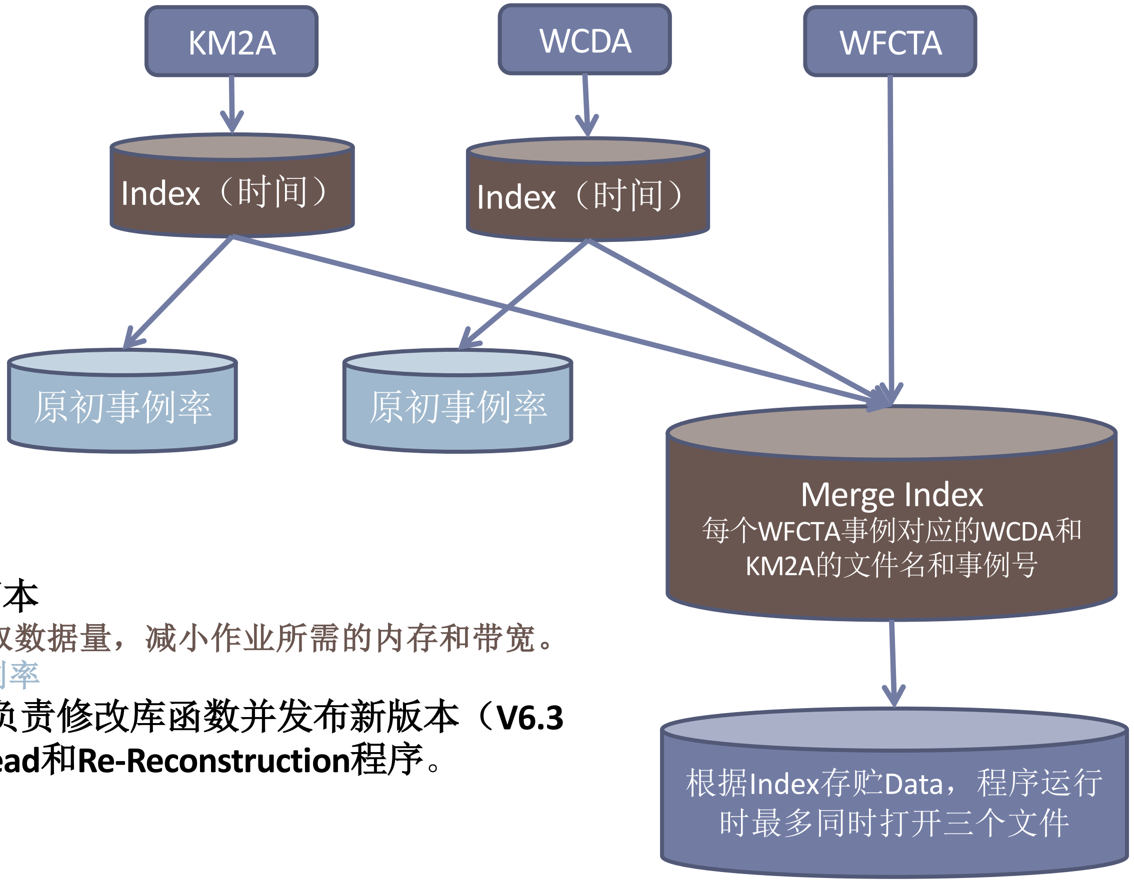
平滑



最终显著性天图

WFCTA 数据符合

◆ **WFCTA和WCDA、KM2A事例符合逻辑框图**



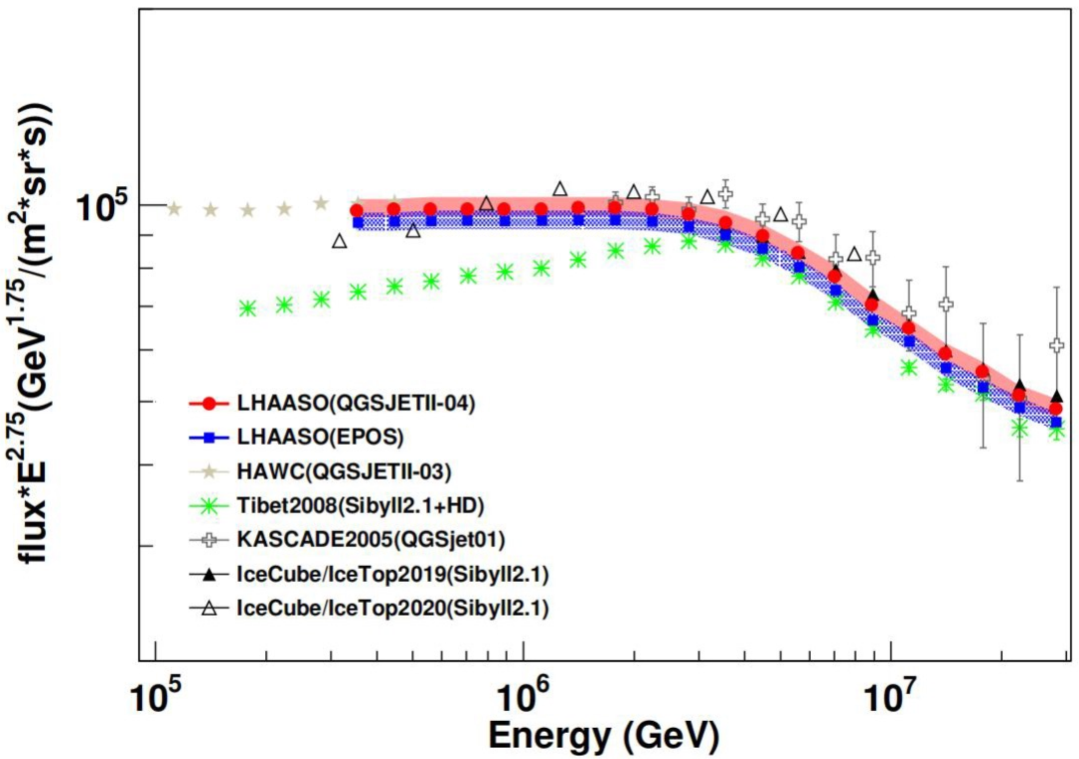
难点1： 数据量庞大（>3TB，~4000个文件/晚），且在LHAASO建设过程中文件格式等发生变化；以及计算资源方面的限制，如磁盘空间大小、eos应用等。

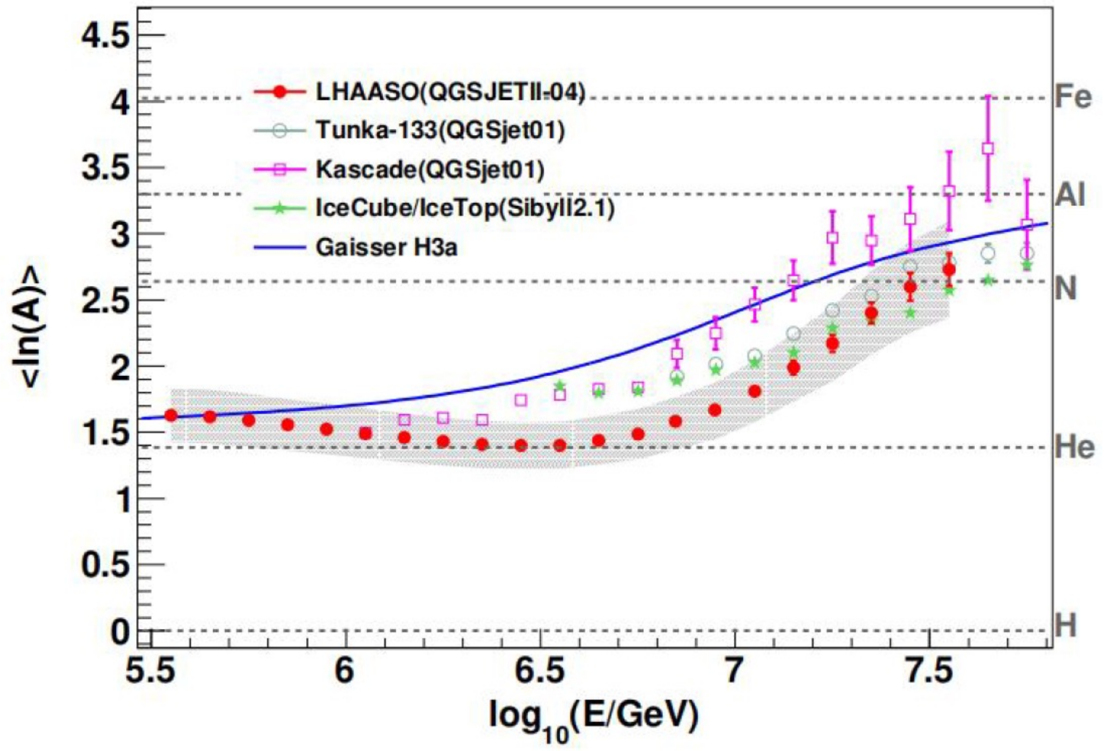
难点2： 随着LHAASO全阵列建设，数据分析全面展开，各物理分析组不断地对各子阵列数据进行优化和更新，如：WFCTA波形积分、KM2A不同标定版本数据等。

◆ **使用索引方法，建立自动运行程序和脚本**
- 使用索引（**Index**）可以压缩符合时读取数据量，减小作业所需的内存和带宽。
- 根据索引检测KM2A和WCDA的原初事例率

◆ **建立Public函数库统一管理程序，专人负责修改库函数并发布新版本（V6.3版本），为其他科学用户提供统一的Read和Re-Reconstruction程序。**

all particle energy spectrum and composition by LHAASO





A complex variable $N_{e\mu}$ is constructed with weak dependent on primary CR mass

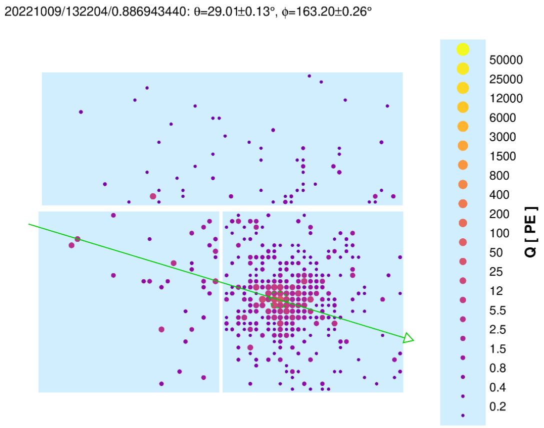
Energy reconstruction with $N_{e\mu}$

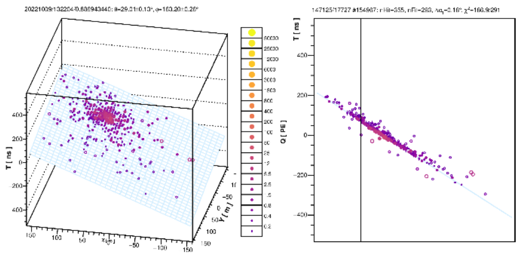
- better resolution, less bias between components + R: 12% + B: <5% @ 1 PeV

  → Systematic uncertainties are sufficiently small

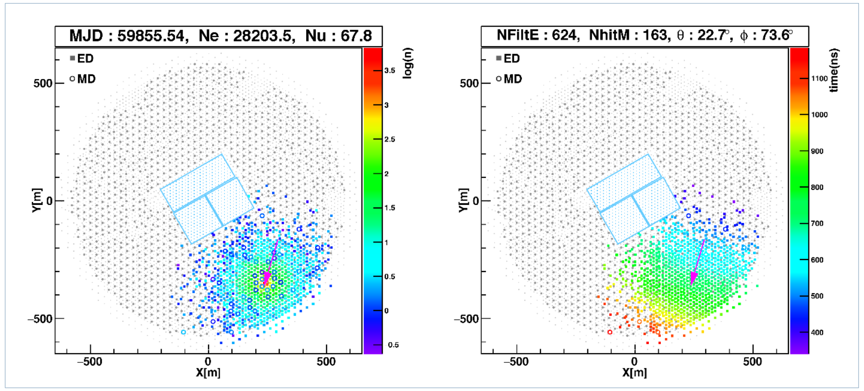
The all-particle energy spectrum knee is dominant by the knee of light components, instead of the medium-heavy components

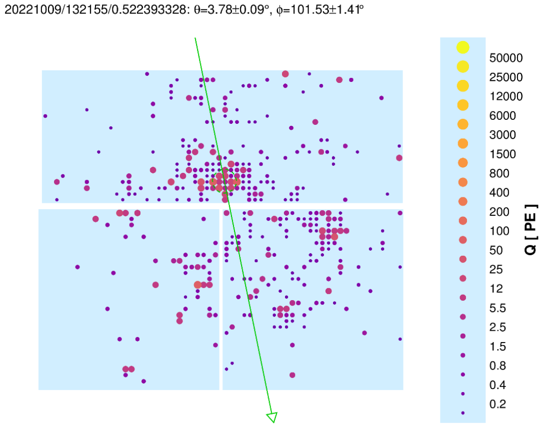
ML or AI @ LHAASO

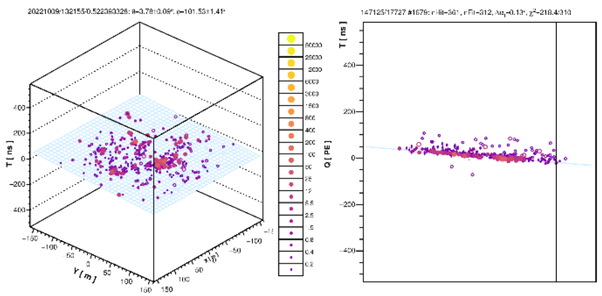




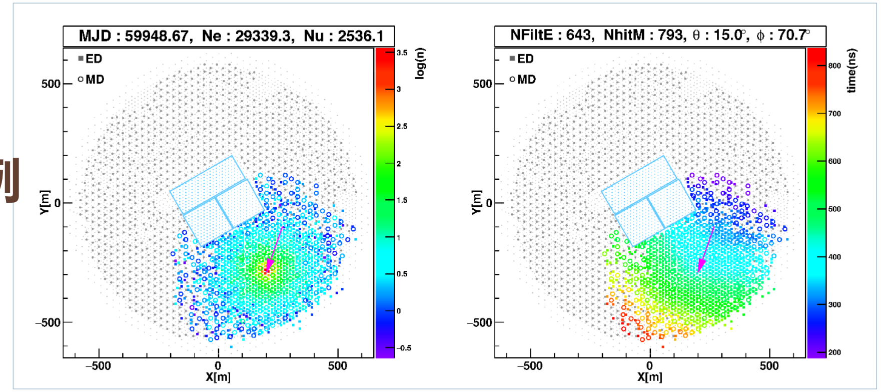
伽马事例







宇宙线事例



Deep learning @ shower reconstruction (geometry + particle identification)

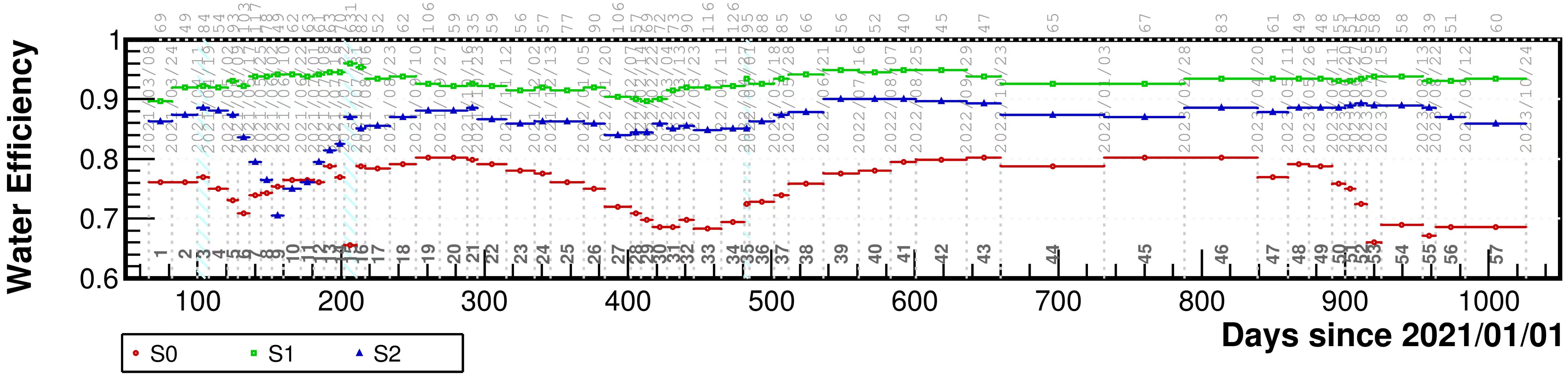
Summary and Prospect

- LHAASO 原始数据经过标定后转变为可以测出簇射信息的物理量；
- LHAASO模拟数据的真实化是后续物理分析中的系统误差的一个主要来源；
- LHAASO将继续在20年内将采用四种探测技术，全方位、多变量地测量来自于北天区的高能天体的伽马射线和宇宙线；
    - 甚高能区（1 TeV – 30 TeV）灵敏度最优的伽马巡天探测器；
    - 超高能区（30 TeV – 1 PeV）灵敏度最好的伽马天文探测器；
    - 能区跨度范围（10 TeV – 1 EeV）最大的宇宙线探测器。
- LHAASO数据分析的优化，更新和升级是LHAASO生命力的重要支撑点。
- 参考文献

    - LHAASO collaboration, Chinese Physics C Vol.45, (2021) 025002;
    - LHAASO Collaboration, Chinese Physics C Vol.45, (2021) 085002;

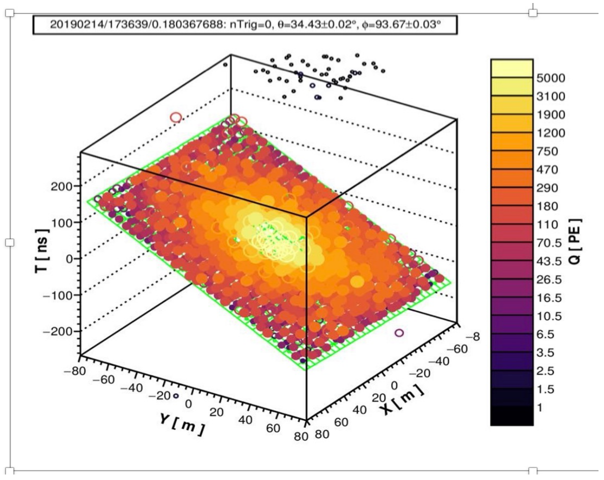
backup

Data and Simulation are divided into periods

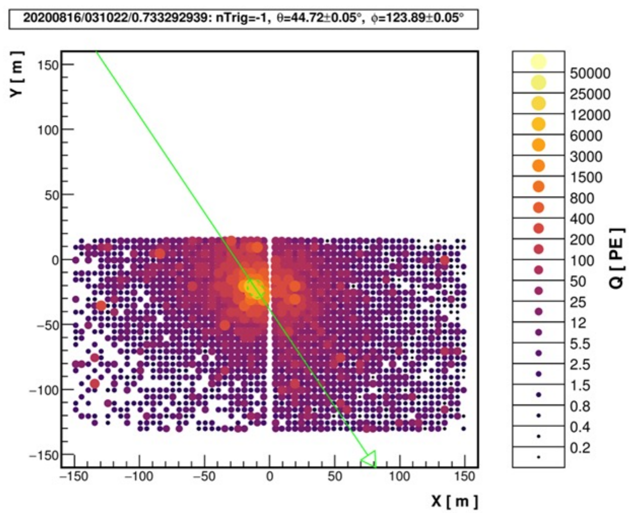


Timeline of LHAASO

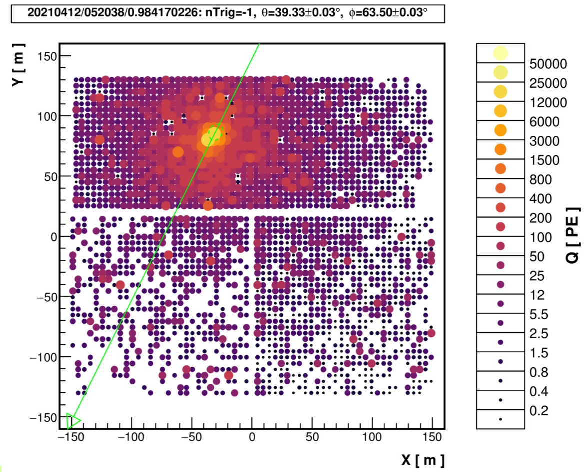
Valentine event @
WCDA-1 20190214



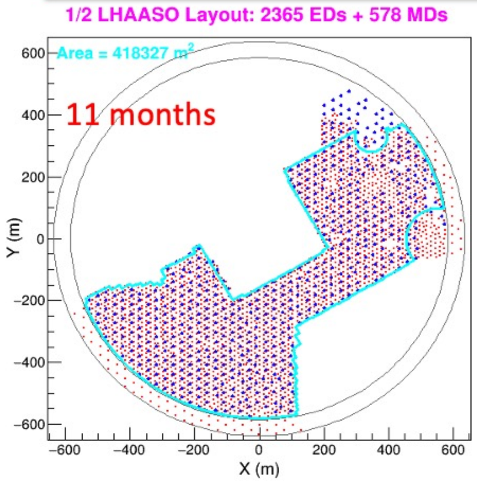
WCDA-U @ 202010



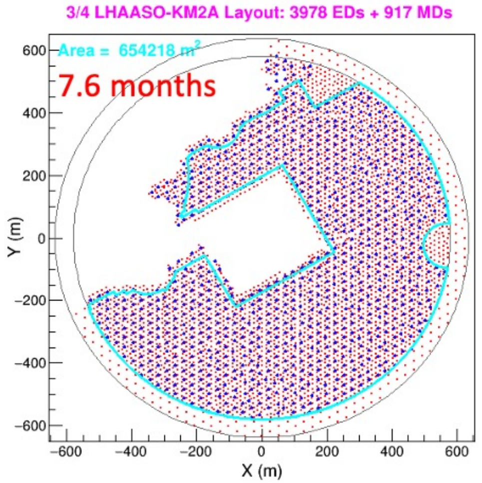
WCDA full array from 202103



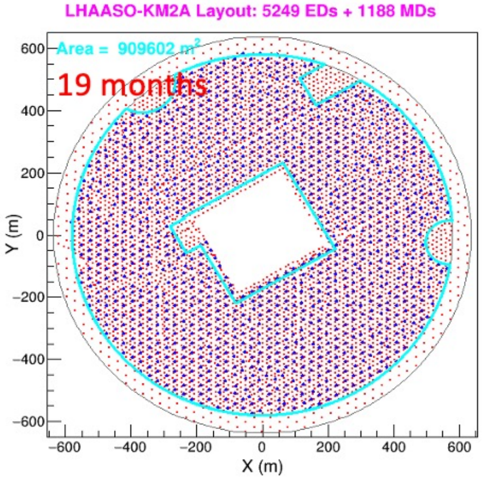
½ KM2A → ¾ KM2A → Full-KM2A



2019-12-27—2020-11-30

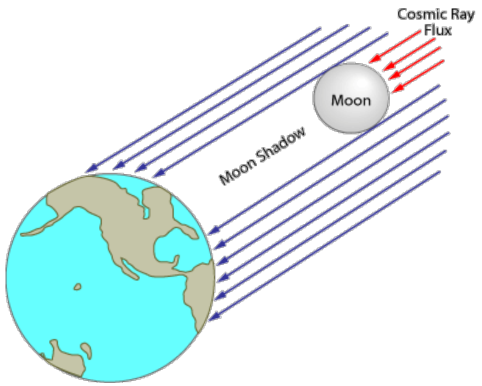


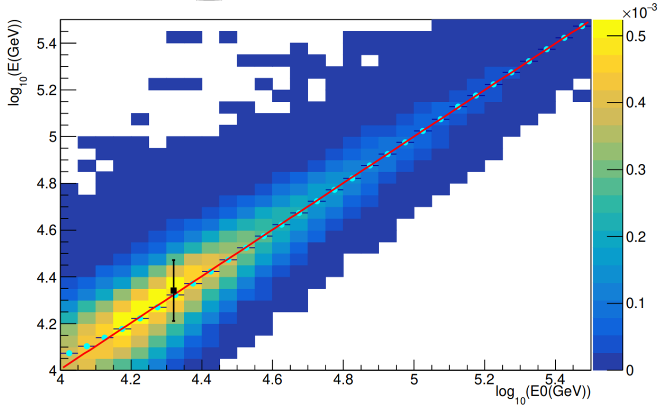
2020-12-01—2021-07-19

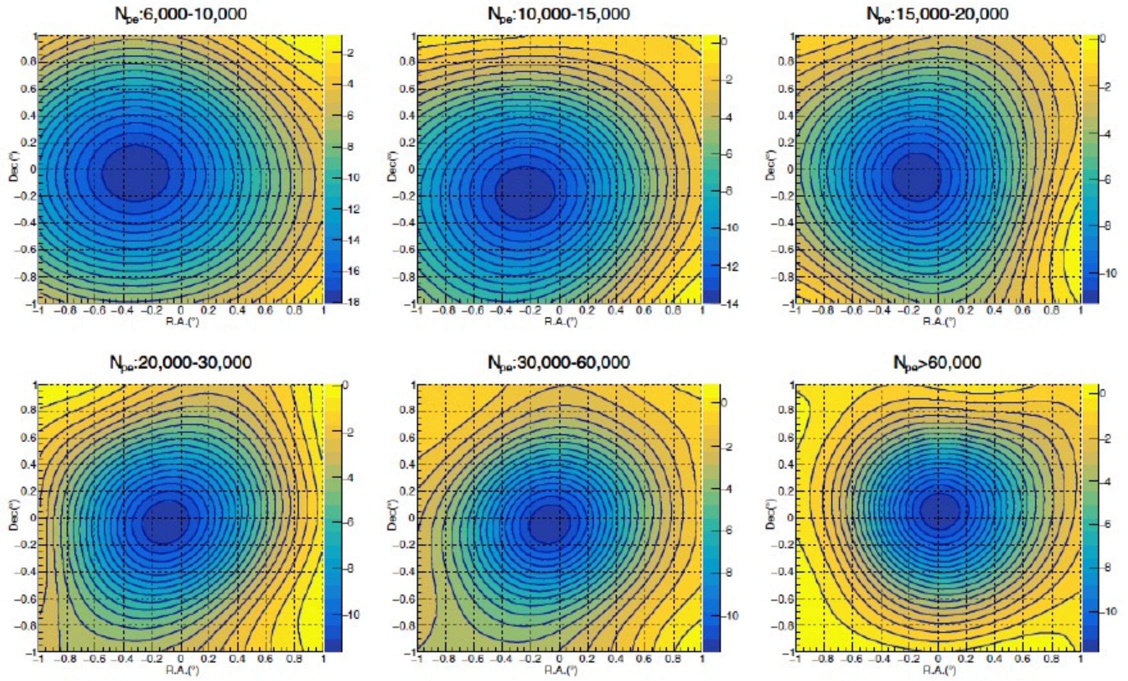


2021-07-20—> now

Energies of the commonly triggered events derived by WFCTA and by the formula of the absolute energy scale







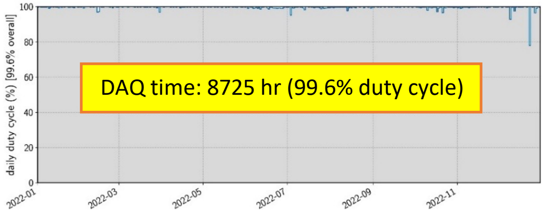
能标结果：

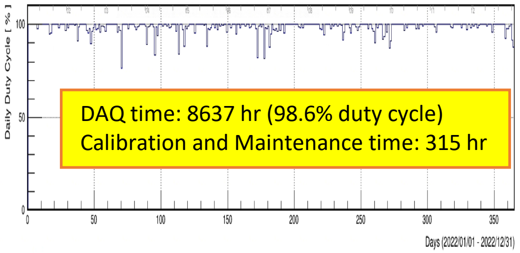
✓ **由 WFCTA 得到：**
21.9±0.1TeV

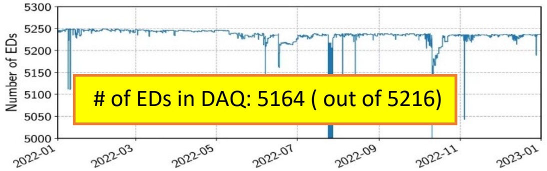
✓ **由绝对能标公式得到：**
23.4±0.1±1.3TeV

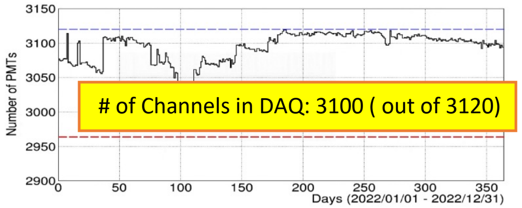
- 得到 WCDA-1 的能标在 6.6 TeV 的不确定度为 12%， 经过 4 年年的统计量量的累计，统计误差将会分别减小小到 3%。

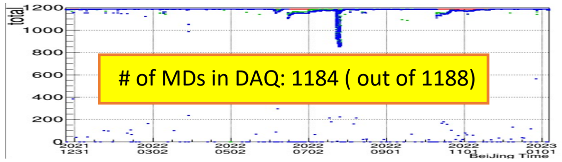
Supper Stable & Fruitful Operation

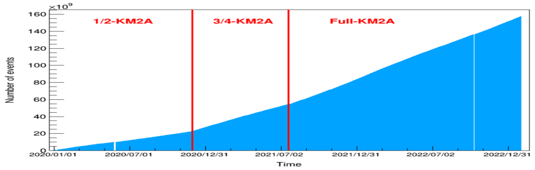


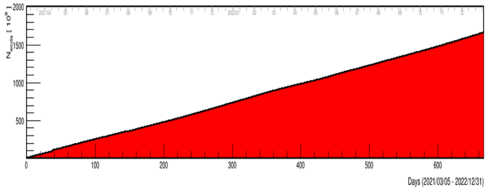


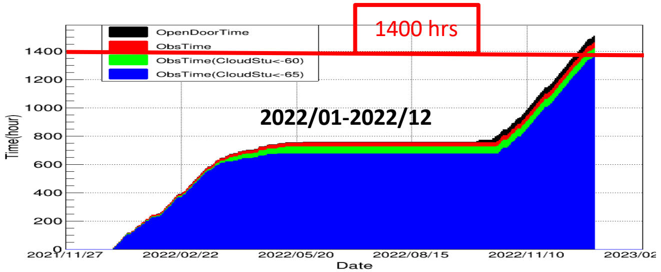








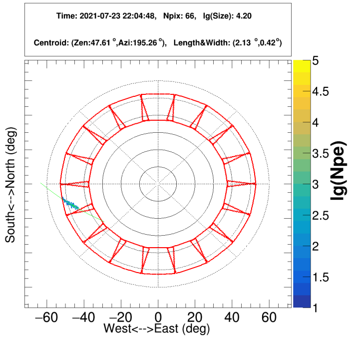


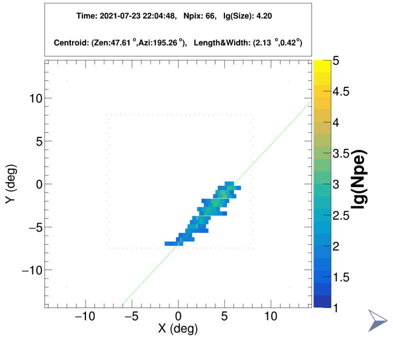


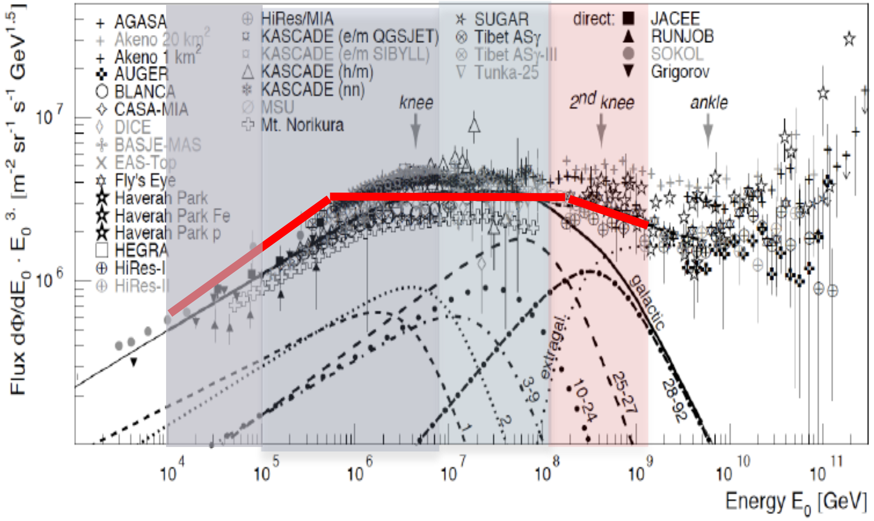
Reconstruction and Analysis

- Data procession
  - # of events: 1.e12 LE、1.5e11 HE、70 million hybrid
  - Amount: 11 PB
- Simulation
  - # of events: 1 billion LE、0.7 billion HE、150 million hybrid
  - Amount: 4 PB
- # of jobs: 10M for data, 50M for simulation

Wide Field of View Cherekov Telescope Array







10TeV-200TeV/ 100TeV-10PeV / 10PeV-100PeV/ 100PeV-2EeV