Data analysis and software of LHAASO

Min Zha on behalf of the LHAASO Collaboration
Institute of High Energy Physcis, CAS, China
第一届高能物理计算用户研讨会@ 19th May - 23rd May 2024

outline

- LHAASO experiment
- Data Analysis Status
- Summary and Prospects


Large High Altitude Air Shower Observatory (LHAASO) “拉索”
Haizi Mountain 4410 m a.s.l. Daocheng, Sichuan Province, China
2021-07 completed built and in full array operation

LHAASO collaboration

280 Researchers 32 Institutions from 6 countries

LHAASO

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

UHE gamma ray astronomy

WFCTA:
18 telescopes

CR individual spectrum...

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

VHE gamma ray astronomy

Some planed detectors

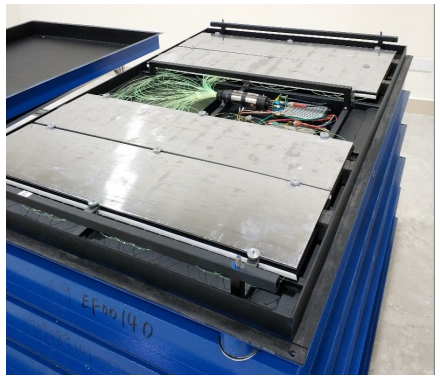
Neutron detectors

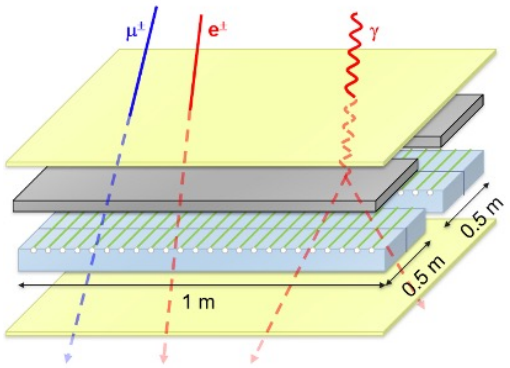
High energy IACTs

...

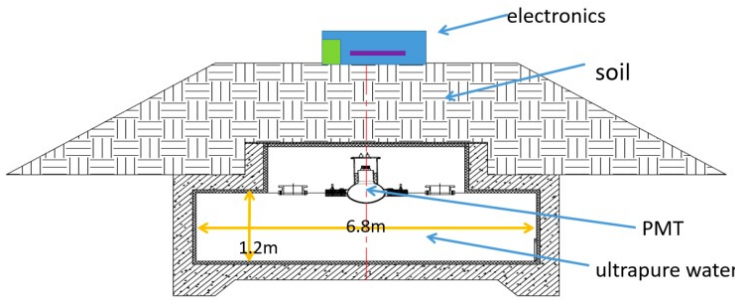
.....

Electromagnetic Detector (ED)

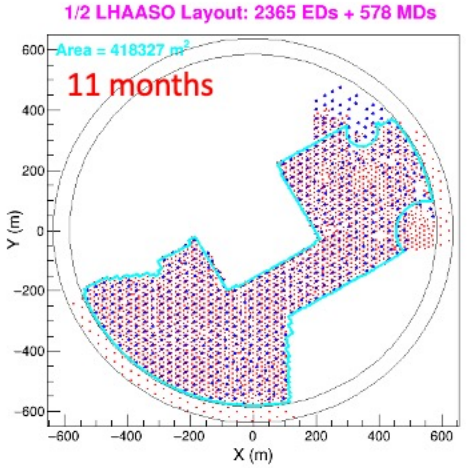


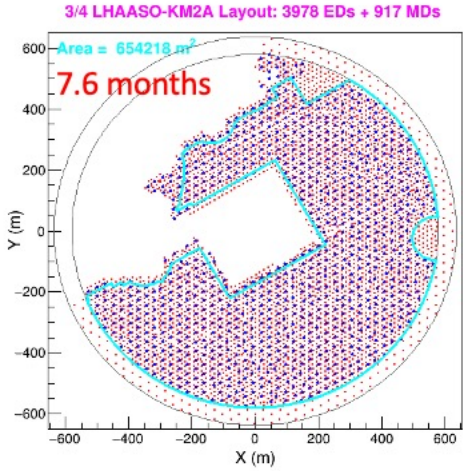


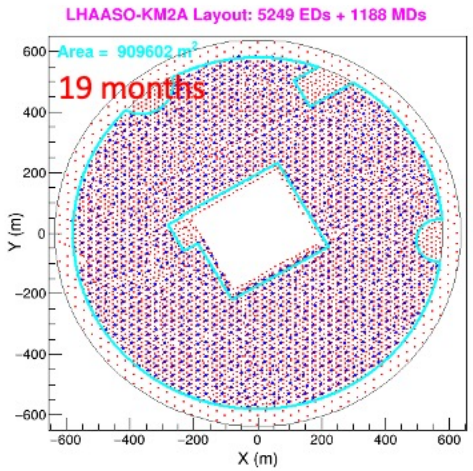
Muon Detector (MD)

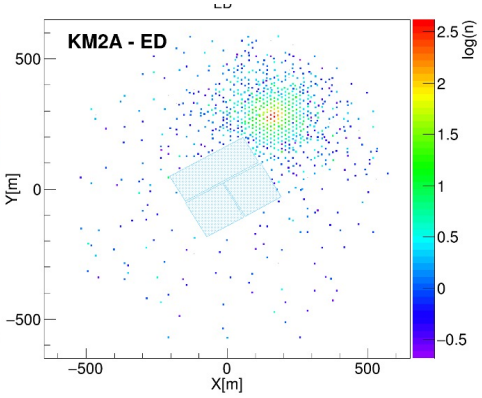


½ KM2A → ¾ KM2A → Full-KM2A

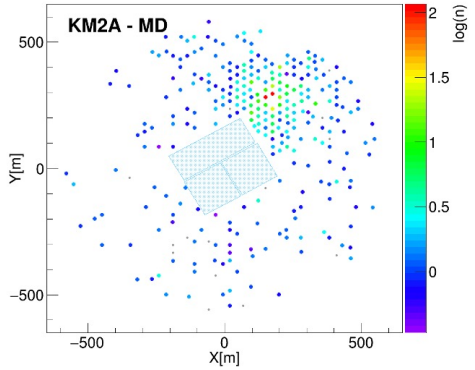








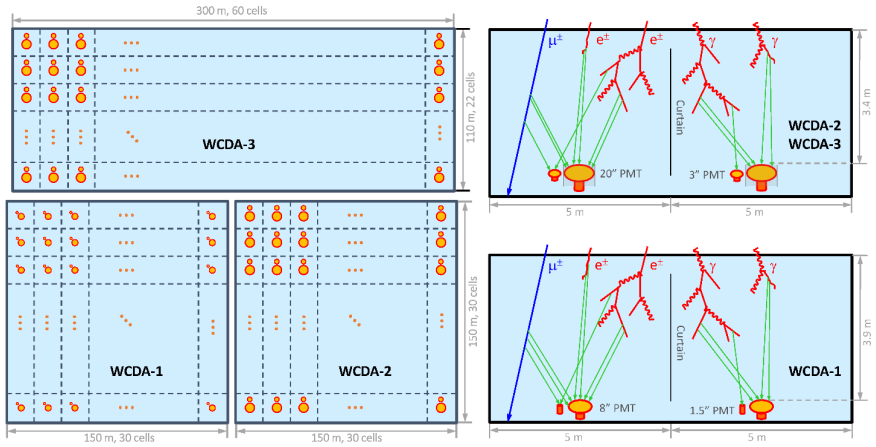
Nμ=853, N_{MD}=279



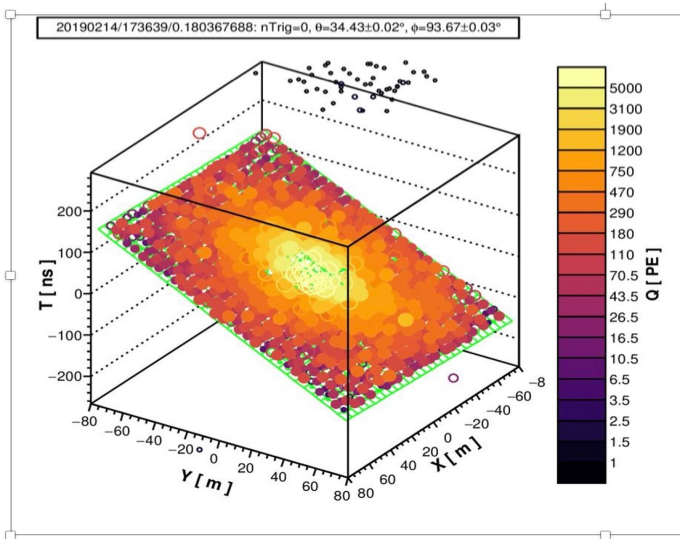
2019-12-27—2020-11-30

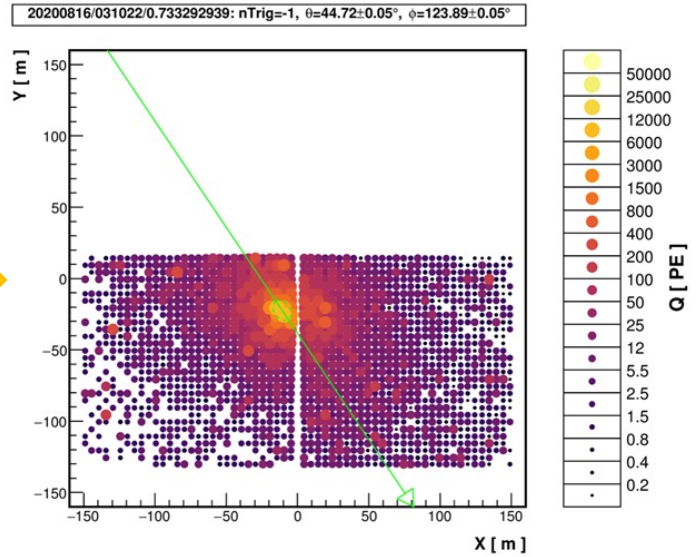
2020-12-01—2021-07-19

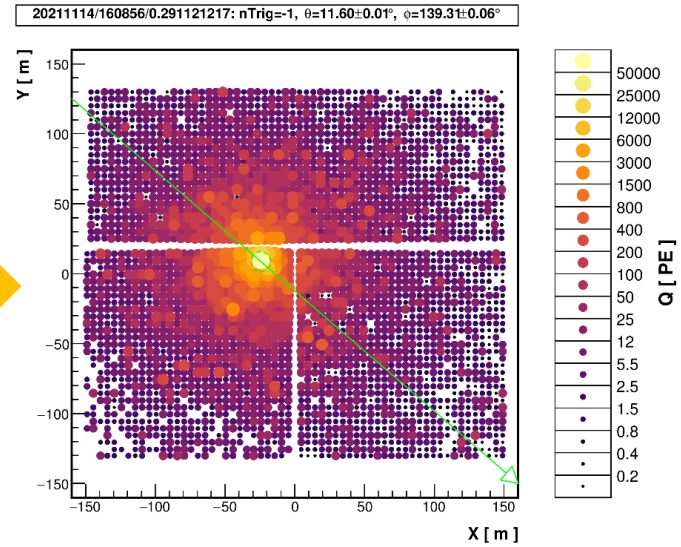
2021-07-20—> now



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







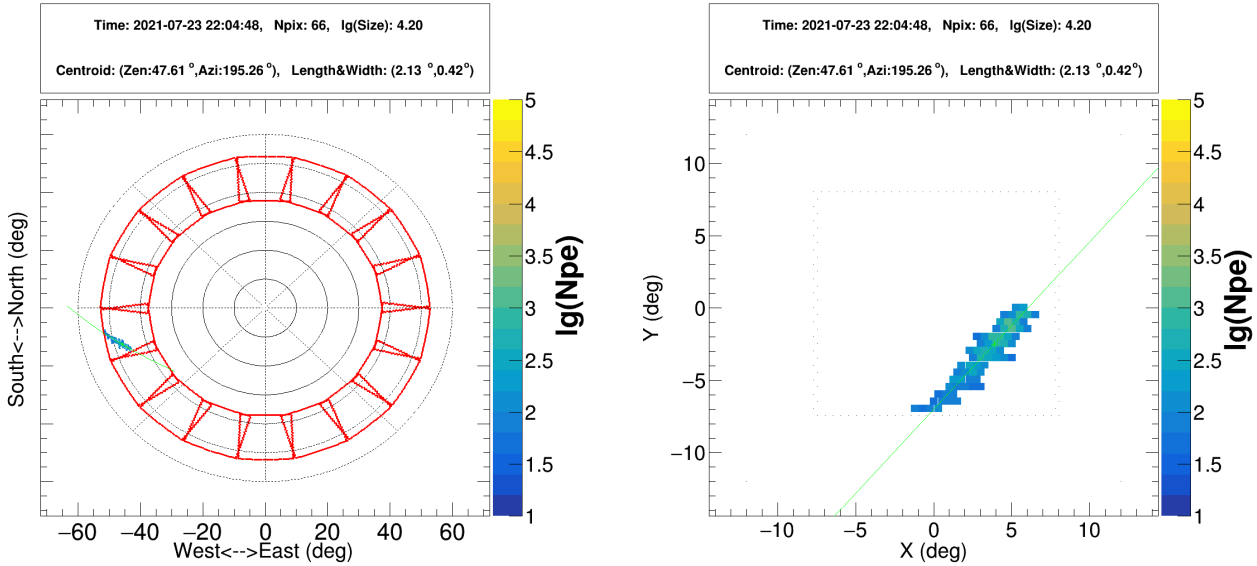
Wide Field of View Cherekov Telescope Array (WFCTA)

Mirror: 5 m^2 spherical mirror

FOV: 16°×16° / telescope

Camera: 32x32 =1024 pixels /telescope

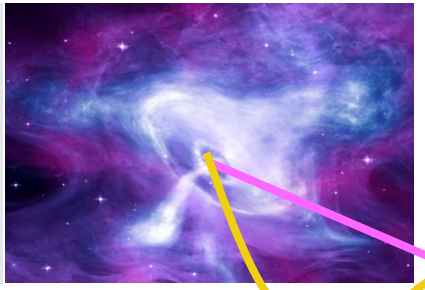
Pixel: 0.5° each



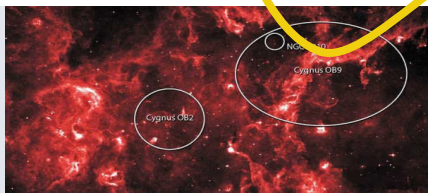
全方位研究宇宙线起源问题

宇宙线起源候选天体

脉冲星风云

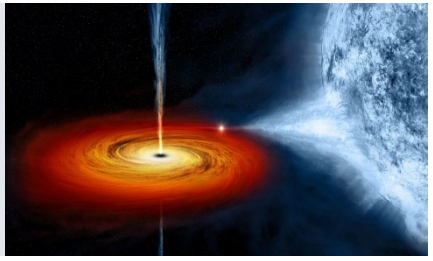


年轻大质量星团

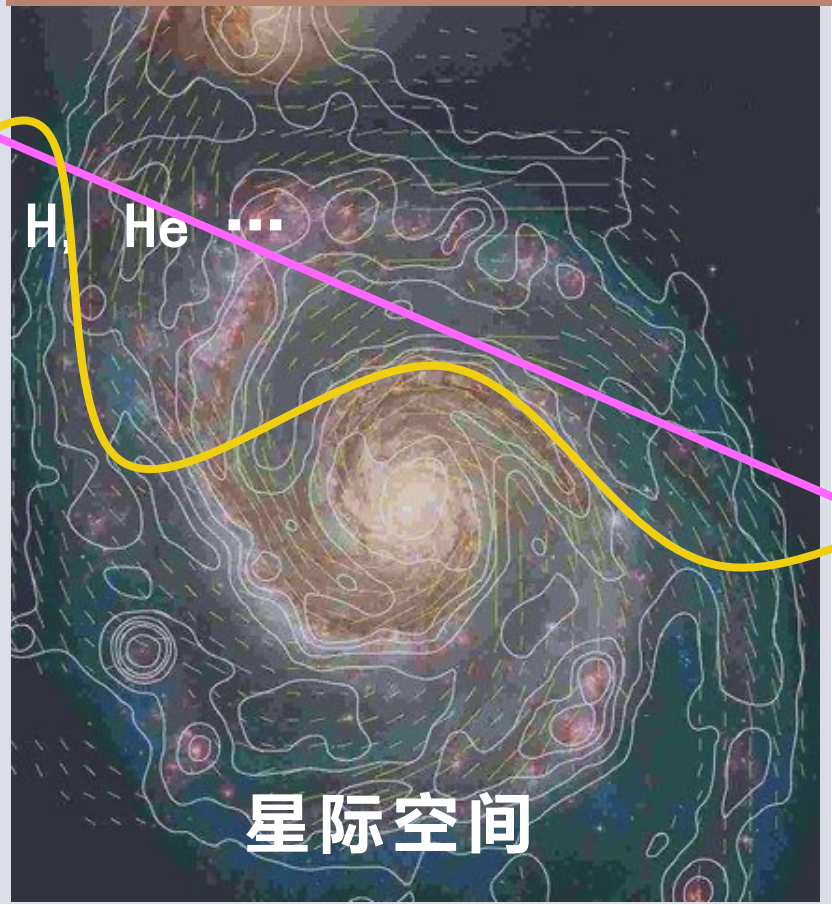


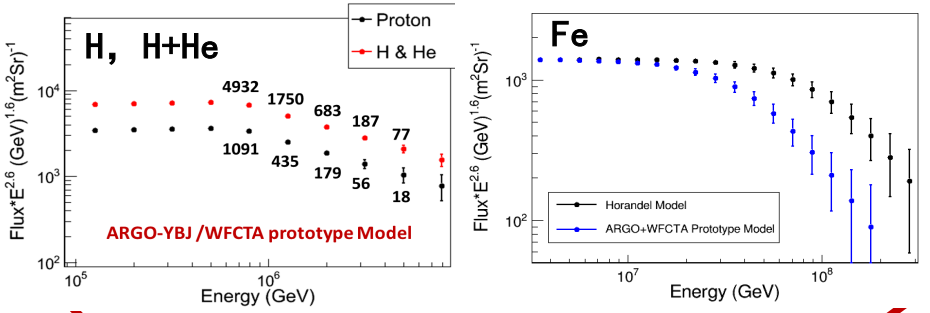
超新星遗迹

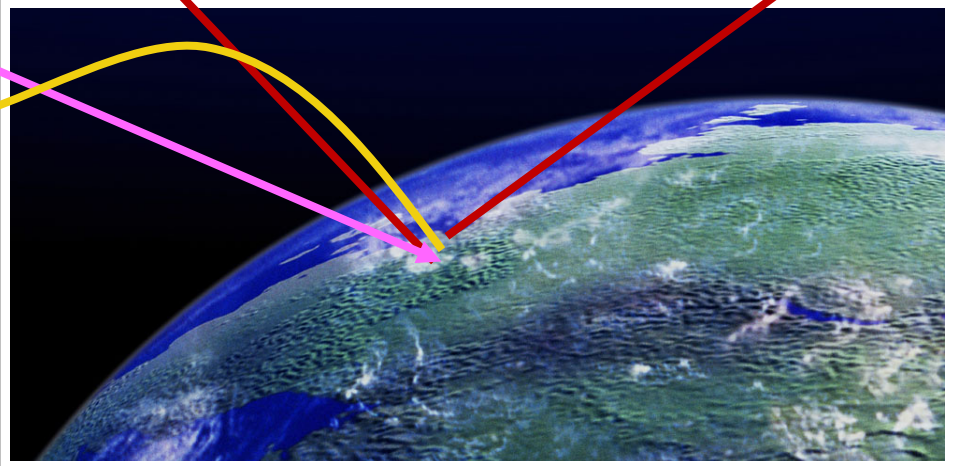
微类星体



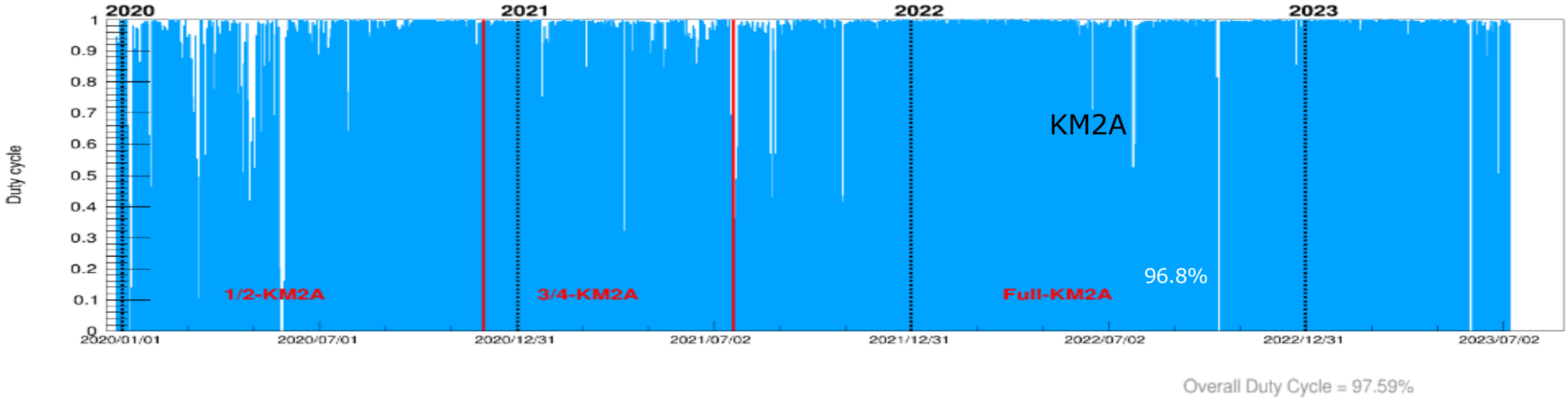
起源天体和加速（伽马射线）→银河系内分布（伽马射线）→地球宇宙线测量（宇宙线）

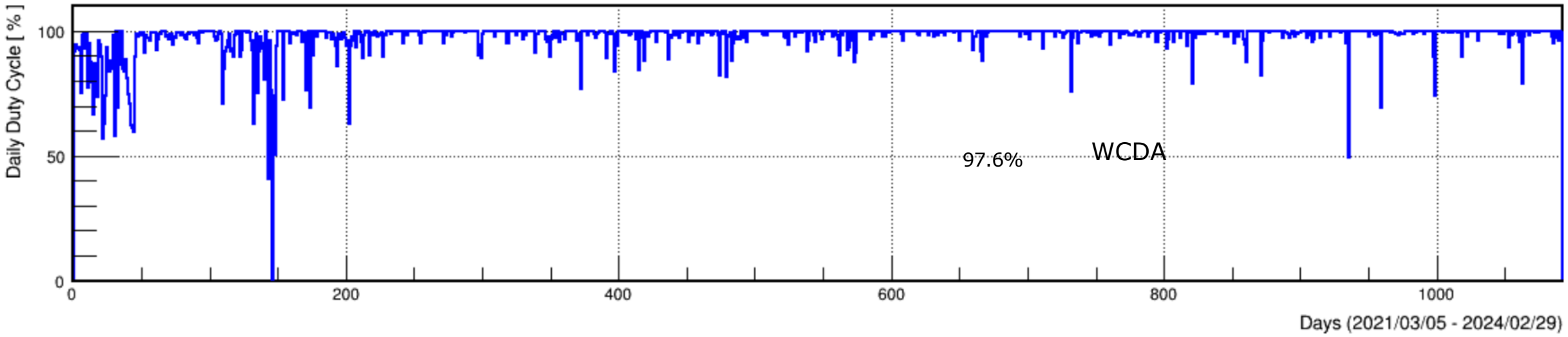






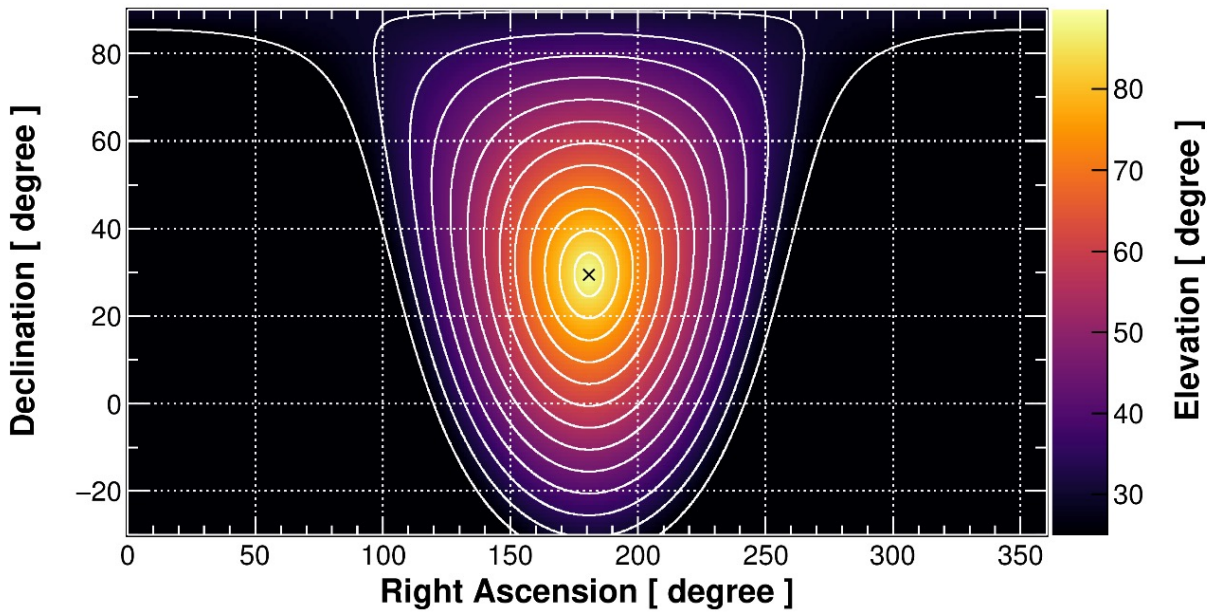
Features: full duty cycle



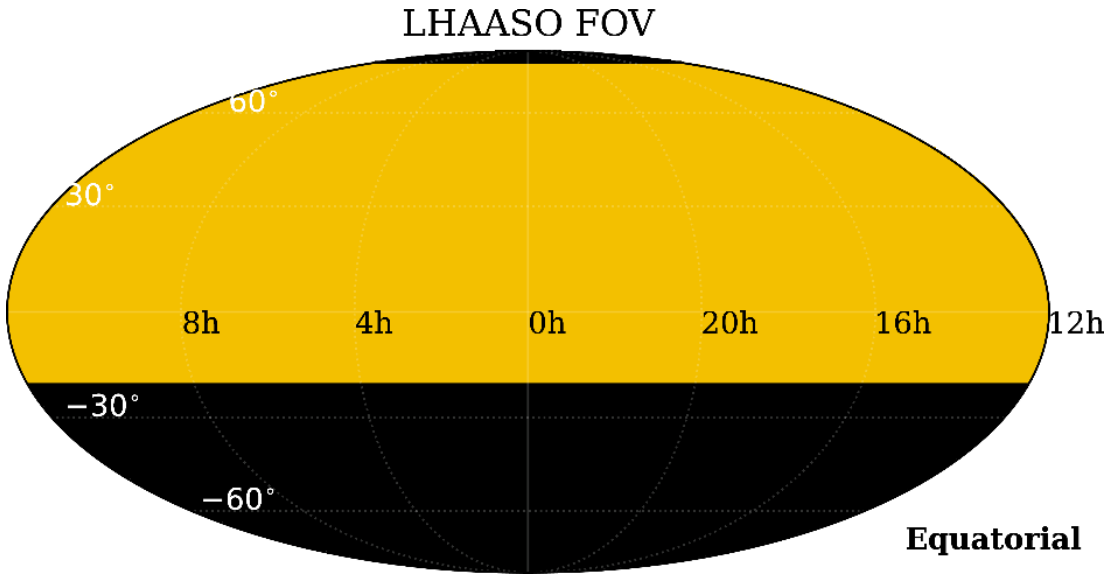


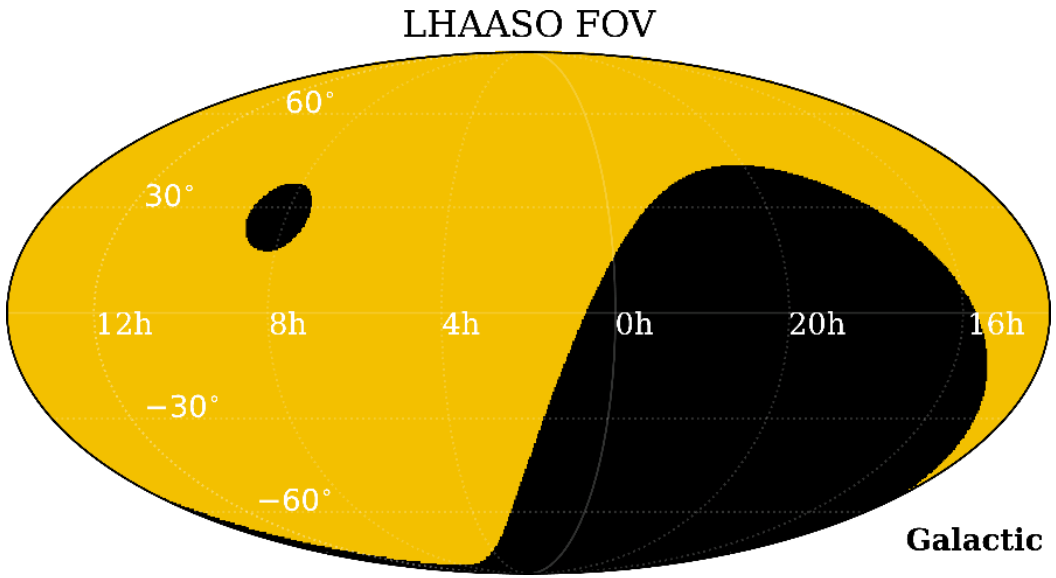
Features: wide field of view

Instant FOV



Daily/yearly FOV

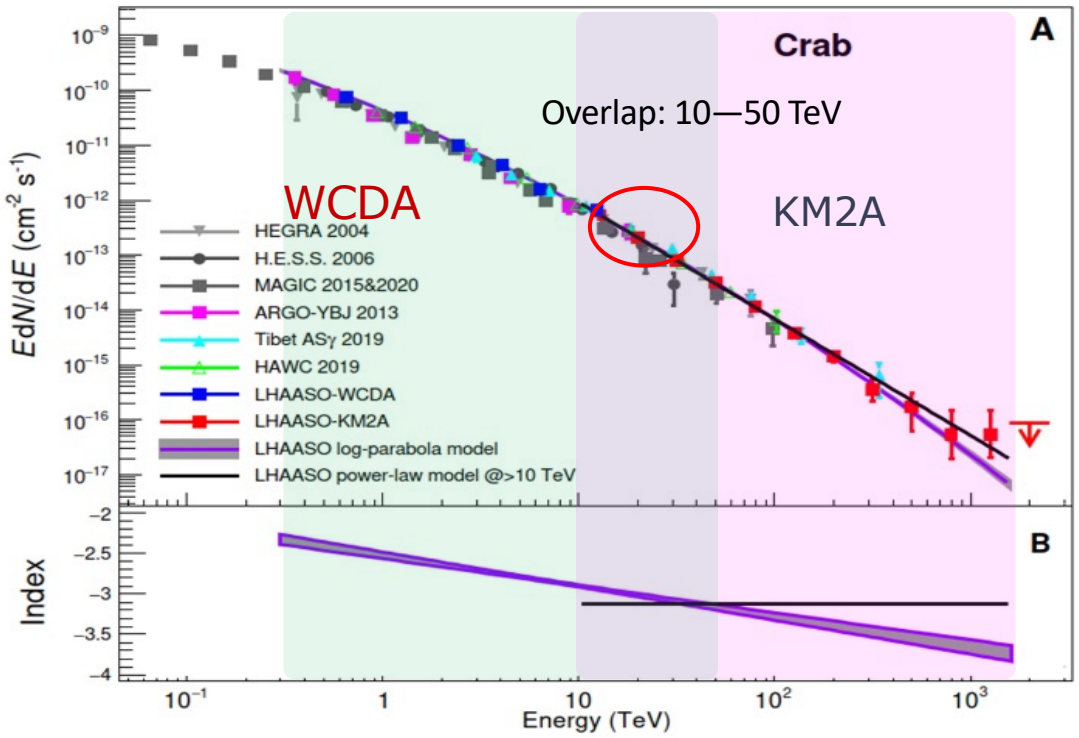




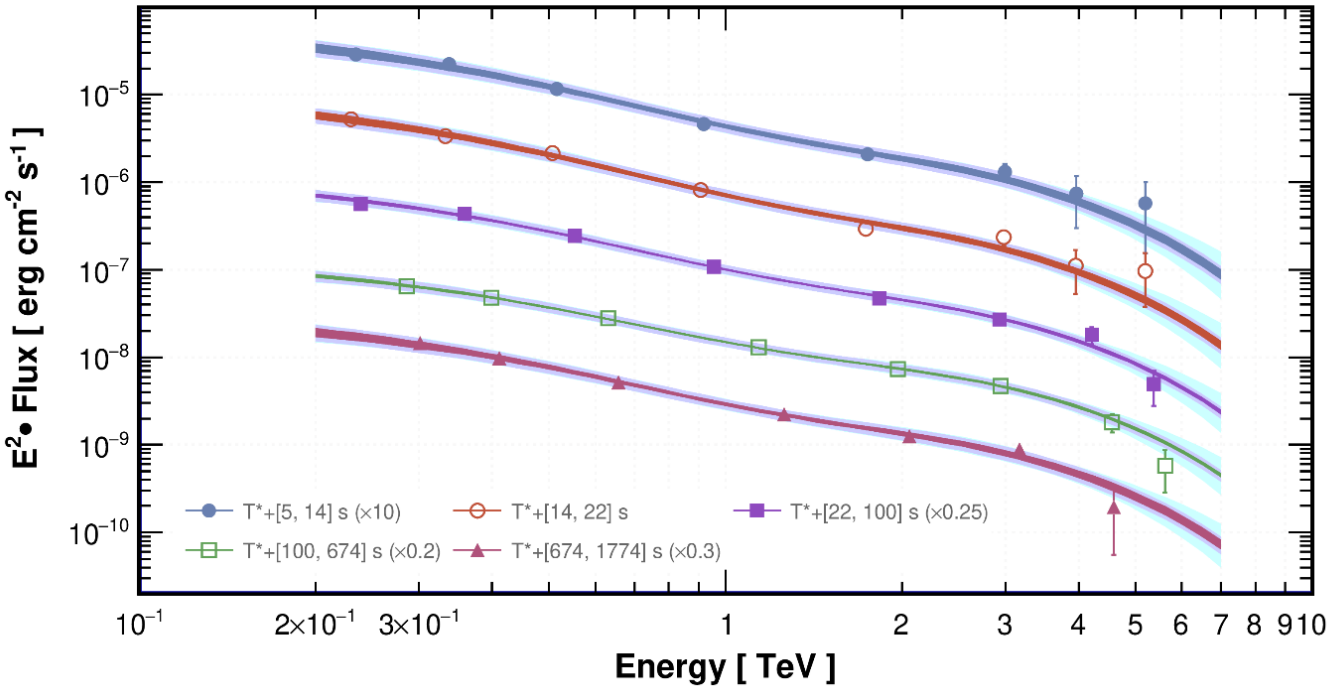
1/6 of the entire sky at any given moment.

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

Features: wide energy range coverage



GRB 221009A



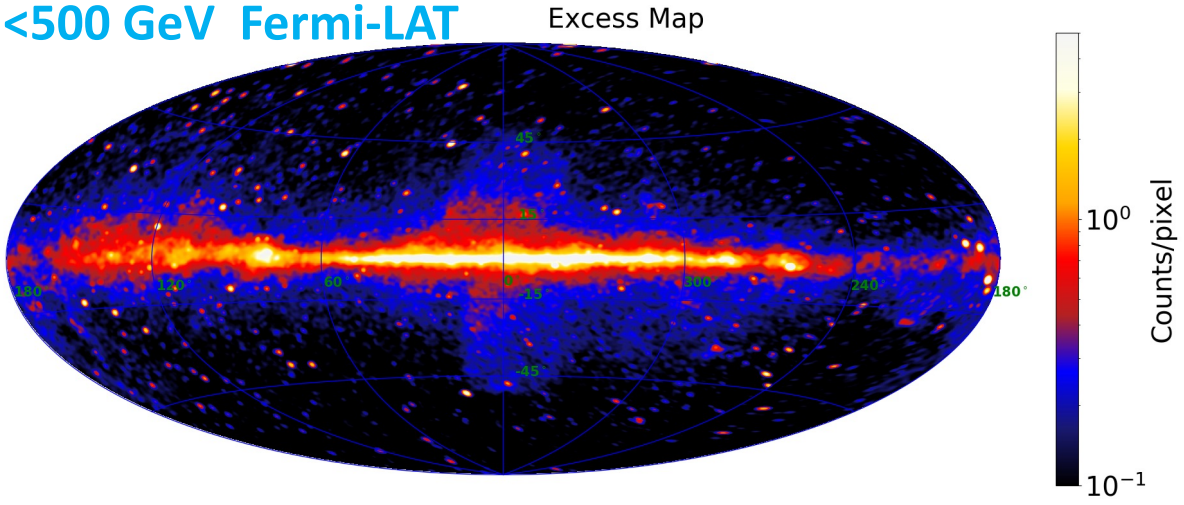
The lowest can reach < ~100 GeV?

- 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

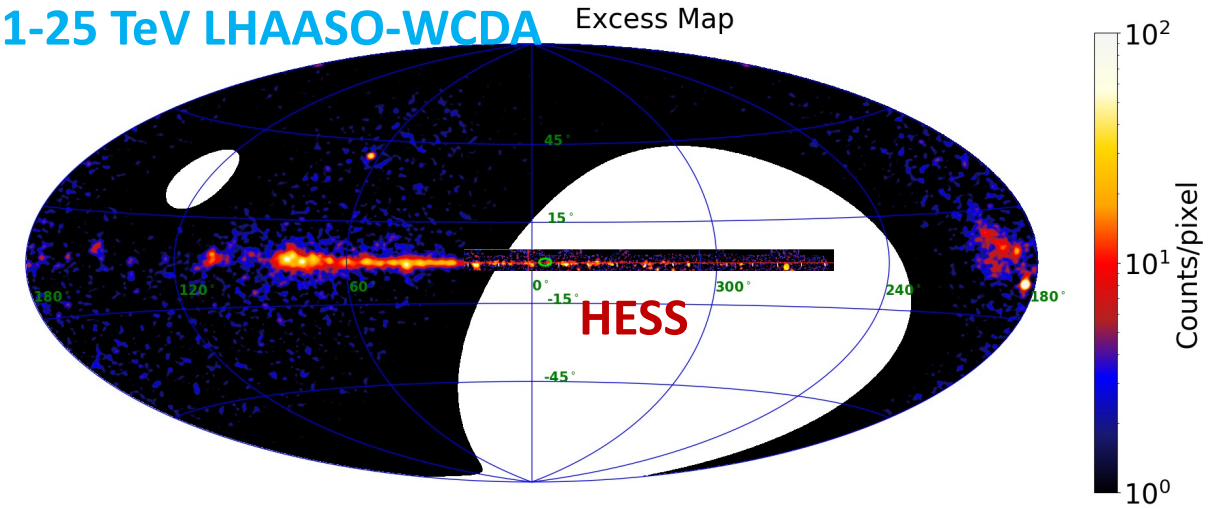
UHE γ-ray Astronomy: sources and diffuse emission

- Survey discovered 30+ new sources, 40+ PeVatrons and diffuse γ-ray emission

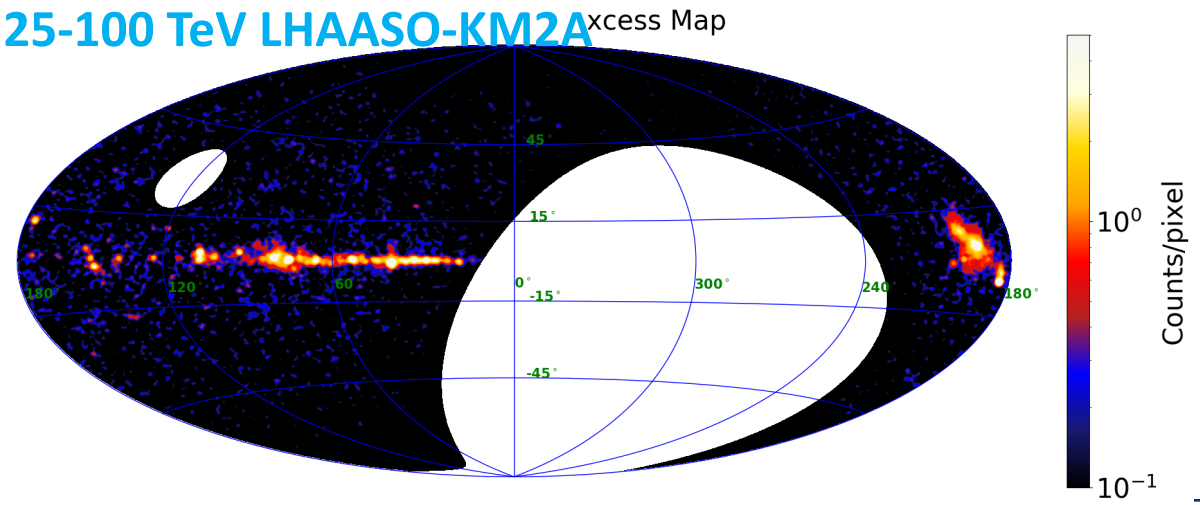
<500 GeV Fermi-LAT



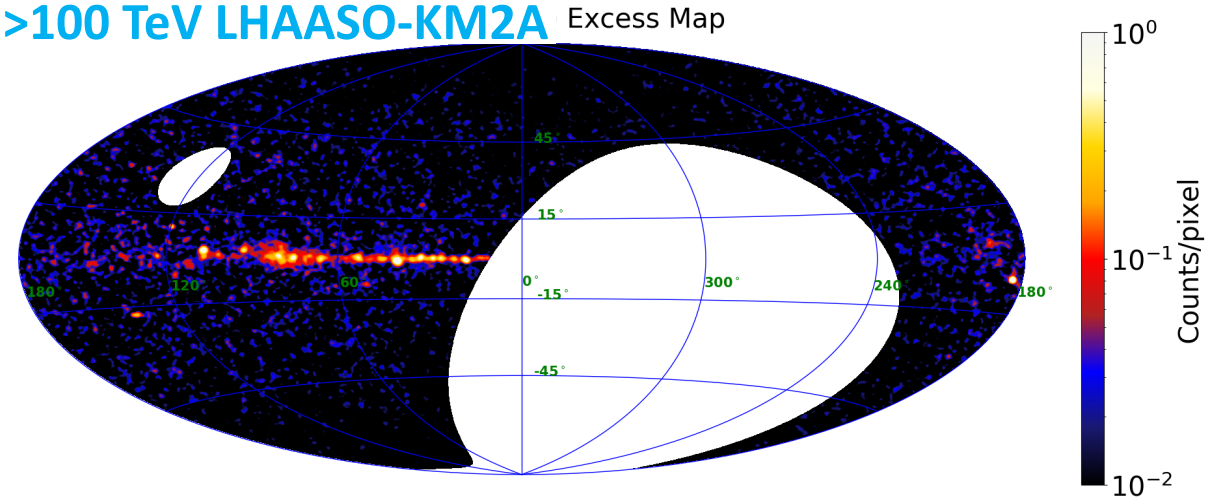
1-25 TeV LHAASO-WCDA



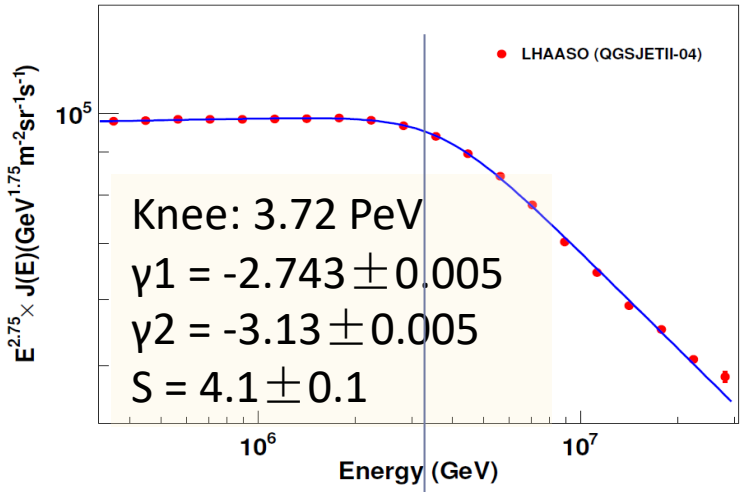
25-100 TeV LHAASO-KM2A

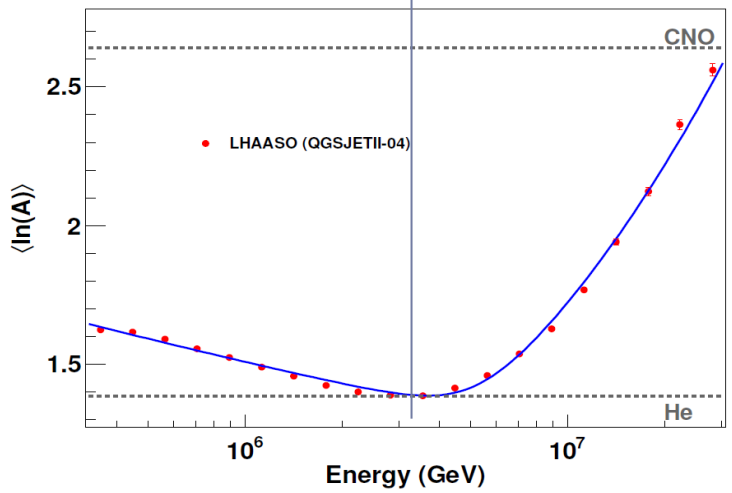


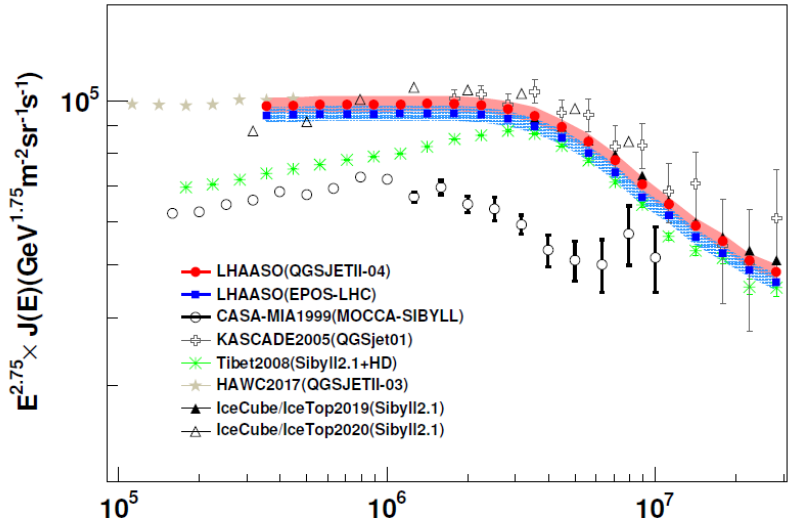
>100 TeV LHAASO-KM2A

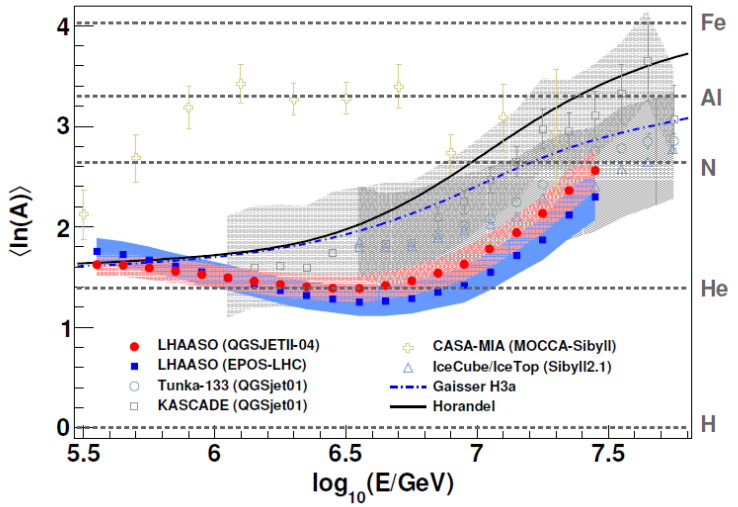


All-particle energy spectrum & composition by LHAASO (from 0.3 to 30 PeV)





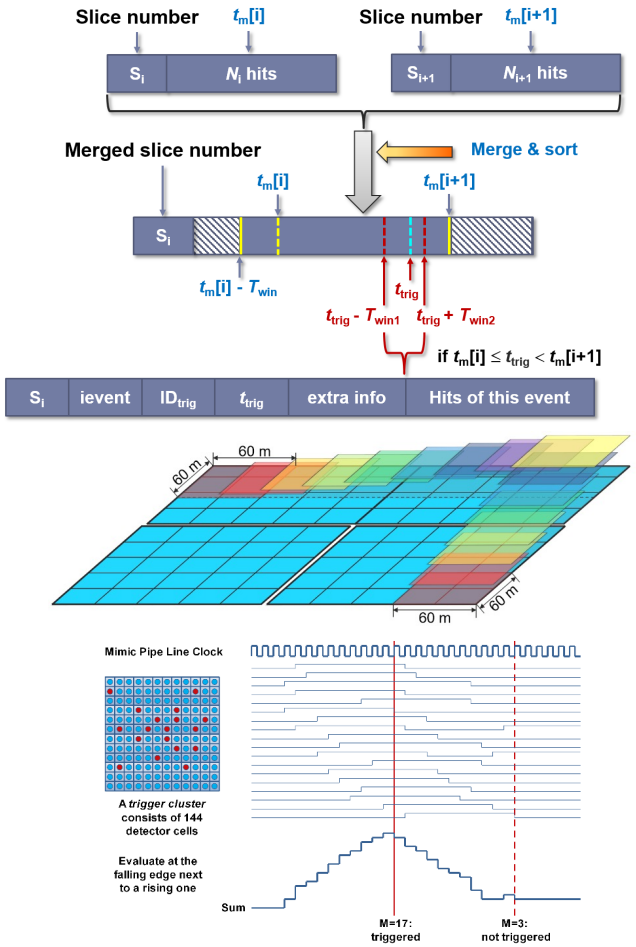




- Systematic uncertainties are sufficiently small
- This unveils a clear correlation between the flux and the composition at the knee

Trigger

- Implemented on a computing cluster:
  - Soft trigger.
- Basic triggers:
  - KM2A (EDA + MDA), WCDA and WFCTA, independently;
  - 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.

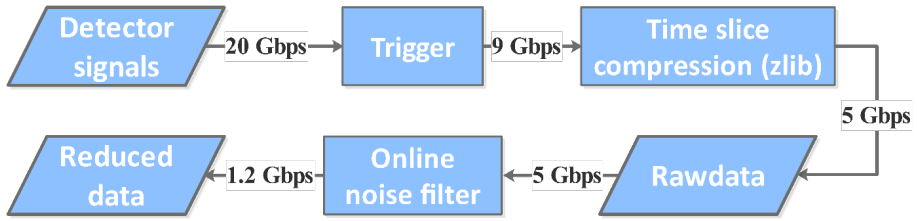


Trigger logic of WCDA

LHAASO数据量：~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



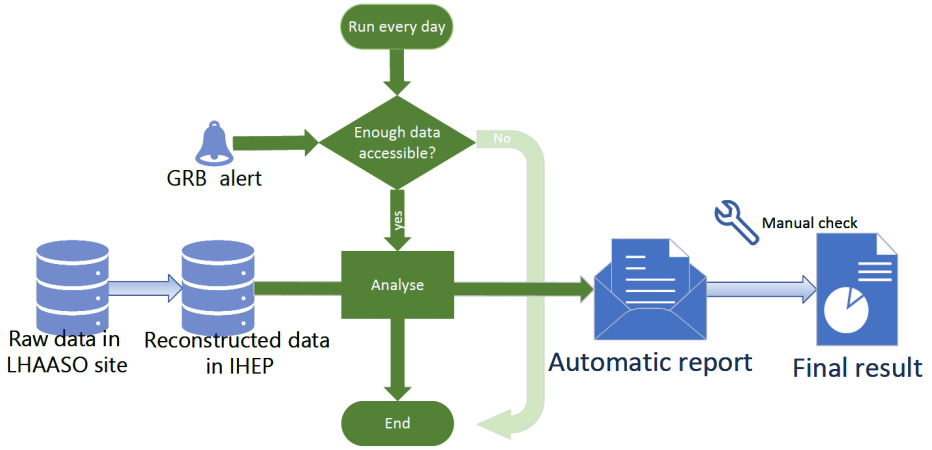
WFCTA原始数据：

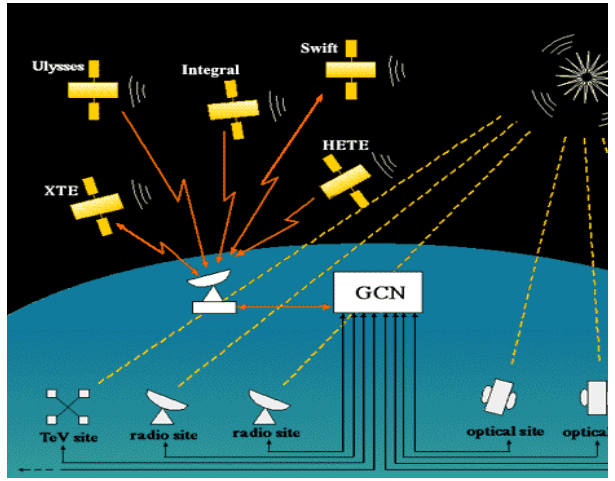
- 触发率：1.1 Hz/telescope * 18 = 20 Hz
- 数据量：100 TB/yr （注意：1400 hour/yr）

WCDA原始数据：

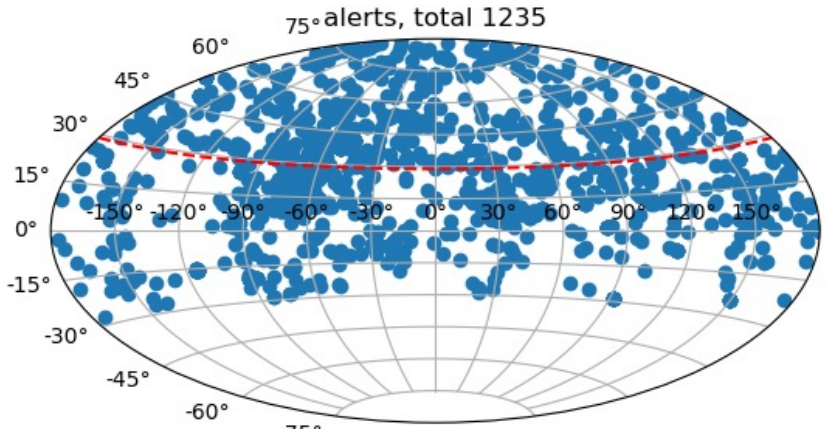
- 触发率：34 kHz ➔ 160 kHz（降低单道阈值及触发多重度阈值）
- 数据量（噪声过滤前）：1.1 Gbps = 12 TB/day = 4.4 PB/yr ➔ 3.9 Gbps = 42 TB/day = 15 PB/yr
- 数据量（噪声过滤后）：0.42 Gbps = 4.5 TB/day = 1.6 PB/yr ➔ 1.2 Gbps = 12 TB/day = 4.3 PB/yr
- GRB数据（~3 triggers/week，LAT GCN only）：8.7 TB/burst = 1.3 PB/yr ➔ 30 TB/burst = 4.6 PB/yr

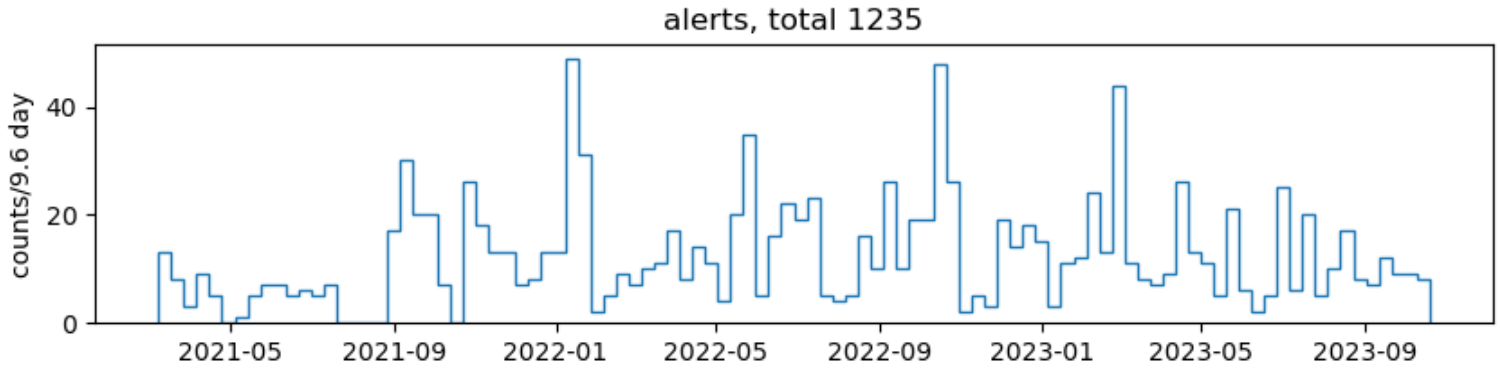
GRB/transient/GW candidate follow-up @ trigger and triggerless data





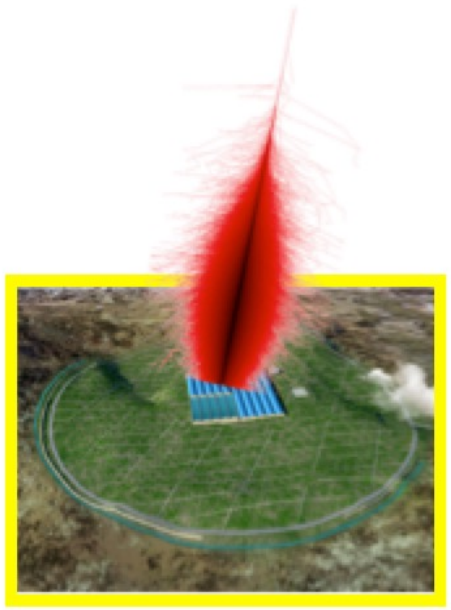
- Receive a GCN alert inside LHAASO FOV
  - Alert rate: 2.5/week
- Save (T0-0.5 h, T0 + 2 h) hours of data
- (Npe, T) of 3120 detector units
- Big data size → 8 TB/alert
- Touching low energy band





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

跨平台软件编译环境ANYSW

含有几乎所有粒子物理相关的软件包

- GCC/G++/GFORTRAN BOOST MYSQL POSTGRESQL CERNLIB CLHEP CASTOR COIN3D JAS SLALIB PAL ROOT CORSIKA GEANT4
- BINUTILS CMAKE AUTOCONF AUTOMAKE READLINE ZLIB XZ SZIP LZ4 TAR OPENSSL SSH2 CURL LIBFFI SQLITE PYTHON2 PYTHON3 OPENJPEG TIFF XMLTO IMLIB2 XML2 XMLTOMAN GIF PNG FREETYPE AFTERIMAGE EXPAT GRAPHVIZ CPPUNIT XERCESC FFTW GSL CFITSIO XROOTD LIBAIO ODBC PYTHIA6 PYTHIA8 DAWN VRML PCRE2 PCRE MONALISA LIBDAEMON INTLTOOL CAIRO AVAHI GL2PS VECCORE VECGEOM HDF5 LAPACK HEPRAPP LIBPAPER GS VIM WT ZEROMQ JSONCPP XSD SPDLOG SOCI REDIS HIREDIS RE2 PROTOBUF LIBEVENT LIBEVENT2 GOOGLETEST
- Most used python3 packages (more than 300)

一个编译脚本，一次完成所有程序包的编译

兼容slc5、slc6、centos7、ubuntu 18.04、almalinux9 (即将) 、应该可以做到支持arm架构

编译环境支持c++14 (gcc 7.3.0)，即将升级为gcc 11.3.1 (>c++20)

多种geant4、root版本

被LHAASO广泛使用

易移植（只要拷贝到一个目录下，然后对环境脚本稍作修改即可）

使用方法（所有平台）：

- source /cvmfs/lhaaso.ihep.ac.cn/anysw/slc5_ia64_gcc73/external/envf.sh

OptParser - a New C++ Class to Parse Configuration & Command Line Options

Setting strings:

- c-string format, like
- optkey = "value|unit=unit1,unit2,…|key=key1,key2,…|text=explanation to this option";

Command line:

- command [cpp options] [parser options] [user options] [parameters]
- The option line can be in one of the forms like: "-optkey value", "--optkey=value", "-optkey+ value", "--optkey+=value", "-optkey", "+optkey", "--optkey=true", "--optkey=false", where the latter few are for switches (options has bool values).

Access to option / parameter values:

- With functions like optchar("optkey"), optint("optkey"), optfloat("optkey"), optlong("optkey"), optlonglong("optkey"), optbool("optkey"), optstring("optkey"), optvdouble("optkey",&n) for options, and optchar(ipos), optint(ipos), … for parameters, where ipos is the parameter position.

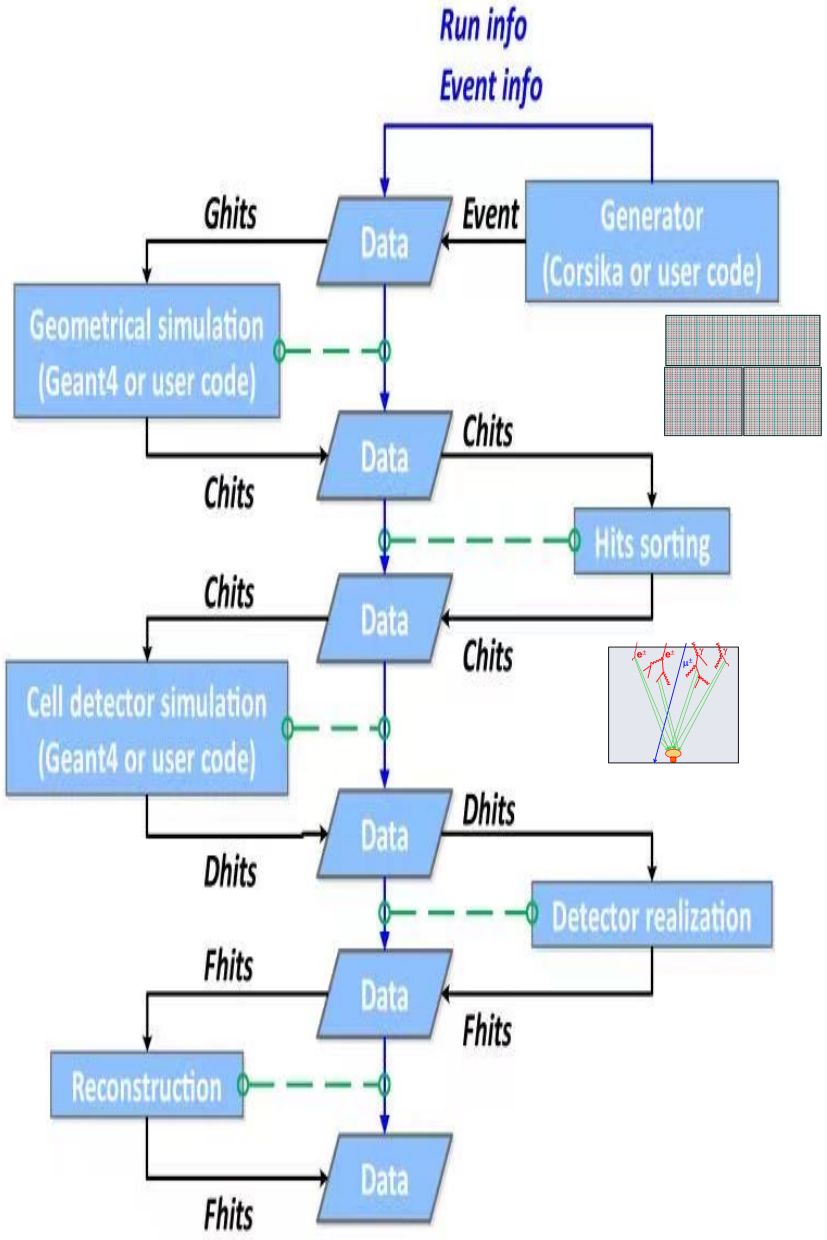
Support cpp in the setting file, where you can include any other files in cpp style and directory control algorithm with "#include", and control part of the setting lines appearing / disappearing with "#ifdef", …

Support physical constants & units that defined by GEANT4/CLHEP;

Support saving all settings to a file, and rerun the program in the completely same conditions later with the saved setting file (as all arguments are saved in the file too);

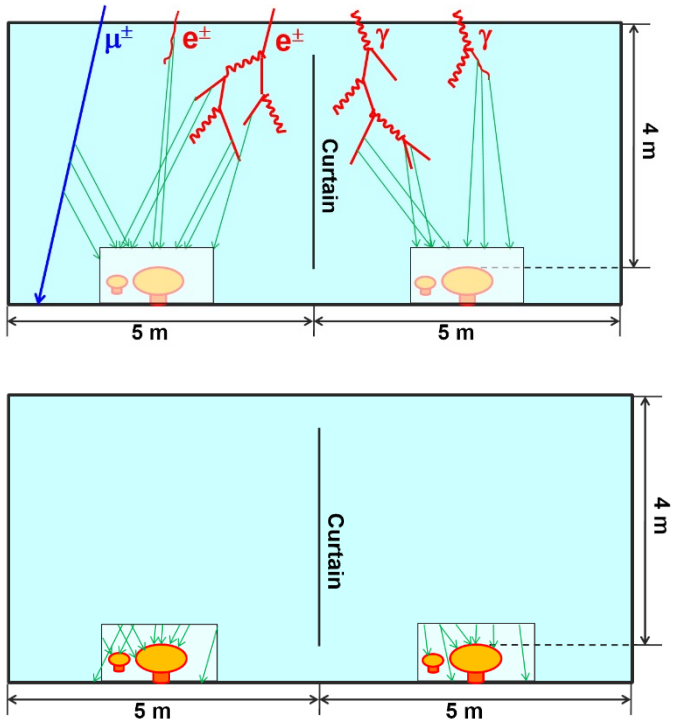
Can be used as a run control tool, and book-keeping the full settings of a job.

G4WCDA: Hits Stream



- Corsika数据的随机读取程（CorsikaReader）
- Hit流处理程序包（HitsReader）；
- 命令行解析、控制参数和数据库调用程序包（OptParser）。

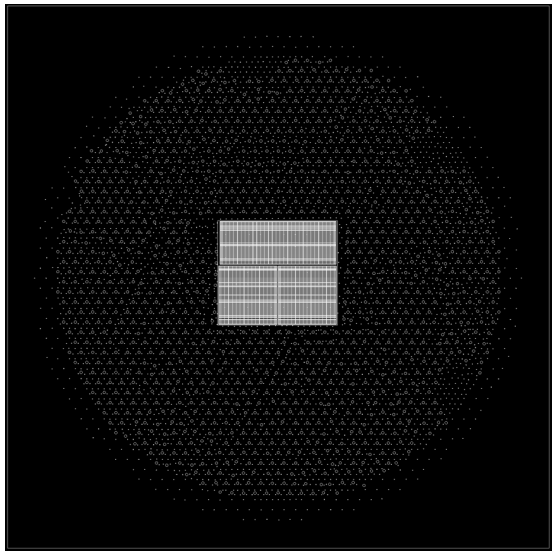
- 4 kinds of hits:
  - Ghit: generator hit;
  - Chit: cell hit;
  - Dhit: detector hit;
  - Fhit: final hit.
- Hit stream:
  - In: a batch of hits;
  - Out: a hit.
- Storage & buffering:
  - ROOT tree

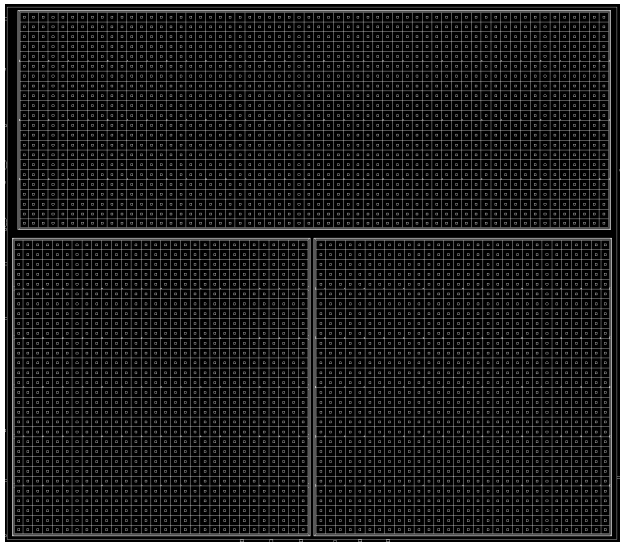


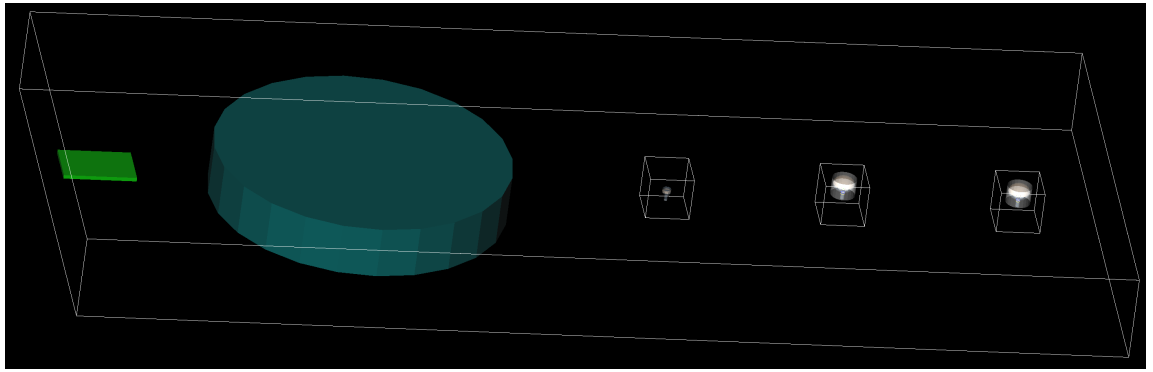
- 解决了内存耗尽
- 优化中间结果的存储
- 易于探测器真实化
- 简化各类探测器的统一模拟。
- Example: version >2.0

G4WCDA

- **Geometry setup**
  - applied with the full & detailed structure of the buildings, roofs, walls, pillars, electronic boxes, curtains, liners and supporting bars;
  - include also ED boxes, MD cylinders
- **PMT model**
  - Muti-film model, shape and optical parameters
- **Water absorption & scattering**
  - No absorption & Mie scattering, only typical/calculated Rayleigh scattering
  - but photon track lengths are recorded.
- **Particle/Photon thinning**
  - Cherenkov lights in the production can be thinned if plenty of lights have been generated;







pipeline of data processing

DAQ触发判选，数据传输至高能所

数据解码为root文件

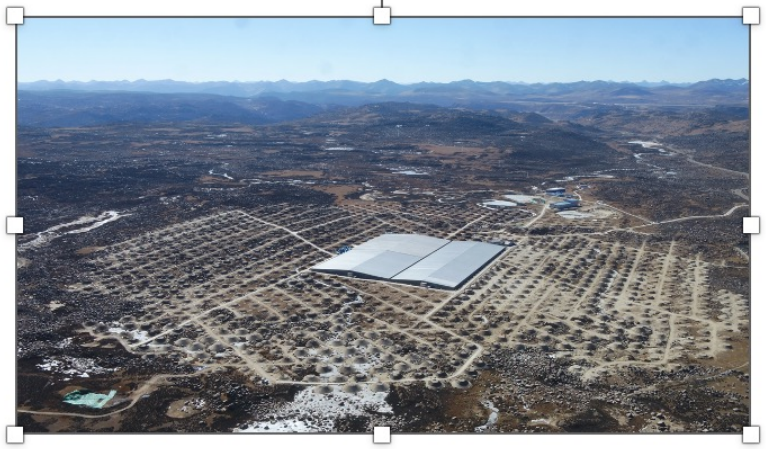
数据标定

单元探测器质量监测

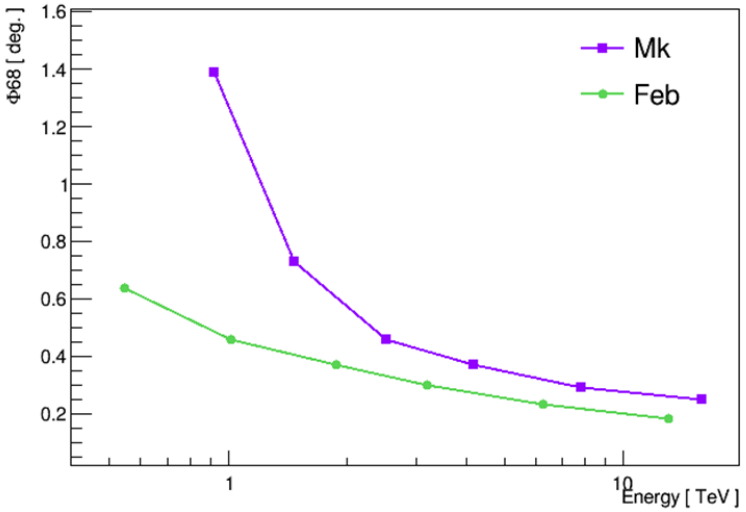
数据重建

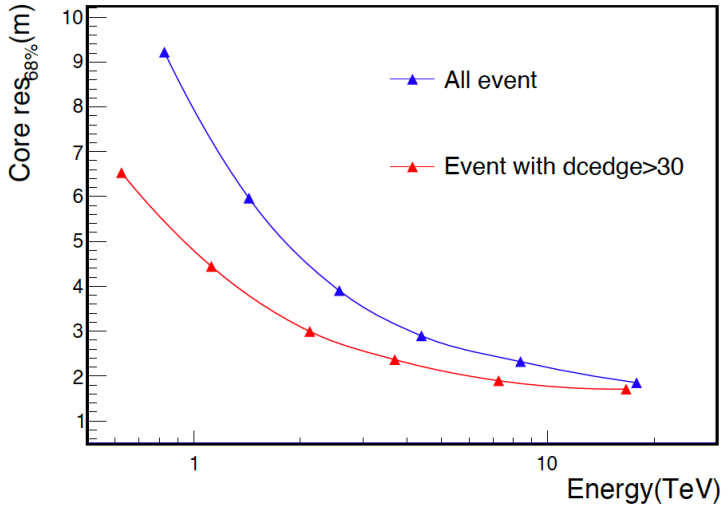
重建文件质量监测

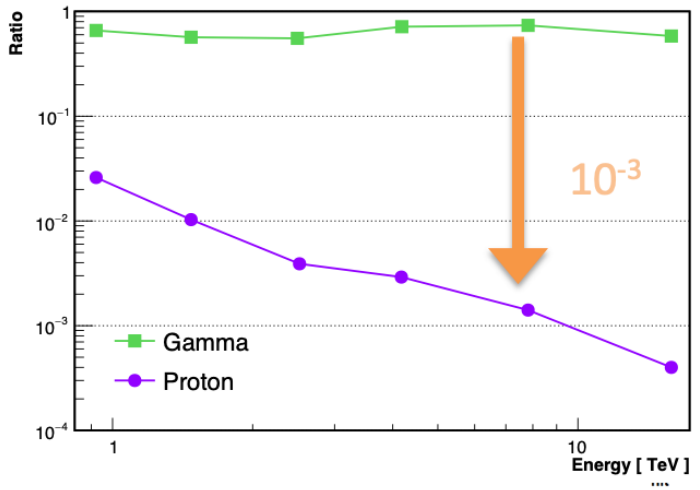
各种物理分析

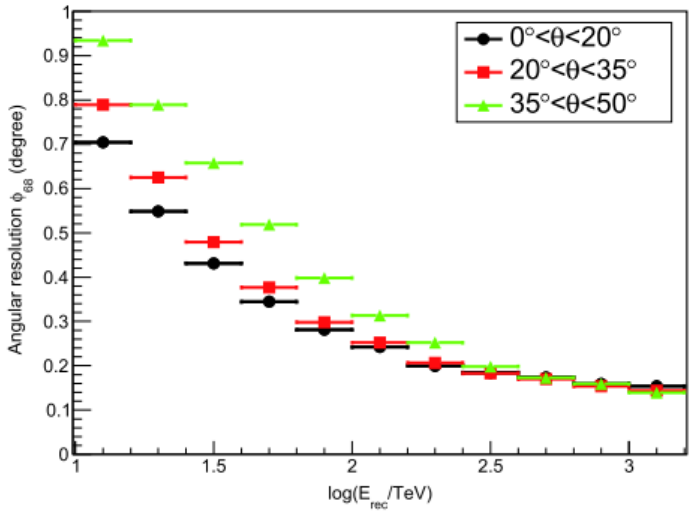


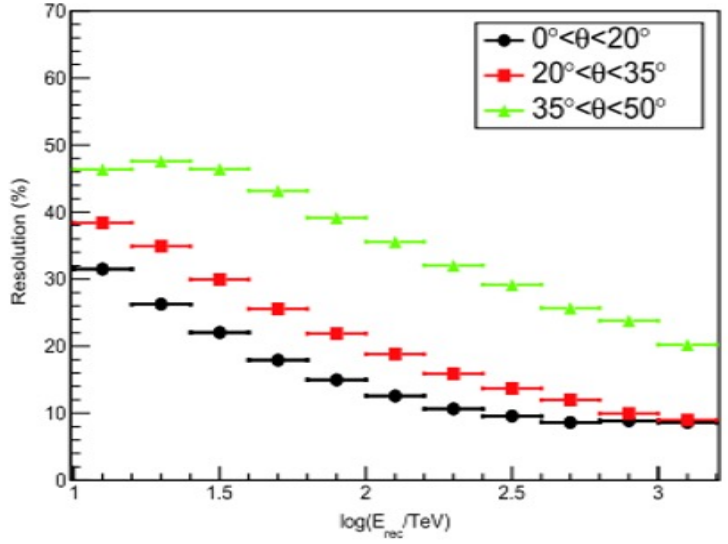
Shower reconstruction resolution

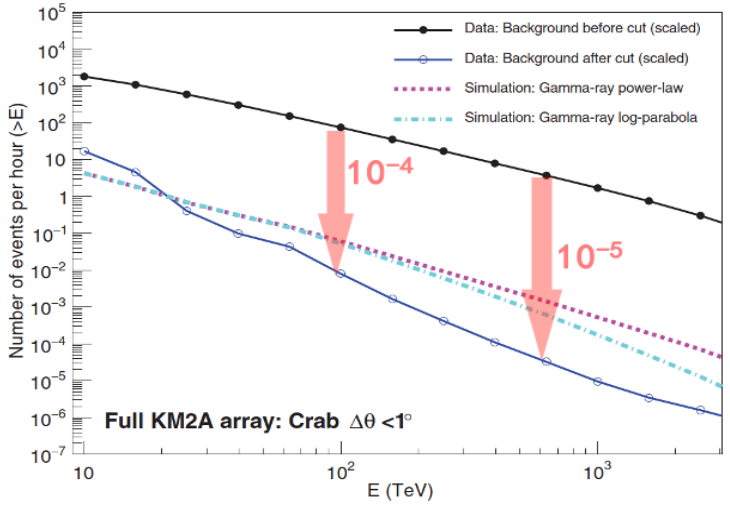




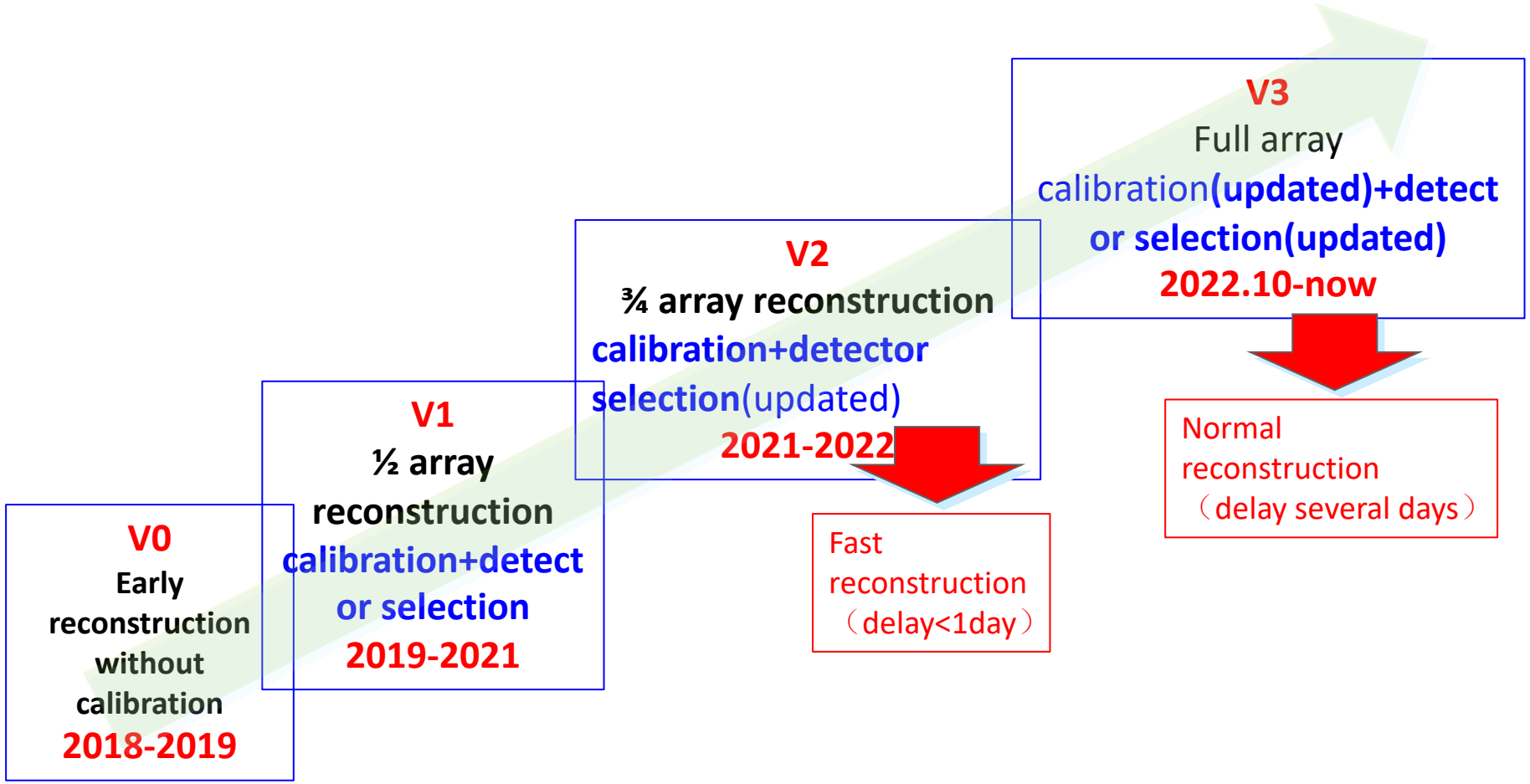








Update history of reconstruction software @ KM2A



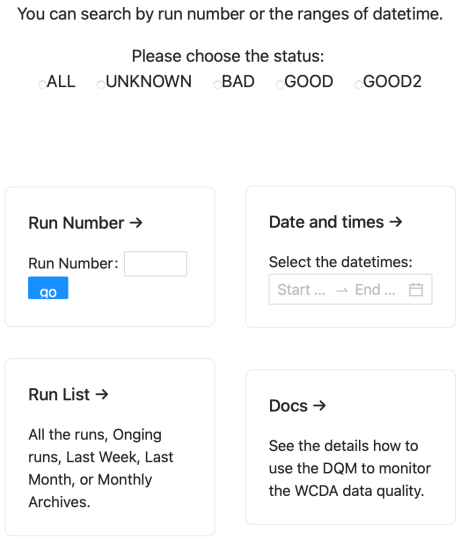
Reconstruction Data Quality Check

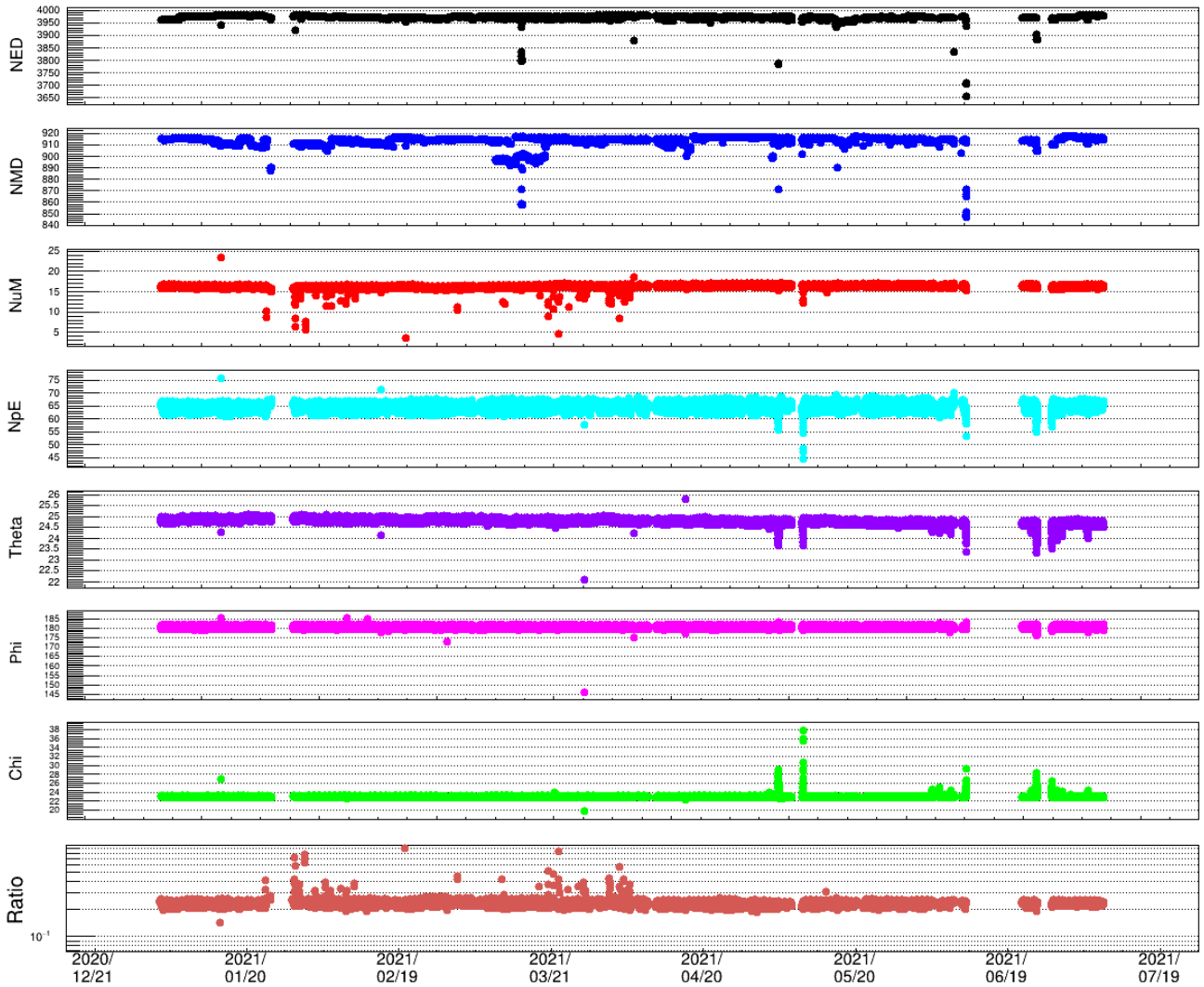
■ 重建文件情况：

- KM2A：2200 files/day；
- WCDA: 4500 files/day

■ 重建文件监测：

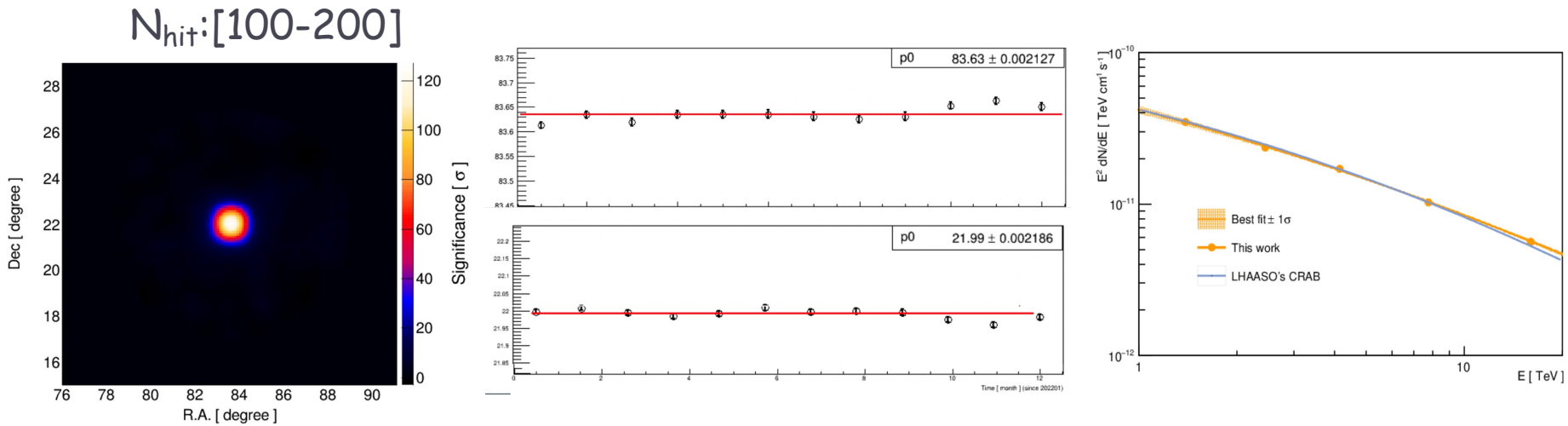
- 参与重建的ED、MD数目、缪子数、电磁粒子数、重建天顶角、方位角 ……



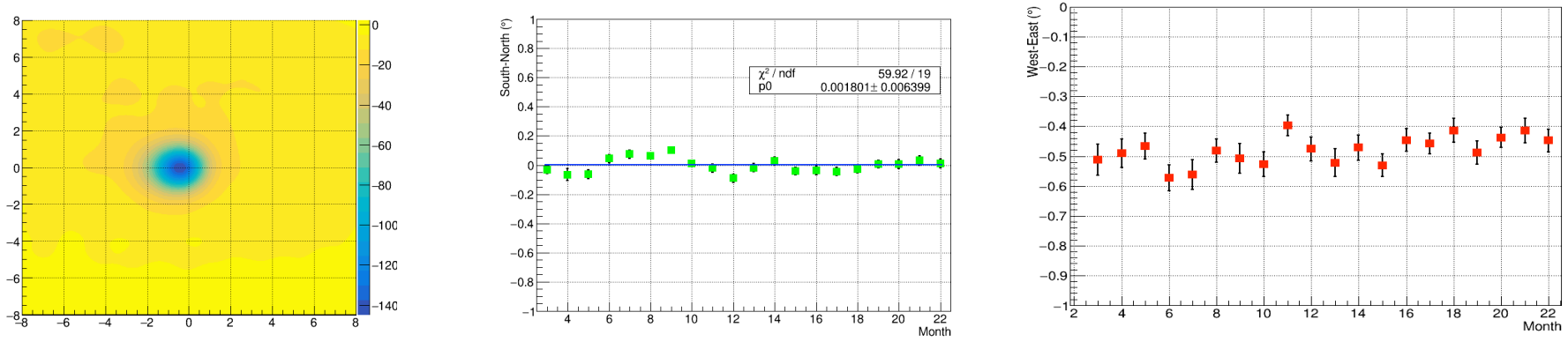


- A good-file-list can be collected;
- Some detailed parameters informaiton can be checked.

Crab and Moon shadow monitoring @ WCDA



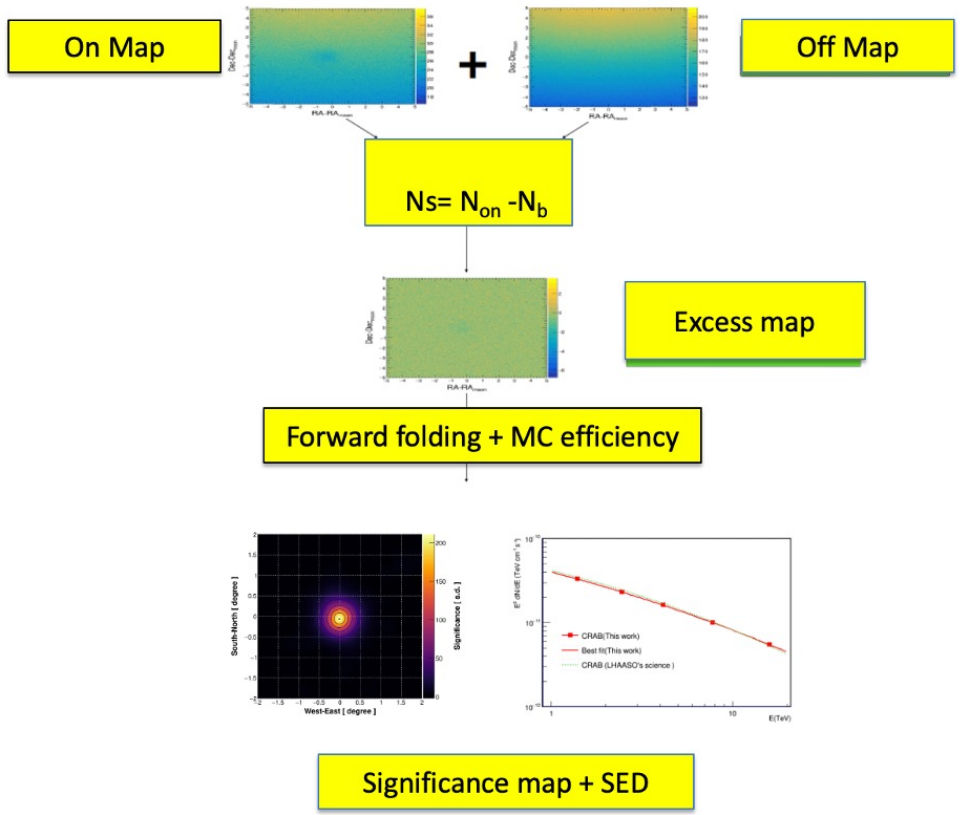
Pointing error <0.02 deg && SED is consistent with LHAASO science result.



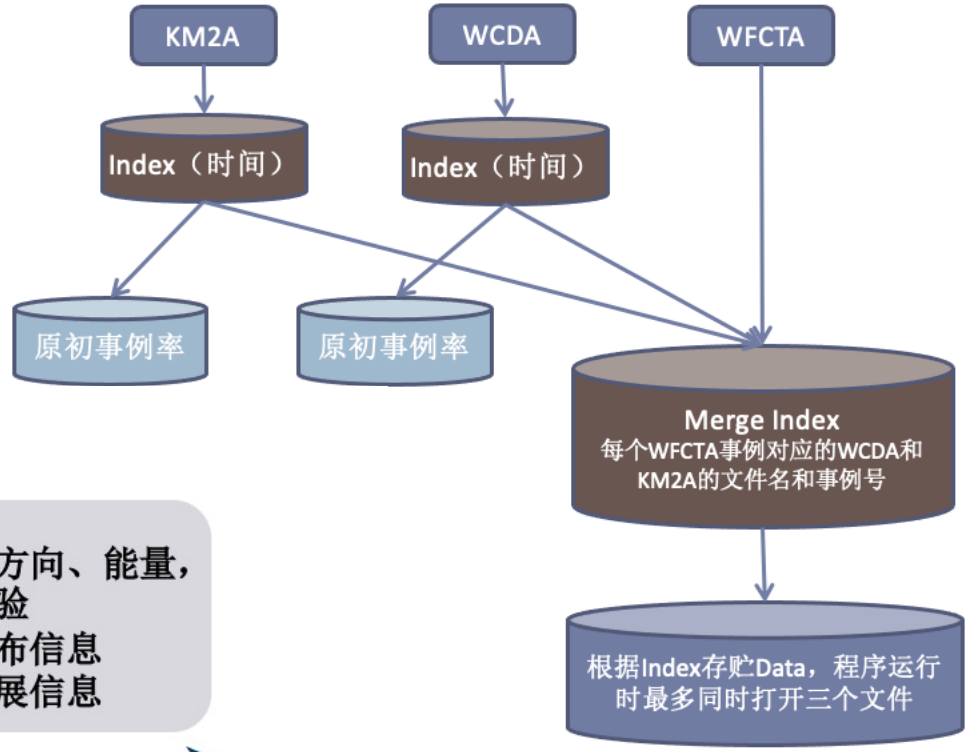
- N_{hit}>100 pointing error <0.01 deg @ moon shadow

LHAASO Gamma-Ray Data Products

- **Releasing version**
  - **1:Code/ 2:Data/ 3:goodlist/ 4:Simulation/ 5:Skymap/**
- **Three reconstruction data products**
  - Recdata/ -→ Standard reconstruction data
  - Recgdata/ → Gamma-like reconstruction data
  - Sampdata/ → specific sample data around the sources(crab)
- **Two scientific data products in root format**
  - One skymap data in root
  - One simulation samples in root
- **One scientific analysis tool.**



WFCTA：hybrid data collection



数据符合

- 三个阵列事例时间索引
- 建立自动化运行脚本

2019年：Merge_v4
2020年：Merge_v5
2021年：Merge_v6

数据标定

- 全面掌握三个子探测器的标定进展(13个文件)

2019年：CaliMerge_v4
2020年：CaliMerge_v5.2.2
2021年：CaliMerge_v5.3.2
2022年：CaliMerge_v5.4

联合重建

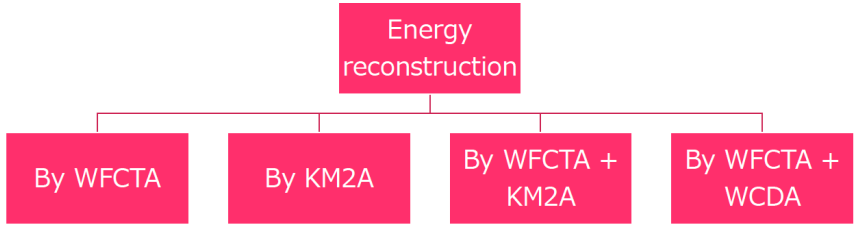
- 芯位、方向、能量，相互检验
- 横向分布信息
- 纵向发展信息

2019年：RecLHAASO_v5
2020年：RecLHAASO_v5.3
2021年：RecLHAASO_v5.4
RecLHAASO_v5.5
2022年：RecLHAASO_v5.6

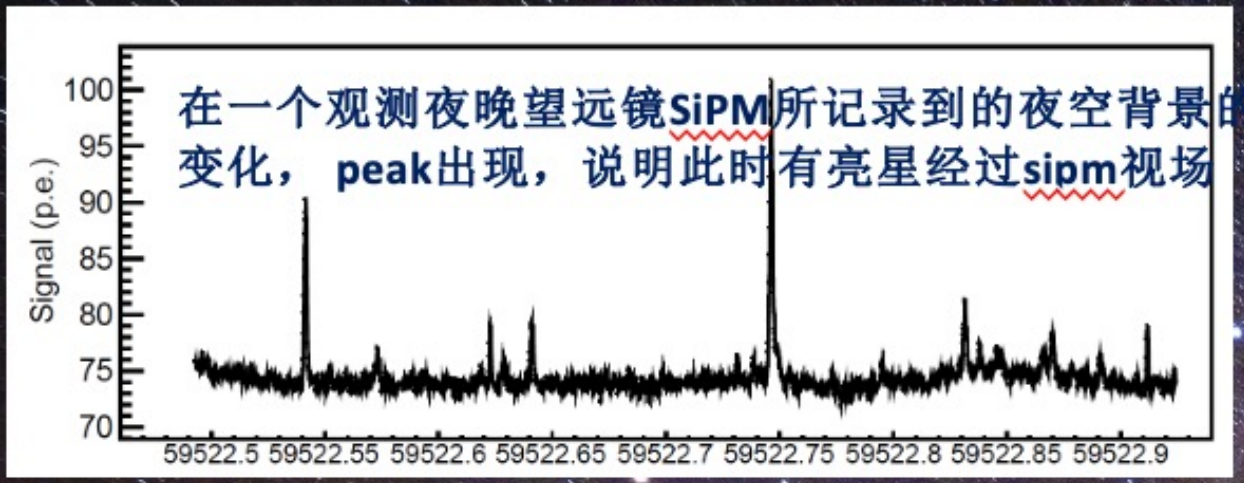
- 标定和重建9个版本的程序中，共40+项程序优化
- 建立Public函数库，提供统一的Read程序和数据说明文档

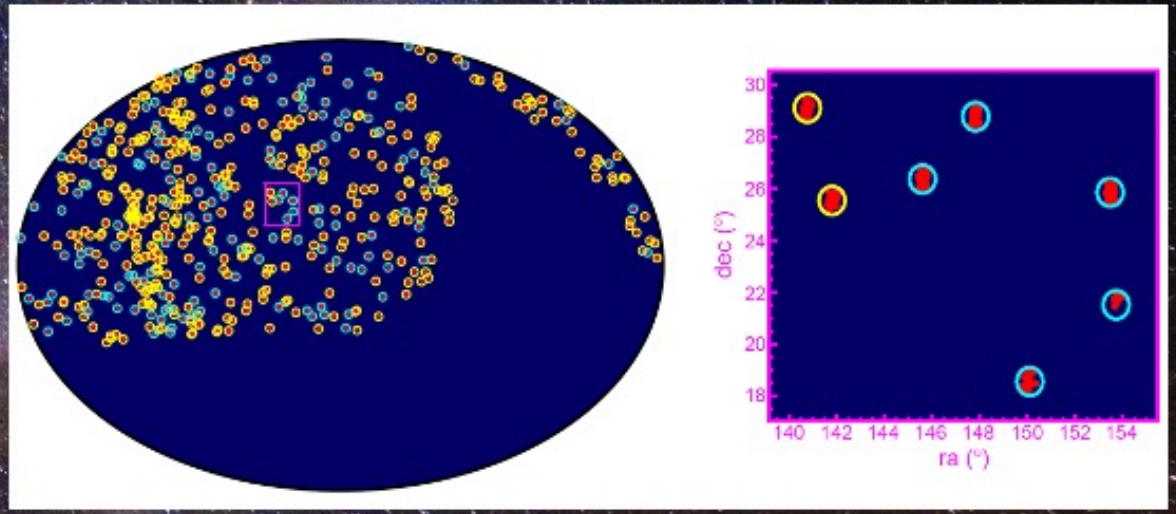
符合数据的模拟

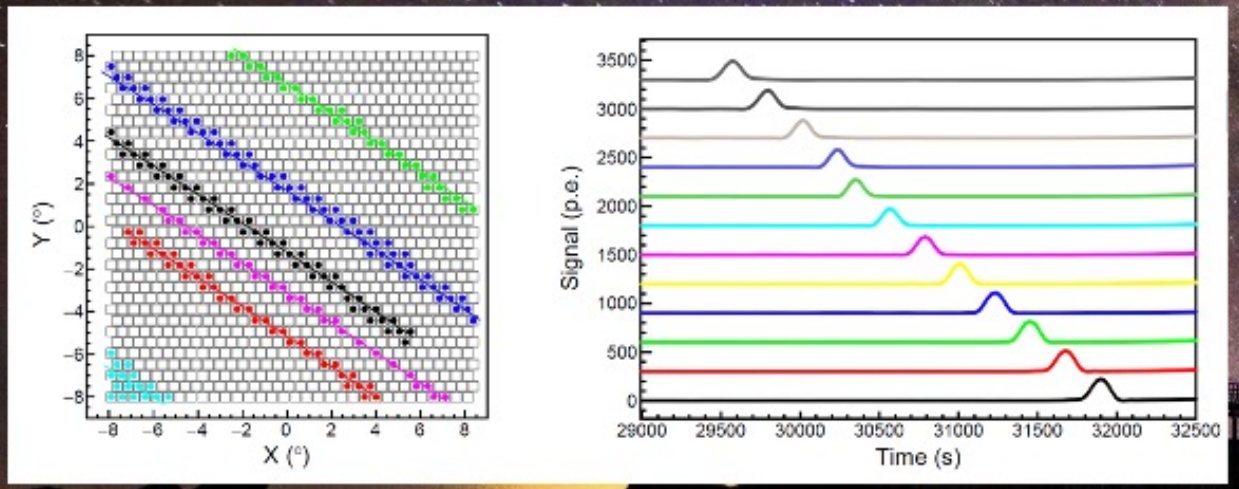
- 物理目标：研究宇宙线的分成分能谱
  - 通过模拟数据研究探测器探测宇宙线的有效面积
  - 通过模拟数据研究对宇宙线的能量重建方法，挑选宇宙线的方法，挑选宇宙线的能力等
- 模拟数据的特点
  - 多种探测器对宇宙线的响应都需要模拟
  - 缩小模拟引入的系统误差，所需统计量大
  - 高能宇宙线事例模拟时长消耗大
  - 成分：质子，氦核，CNO，MgAlSi，Iron
  - 宽能量范围：分能量段模拟，10TeV-100TeV，100TeV-1PeV, 1PeV-10PeV
  - 相互作用模型：分别模拟了QGSJETII-04,EPOS-LHC 两个相互作用模型，以研究强子相互作用模型引入的系统误差
  - 根据望远镜的观测阶段产生了两批模拟数据
    - 第一阶段：天顶角：20°~40°， 777TB，3900000文件数
    - 第二阶段：天顶角：35°~55°， 559TB，1500000文件数



望远镜指向标定: 夜空中最亮的星







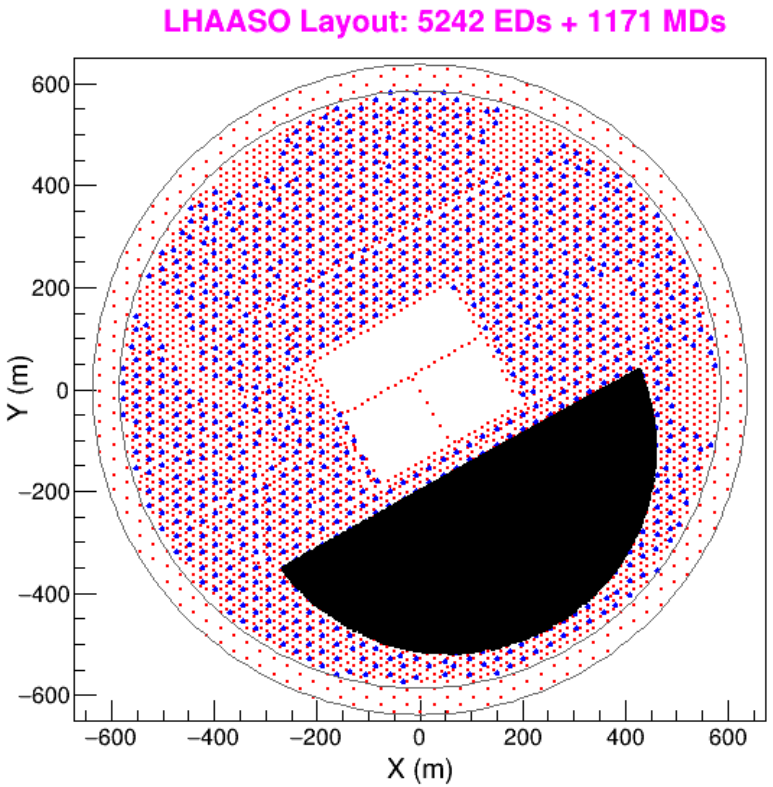
WFCTA记录下记录下的6颗星的径迹（点）
与恒星的在某一望远镜指向下的径迹

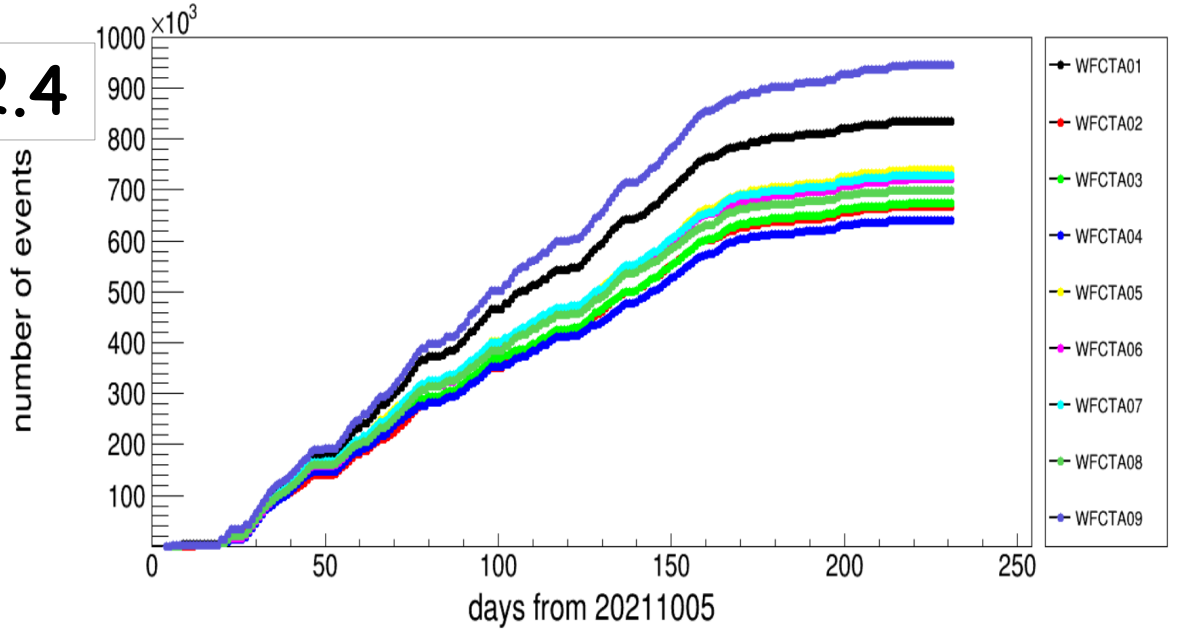
方法特点：

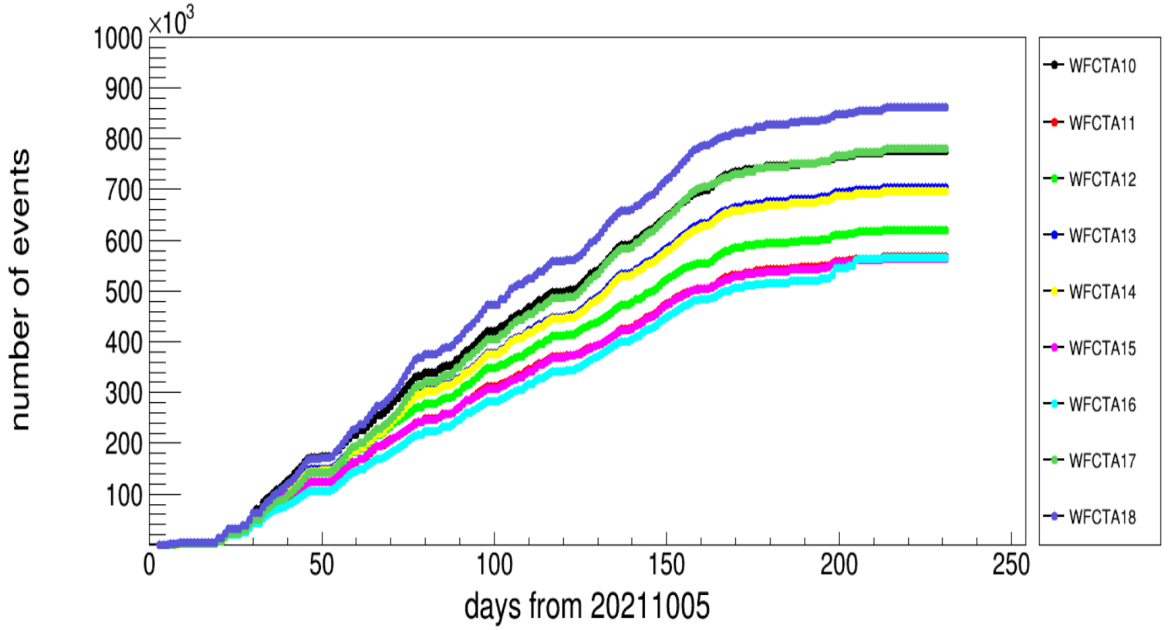
用望远镜自己的观测数据，一个观测夜晚的数据可以完成标定，用时约10分钟；
有5颗星时，指向精度约0.02度
有15颗星时，指向精度可以达到0.01度

- 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

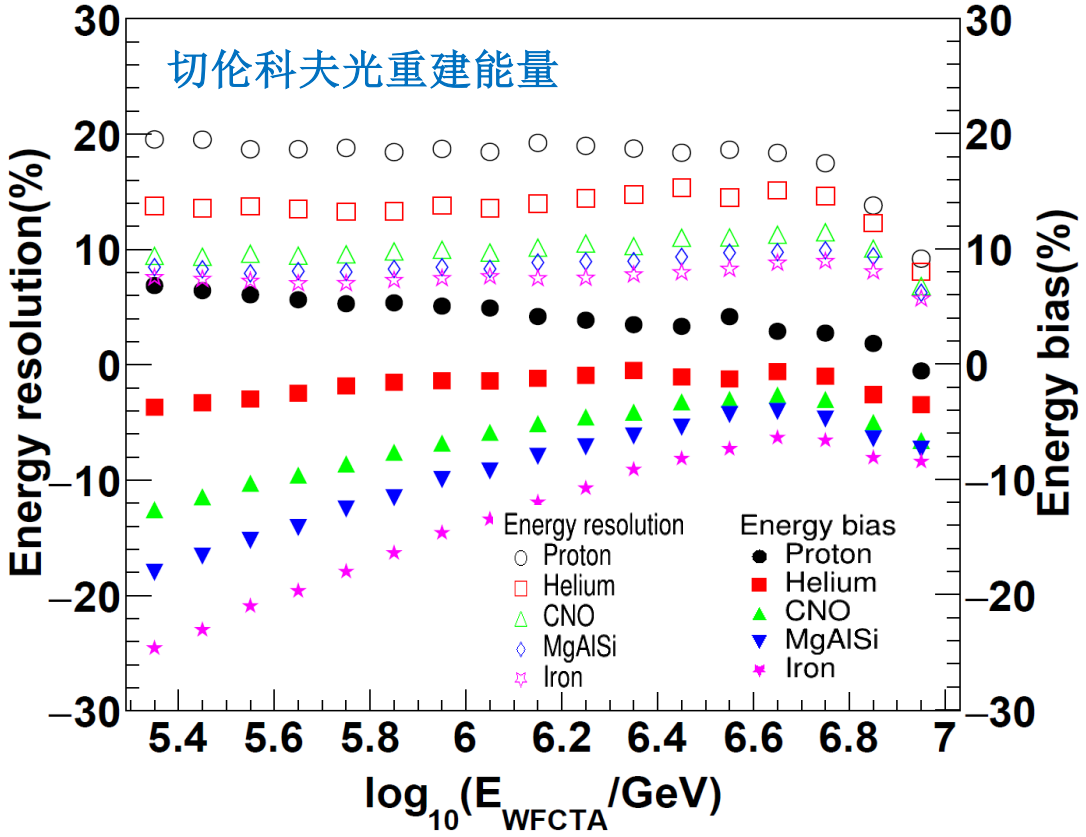
Observation time: 2021.10 - 2022.4

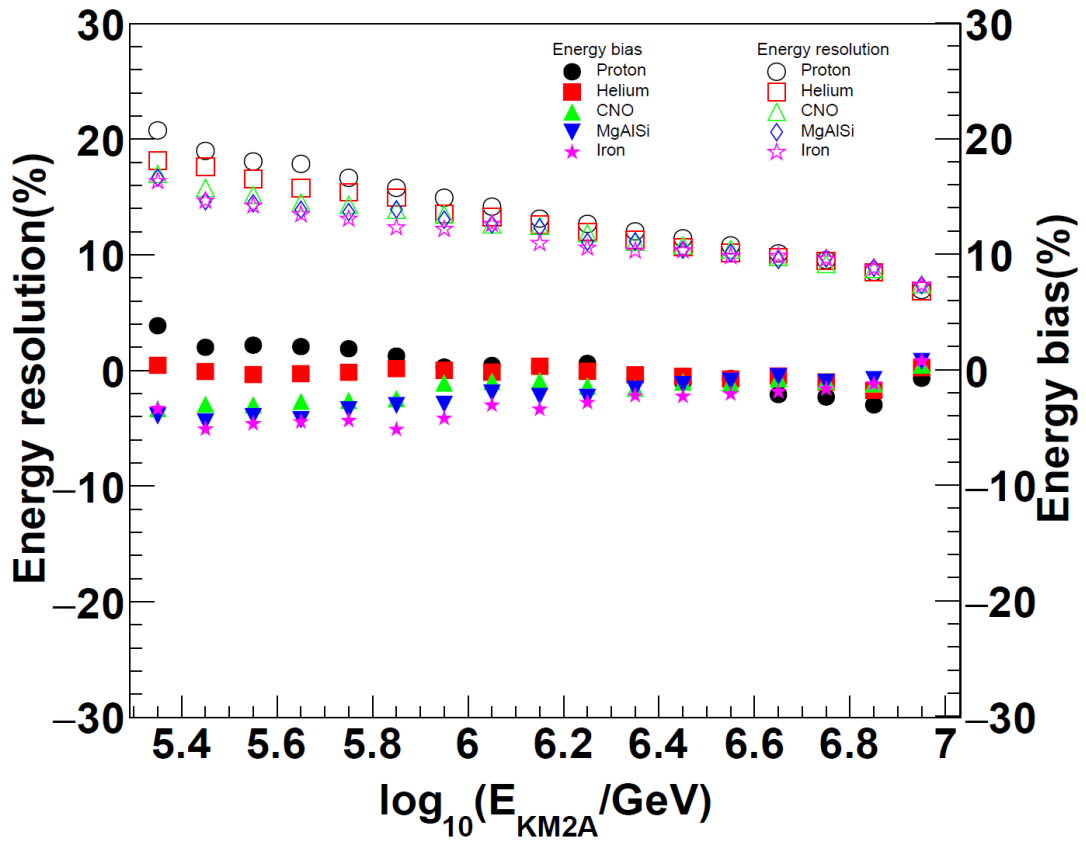




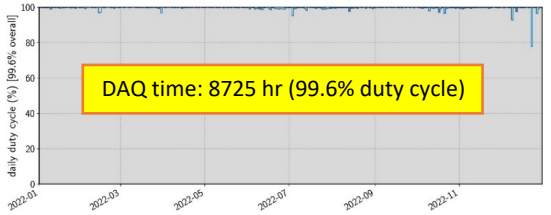


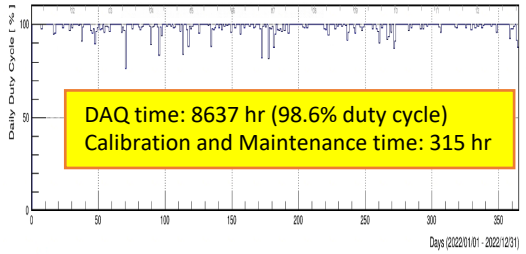
Energy reconstruction

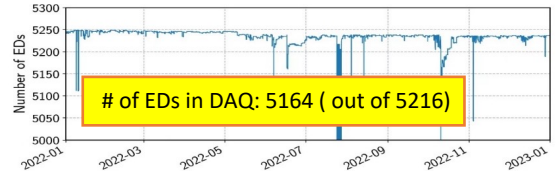


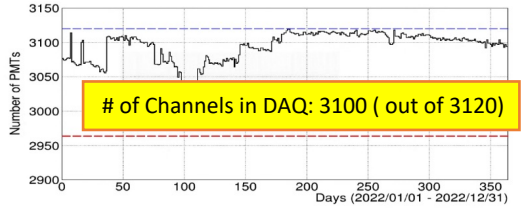


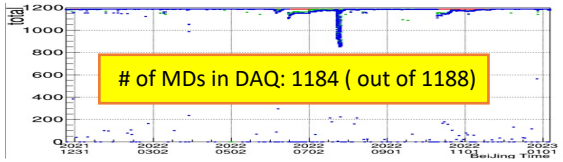
Supper Stable & Fruitful Operation

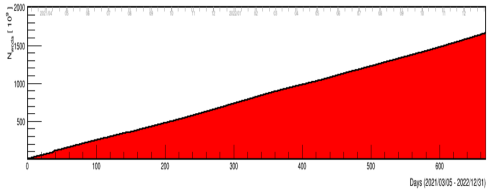


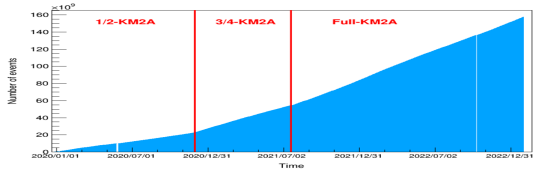


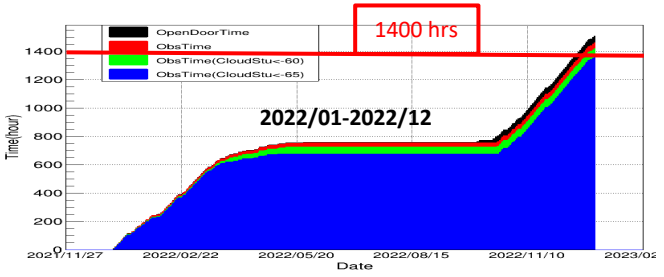








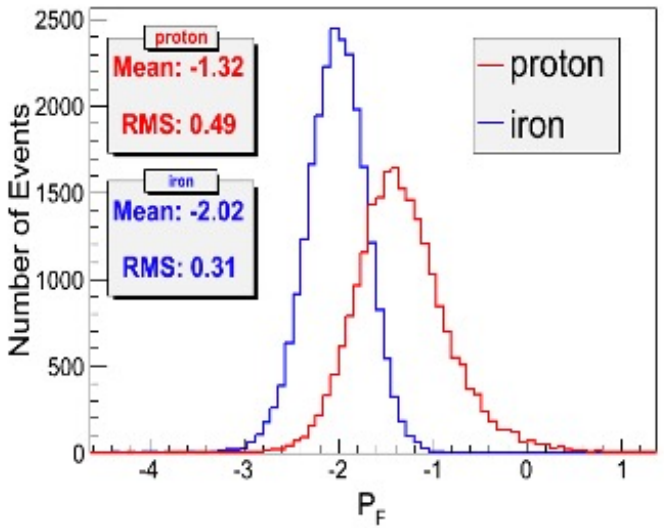


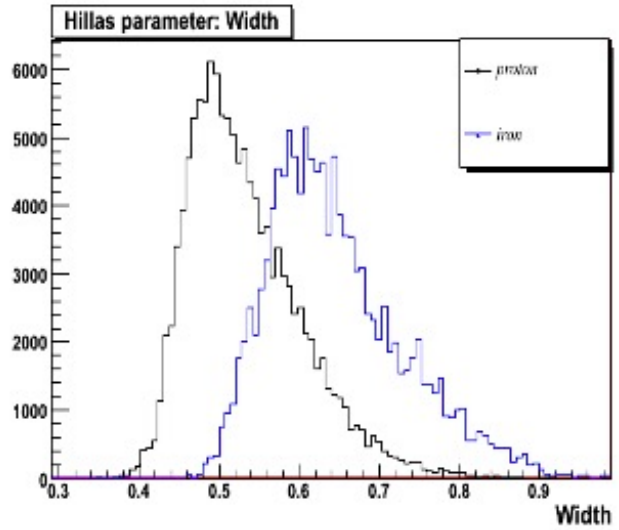


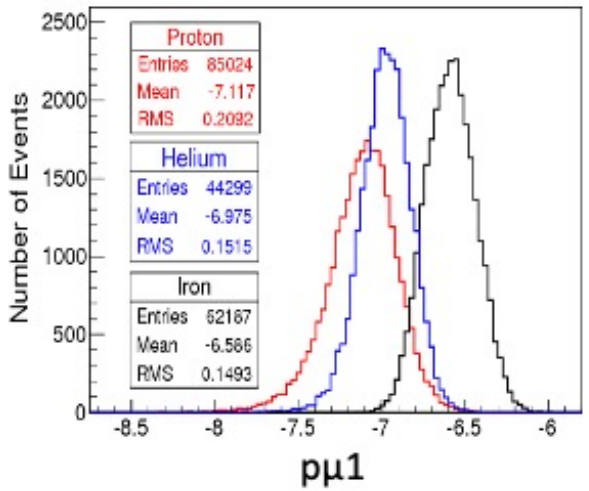
Reconstruction and Analysis

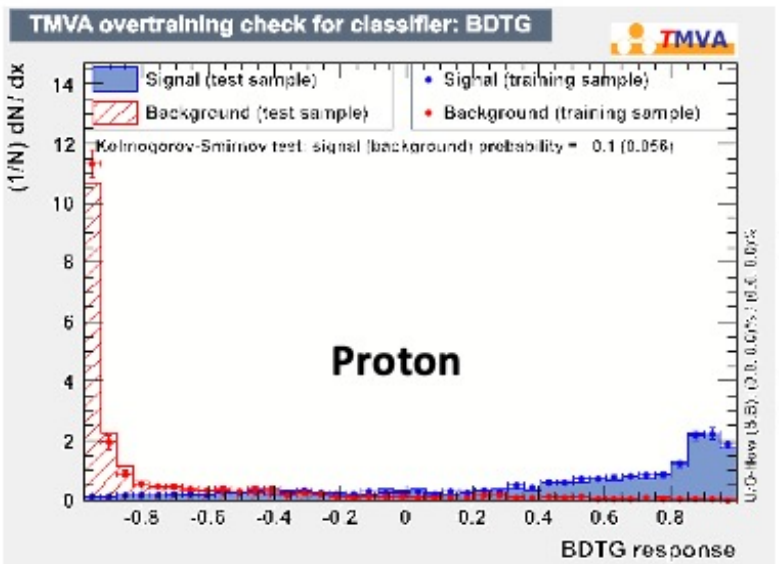
- Data procession
  - # of events: 1.e12 LE、1.5e11 HE、70 million hybrid
  - Amount: 11 PB
- Simulation
  - # of events: 1 billion LE、0.7 billion HE、150 million hybrid
  - Amount: 4 PB
- # of jobs: 10M for data, 50M for simulation

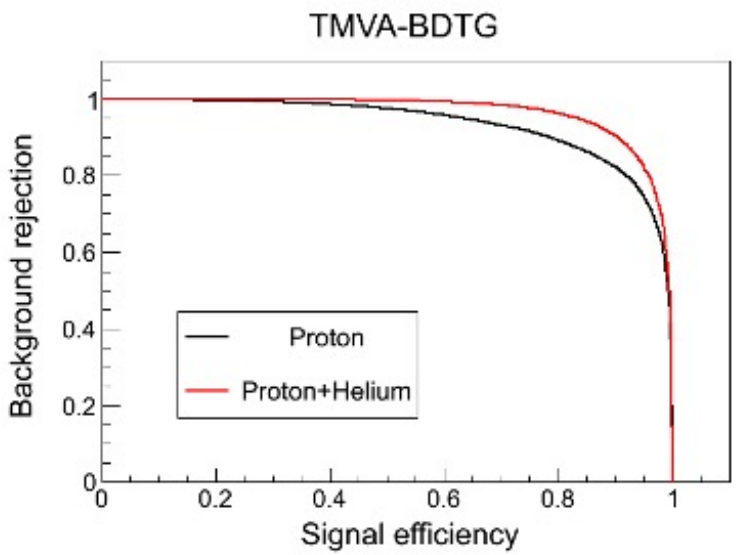
Particle Identification @ WFCTA

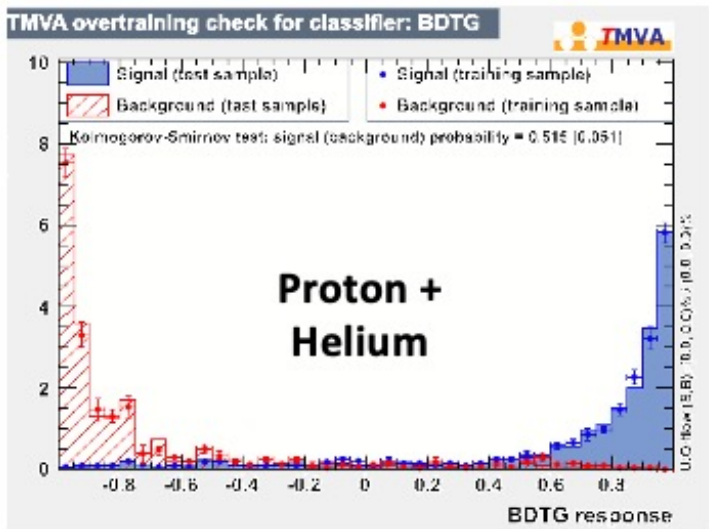








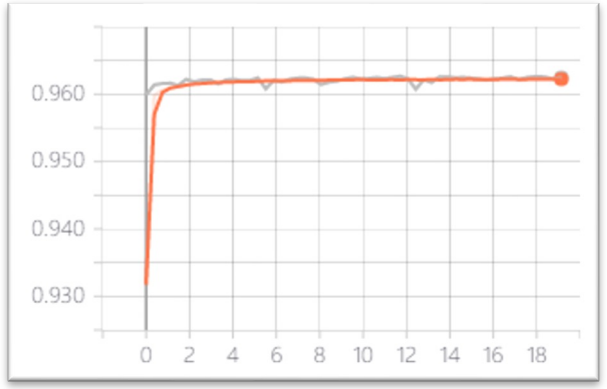


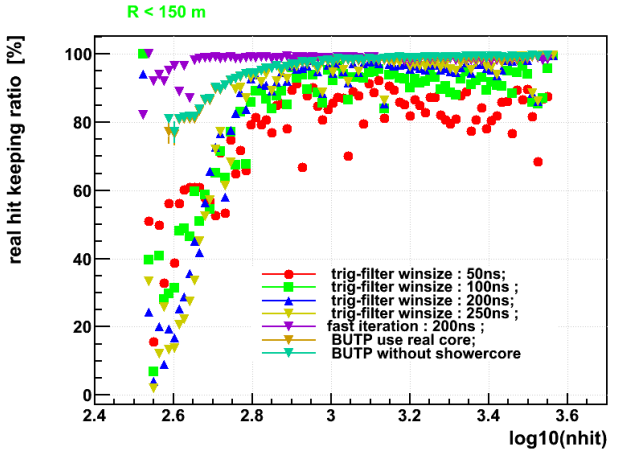


WCDA信号监测--基于CNN的分割方法概述

- CNN是一种数据驱动的算法，数据和计算力是CNN的关键。基于大量数据，CNN利用反向传播算法优化网络中的参数，自动学习到输入数据的特征表达，避免了显示地设计和提取特征的过程。

- 15757个事例，随机选取10000个事例进行实验，按照8:1:1的比例分为训练集、验证集和测试集。





- 在(x,y,t)三维坐标系下，借鉴图像实例分割的思路，通过粒子邻域内的空间结构，对粒子类别进行判断。

- 共训练了50轮次，总共19个小时。橙色曲线为训练集Accuracy，灰色曲线为验证集Accuracy，可以发现模型很快收敛。
- 选择在验证集Accuracy最高的模型，对测试集中1000个事例进行测试，总共需要11s左右,平均每秒处理91个事件。

实验机器配置信息	
CPU	Intel(R) Core(TM) i7-4790K CPU @ 4.00GHz
内存	32G
GPU	GeForce GTX TITAN

- 测试集所有粒子统计结果
  - Accuracy=96.27%　　Precision=96.94%　　Recall=97.07%

Summary and Prospects

- LHAASO从2021年7月进入全阵列模式，在此后的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；

- 定期释放数据产品，面向合作组提供三种数据产品、两类科学数据产品和相应的科学数据分析软件；

- 在重建和数据分析等方面，中小模型AI，获得比传统方法更优的结果；
- 开展大模型AI的使用，最终在探测器研发、运行控制、标定与数据分析、多波段多信使联合研究等多方面获得重大突破