高能物理软件框架的发展与展望

黄性涛

山东大学

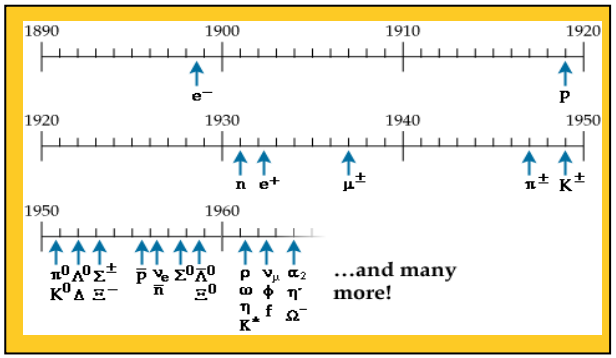
2017年7月4-7日，威海

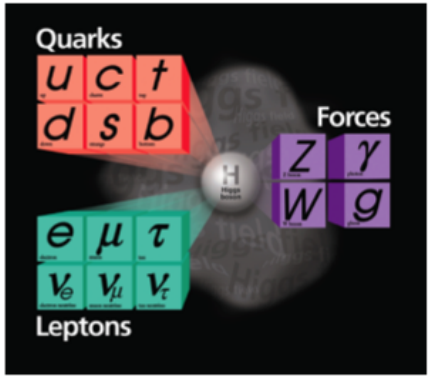
报告提纲

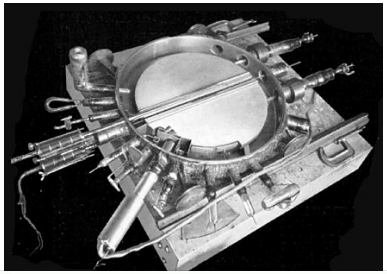
- 产生背景
- 发展过程
- 功能作用
- 发展趋势
- 总结

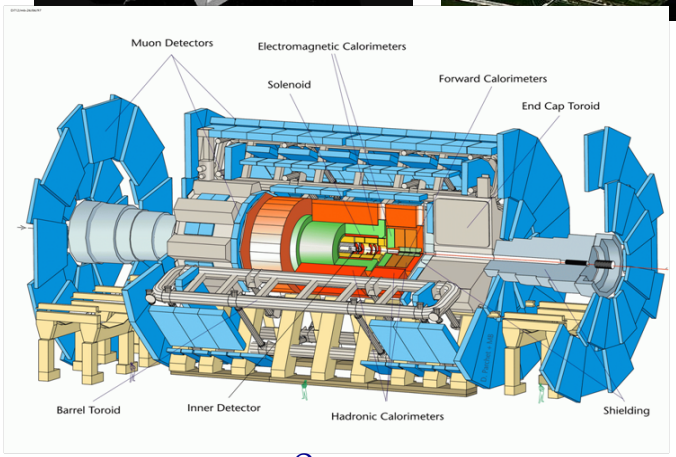
背景

- 高能物理
  - 研究物质最深层次结构
  - 相互作用的前沿基础科学
- 高能物理的发展密切依赖于
  - 加速器
  - 粒子探测
  - 电子学
  - 在线系统
  - 离线软件系统









- 高事例率
- 高信息量
- 高统计量

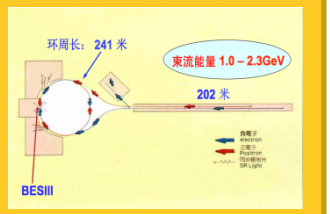
作用

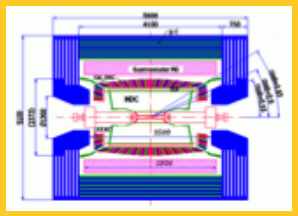
- 离线软件系统是实现从海量实验数据到物理成果转化的关键技术和重要环节

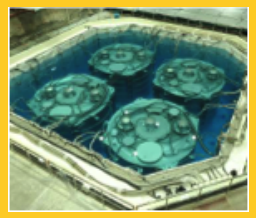
强子物理
BESIII

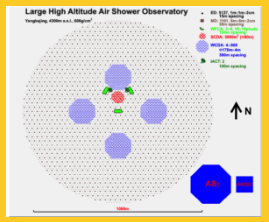
中微子
DayaBay
JUNO

宇宙线
LHAASO





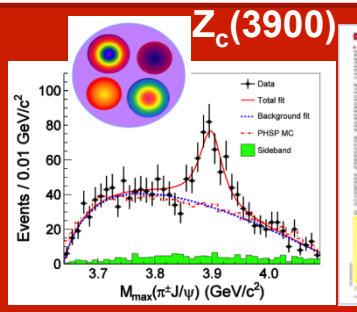




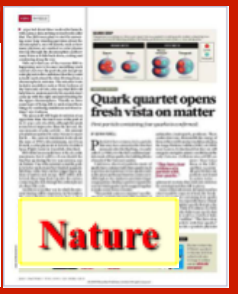
粒子源

探测器

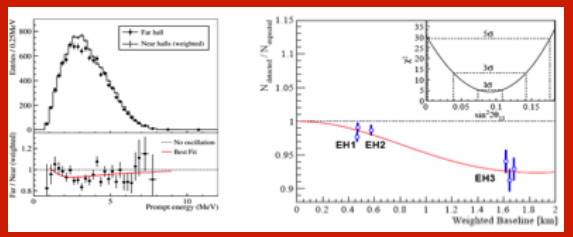
离线软件系统



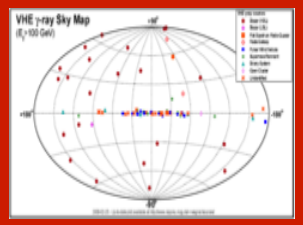
$Z_c(3900)$

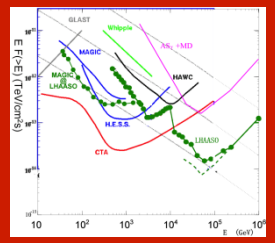


被美国《物理》杂志评为2013年物理学11项重大成果之首

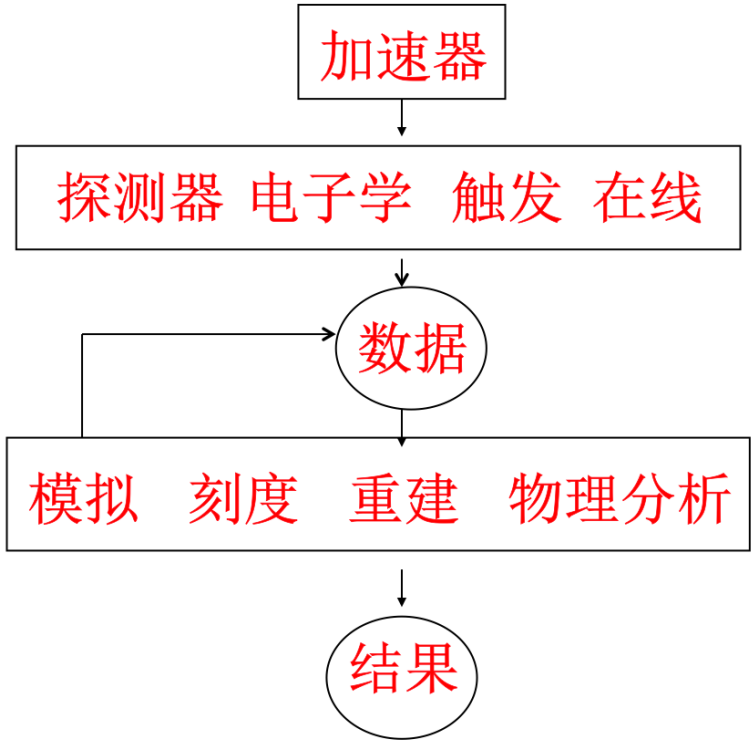


入选《Science》2012年十大科学突破



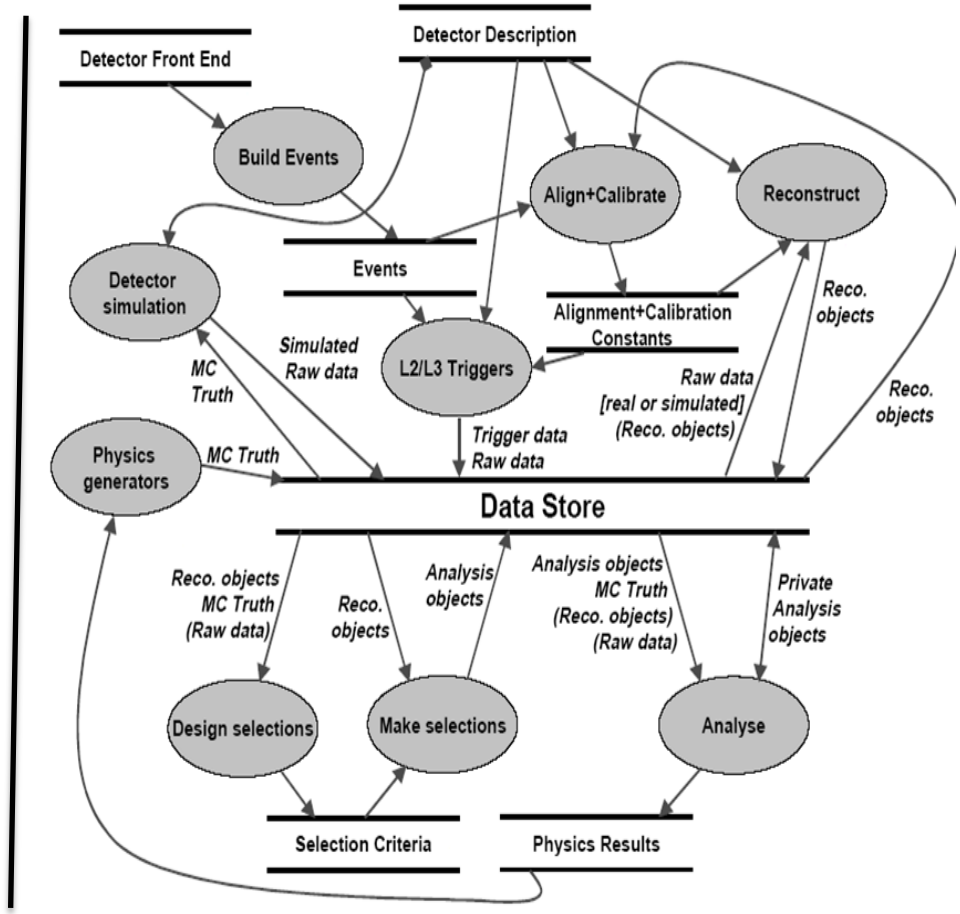


离线数据处理和物理分析过程



硬件

软件



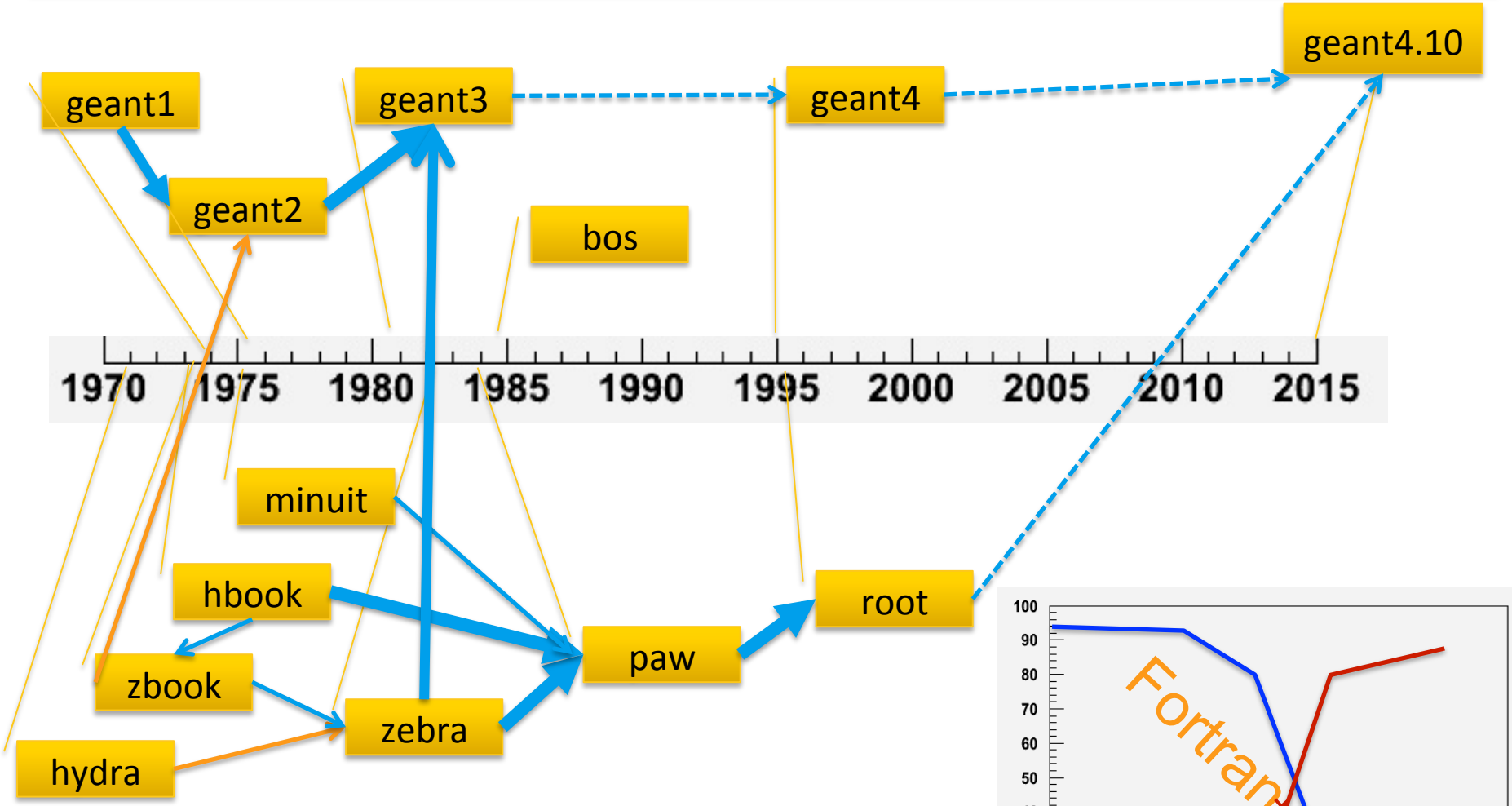
探测器模拟专用软件

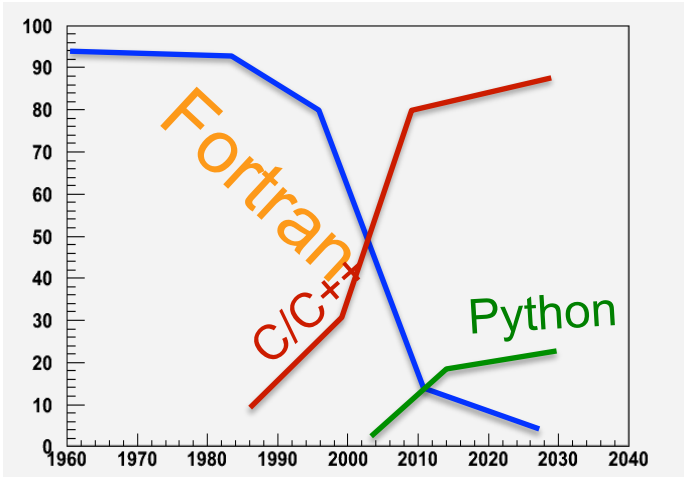
- EGS3/EGS4 （ Electron Gamma Shower ）at SLAC
- FLUKA (FLUktuerende KAskade) at INFN and CERN
- MCNP (Monte Carlo N-Particle ） at Los Alamos NL
- GEANT3/4 (GEometry ANd Tracking) at CERN
  - GEANT3(FOTRAN)
  - GEANT4(C++)
- VMC (Virtual Monte-Carlo System)
  - GEANT3
  - GEANT4
  - FLUKA

数据管理、存储和分析专用软件

- HYDRA (Memory and Data Structure Manager part of CERNLIB)
- ZBOOK (Memory and Data Manager programme part of CERNLIB)
- ZEBRA (CERNLIB package for dynamic memory management)
- HBOOK (Histogram BOOKing, histogramming and statistical analysis package part of CERNLIB)
- MINUIT(Function Minimization and Analysis part of CERNLIB)
- PAW (Physics Analysis WorkStation)
- ROOT (Data Analysis Framework)

发展演化过程


geant1
geant2
geant3
bos
geant4
geant4.10
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
minuit
hbook
zbook
hydra
zebra
paw
root


100
90
80
70
60
50
40
30
20
10
0
1960
1970
1980
1990
2000
2010
2020
2030
2040
Fortran
C/C++
Python

BESI，BESII离线软件系统

Huimin Liu

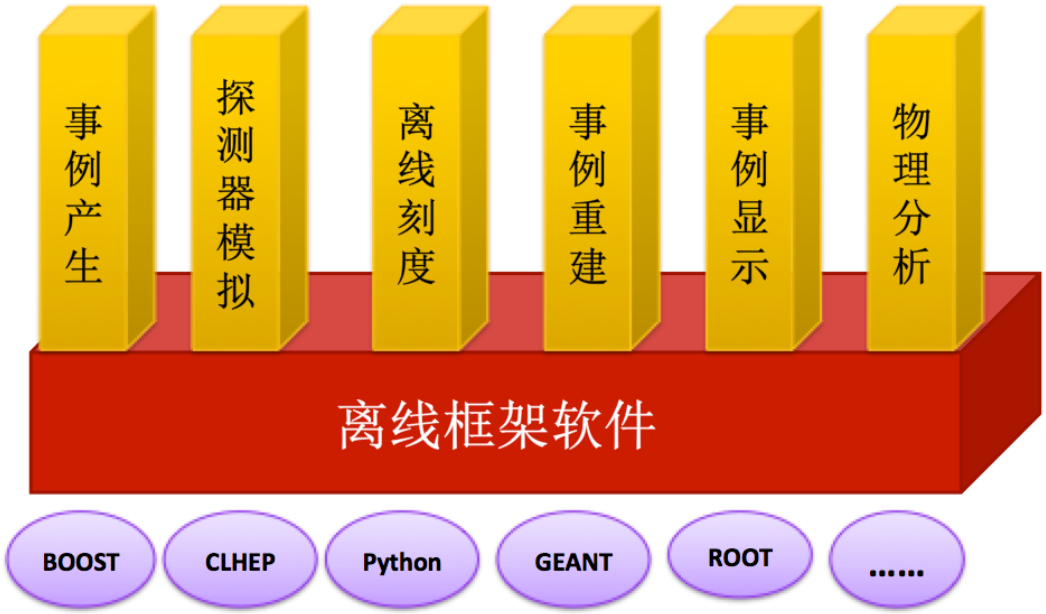
MARKIII	…	BESI	BESII
MOAN	**…**	**DRUNK**	**DRUNK**
GASP	**…**	**SOBER**	**SIMBES**
EGS	*…*	*EGS*	*GEANT3*
F(M)ORTRAN…		*FORTRAN*	*FORTRAN*
1980s		*1980s*	*1990s*

离线软件框架存在的必然性

- 技术：领域内大量专用成熟软件
- 物理：多种探测器信息的综合利用
- 合作：周期长、规模大、专用人才

离线软件框架的功能

- 核心功能
  - 数据管理
    - 数据类型、数据存取、
    - 内存数据、存储数据、输入
  - 流程控制
    - 顺序执行、嵌套执行
    - 串行、平行
  - 公共服务
  - 用户接口
- 统一计算环境和软件平台
  - 编程语言
    - FORTRAN / C / C++ / Python
  - 代码管理
    - PATCHY / SCRAM / CMT/CMake
  - 开发管理
    - CVS / SVN / Git
  - 安装、运行



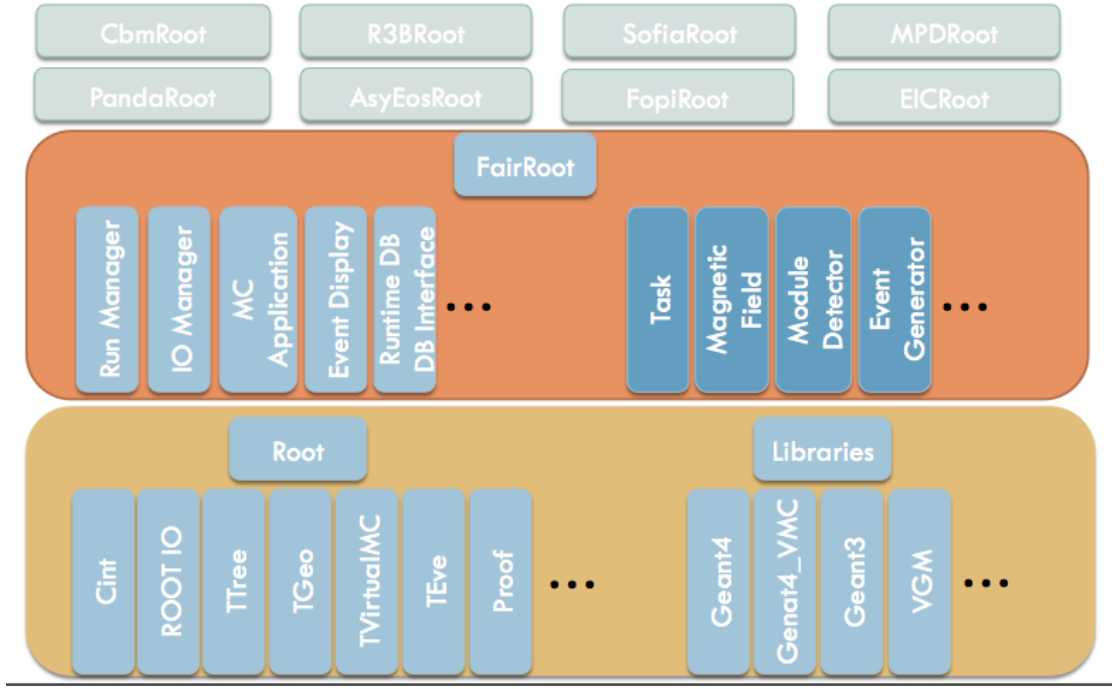
软件框架决定着整个离线软件系统的设计理念、实现方式、性能以及简易程度等。

Main Frameworks

- FairRoot (Germany)
- BASF2 （ Janpan ）
- Art (U.S.)
- Gaudi （ Cern ）
- BOSS （ China ）
- NUWA （ China ）
- SNiPER （ China ）

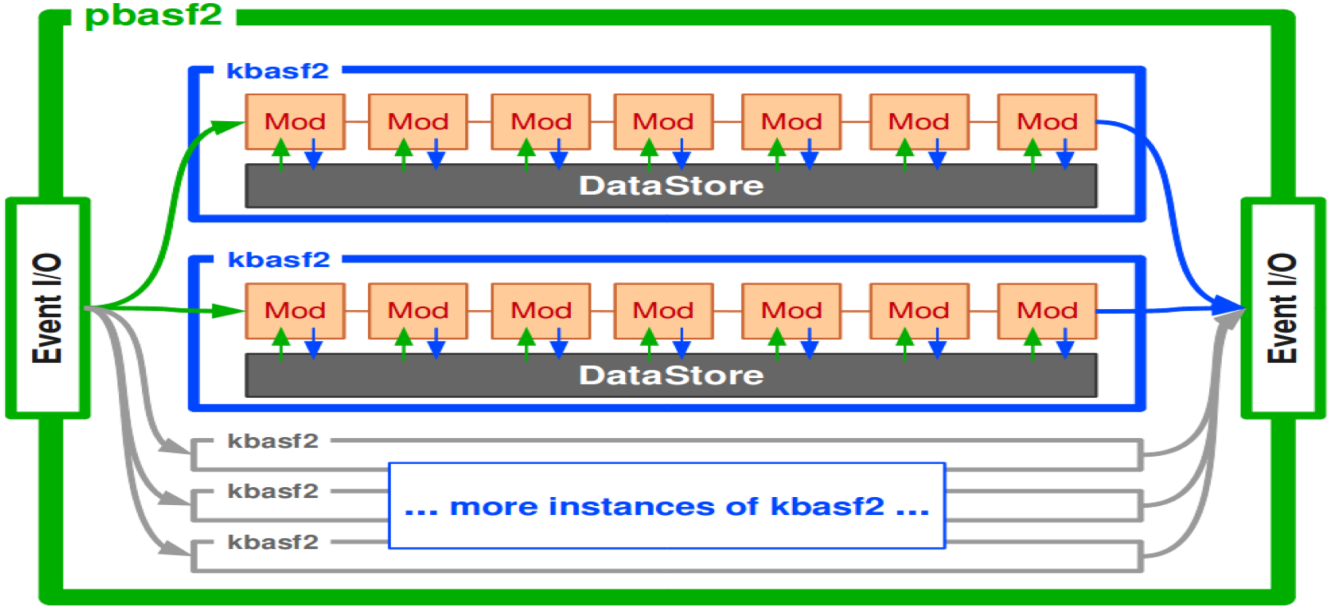
FairRoot

- Developed by GSI-IT fully based on ROOT
- Use plug-in mechanism from Root to load libraries
- Use a dynamic event structure based on Root TFolder and Ttree
- Task , TGeoManager, TEve, TSQLServer are used
- Root macros for the configuration



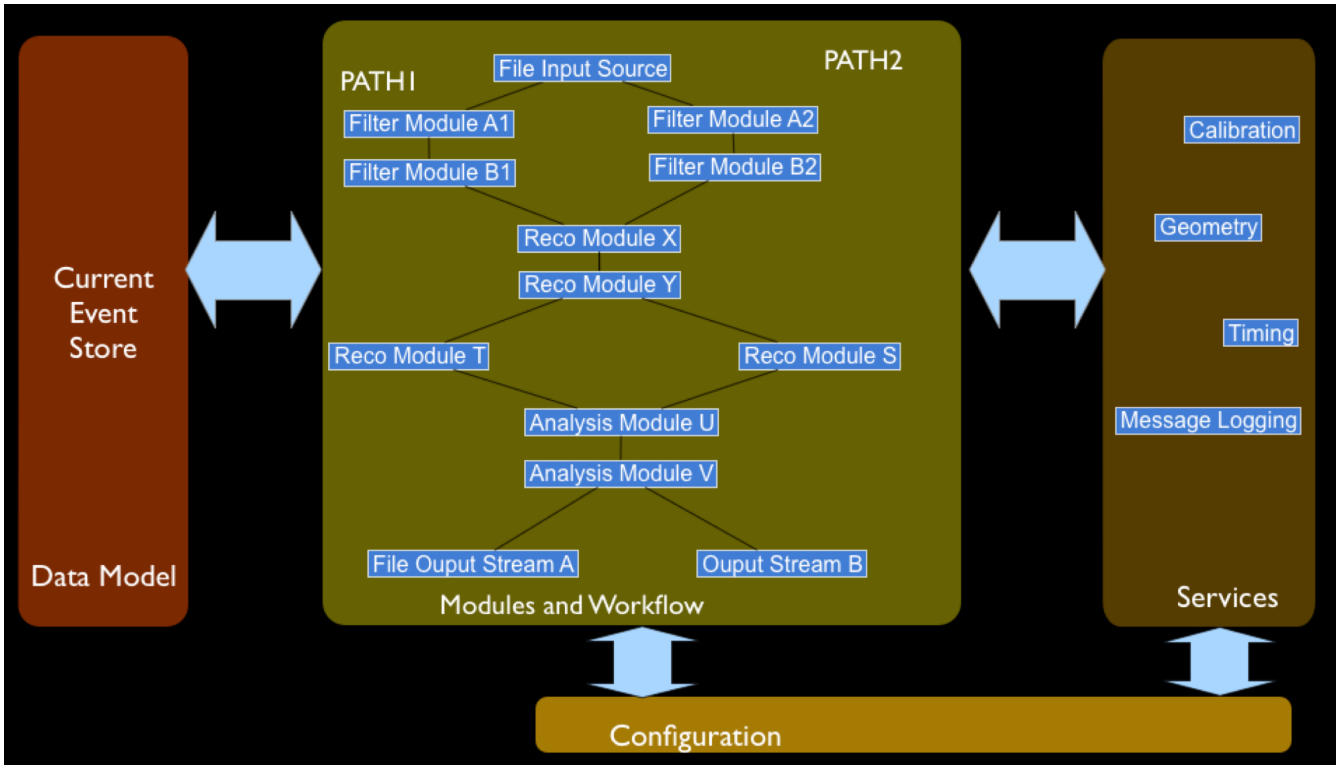
BASF2

- Developed for Belle II for both online and offline
- Data are managed by ROOT and shared with DataStore
- ROOT IO for (de)serialisation of Objects
- Path is a linear arrangement of modules
- Multipaths (processes) are controlled with conditions
- Python for configuration



Art

- Grew from CMSSW and used by g-2, Mu2e, NovA and LArSoft
- Modules include inputs, producers, filters, analyzers and outputs
- I/O and work schedule are handle by a state machine
- Products are managed with DataStore and shared between modules



GAUDI

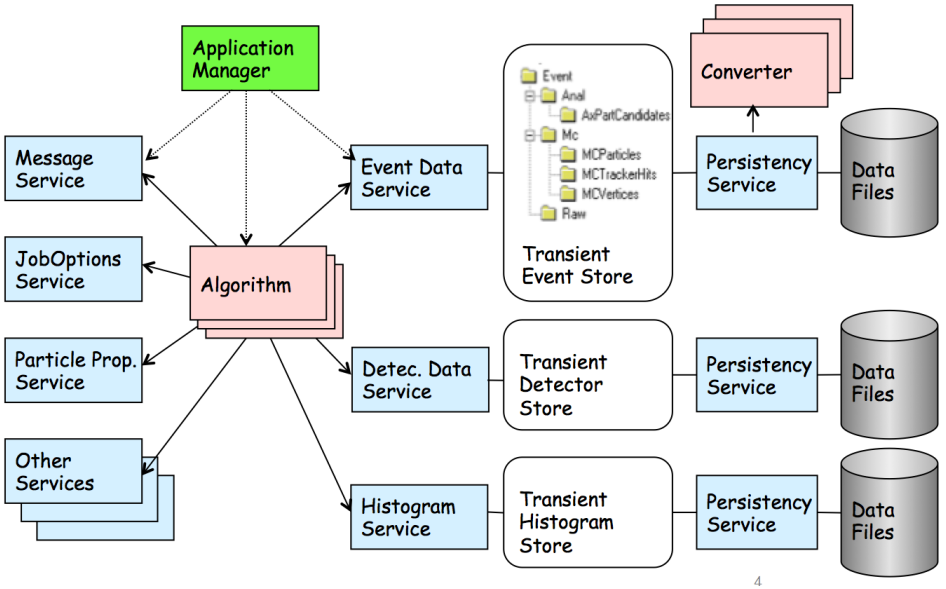
- Developed by LHCb in 1999
- Used by ATLAS, Fermi, BESIII, Daya Bay, MINERVA,LBNE

LHCb

GAUDI

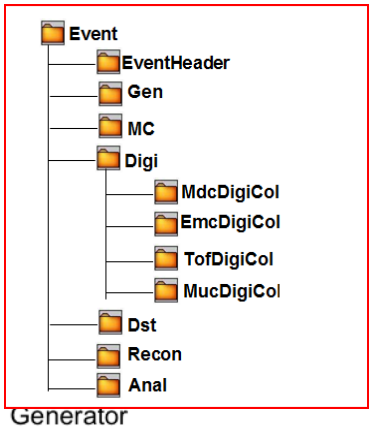
LHCb Data Processing Applications Framework

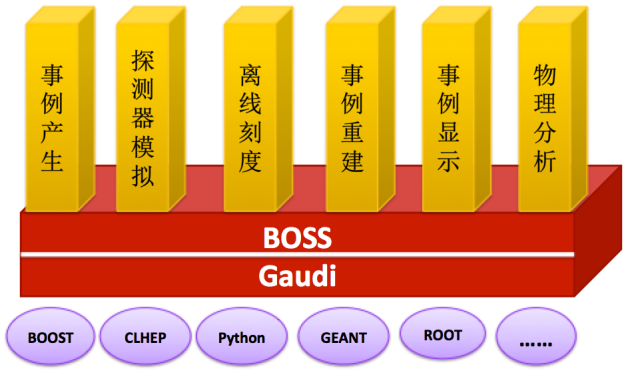
- Advantages
  - Data Store-centered architectural style
  - Clear separation between data and algorithms
  - Clear separation between persistent data and transient data
  - Encapsulated User code
  - Dynamic loading libraries
  - Well defined generic interfaces

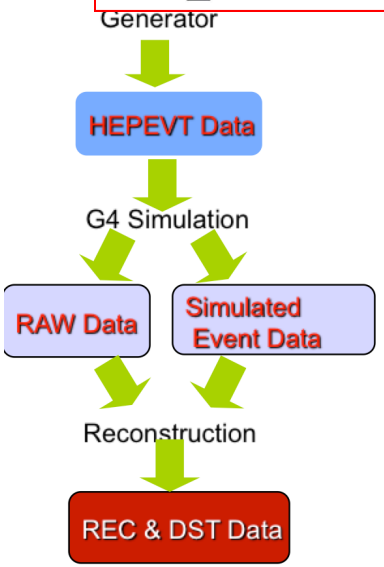


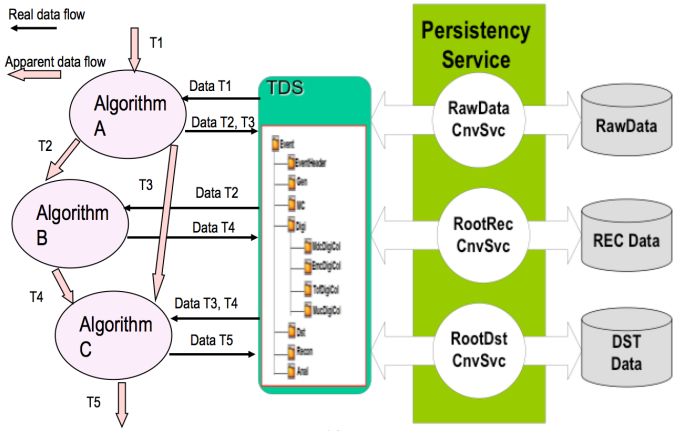
BOSS

- BESIII Framework
- BESF based on Belle/BASF, LHCb/Gaudi ,BaBar/ProxyDict
- BOSS based on Gaudi in 2003
  - Event Data Model
  - Raw ,MC, REC storage
  - File I/O
  - Services
  - ……
- Published > 160 papers
- China 1st Framework designed with OO and C ++ for the whole offline data processing and analysis



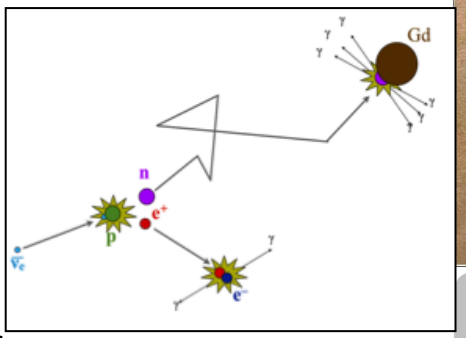




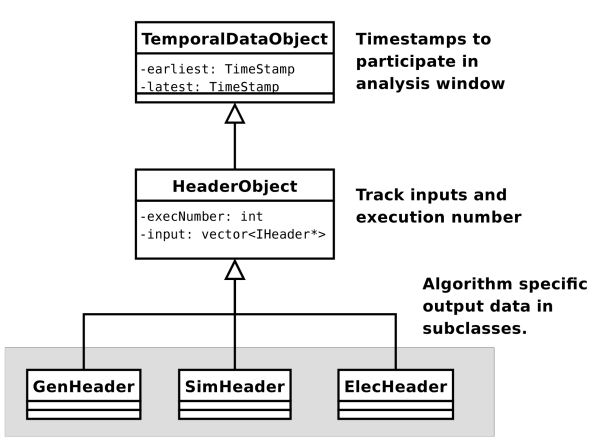


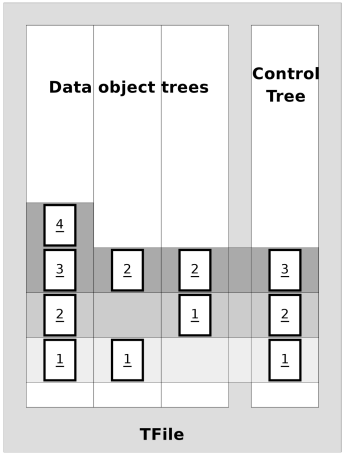
NuWa

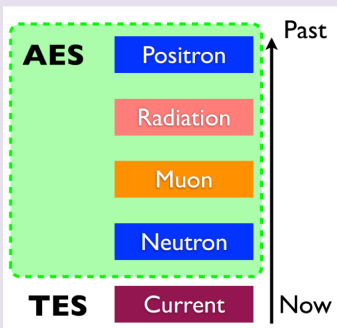
- Daya Bay Framework
- Developed based on Gaudi in 2006 after discussion
  - New challenge :Correlation Analysis
  - Design new Event Data Model
  - Design Archive Event Store
  - Keep data relationship in diff. stages
  - design Lightweight Analysis Framework
- discovered a new oscillation mode of ν



NuWa: Neutrino at Daya Wan

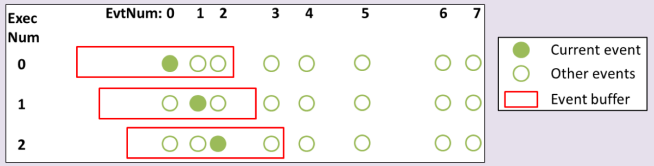






Extending Gaudi Transient Event Store (TES) to Archive Event Store (AES) for prompt-delayed analysis

- Keeps data objects in memory across execution cycles.
- Allows users to look for correlated events in past.
- Configurable based on TES location.

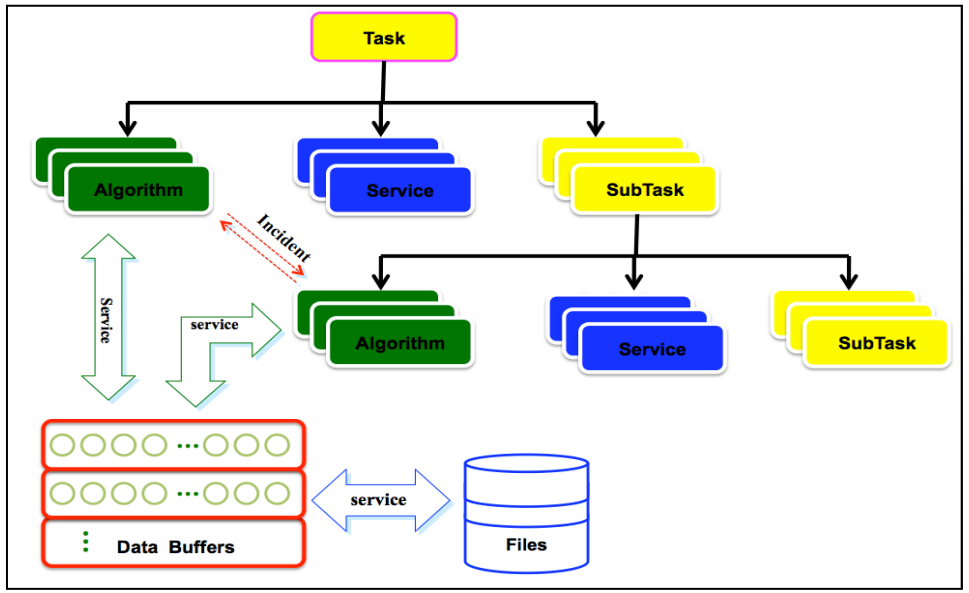


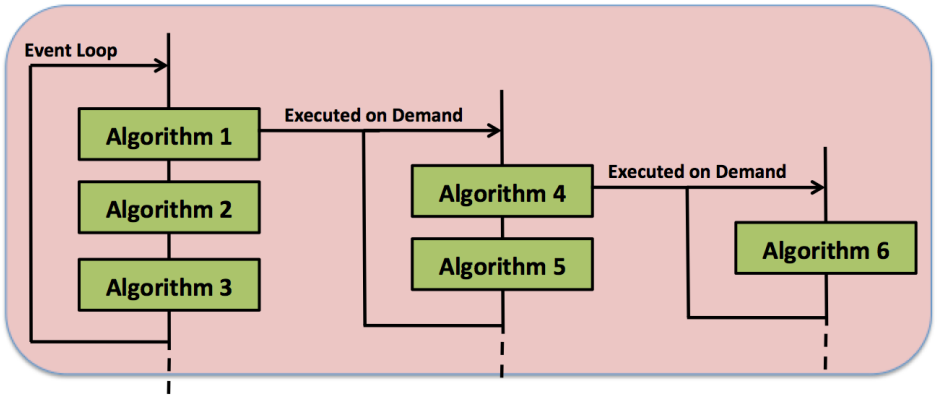
A Lightweight Analysis Framework (LAF) with high I/O performance and flexible event buffer

SNiPER

- SNiPER (Software for Non-collider Physics ExpeRiment)

  - The 1st designed from scratch in China
  - Hybrid C++ and Python
  - Modular design+ dynamically loading
  - Event Buffer for multiple events
  - flexible processing
  - Common Utilities, Services and Tools
  - Friendly User Interfaces





- A lightweight framework
- Successfully used by JUNO
- Being used by LHAASO, others

新挑战

种类	其它实验	BESIII实验	目标
J/ψ	BESII: 58 M	1.2 B （BESII 的20倍）	10 B
ψ’	CLEO-c: 28 M	0.5 B （CLEO-c的20倍）	3.0 B
ψ”	CLEO-c: 0.8 fb^{-1}	2.9 fb^{-1}（CLEO-c的3.5倍）	10-20 fb^{-1}
Above open charm threshold	CLEO-c: 0.6 fb^{-1} @ ψ(4160)	0.5 fb^{-1} @ψ(4040) 2.3 fb^{-1}@4260, 0.5 fb^{-1}@4360 0.5 fb^{-1}@4600, 1.0 fb^{-1}@ψ(4415)	5-10 fb^{-1}

BESIII：

- 目前各类数据已达 3 PB
- 每年新增约 0.5 PB
- 10年后总数据量超 8 PB

LHC Raw Data：

- RUN2每年 ~15PB
- RUN3每年 ~140PB
- RUN4每年 ~400PB

→ 计算量极大

多核、众核处理器迅猛发展 → 追求更高的性价比

→ 现有离线软件平台面临巨大压力和挑战！

技术挑战：并行计算居首

- 国际研究动态
  - 高能物理软件联盟首要技术挑战
  - 部分常用软件工具等已支持并行
    - Geant4, ROOT等
  - 大型国际合作实验已开始研究
    - ATLAS、CMS、LHCb 等

高能物理联盟白皮书

HEP Software Foundation (HSF)
White Paper Analysis and Proposed Startup Plan

The HSF Startup Team
Version 1.1, January 7 2015

Technology challenges

There is general agreement among the White Papers about the technology challenges faced by HEP software, i.e. the reasons why it is essential to evolve it to optimize performance for the present and future experimental programmes:

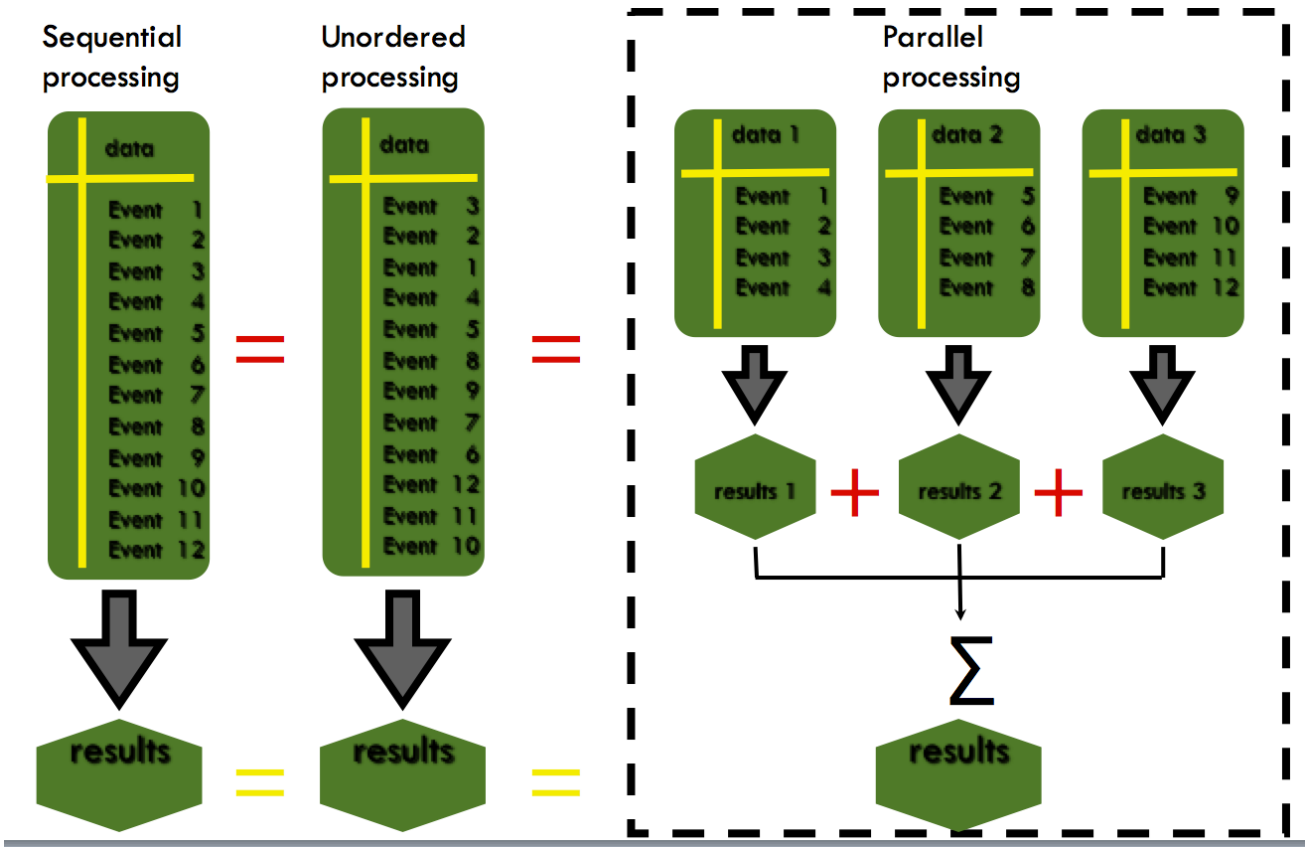
1. The main challenge that HEP software is already facing today is widely recognized as the paradigm-shift resulting from the evolution of CPU architectures towards the use of **parallelism (multi-core CPUs and vector units).** Our current code is sub-optimal in this respect and has to be reengineered to fully exploit the available performance.
2. Some White Papers such as that from Openlab [12] additionally point out the challenges that are coming from the emergence of **GPU accelerators, low power computing cores (e.g. ARM) and heterogeneous architectures**, which may eventually lead to another paradigm shift away from x86 computing.
3. One White Paper [7] explicitly mentions the challenge of **efficient access to large volumes of distributed data**.
4. The need to adapt our software and computing models to new resource provisioning technologies such as **Cloud computing (including commercial providers), HPC facilities and volunteer computing resources** has also been mentioned [1,10].
5. More generally, all White papers explicitly or implicitly recognize that the HSF would be an asset in facing **any new technological challenges that may arise in the future**.

Parallelism of FairRoot

- Use PROOF
- Only change one line in the macro to use it:

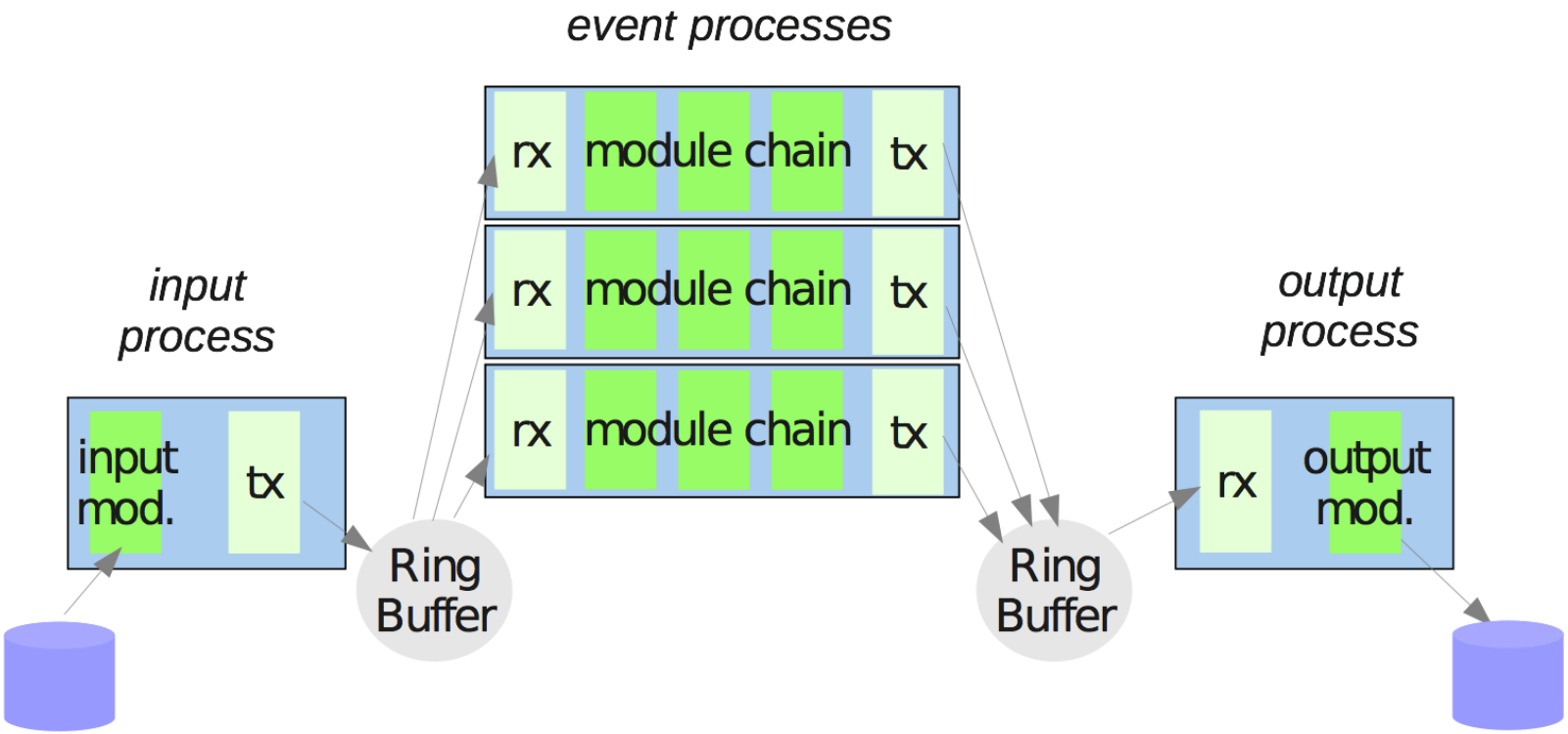
```
FairRunAna *fRun = new FairRunAna();
FairRunAna *fRun = new FairRunAna( "proof" );
```

- CUDA is fully integrated into the FairRoot build system



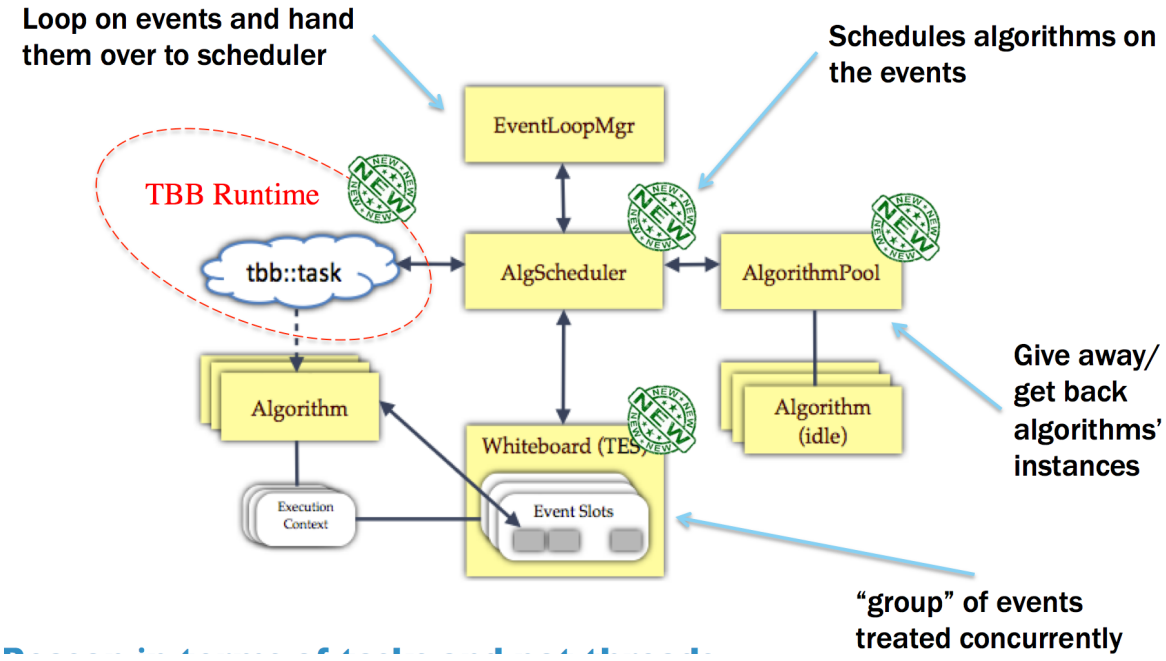
Parallelism of BASF2

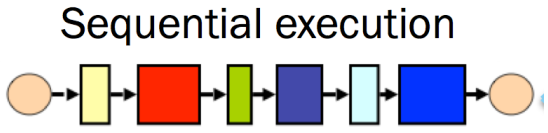
- Event by Event Parallel process via fork()
- Ring buffer (Linux IPC) is shared memory for multiple processes.
- A transmitter(tx) module places objects in the ring buffer.
- A receiver module(rx) picks up objects from the ring buffer
- The load balancing of event processes is ensured by the ring buffer automatically.

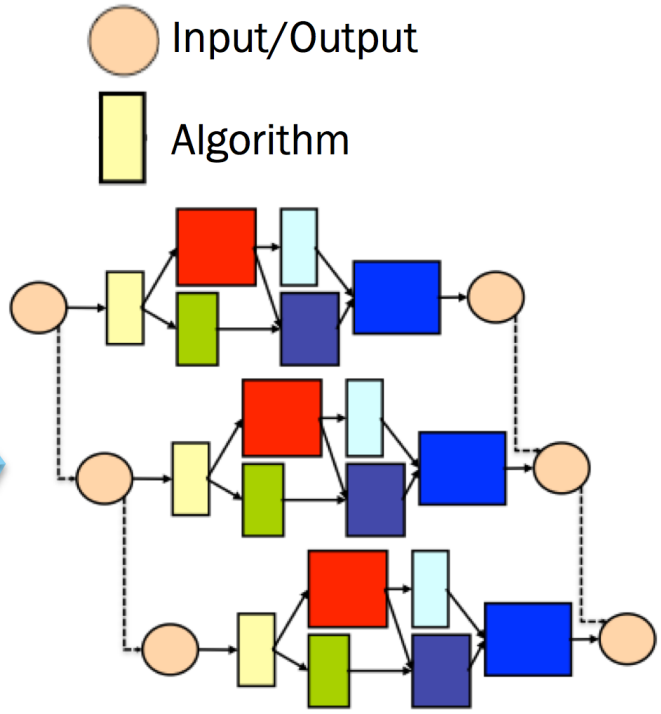


Parallelism of Gaudi

- GaudiHive: multi-threaded, concurrent extension to Gaudi
- Adopted Intel Thread Building Blocks (TBB)
- Support multi-levels parallelism





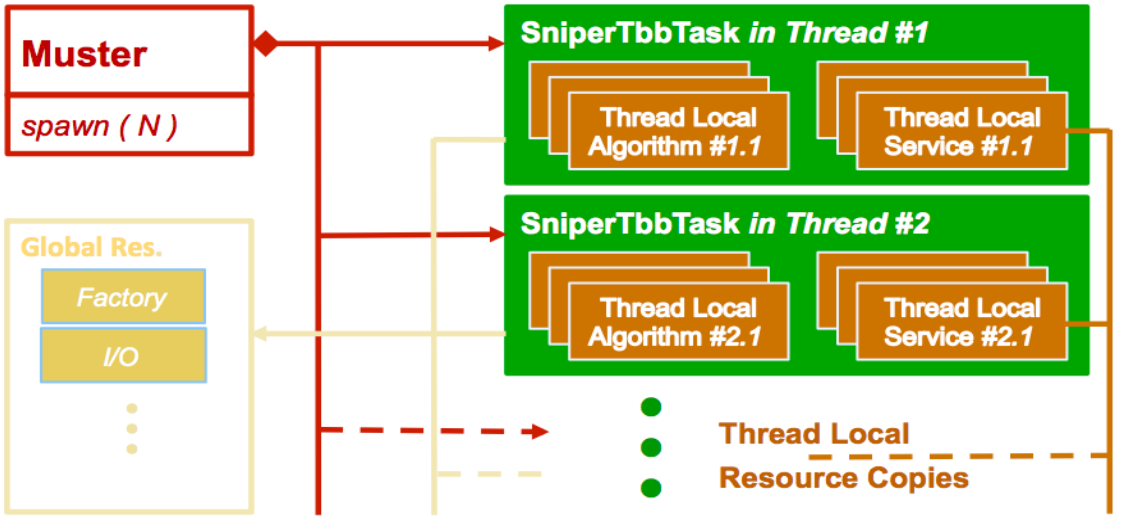


Parallelism at all levels

Parallelism of SNiPER

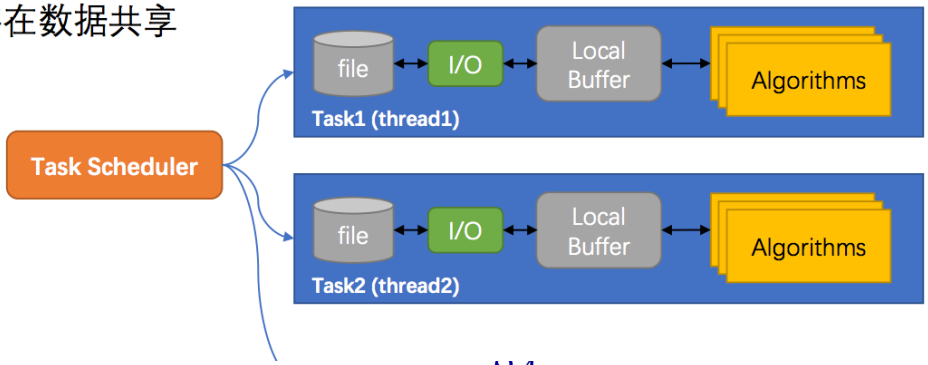
■ 多线程平行

- 发挥SNiPER多Task 机制
- 利用高层平行软件技术 Intel TBB
- 建立SNiPER Task 与TBB Task的映射关系
- 获得合理的平行度和高效的线程调度策略



方案一：简单方案

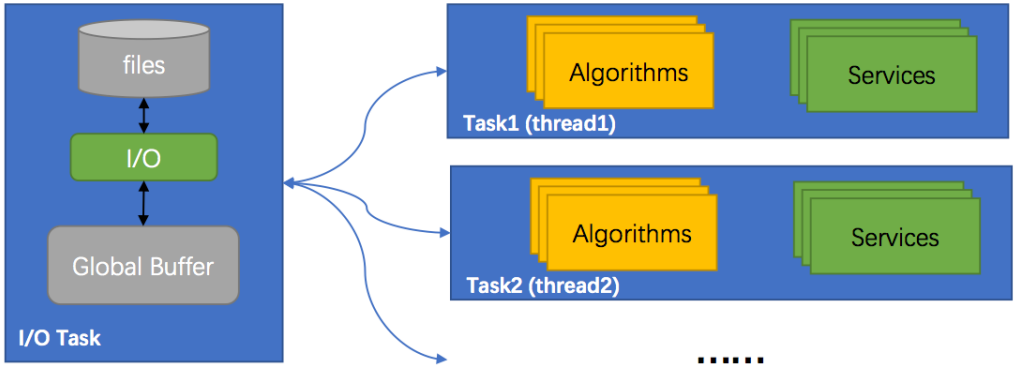
- 线程（Task）各自拥有独立的数据管理系统（输入输出文件、输入输出流和内存管理）
- 线程间不存在数据共享



Parallelism of SNiPER

方案二：

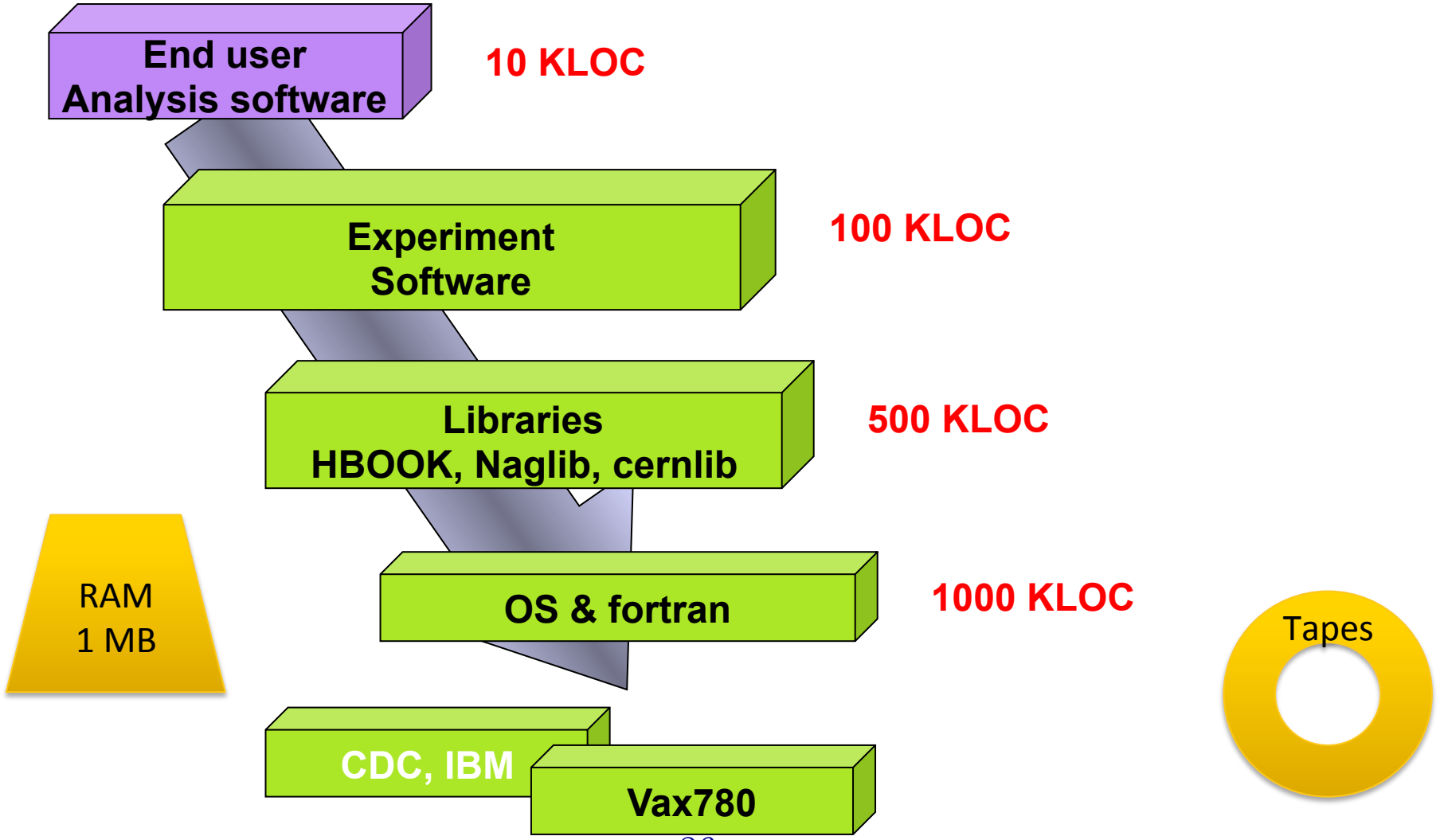
- 由全局提供数据管理功能（输入输出文件、输入输出流和内存管理）
- 全局内存管理模块向各个线程转发事例
- 各线程处理后将输出数据归还统一输出



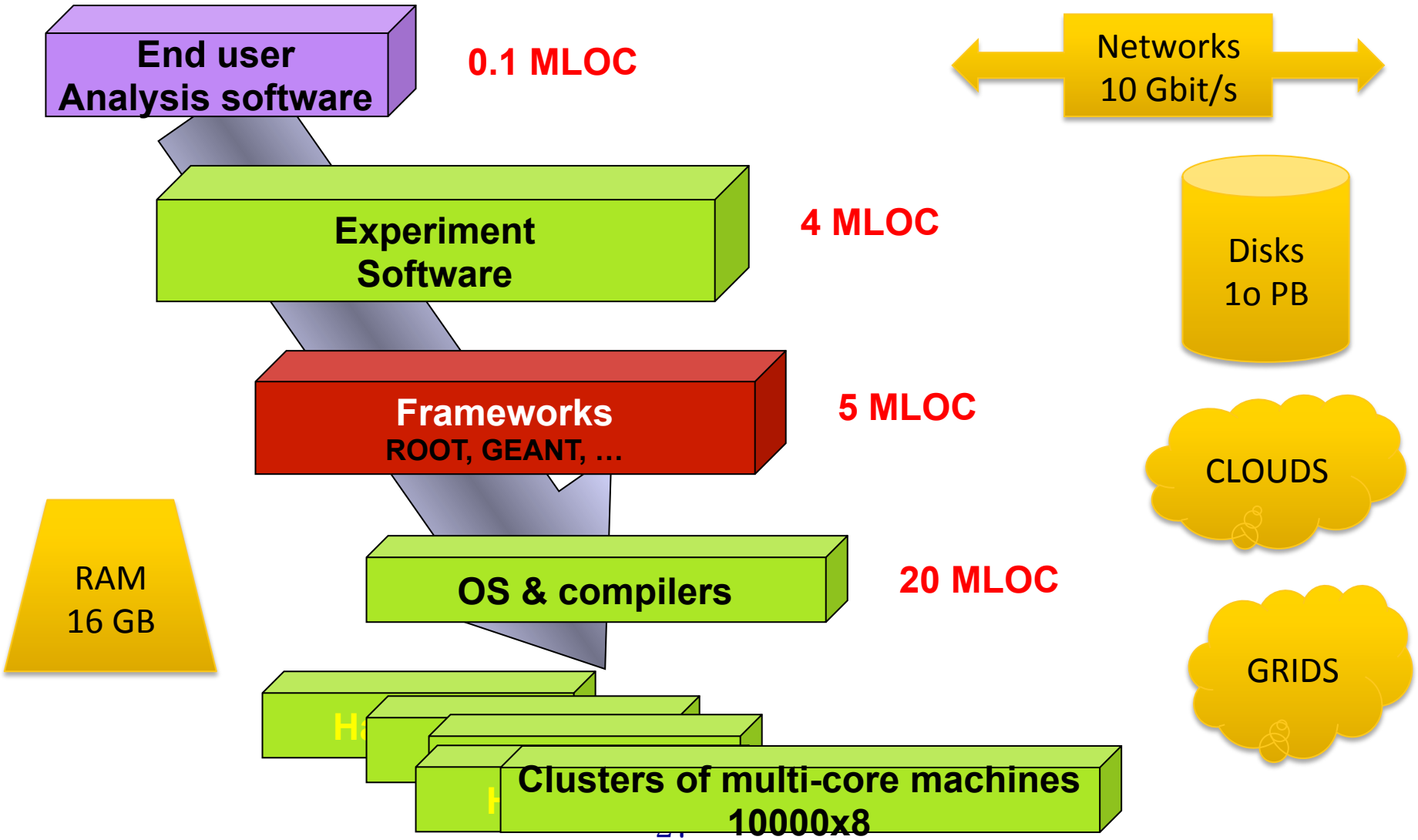
- 利用MPI技术，开展跨节点的并行
- 利用CUDA，开展异构平台的并行（CPU&GPU）

目前处于研发阶段，旨在实现多种层次的并行，推出高性能软件框架HP-SNiPER

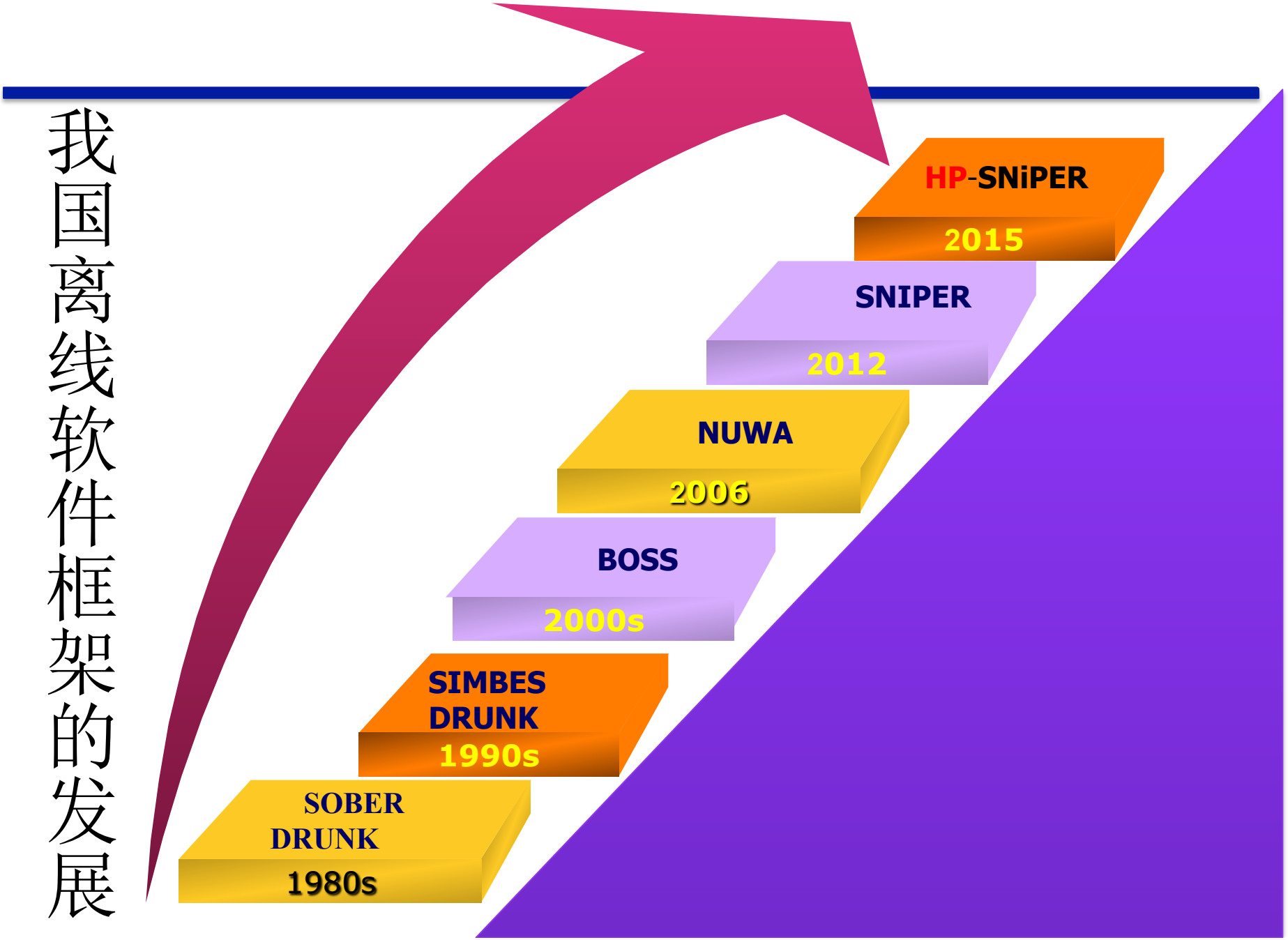
八十年代离线软件系统


End user
Analysis software
10 KLOC
Experiment
Software
100 KLOC
Libraries
HBOOK, Naglib, cernlib
500 KLOC
RAM
1 MB
OS & fortran
1000 KLOC
Tapes
CDC, IBM
Vax780

当前离线软件系统


End user
Analysis software
0.1 MLOC
Networks
10 Gbit/s
Experiment
Software
4 MLOC
Disks
1o PB
Frameworks
ROOT, GEANT, …
5 MLOC
CLOUDS
RAM
16 GB
OS & compilers
20 MLOC
GRIDS
Clusters of multi-core machines
10000x8

我国离线软件框架的发展



软件研究力量和组织活动

- 定期召开高能物理计算与软件会议
  - 2013.7.4-6， 粒子物理实验计算软件与技术研讨会（威海，31人）
  - 2015.8.11-13， 粒子物理实验计算软件与技术研讨会（威海，38人）
  - 2016.6.5-8， 高能物理计算与软件会议（东莞，？ 人）
  - 2017.6.4-7， 高能物理计算与软件会议(成都，50人）
- 软件队伍规模逐渐扩大,各方面的水平快速提升
- 中国加入了国际高能物理软件联盟（**HSF**），并于**2016**年**12**月受邀全面介绍了中国离线软件方面的研究工作

总结

- 简述了离线软件框架产生、发展演化及其功能和作用
- 概括介绍了国际上几个主要的离线软件框架的设计架构以及对并行计算实现的情况
- 我们自主研发了全新的离线软件框架SNiPER，并在JUNO 、LHAASO等实验获得成功的应用，在并行计算方面，我们也提出了自己的设计方案，预计2019年完成。

谢谢大家！