



基于异构框架的STCF MDC HLT算法研究进展

彭亮 李慧霖 方竹君 封常青

郭兆立 张宝林 赵佳睿





目 录

1

任务目标

2

CPU+GPU实现MDC二维径迹在线重建

3

GPU加速Hough变换实现MDC在线径迹重建

4

其他工作

5

未来工作



里程碑节点

	任务名称	开始时间	完成时间	任务指标
1	HLT算法软件编写和计算加速研究	2025/1/1	2025/8/29	算法运行通过，计算加速性能验证通过
2	HLT算法演示样机系统搭建	2025/4/1	2025/8/29	搭建HLT算法演示样机系统
3	HLT算法演示样机测试	2025/9/1	2025/9/30	完成HLT算法演示样机测试
4	HLT算法演示样机测试总结	2025/10/1	2025/11/28	完成HLT算法演示样机测试总结
5	关键技术攻关项目结题	2026/4/1	2026/8/1	完成项目结题，文件归档

□ 背景

- STCF高亮度带来高事例，需要HLT在线精确筛选物理事例
- MDC在线重建是HLT较为复杂和重要的环节

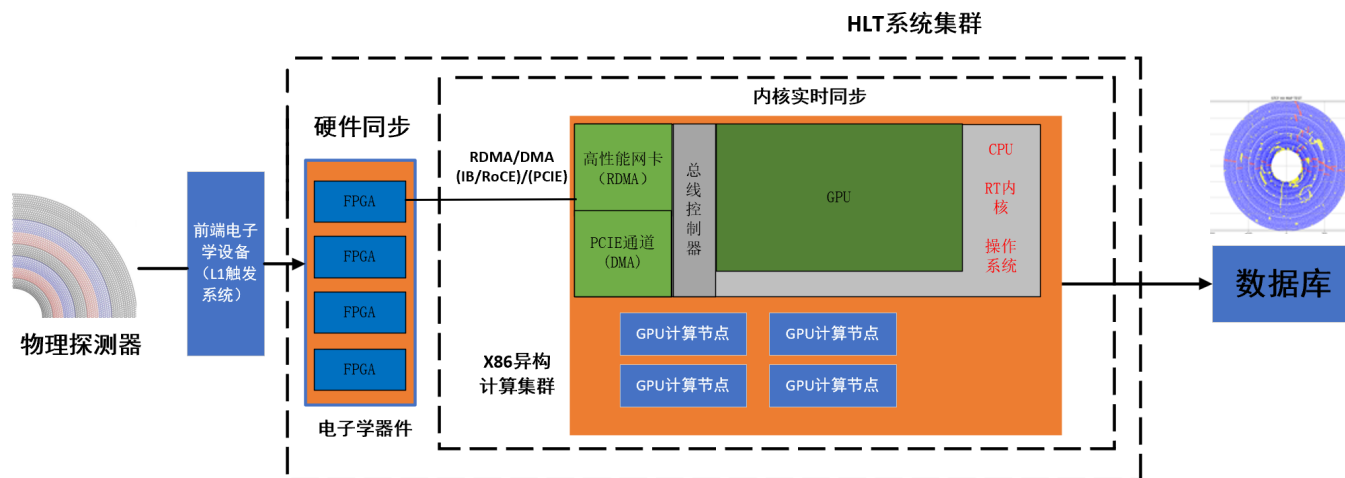
□ 工作目标

- 径迹识别
- 本底去除
- 径迹重建

□ 已完成工作

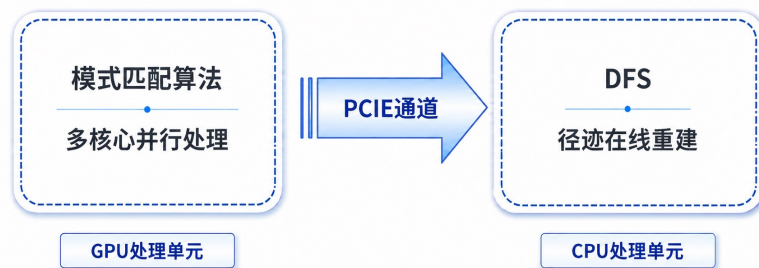
- 异构计算框架实现组合算法运行
- 在线MDC数据压缩与本底去除
- 在线MDC 2D轨迹重建

HLT集群设计图



□ 组合算法设计

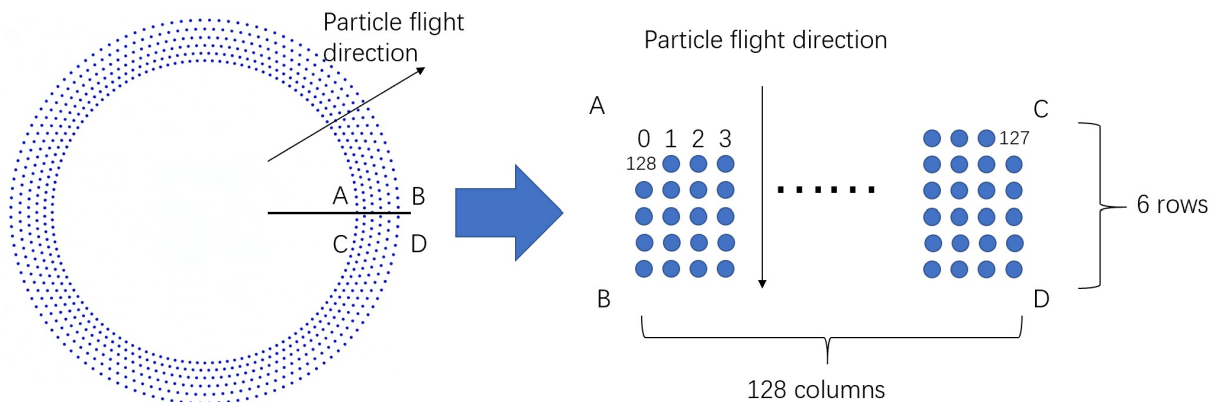
- GPU执行模式匹配算法，筛选并保留候选径迹段的信息
- CPU执行DFS（深度优先搜索）重建算法，输出完整的粒子轨迹信息
- 针对横动量 $>160\text{MeV}$ 的径迹



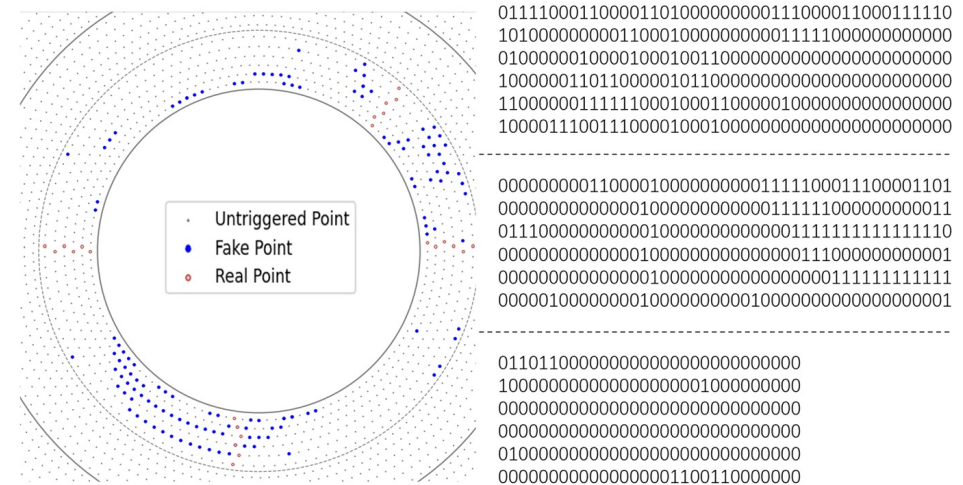
组合算法计算框架图

□ 模式匹配算法并行加速设计：将模式匹配转化成数学矩阵求解

- 数据重构：将每个超层的探测丝视为由0/1构成的二位矩阵
- 模板重构：轨迹信息映射为6x8的数学矩阵作为模板，构成模式矩阵
- 算子重构：匹配机制转变为矩阵计算（相似度计算）
- 信息获取：根据模式库和匹配位置矩阵计算轨迹段的位置、曲率方向和动量区间



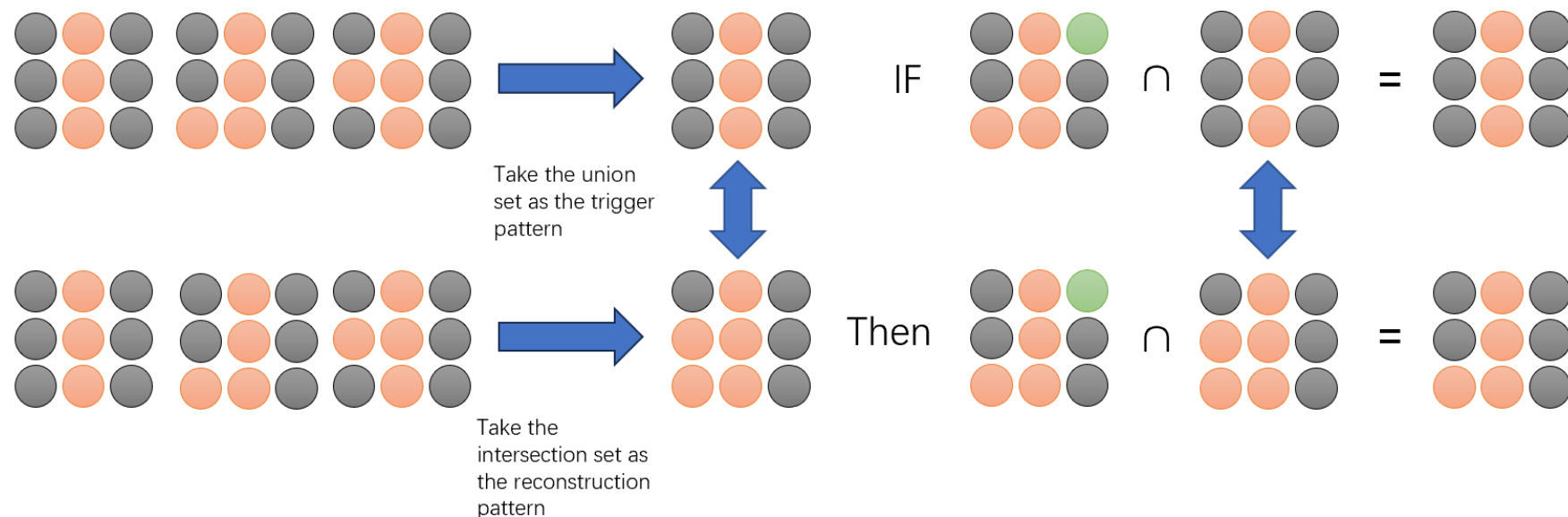
MDC各超层映射为二维矩阵示意图



MDC 超层对击中响应后对应的二维bool矩阵

□ 模式匹配算法采用算法-架构协同设计

- 逻辑运算优化提升运算速度
- 触发-还原模板分离，减少模板数量
- 向量化技术将二维矩阵压缩成，减少显存占用
- SIMT技术消除显式循环以及掩盖访存延迟，降低延迟



模板构建与模板分离后的模式匹配流程

□模式匹配算法：适合作为HLT第一级算法

- 单事例平均耗时2-3us
- 单卡每秒事例处理数量达到数百k
- 专业卡与非专业卡均可运行

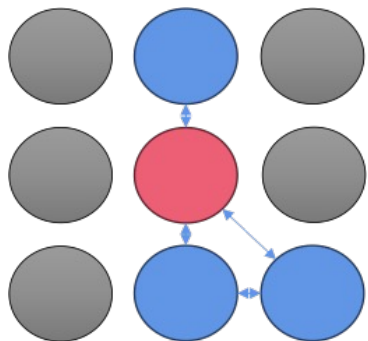
专业卡与非专业卡吞吐量以及可以容纳的事例数量

设备	架构	显存容量	带宽	PCIe速率 (单向)	显存容纳 事例	CUDA计算 单元数量	设备平均 每个事例 处理时间	设备每秒 处理的事 例数量
RTX5060TI	Blackwell	16GB	448GB/s	32GB/s	2400K	4608	1.9us	505K
RTX4060	Ada Love	8GB	272GB/s	16GB/s	1200K	3072	2.8us	350K
Tesla A16	Ampere	16GBx4	800GB/s	32GB/s	9600K	1280x4	1.35us	740K

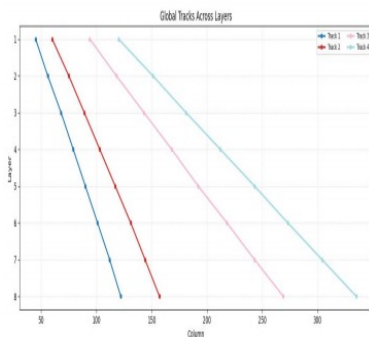
*已投稿至IEEE Transactions on Nuclear Science 期刊

DFS结合最小二乘法实现径迹重建算法步骤

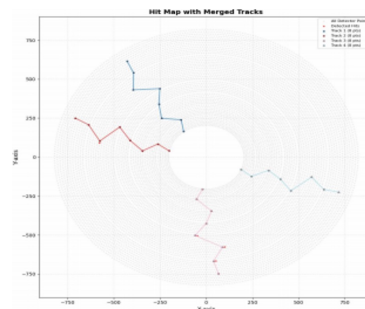
- 在候选径迹段的每一超层选取种子点，构建DFS递归矩阵进行遍历，得到击中点状态
- 若值为0，表示探测丝没被击中，则跳过，不再继续扩展
- 若值为非0，表示探测丝被击中，则计算距离方向，进行压栈处理
- DFS跨层合并，完成轨迹段合并
- 利用直丝层所有点进行圆拟合，过滤错误径迹，完成二维径迹重建



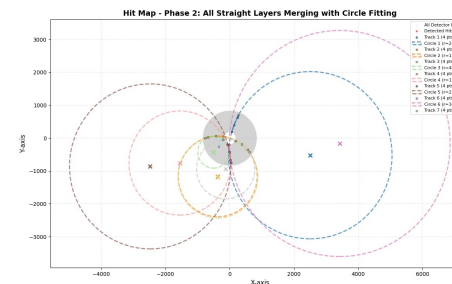
矩阵遍历示意图



DFS在超层矩阵中遍历



DFS遍历对候选轨迹段进行组合



最小二乘对径迹二维圆拟合

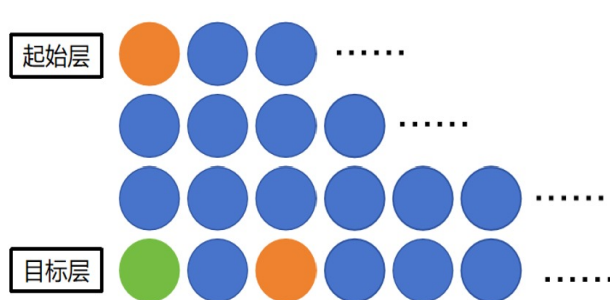
DFS遍历算法并行加速设计：将DFS遍历转化成数学矩阵求解

- 数据重构：数学矩阵设计，将 MDC 8个超层视为一个矩阵，其中非0值表示被击中，0表示未被击中
- 矩阵结构：矩阵上的击中点满足探测器结构的探测丝的相对空间位置，呈梯形状态
- 径迹归类：当对矩阵进行遍历时，从内部节点出发向外递归，不同的起始点会导致遍历留下不同的特征，即在矩阵中用不同的数字表示

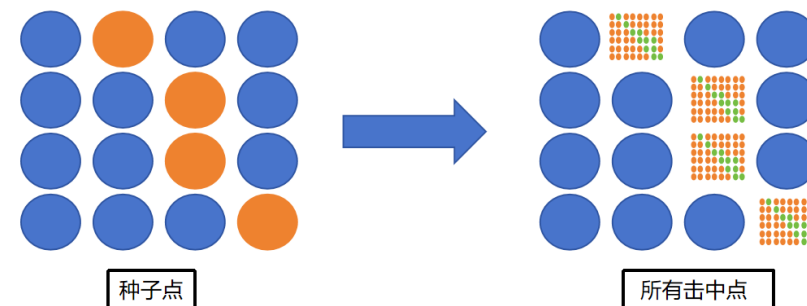
00000000000100000000	00000400000000000000
00000000010000000000	00000400000020000000
00000000000000000001000000	00000400000020000000
0000000000000100000000	00040000000020000000
0000000000000000000010000	00000400000020000000
00000000000000000001000000	00000000000020000000
000000000000000000000100	0000000020000000000000
0000000000000000000010000	0000000020000000000000

单条径迹矩阵结构

不同径迹在矩阵中的归类



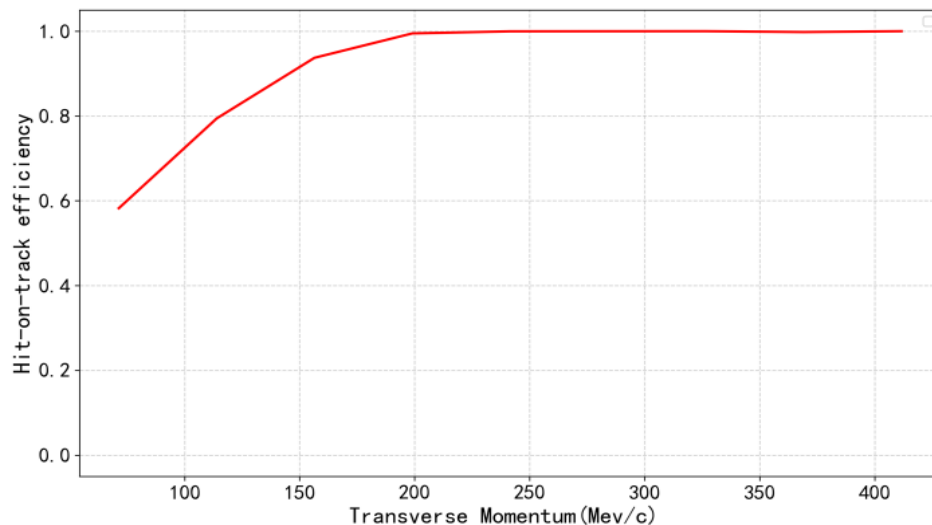
击中点映射关系图



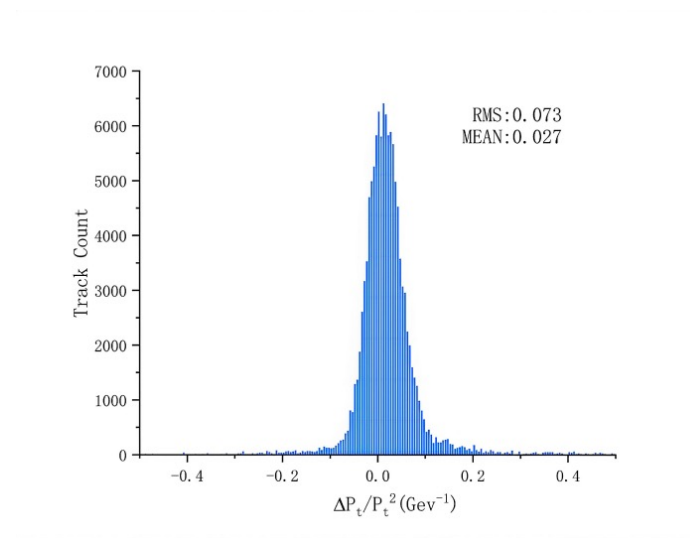
超层矩阵中的每个元素与轨迹段的映射

□ 异构框架下组合算法单粒子性能测试

- 测试数据集采用软件组Oscar生成的单粒子Digi数据集（版本号20251121），使用8组数据，每组数据包含正负粒子各1万个测试数据集
- 当横动量在大于160Mev左右时，寻迹效率可以达到95%左右。且横动量越高，寻迹效率趋近于100%
- 单粒子的RMS在0.07-0.1的范围内*



寻迹效率随横动量的变化



单粒子径迹重建横动量分辨率

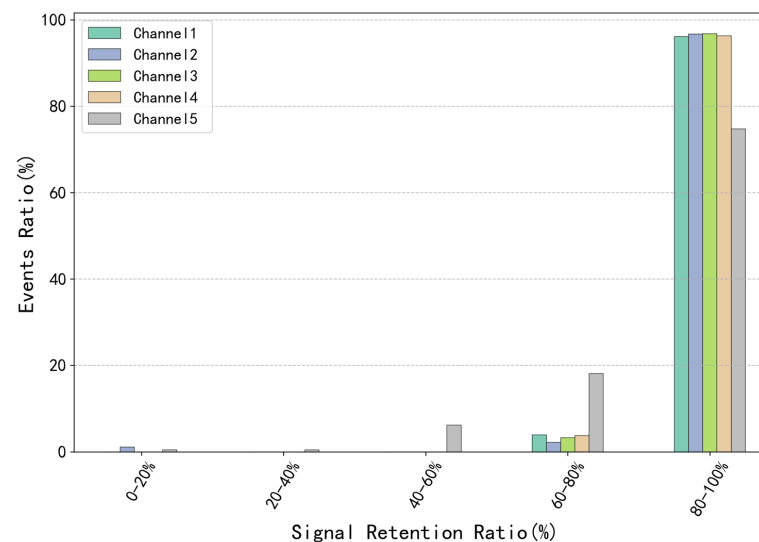
□ 异构框架下组合算法性能测试

- 测试数据集采用软件组Oscar生成的一倍本底下的Digi数据集用于测试，包括 pipiee、pipimumu、KKJpsi、DODO等等物理过程，其中每个事例数据都包含有径迹。
(版本号：202412)

$$\text{信号留存率 } SRR = \frac{N_{true\ select}}{N_{true\ total}}, \quad \text{信号压缩率 } SCR = \frac{N_{select}}{N_{total}}$$

组合算法在一倍本底下典型物理反应道的寻迹性能指标

通道编号	通道	信号压缩率	信号留存率
Channel 1	Pipiee	0.381	0.956
Channel 2	pipimumu	0.381	0.965
Channel 3	Kkjpsi	0.391	0.950
Channel 4	taupair	0.260	0.974
Channel 5	Lambda	0.362	0.851



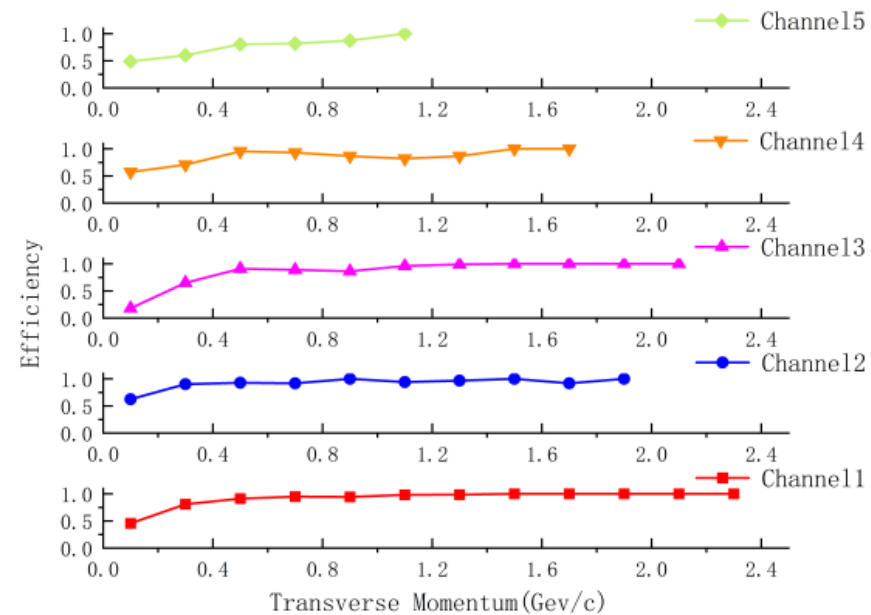
5个典型物理反应道信号留存率分布的事例占比统计

□ 异构框架下组合算法重建性能测试

- 定义重建横动量与真实横动量的差值 Δp_T 满足 $\Delta p_T/p_T^2 < 0.3$ 的粒子径迹重建成功*
- 重建效率随着横动量的变大而逐渐升高，且横动量越高，重建效率趋近于100%

通道编号	物理分析道	事例平均重建失败数目	重建效率
Channel1	$e^+e^- \rightarrow \pi^+\pi^-J/\psi$	0.23	0.926
	$J/\psi \rightarrow e^+e^-$		
Channel2	$e^+e^- \rightarrow \pi^+\pi^-J/\psi$	0.20	0.934
	$J/\psi \rightarrow \mu^+\mu^-$		
Channel3	$e^+e^- \rightarrow K^+K^-J/\psi$	0.41	0.866
	$J/\psi \rightarrow l^+l^-$		
Channel4	$e^+e^- \rightarrow \tau^+\tau^-$	0.23	0.869
Channel5	$J/\psi \rightarrow \Lambda \bar{\Lambda}$	0.43	0.722

一倍本底下典型物理反应道的重建性能



重建效率随横动量的变化

*Dong M Y, Xiu Q L, Wu L H, et al. Aging effect in the BESIII drift chamber[J]. Chinese Physics C, 2016, 40(1): 016001.

*Dong W, Feng C, Hao Y, et al. FPGA-based fast track reconstruction for the conceptual design of STCF MDC trigger[J]. Journal of Instrumentation, 2022, 17(10): P10027.



□ 异构框架下组合算法耗时

- 矩阵化DFS提速明显
- 单事例在线重建平均耗时20ms以内，满足在线重建的需求
- 算法优化在内存占用空间以及算法的速度都有明显的提升

在线重建平均耗时

通道编号	DFS 全局拟合耗时 (ms)	2D 重建平均耗时(ms)	在线重建总体耗时(ms)
Channel 1	6.60	9.12	15.94
Channel 2	7.77	7.93	15.72
Channel 3	6.92	8.48	15.45
Channel 4	4.33	6.34	10.69
Channel 5	6.30	9.08	15.42

优化前后DFS遍历算法平均耗时对比

物理分析道	优化前 DFS 平均耗时(ms)	优化后 DFS 平均耗时(ms)	加速比
$e^+e^- \rightarrow \pi^+\pi^-J/\psi$ $J/\psi \rightarrow e^+e^-$	17.20	6.60	2.6
$e^+e^- \rightarrow \pi^+\pi^-J/\psi$ $J/\psi \rightarrow \mu^+\mu^-$	19.39	7.77	2.5
$e^+e^- \rightarrow K^+K^-J/\psi$ $J/\psi \rightarrow l^+l^-$	19.78	6.92	2.85
$e^+e^- \rightarrow \tau^+\tau^-$ $J/\psi \rightarrow \Lambda \bar{\Lambda}$	17.56	4.33	4.05
	17.61	6.30	2.79

优化前后DFS遍历算法的速度以及每秒能处理事例的数量

	单个事例占用空间(KB)	单个事例处理平均开销(ms)	单个核心每秒处理事例个数
优化前 DFS 遍历	45	18.32	55
优化后 DFS 遍历	25	6.23	161

□ Hough变换算法加速设计：矩阵化并行加速

- 矩阵设计：将逐点、逐角度的离散计算过程显式重构为规则的矩阵运算
- Hough参数空间矩阵设计：设计独立列向量，便于GPU中快速累加操作，并以此构建完整的Hough参数矩阵

① 击中点矩阵 \mathbf{X}

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_i \\ y_1 & y_2 & y_2 & \cdots & y_i \end{bmatrix}$$

$$\xrightarrow[u = \frac{2x}{x^2+y^2}, v = \frac{2y}{x^2+y^2}]{\text{共形变换}}$$

② 共形矩阵 \mathbf{Y}

$$\mathbf{Y} = \begin{bmatrix} u_1 & u_2 & u_3 & \cdots & u_i \\ v_1 & v_2 & v_3 & \cdots & v_i \end{bmatrix}$$

③ 角度基矩阵 \mathbf{H}

$$\mathbf{H} = \begin{bmatrix} \cos \theta_1 & \sin \theta_1 \\ \cos \theta_2 & \sin \theta_2 \\ \vdots & \vdots \\ \cos \theta_M & \sin \theta_M \end{bmatrix}$$

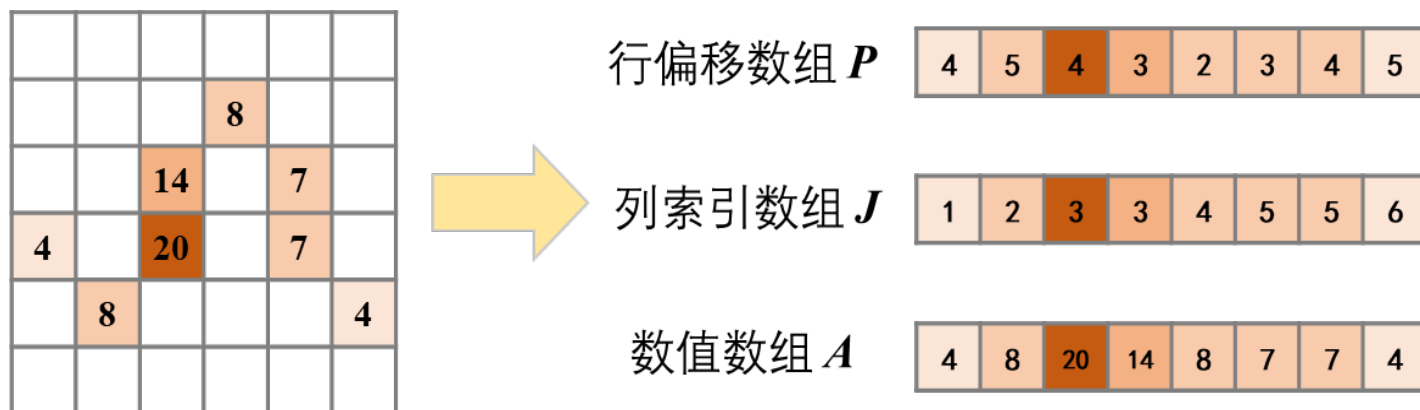
$$\longrightarrow$$

④ $\rho - \theta$ 响应矩阵

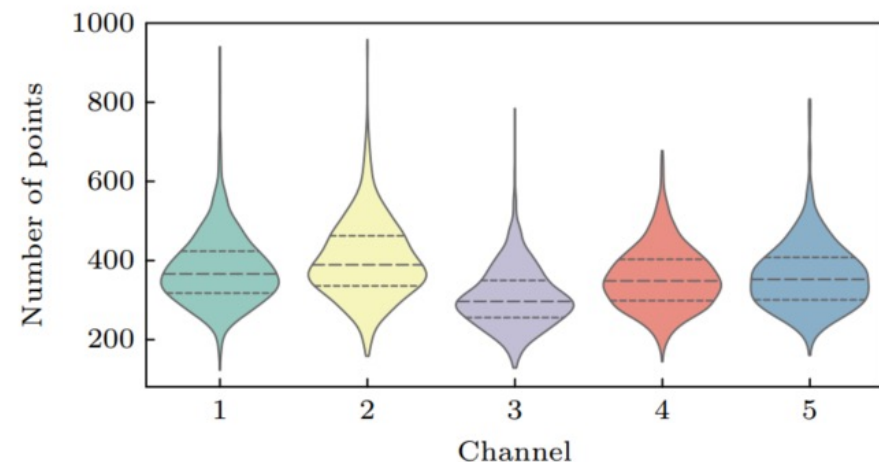
$$\rho = \mathbf{HY} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,N} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,1} & \rho_{M,2} & \cdots & \rho_{M,N} \end{bmatrix}$$

□ Hough变换算法优化设计：稀疏阵

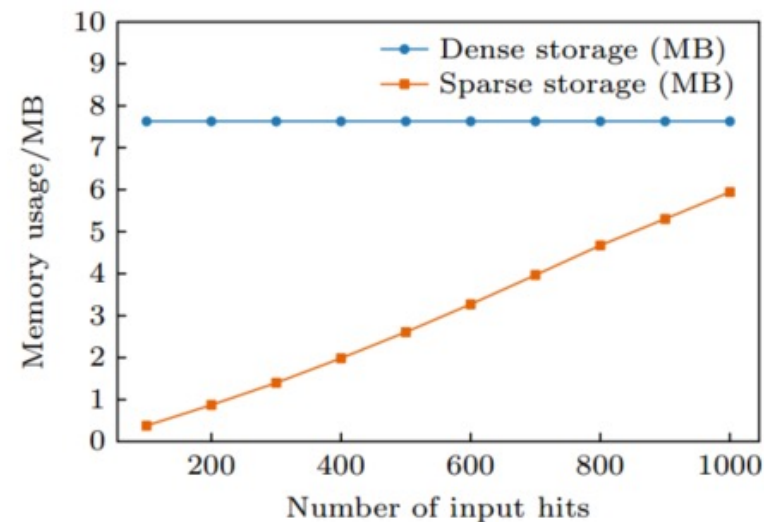
- 提供更高的计算效率与更低的内存消耗
- 采用压缩系数行存储结构，仅存储非零元素及其索引
- 1bkg下各物理反应道（版本号：202412）的有效输入信号点主要分布在200~600个点之间，平均内存占用率与稠密存储相比降低约70%



压缩稀疏行结构存储示意图



1bkg下各物理反应道信号击中点数量分布



内存占用对比

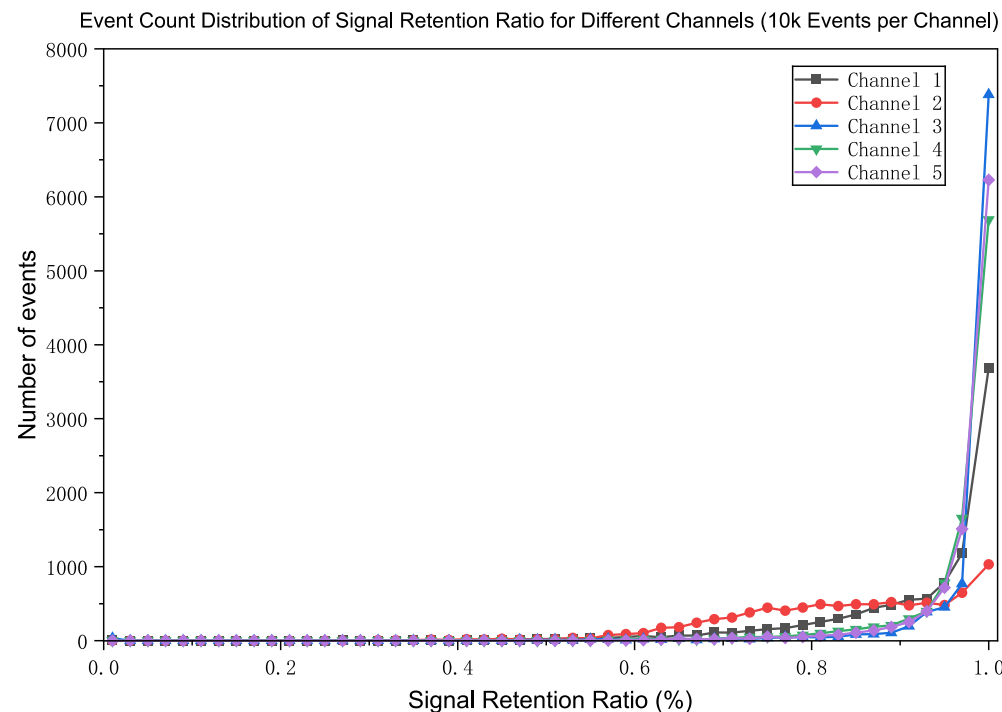
□ Hough变换算法性能测试

➤ 测试数据集：digi数据集5个典型物理反应道通道各10000个事例（版本号20251205）

➤ 降低后续分析数据规模与存储占用：数据压缩约50~60%；平均信号留存率90%以上

Hough变换在1bkG下典型物理反应道的性能指标

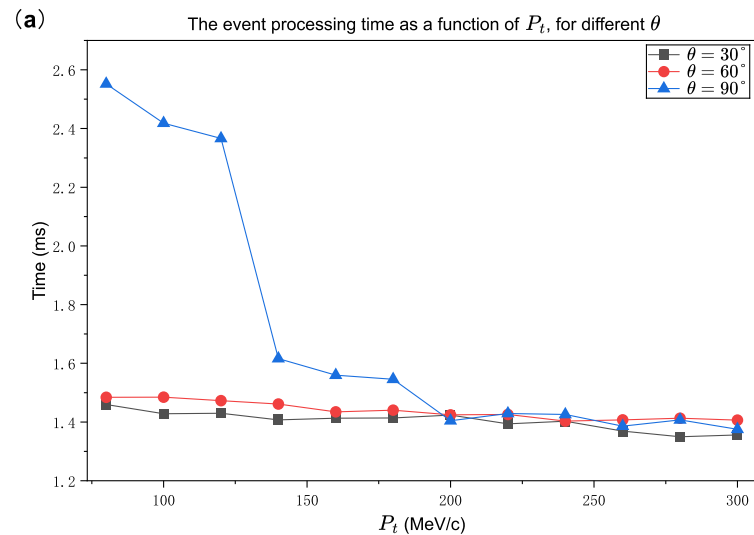
通道编号	物理分析道	数据压缩率	信号留存率
Channel 1	$e^+e^- \rightarrow K^+K^-J/\psi$ $J/\psi \rightarrow l^+l^-$	0.494	0.912
Channel 2	$J/\psi \rightarrow \Lambda\bar{\Lambda}$	0.469	0.831
Channel 3	$e^+e^- \rightarrow \tau^+\tau^-$	0.387	0.966
Channel 4	$e^+e^- \rightarrow \pi^+\pi^-J/\psi$ $J/\psi \rightarrow e^+e^-$	0.495	0.960
Channel 5	$e^+e^- \rightarrow \pi^+\pi^-J/\psi$ $J/\psi \rightarrow \mu^+\mu^-$	0.496	0.967



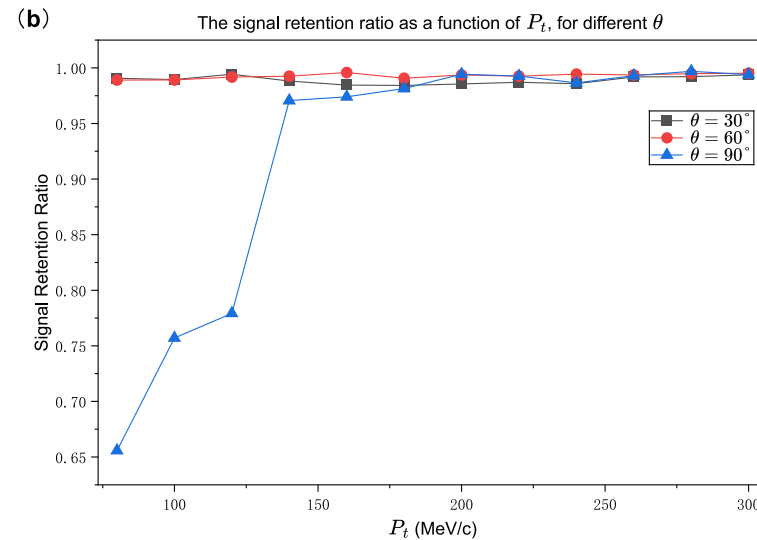
五个典型物理反应道在10000事例规模下信号留存率分布的事例数统计

□ Hough变换算法单粒子性能分析

- Hough变换算法在多数极角条件下能够兼顾实时性与高信号留存率
- $\theta = 30^\circ$ 和 60° 时，信号留存率始终维持在99%左右

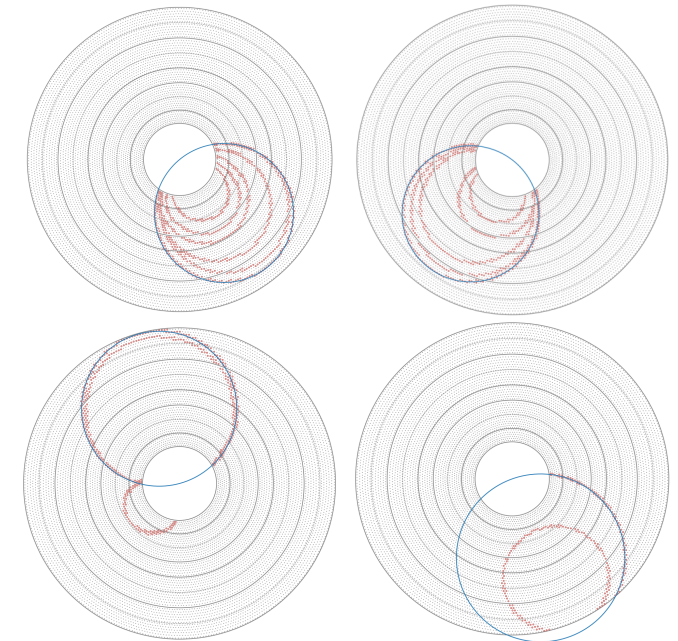


(a) 不同极角条件下算法处理时间与横动量的关系



(b) 不同极角条件下信号留存率与横动量的关系

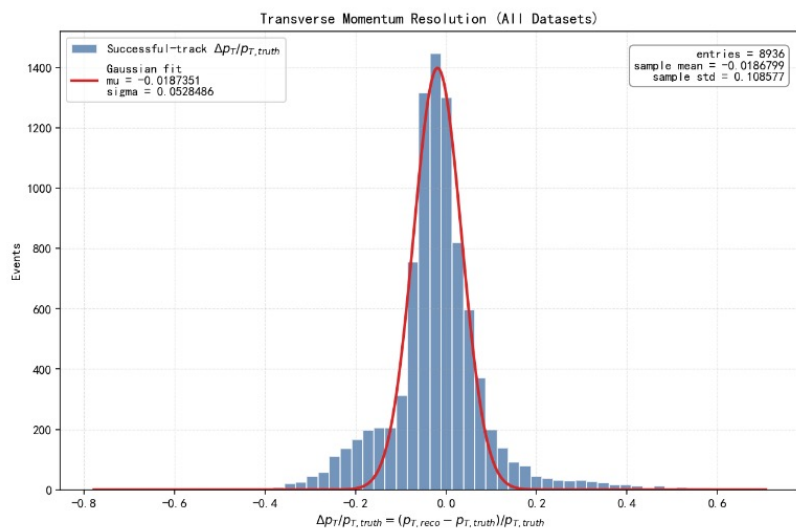
不同极角条件下霍夫变换算法性能随横动量的变化



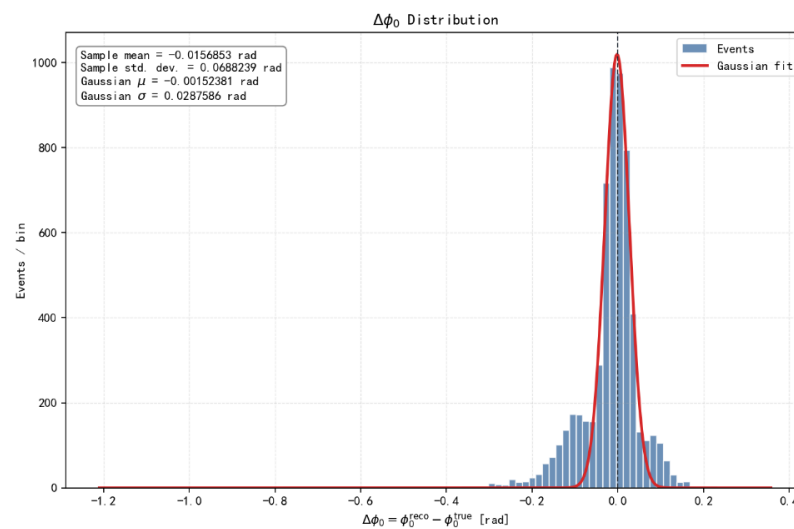
$\theta = 90^\circ$ 时一些低信号留存率事例
Hough寻迹结果

□ Hough变换算法单粒子性能分析

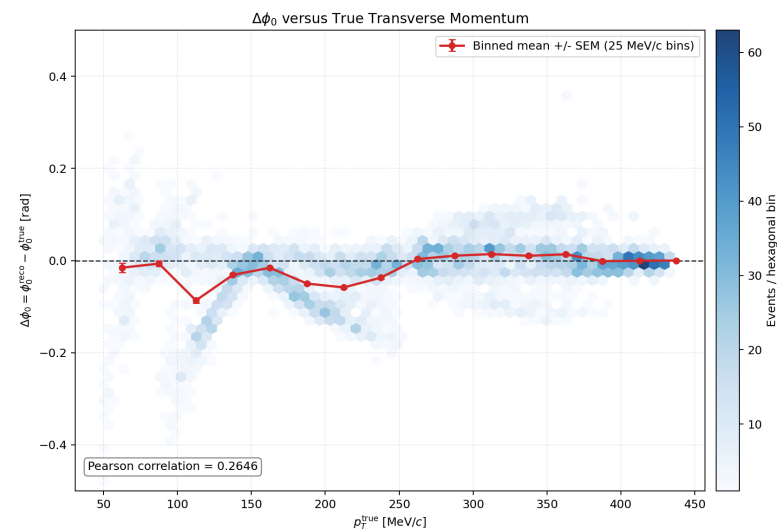
- 横动量残差分布与方位角 $\Delta\phi_0$ 分布主要集中在零值附近，主体分布近似高斯，存在非高斯尾部，主要源于中低横动量径迹的偏转较大
- 随真实横动量变化时， $\Delta\phi$ 偏差整体接近零，在中低动量区存在小幅波动



横动量残差分布



$\Delta\phi_0$ 分布



$\Delta\phi_0$ 与粒子横动量分布关系



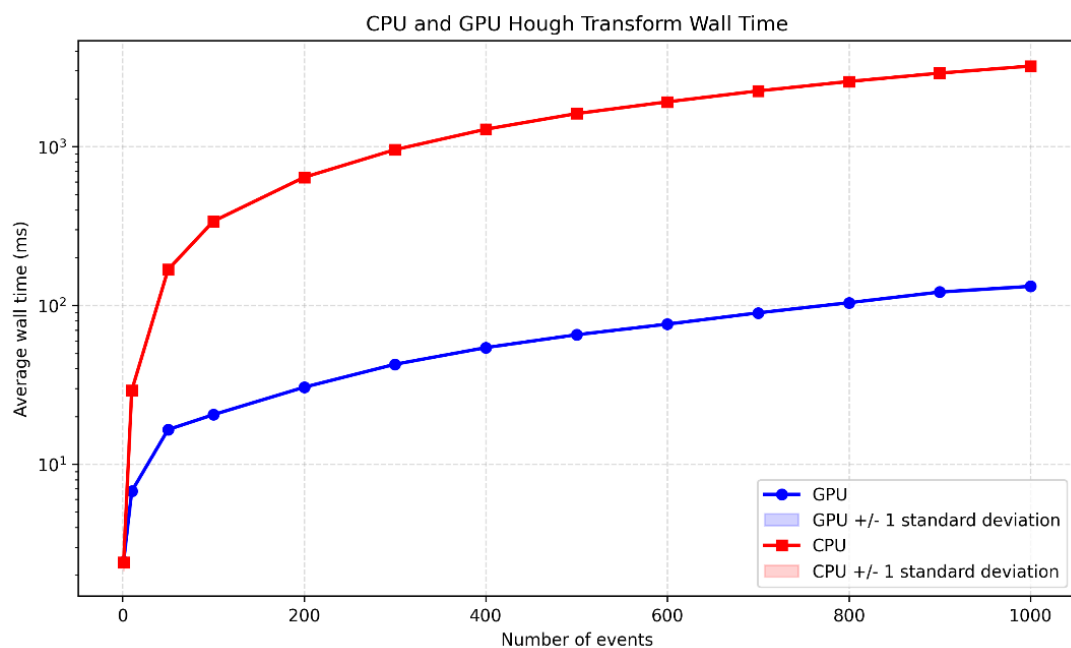
□ Hough变换算法耗时分析

- 测试数据集 1000个事例（版本号：202412）
- GPU平台单事例平均耗时仅0.14~0.35ms，加速比达60.63~151倍
- 算法-架构协同设计，从数学结构出发对算法进行矩阵化重构，释放算法并行加速潜力，为高亮度粒子对撞装置在线算法的并行加速设计提供了一个全新的研究思路

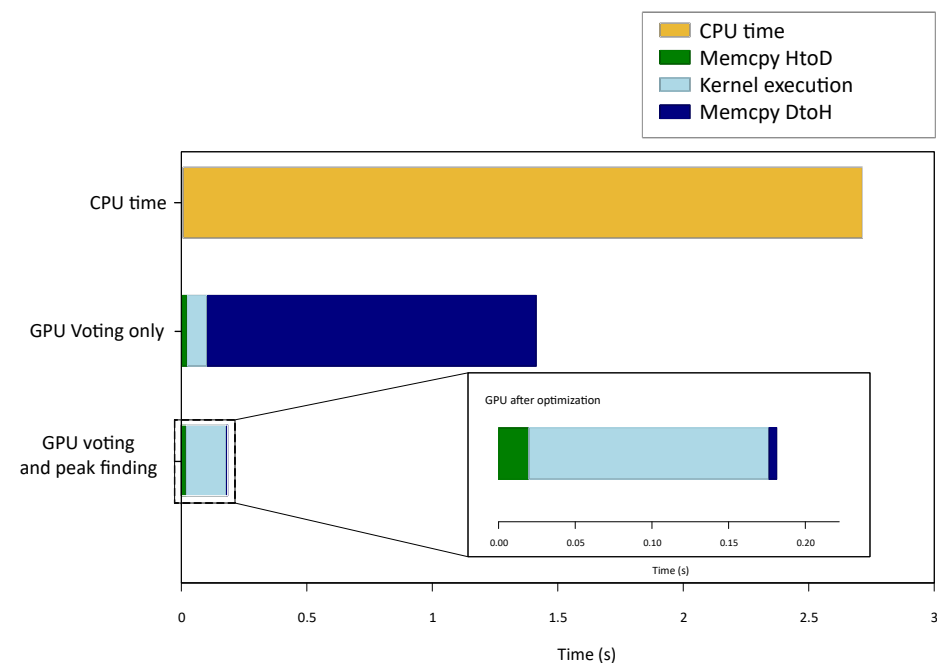
平台	单事例平均耗时 (ms)	吞吐量	加速比	显存峰值
CPU 未优化Hough	21.22	47 events/s	—	—
CPU 矩阵化Hough	1.81	552 events/s	11.72x	—
RTX5060TI	0.14	7143 events/s	151.57x	3.92GB
RTX4060	0.20	5000 events/s	106.1x	3.92GB
Tesla A16 (1/4卡, 单GPU)	0.35	2857 events/s	60.63x	3.92GB

□ Hough变换算法耗时分析

- CPU与GPU均运行矩阵化Hough算法
- 事例规模增加，GPU耗时增幅明显低于CPU端，1000事例规模下，CPU端耗时2.77s，而GPU端耗时约为0.14s，处理速度提升约20倍

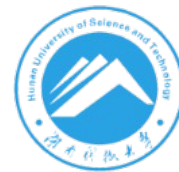


CPU与RTX 5060 Ti平台在不同事例规模下的端到端运行时间



GPU与CPU平台耗时拆解

*成果已投稿The European Physical Journal C期刊



□ Hough变换算法300k事例率下GPU中实时处理规模评估

- 在固定测试条件下，单批次处理耗时影响单卡实际吞吐，300k events/s 的实时处理目标需要多 GPU 协同完成
- 面向STCF在线实时处理部署时，需要综合权衡现存规模、处理效率和硬件成本，选择合适的GPU平台

平台	单卡显存	有效显存 (单卡显存的80%)	单事例耗时 (ms)	估算最大 batch	估算 单卡吞吐	满足300k事例率 所需卡数
RTX5060TI	16GB	12.8GB	0.14	≈3265	≈23.3k events/s	≈13 张
RTX4060	8GB	6.4GB	0.20	≈1632	≈8.16k events/s	≈37 张
Tesla A16	64GB	51.2GB	0.35	≈13061	≈37.32k events/s	≈9张



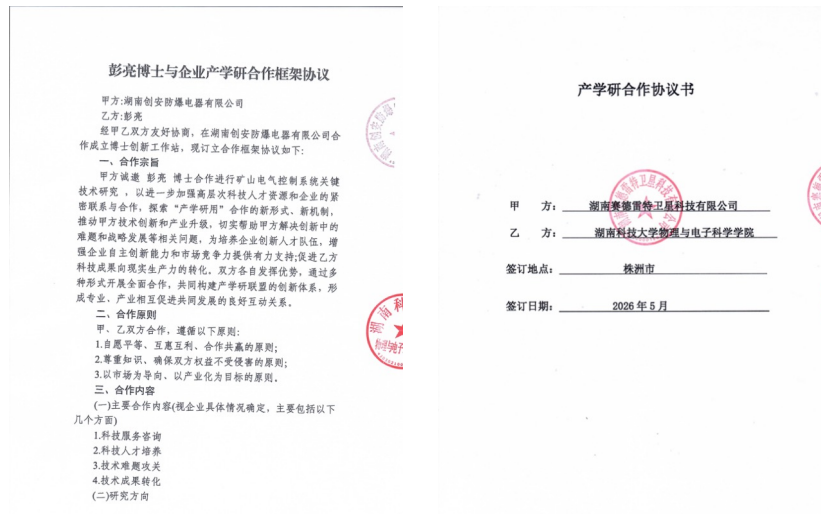
- 进一步完善算法库（卡尔曼滤波等）
- 算法进一步优化和适配（结合AI人工智能等）
- 建立和规范GPU平台算法组合机制



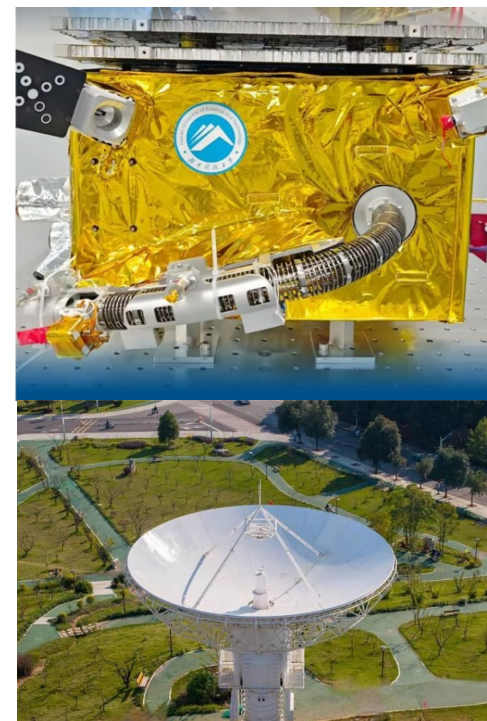
- 感谢STCF项目的支持!
- 湖南省芙蓉计划科技领军人才项目、湖南省科技创新计划、湖南省科学技术协会国际与区域科技交流类项目
- 湖南省军民融合专项基金项目（航空航天）



湖科大团队



产学研协议



湖科大卫星与地面接收站

谢谢
请多提宝贵意见！