# The scientific data analysis software framework for HEPS

# Yu Hu (on behalf of HEPSCC) Institute of High Energy Physics, CAS



2024/07/19 ICHEP @ Prague



1



- 2. Demand and Challenges of scientific data and software system
- 3. The architecture and design of the framework
- 4. The progress of the framework
- 5. Summary

2

# **High Energy Photon Source (HEPS)**

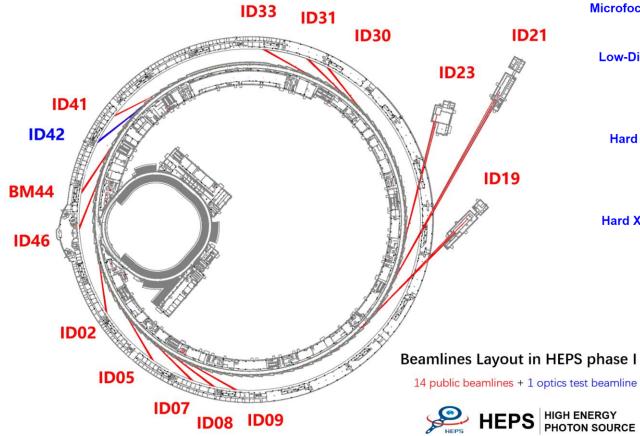
- New light source in China High energy, high brightness
- Located in Beijing about 80KM from IHEP
- Officially approved in Dec. 2017
- The construction was started in the mid-2019
- The whole project will be finished in mid-2025

Main parameters	Unit	Value
Beam energy	GeV	6
Circumference	m	1360.4
Emittance	pm∙rad	< 60
Brightness	phs/s/mm <sup>2</sup> /mrad <sup>2</sup> /0.1%BW	>1x10 <sup>22</sup>
Beam current	mA	200
Injection		Тор-ир





# **Beamlines in HEPS phase I**



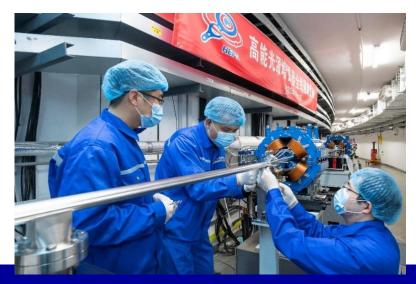
14 public beamlines + 1 optics test beamline in Phase I Can accommodate over 90 beamlines in total

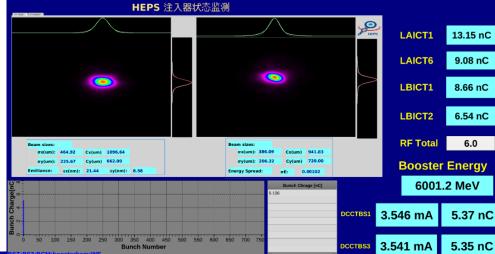
ofocusing X-Ray Protein Crystallography-ID02 Beamlin <u>e</u>	ID30 <u>-</u> Transmission X-Ray Microscopic Beamline
w-Dimensional Structure Probe Beamline-ID05	ID31 <u>-</u> High Pressure Beamline
Engineering Materials Beamline <u>-</u> ID0 <u>7</u>	ID33 <u>-</u> Hard X-Ray High Resolution Spectroscopy Beamline
ard X-Ray Coherent Scattering Beamline <u>-ID09</u>	BM44 <u>-</u> Tender X-Ray Beamline
Pink Beam SAXS Beamline_ID08	ID41-High Resolution Nanoscale Electronic Structure Spectroscopy Beamline
rd X-Ray Nanoprobe Multimodal Imaging_ID19 Beamline	ID42 <u>-</u> Optics Test Beamline
Hard X-Ray Imaging Beamline_ID21	ID46 <u>-</u> X-Ray Absorption Spectroscopy Beamline
Structural Dynamics Beamline-ID23	



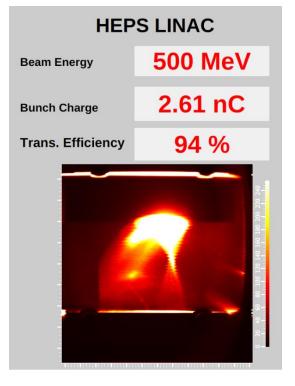
# **Progress of the HEPS project**

- The construction of the civil structure completed. Now at the stage of equipment installation
- □ 2023.01, HEPS booster installation completed
- 2023.03, HEPS achieved the first electron beam accelerated to 500 MeV
- □ 2023.11, Electron beam ramped up to 6 GeV
- □ 2024.07, HEPS storage ring installation completed
- □ 1st SR X-ray to be emitted in the near future











# 2. Demand and Challenges of scientific data and software system

# 3. The architecture and design of the framework

# 4. The progress of the framework

5. Summary

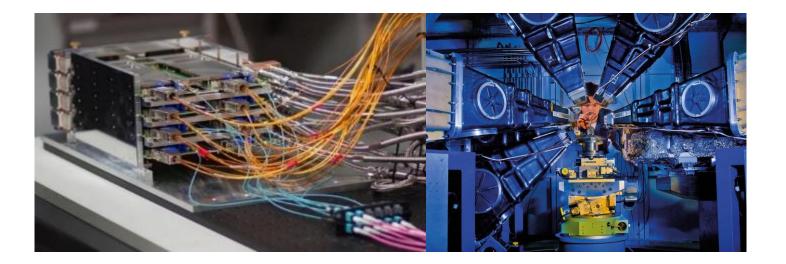
6

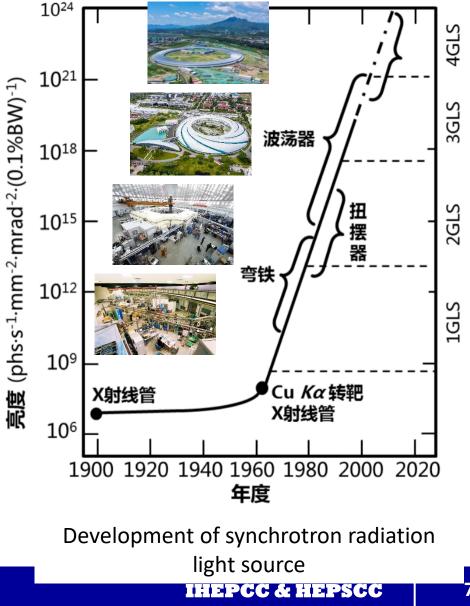
□ 10<sup>2</sup>~10<sup>3</sup> higher brightness than 3<sup>rd</sup> gen. SR sources

• More raw data in greater detail and less time

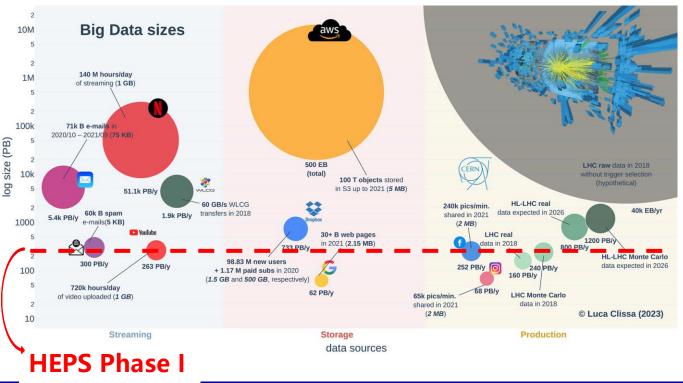
Detector capabilities constantly improving:

- Increased dynamic range, faster readout rates, larger pixel arrays (e.g. 32bits, 20KHz, 30kx30k)
- Bigger frames, higher frame rates => more raw data





- >200PB raw data per year for HEPS Phase I (15 beamlines)
- $\hfill\square$  More than 90 beamlines volume in total
- Data volume will soon reach the EB scale



#### **Data volume of HEPS Phase I Beamlines:**

Beamlines	Burst output (TB/day)	Average output (TB/day)	
Engineering Materials	600.00	200.00	
Hard X-ray Multi-analytical Nanoprobe	500.00	200.00	
Structural Dynamics	8.00	3.00	
Hard X-ray Coherent Scattering	10.00	3.00	
Hard X-ray High Energy Resolution Spec.	10.00	1.00	
High Pressure	2.00	1.00	
Hard X-Ray Imaging	1000.00	250.00	
X-ray Absorption Spectroscopy	80.00	10.00	
Low-Dimension Structure Probe	20.00	5.00	
Biological Macromolecule Microfocus	35.00	10.00	
pink SAXS	400.00	50.00	
High Res. Nanoscale Elec. Struc. Spec.	1.00	0.20	
Tender X-ray beamline	10.00	1.00	
Transmission X-ray Microscope	25.00	11.20	
Test beamline	1000.00	60.00	
Total average:		805	

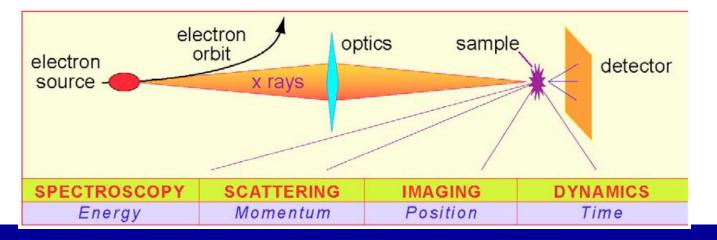
□ Multi-disciplinary, e.g. spectroscopy, imaging, diffraction & scattering .....

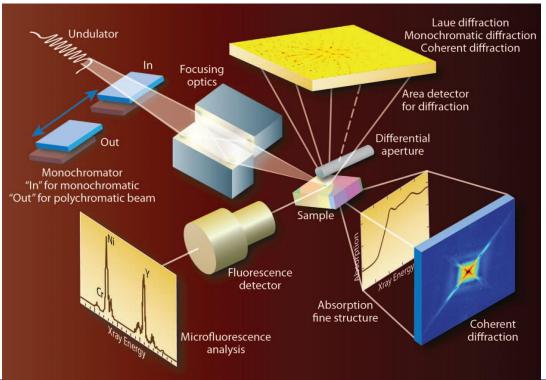
New and more complex experiments, new algorithm and tools

□ Multi-modal experiments that combine data from multiple samples, techniques, and facilities

□ In situ and in operando experiments require real-time feedback and autonomous control

- Data throughput and volume vary greatly with beamlines and scientific goals
- New users from a wide range of domains and backgrounds





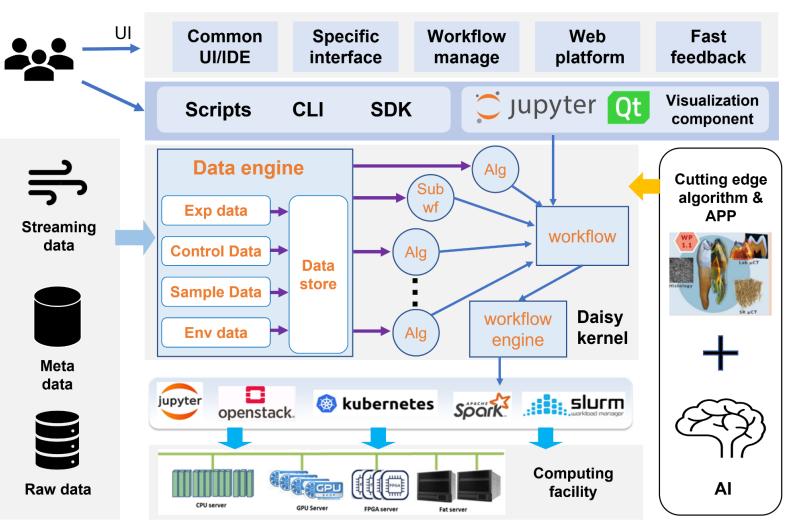
- Analysis and management of large datasets at advanced SR sources is becoming progressively more challenging
- Development and integration of advanced analysis and management tools is needed
  - Provide storage, organization and management of massive scientific data
  - During the experiment, provide real-time analysis and fast feedback to guide the experiment steering and optimize the data acquisition
  - After the experiment, process the massive offline data, accelerate the scientific discovery
  - Provide the scalable distributed heterogeneous computing power, meet the diverse computing requirements of different scientific goals



- 2. Demand and Challenges of scientific data and software system
- 3. The architecture and design of the framework
- 4. The progress of the framework

5. Summary

# Data analysis software framework—Daisy

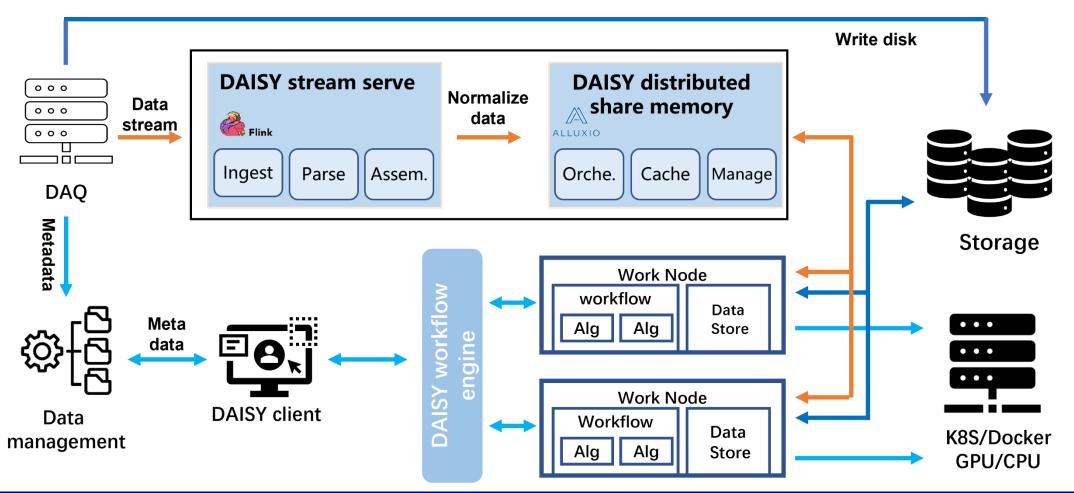


#### • Kernel of the framework

- Derivative modules to meet the requirements of advanced SR sources
  - Data object management module for high-throughput data I/O, multimodal data exchange, and multi-source data access
  - Scalable cluster computing power support for data processing with different scales, different throughputs, and low latency
  - Interface and developing environment for scientific software integration and development
- Domain specific App and flexible general workflow management system based on the framework

# **Daisy data processing flow**

- Supports stream processing and batch processing
- Supports interactive processing via GUI and automated processing





🔀 Daisy Workbench		<u>I</u> nterfaces	<u>H</u> elp						- 🗆 🗙	Per jupyterhub Home huy 🕒 Logout	
<u>File View</u> Interfaces	<u>H</u> elp	Integra	tion								-
Workspaces Load Delete	Cle	MaxMir	ı		IPyth	on 4]: load start:tooth	System Men	nory Usage			
	CIE	PyFAI c	alib	- C	ptions * work	flow:LoadHDF5.config		1.72/62.76 GB (	2%)	。 1993年1月1日日日(1993年1月1日日) 1993年1月1日日日(1993年1月1日日日) 1993年1月1日日日(1993年1月1日日)	
Sort Save		H '	tch Fitting	-rt Dasiy algorit	HDF5	: initialized, read File: /root/tooth.h	5 Plots		0 8		
			_	_ api import	* work	flow:LoadHDF5.execut	e inots	Hide C	Close	Availteetteve	
name v tooth		Spectra	a Matching	∫rt matplotli	b.pypl INFO	: Load data /exchang as tooth from /root	colort All			Applications	
tooth		daisyworkbench	- D	umpy as	💌 tooth			× –			
		1	2	*		+ Q ≆ ∠ (	<b>D</b>			分析环境1	
	2	-	[27098.75		<b>n v</b> 7	∓≺≠≝		Plo	ot Name	分析环境2	
	2							tion Pla	ot 😽 🖉 🗙		
	3	[27051.75	[26986.25						ΦX	CT 3D reconstruction	
	4	[27192.75	[27107.5							CT 3D reconstruction service based on tomopy.	
	5	[27208	[27020.25		0						
	6	[27181.75	[26995							0	
	-	-	[27033.5		100 -					alphafold-with-40g	
	/	[27190	-							alphafold-with-40g	
	8	[26869.75	[27169.25		0 1	100 200 300	400 500 6	config 500 DF5 Fil	INFO:		
Algorithms	9	[27142.25	[26977.75					execute	e INFO:	0	
Aigoritims	10	[27407.75	[27282.5	-					as tooth from	cumopy	
Excute AlgMatrixTra	nspos	-								cumopy	
<ul> <li>Daisy LoadHDF5</li> </ul>			▲								
LoadTIFs								.9897460938s cation::exec: Th	he event loop	开发者环境	
SaveHDF5 SaveH5VDS							is already			0	
AlaMatriyTranspace			▼ 4		•		Ŧ		v		

#### Workbench client

- General-purpose GUI based on PyQt5
- Include data object list, algorithm list, data viewer tools, and IDE for developers
- Interfaces of customized GUIs for scientific applications

#### Web data analysis platform

- Based on JupyterLab
- Advantageous for remote data access and analysis
- Terminal and dedicated scientific App

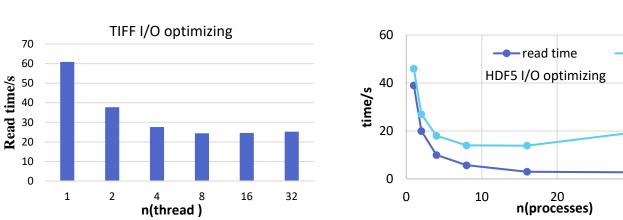


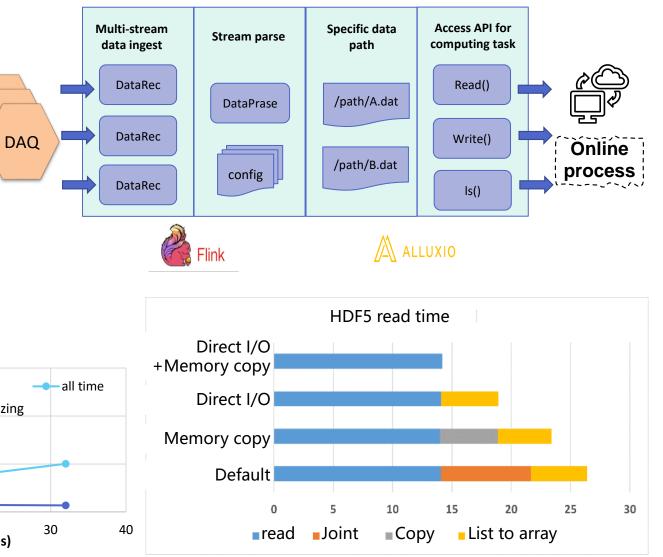
- 2. Demand and Challenges of scientific data and software system
- 3. The architecture and design of the framework
- 4. The progress of the framework

# 5. Summary

# **Daisy I/O module**

- Designed a unified I/O interface to shield the difference of underlying architecture and data structure
- Data I/O optimization
  - HDF5 parallel I/O based on multi-process. Memory copy, asynchronous I/O, direct I/O also employed
  - TIFF parallel I/O based on multi-thread
- R&D on stream data processing is under way

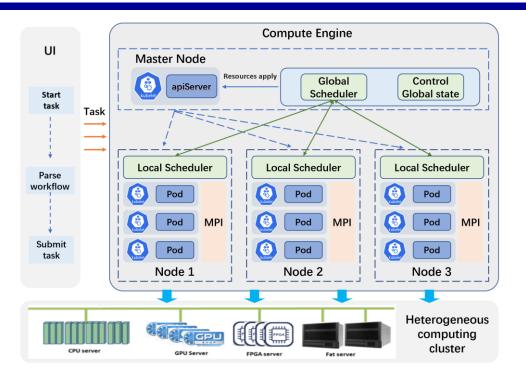


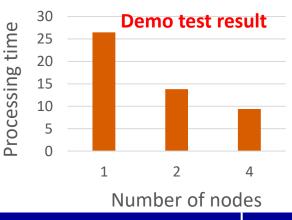


# **Daisy distributed computing engine**

- A single dataset of HEPS imaging experiment will reach the TB scale
- Scientists expect data processing time at the scale of DAQ time
- **D** A distributed data processing system is developing
- □ Support heterogeneous distributed computing power
- Provide a unified flexible programming interface API for computing models, to reduce the complexity of parallel programming

Mode	Detector pixel	Frame rate	Projections number	Data rate	Dataset (TB)	DAQ time	Daily data (TB/d)	Annual data (PB/y)
Powder CT	6k×6k	19fps@16bit	6k	1.08 GB/s	0.432	6.3 min.	78	9.4
High voxel CT	28kx10k	2pfs@16bit	28k	1.1 GB/s	15.68	240 min.	87	10.4
Fast CT	5k×4k	595fps@8bit	5k	1.7 GB/s	0.1	1 min.	98	5.9



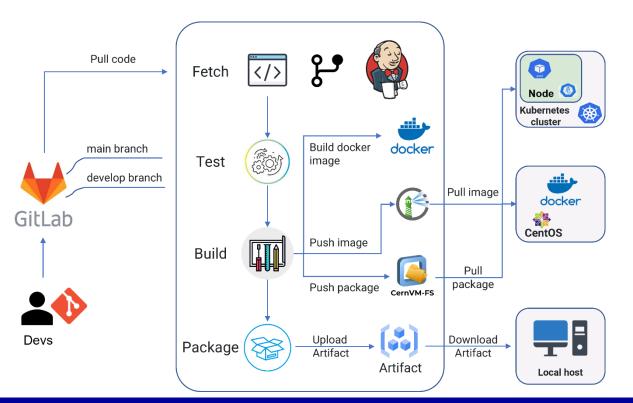


17

# **Developer user support**

Continuous integration/delivery/deployment system for software development

- CI/CD platform, implement full-process automation of code integration -> testing -> building -> deployment for multi-developer
- Simplify and accelerate the software development lifecycle



~											
Daisy	r_main ☆				活动	分支	Pull Requests				
状态	运行	提交	消息	持续时间	完成	12	- 1				
	36	_	test new LoadD	Dashboard > Daisy_main >							
				🖹 Status		P	peline Daisy_ma	ain			
	35	-	Started by user	Changes		Dai	sy项目的main分支构建和测试部				
0	34	-	imp: refine setu	Q Full Stage View					est Result T		(j)
0	33	-	remove Jenkins	🧼 Open Blue Ocean			15				
	32	_	Started by user		<u>trend</u> ~		12				
				Q Filter builds	1		8				
	31	-	Restarted from	2023年12月1日上午10:46					#29 #30 #	31 #32 #33	#34 #35 #3
				<ul> <li>#35</li> <li>2023年11月15日上午9:59</li> </ul>			Daisy.tar.gz 73		v		
				⑦ <u>#34</u> 2023年11月14日下午4:32		St	age View				
				⑦ #33 2023年11月14日下午4:16				Checkout	Test	build-	Declarative: Post Actions
				2023年11月14日下午4:04			Average stage times:		28s	331ms	909ms
				⊘ <u>#31</u>			(Average <u>full</u> run time: ~34s)	•	_		
				2023年11月14日下午3:45		1	12月 01 49 ④ 日 commits 10:46	2s	34s	397ms	939ms
				<u>2023年11月14日下午3:41</u>							
ALLURE REPORT 11/ Minister Vision (Control of State	42023		TRENC			√ Daisy_mai 91≹: -	n < 27 02 251	100188-04	1	31K 525 BH4	*** S ¢ Ð
3		66.66%	2 10 2			HR: -	⊂ ⊙ađayago	Triggered by refulhead	shnain, posted by Rujianii		
		$\mathbf{\bigcirc}$					Start getParameters dr	eckeut codeChecking	bold test		d D
SUITES there are latered	_	1 2						Unit feeting	Performance	he	
	Arrest.		4 4 4						ullez.		
	These excessions and	condite.	CATEGORIES treated	,	-	checkout - 17s					
ENVIRONMENT	S 2 do marcelo		Territes	H or of	-1	V G2	reconstended git tool in: 606				
PEADURES BY STORE											
FEATURES BY STORE	Start		EXECUTORS	Mineral Cany June		1 80 60 1 Ca	g creating light set in the set of the set o		nerixecbecarator\$385457865d am is the Git enecutable no		dsan.reating.Chanei@i2020079:268 ad container)

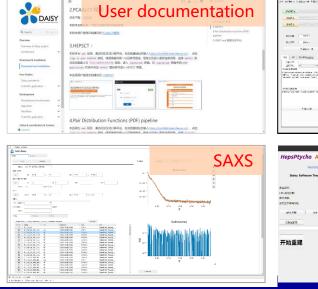
# Daisy applications for synchrotron radiation

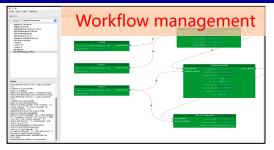
# Serving multiple disciplines: imaging, diffraction/scattering, spectroscopy .....

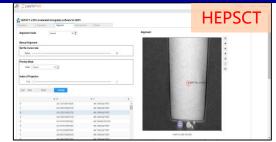
- Daisy-BMX for Biological Macromolecule Crystallography
- XRF-mapping for fluorescence spectrum batch processing
- Hepsptycho for ptychography phase retrieval
- Daisy-PDF for pair distribution function
- XASMatch and PCA&LCF for X-ray absorption spectrometry component analysis
- HEPSCT for X-ray tomography
- Daisy workflow management sys.

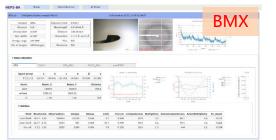
#### More under development:

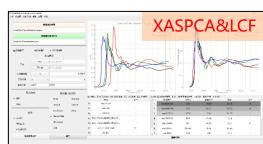
- SAXS
- Holotomography
- XPCS
- Bragg CDI

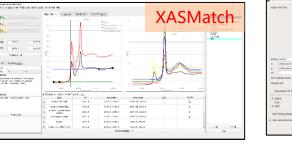




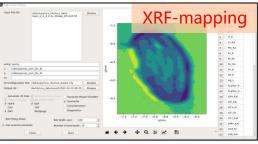












integrate	PDF	
Data: > /oj	transform Zz pipetine pt/jupyter_app_launcher/entries/PDFgul/figs Output:      Same Path as Data	
► Data Files	s Filter & Preview	
	NI: > /opt/jupyter_app_launcher/entries/PDFgui/figs PDF_CFG: > /opt/jupyter_app_lau	ncher/entries/PDFgui/figs
O json	NI: > /opt/jupyter_app_launcher/entries/PDFgui/figs PDF_CFG: > /opt/jupyter_app_lau ings for integration	ncher/entries/PDFgui/figs
⊖ json > Extra sett	+ lobsthabies "abb"-second second se second second sec	ncher/entries/PDFgui/figs
⊖ json > Extra sett	* ropr/papyrer_app_annerer/renormer/ronganings	ncher/entries/PDFgui/figs

# **Heterogeneous computing support**

#### Migration scientific software on ROCm platform (AMD GPUs)

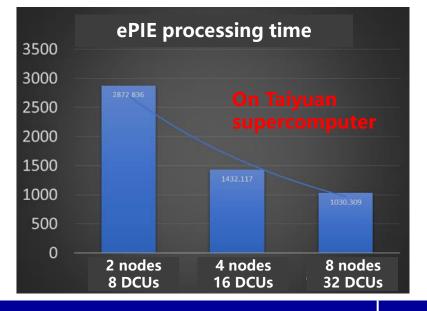
- X-ray tomograpgy: ASTRA, UFO and Tomocupy migrated on Sugon DCU Z100
- X-ray Ptychography: ePIE, DM migrated on Sugon DCU Z100, and deployed on Taiyuan supercomputer(Sugon DCU)
- Deep learning: W1-Net, migrated on Sugon DCU Z100 and HUAWEI Ascend 910a

#### Will be integrated into the framework Daisy

Software	UFO	ASTRA	Tomocupy	ePie, DM	W1-Net
Platform	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100	Sugon DCU Z100、 HUAWEI Ascend 910a
Performance	~ A100	_	_	~25% A100	DCU ~ 66% A100, Ascend: unoptimized









- 2. Demand and Challenges of scientific data and software system
- 3. The architecture and design of the framework
- 4. The progress of the framework

# 5. Summary



- Scientific data analysis framework: Daisy
- Daisy has been applied to light sources and other facilities
- Optimization on stream data processing and distributed computing is in progress
- Serval scientific applications have been developed, more are on the way
- Hope for more cooperation with other facilities and communities

### https://daisy.ihep.ac.cn

# Thank you for your attention!



