

异构计算软件栈进展

张先轶

PerfXLab 澎峰（北京）科技有限公司

xianyi@perfxlab.com

2023.6

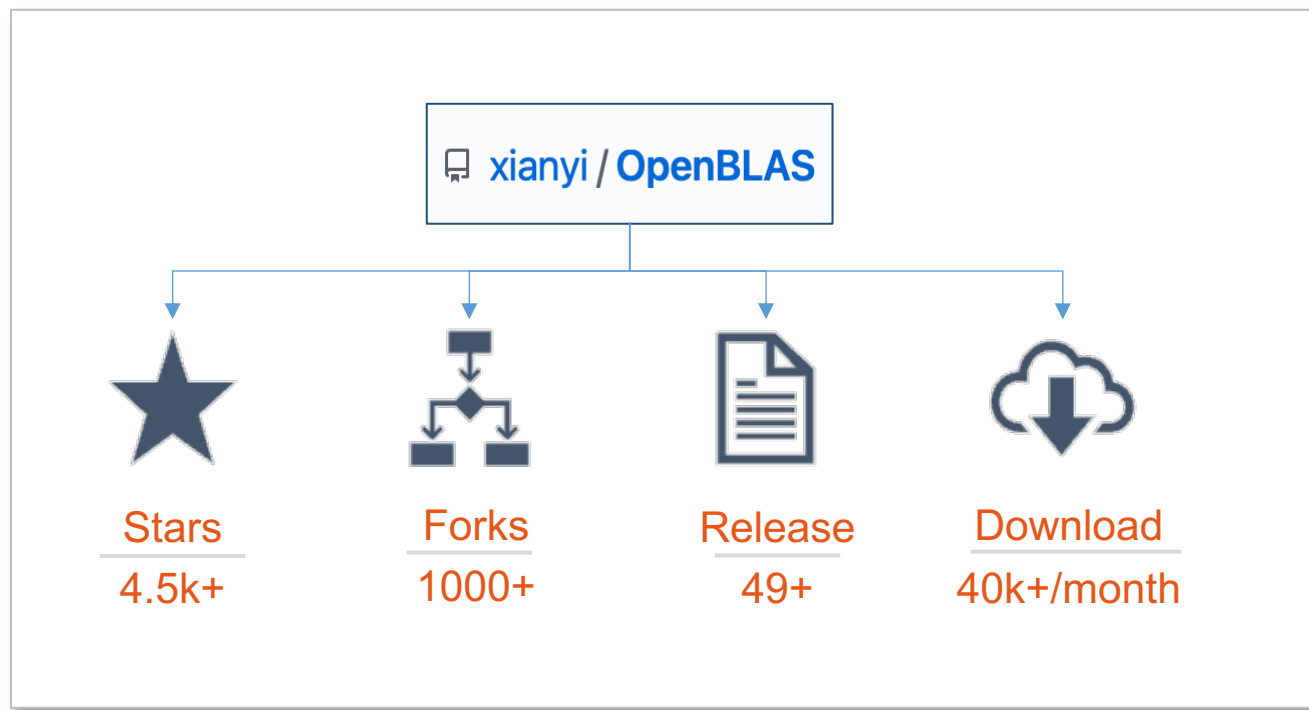
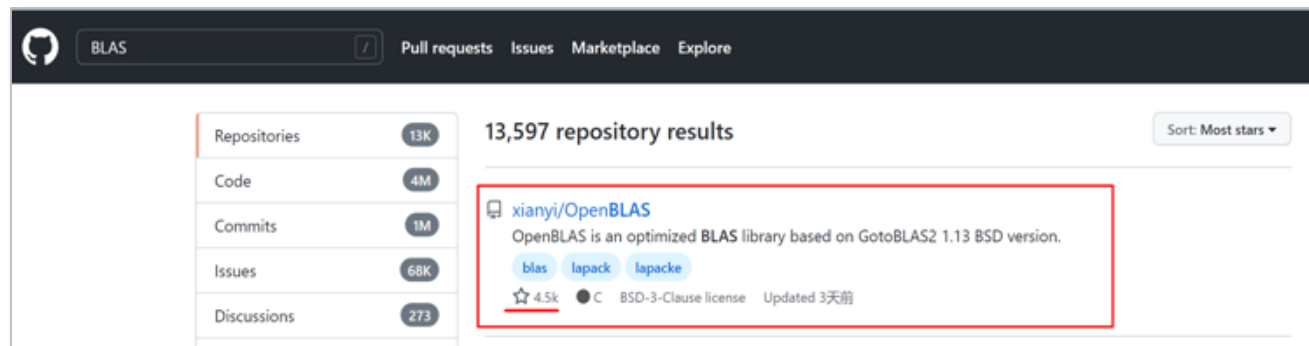
彭峰介绍

- 公司由张先轶博士于**2016年**在北京创立。
- 团队致力于**算力基础软件**研究，十多年的领域积累，国内领先。
- 公司为行业伙伴提供**加速计算解决方案**，为各类计算平台提供**算力软件基座**。

- 2016年，中国计算机学会科技进步二等奖
- 2017年，中国科学院杰出科技成就奖
- 2018年，北京雏鹰人才计划，国家高新企业
- 2021年，数字中国·**集成电路赛道**特等奖
- 2021年，创芯中国·决赛一等奖
- 2021年，CRVA联盟，软件工作组副组长单位
- 2022年，OpenCAX SIG10组长单位
- 相关著作：《高性能科学与工程计算》，《OpenCL异构计算》，《并行编程模式》，《视频图像处理与性能优化》，《人工智能三驾马车》等



全球开源矩阵计算库OpenBLAS介绍



- OpenBLAS 是国际知名的开源矩阵计算库，同领域全球排名前三，由公司创始人张先轶博士（中国）在2011年发起并管理维护，是极少有的由中国人发起并管理的全球开源项目。也被下述项目或公司核心产品作为基础组件使用：



Caffe

GNU Octave



- OpenBLAS 是科学计算、人工智能等领域的基础软件和根技术。在支持国产CPU、GPU、DSA“大芯片”做出了卓越贡献，并得到了行业的认可。

异构计算语言OpenCL顾问委员会成员

Khronos Group Advisory Panel Participation Agreement

This agreement enables membership on
Advisory Panels for one or more Khronos Working Groups
**PLEASE TYPE OR PRINT CLEARLY: THIS IS A LEGAL DOCUMENT
ILLEGIBLE AGREEMENTS CANNOT BE PROCESSED**

Contact for processing this agreement:

NAME:	Xianyi Zhang
COMPANY:	PerfXLab Technology Co., Ltd.
EMAIL:	xianyi@perfslab.com
PHONE:	+86-13466545921

Guidelines for completing this agreement:

1. This Advisory Panel agreement is normally executed by invited individuals, however if you wish to execute on behalf of a company/institute which is a legal entity that you are able to represent, and has been invited to participate in the Advisory Panel, you may execute this agreement on behalf of that entity and designate colleagues within that entity that will also be bound by this agreement and so able to access Advisory Panel resources.
2. Please enter your name, email address on this cover page and select the working groups to which you have been invited to advise on the signature page.
3. Select the tick box on the signature to confirm whether you are executing this agreement as an individual or on behalf of a legal entity. If you are executing this on behalf of an entity, then enter the name of the legal entity on this cover page and as the Signatory on the signature page (instead of your own name) and enter a list of Designated Colleagues in Attachment B. If the legal entity requires more than one authorized signatory, duplicate the signature page for each required signatory.
4. Mail a copy OR email a PDF of the executed agreement to the contact address below. If mailed, send two signed copies and one completed copy will be returned for your records

Contact Details:

Khronos Group Member Services
Khronos Group Inc, 9450 SW Gemini Drive #45043, Beaverton, OR 970088-6018, USA
memberservices@khronos.org
Voice mail: +1 (415) 869-8627

Your **Advisory Privileges** will commence when Khronos has acknowledged receipt of the executed agreement.

Advisory Panel Participation Agreement Signature Page

For Khronos

9450 SW Gemini Drive #45043,
Beaverton, OR 97008-6018 USA

For Signatory

Xianyi Zhang

Signatory name

Room 304, Cuihu Building 9#, No.55 Zique
Road

Street address

Beijing, 100190, China

City, State, ZIP, Count

Emily Stearns

Printed Name

Managing Director

Title

emily@khronosgroup.org

Email Address

March 16, 2023

Date of Signature

Xianyi Zhang

Printed Name

Title (if applicable)

xianyi@perfslab.com

Email Address

Mar 16, 2023

Date of Signature

Tick ONE of the following boxes:

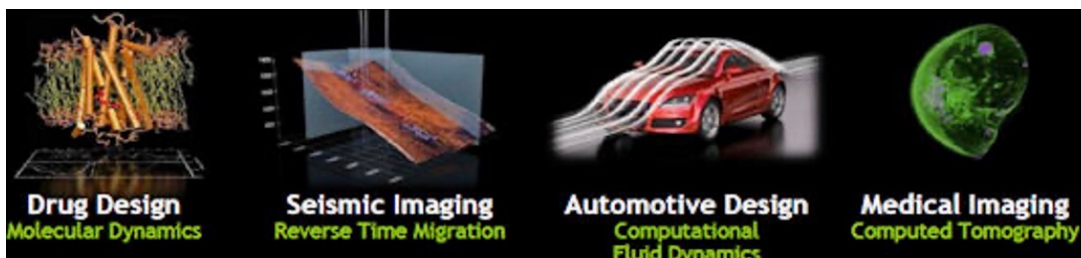
- ☐ Signing on behalf of legal entity as entered as Signatory above (and see Attachment B)
- ☒ Signing as individual as entered as Signatory above

Selected Working Groups:

Working groups in the Graphics and Compute IP Zone:

☐ Vulkan ☒ OpenCL ☐ SYCL ☐ Vulkan SC

今年三月，张先轶博士
受邀加入异构计算语言
OpenCL顾问委员会。



异构计算编程

异构计算库与框架

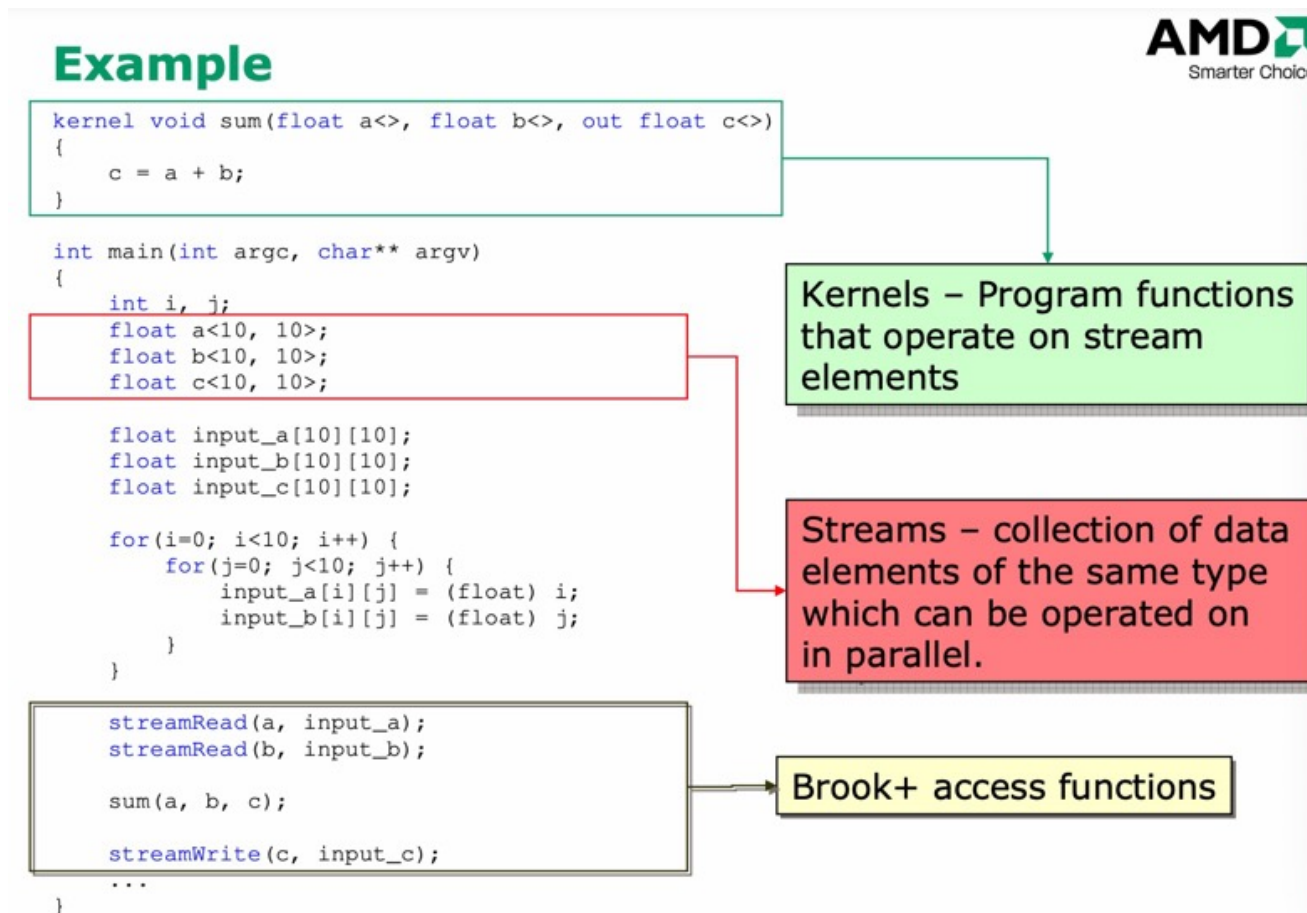
开发工具

高性能异构计算平台

x86/ARM/RISC-V CPU, GPU, NPU/DSA, FPGA

异构计算编程模型

- 异构计算语言
- CUDA (2007)
 - 异构计算代名词
- Brook+ (2007)
 - BrookGPU
 - 有事烧纸
- OpenCL (2009)
 - 试图与CUDA竞争
- SyCL (2014)
 - Intel新宠
- OpenACC (2012)
 - 类似CPU编程, 类似CPU的性能



CUDA vs OpenCL 编程模型差别不大

- CUDA和OpenCL 编程模型相差不多
 - OpenCL相对繁琐一些
- 开放标准和开源就能胜利吗？
- OpenCL 标准自身的弯路
 - OpenCL 2.0标准，大量芯片公司不支持
 - 编程模型与硬件体系结构密切相关
 - 2.0加入的大量功能特性，不实用
 - 基于广泛使用的1.2，发展3.0

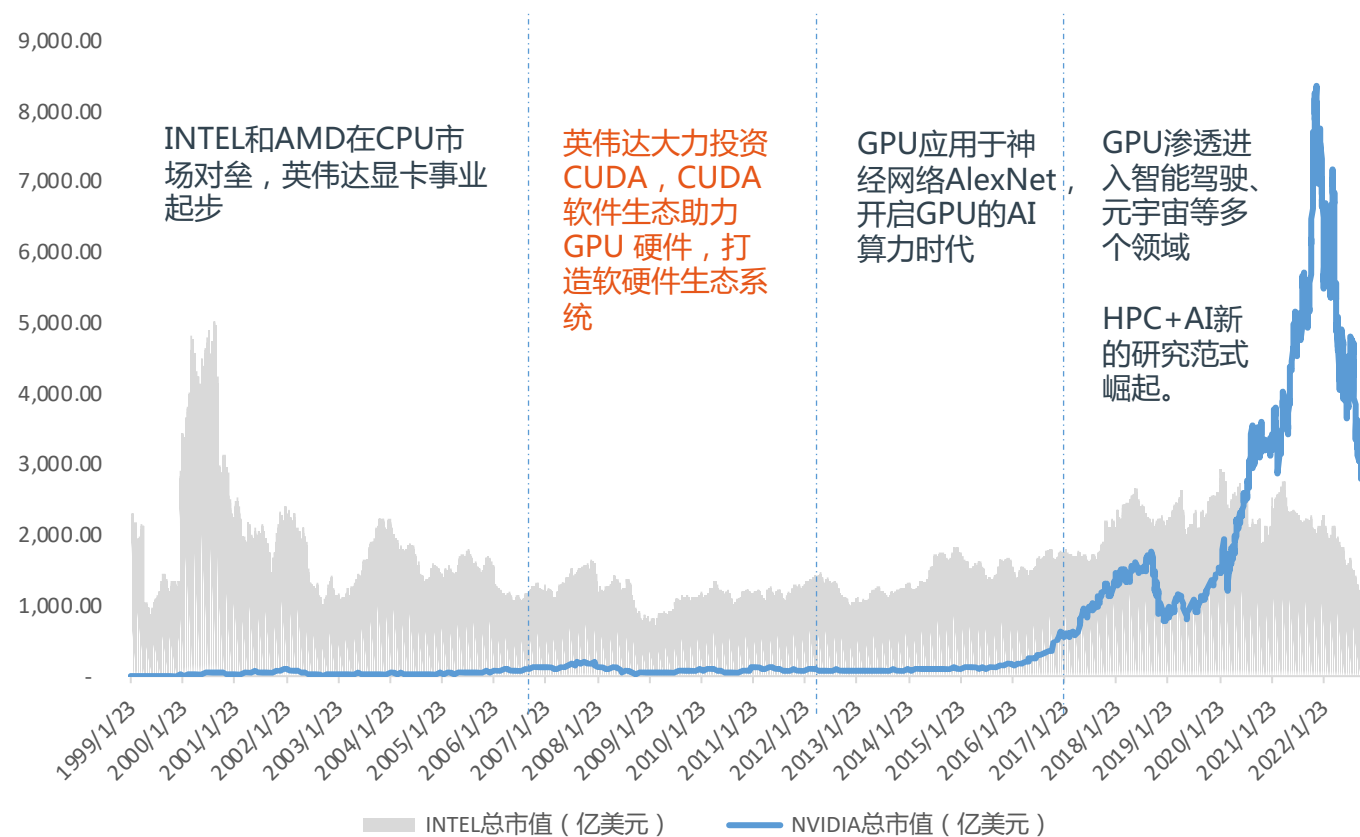
OpenCL 3.0 Specification Finalized and Initial Khronos Open Source OpenCL SDK Released

🕒 September 30, 2020 by OpenCL Working Group 🔗 [opencl](#)

Today, the Khronos® OpenCL™ Working Group is happy to announce the release of the finalized [OpenCL 3.0 specifications](#), including a new unified OpenCL C 3.0 language specification, together with an early initial release of a Khronos OpenCL SDK to enable developers to quickly get up to speed using OpenCL.

CUDA vs OpenCL的库和框架生态差距大

- 库/框架差距巨大
 - LAPACK/Magma
 - PETSc
 - OpenCV
 - FFmpeg
 - AI框架等
(PyTorch/PaddlePaddle/Oneflow...)
 - ...
- 直接写CUDA的人不多，但是用支持CUDA库和框架的人很多

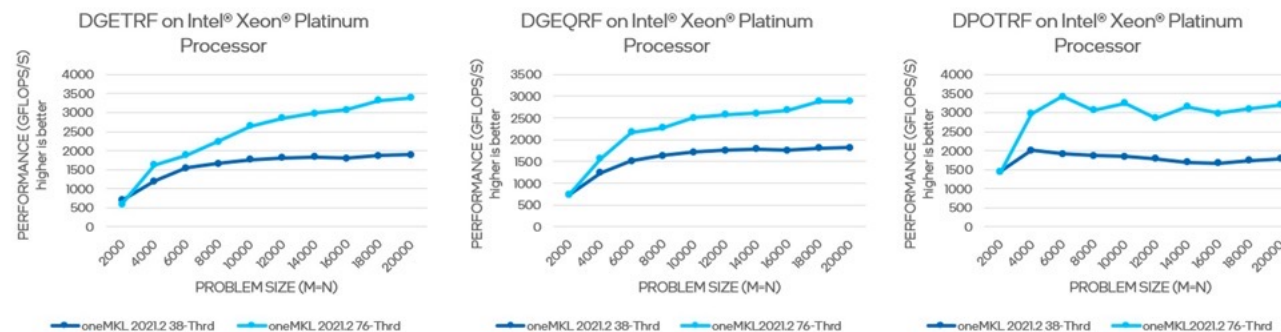


已有计算领域软件生态（库/框架）

- Intel公司
 - MKL库（数学库、几千个函数）
 - 稠密矩阵计算BLAS/LAPACK
 - 快速傅里叶变换FFT
 - 稀疏矩阵计算 SparseBLAS
 - 稀疏解法器（直接、迭代）
 - 随机数发生器
 - 向量数学库VML
 - 统计加速..
 - DAAL库（数据分析）
 - IPP库（图像、信号处理）
 - MKL-DNN库（深度学习）
 - 持续增长、持续优化

oneMKL Linear Algebra Package (LAPACK) Scales Best at 76 Threads

Intel® oneAPI Math Kernel Library (oneMKL) 2021.2 LAPACK on Intel® Xeon® Platinum Processor



Testing Date: Performance results are based on testing by Intel as of April 26, 2021 and may not reflect all publicly available security updates.

Configurations: 1 node, 2x Intel® Xeon® Platinum 8368 processor on Coyote Pass platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0xd000270, HT off, Turbo off, Red Hat 8.3, 4.18.0-240.el8.x86_64, 1x Intel® SSD 960GB OS Drives, Intel® oneAPI Math Kernel Library (oneMKL) 2021.2.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at www.intel.com/PerformanceIndex. Your costs and results may vary.

已有计算领域软件生态（库/框架）

- ARM公司
 - APL库
 - 稠密矩阵计算BLAS/LAPACK
 - 快速傅里叶变换FFT
 - 64-bit ARM，包含ARM Neoverse (SVE指令集)
 - ACL库
 - ML应用
 - ARM CPU/GPU
- NVIDIA公司
 - CUDA-X，GPU加速库、框架和工具等
 - 数学库、并行算法库、图像视频库
 - 通信库、深度学习库等

Math Libraries

GPU-accelerated math libraries lay the foundation for compute-intensive applications in areas such as molecular dynamics, computational chemistry, medical imaging, and seismic exploration.



cuBLAS

GPU-accelerated basic linear algebra (BLAS) library

[Learn More](#)



cuFFT

GPU-accelerated library for Fast Fourier Transforms

[Learn More](#)



CUDA Math Library

GPU-accelerated standard mathematical function library

[Learn More](#)



cuSOLVER

GPU-accelerated dense and sparse direct solvers

[Learn More](#)



cuSPARSE

GPU-accelerated BLAS for sparse matrices

[Learn More](#)



cuTENSOR

GPU-accelerated tensor linear algebra library

[Learn More](#)

国内计算芯片的生态现状

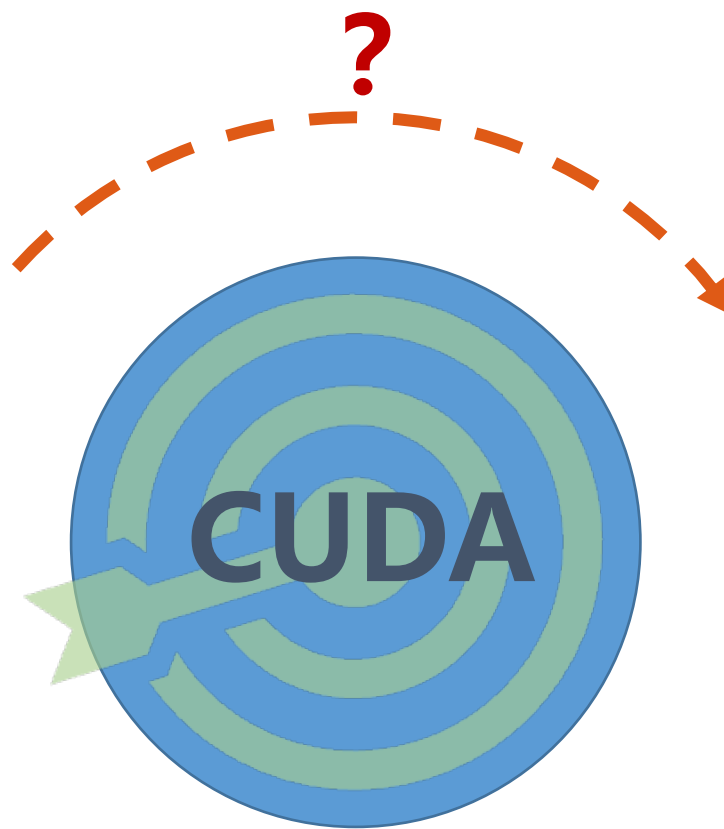
- 老牌HPC强队
 - 科学计算领域已有一定积累
 - 异构编程方式和库API等，各家一套
- 新兴AI芯片公司
 - 兼容CUDA
 - 独立自研

2022年相继收购
Codeplay, Arrayfire

Intel
OneAPI

AMD
ROCm

中国芯



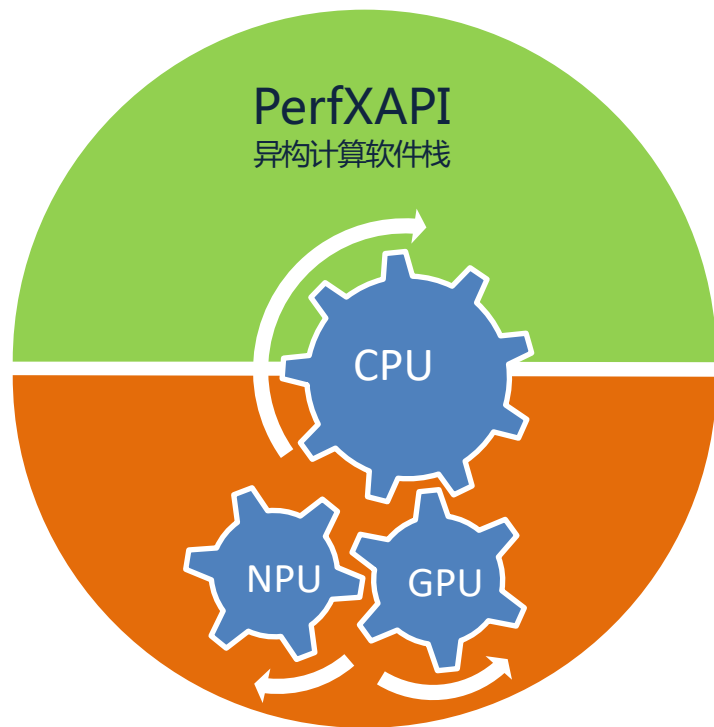
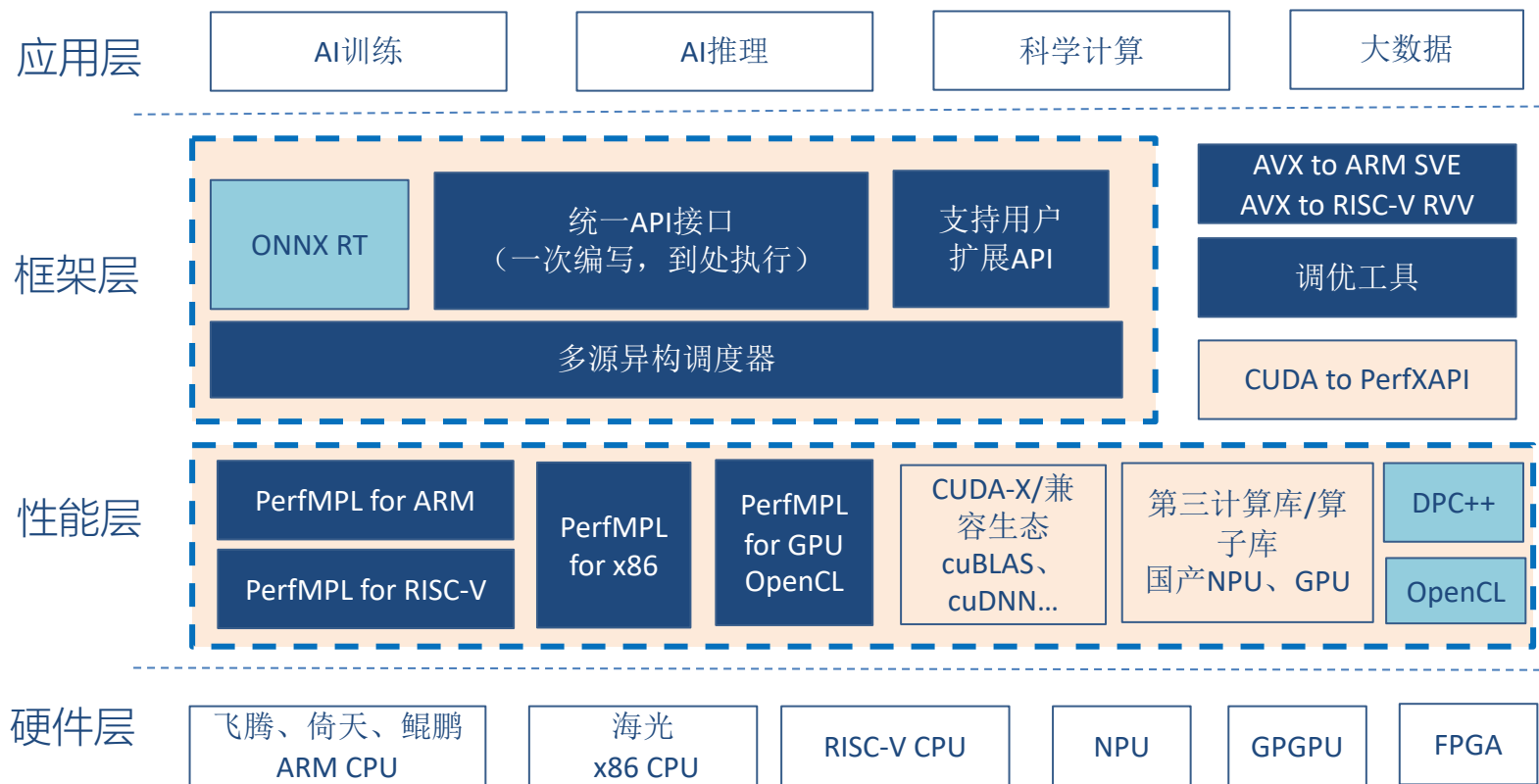
算力时代的开启与爆发

- 人工智能
- 工业软件
- AI for Science

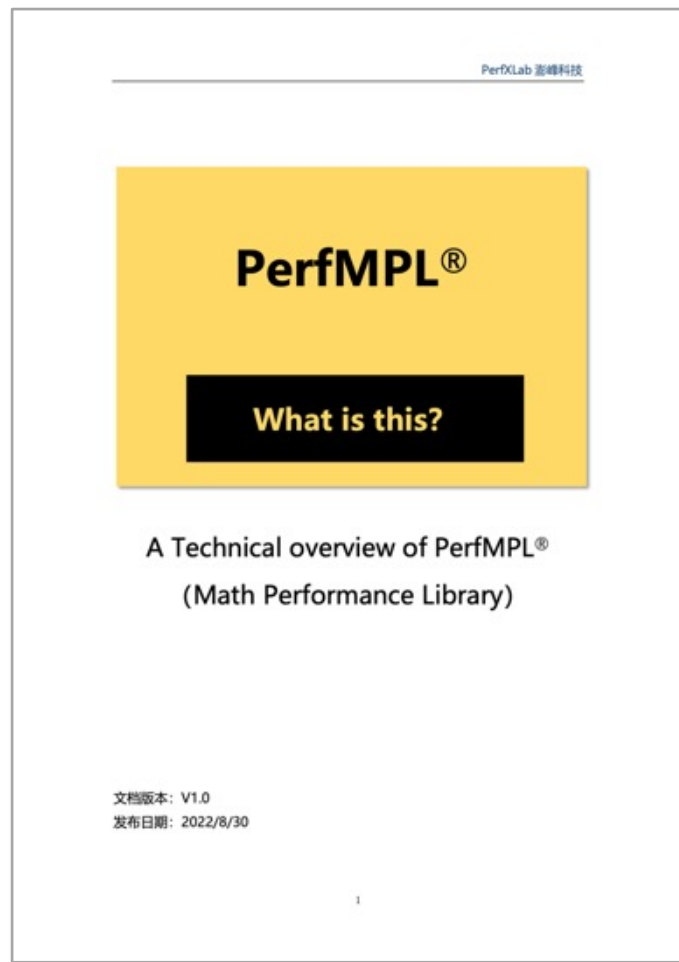
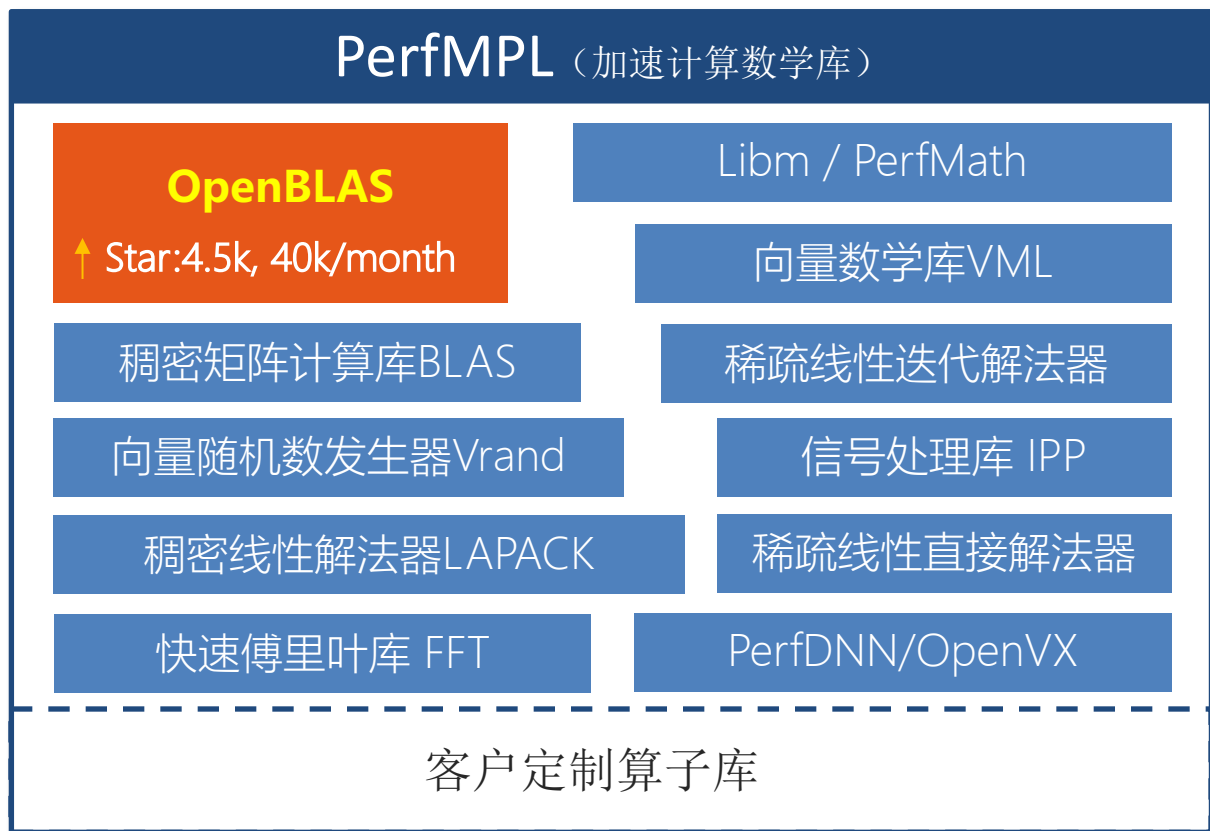
参与开源，共建生态，立足国内，走向国际

PerfXAPI异构计算软件栈

PerfXAPI 围绕应用侧需求，是旨在创建一个开放、基于开源标准的**跨架构**API编程模型，在面对大量跨各种架构的硬件和复杂工作负载时简化开发工作。特性：1) 支持多种异构设备；2) 统一API调用接口；3) 高性能



高性能数学库PerfMPL



- **标准版本**：PerfMPL for ARM , PerfMPL for RISC-V (RVCL)
- **领域版本**：为应用领域提供定向优化。例如EDA、CFD、VSIPL、CV、Big Data等。
- **定制版本**：为国产GPU/NPU厂商提供PerfMPL的定制服务，对标 cuBLAS , cuFFT , cuMath...。

OpenBLAS项目

- 开源矩阵计算库，12年历史
 - <https://github.com/xianyi/openblas>
- 矩阵和向量基本计算
 - 当前版本：0.3.21
- 支持主流处理器
 - X86/ARM/Power/龙芯
- 支持多种操作系统
 - Linux/Windows/Mac OSX/Android
- 目前支持RISC-V架构
 - 已经支持C910 RVV指令优化（RVV 0.7.1，1.0进行中）
 - RVV 1.0进行中



Caffe

IBM

Ubuntu
openblas packagedmlc
mxnet

julia



GNU Octave

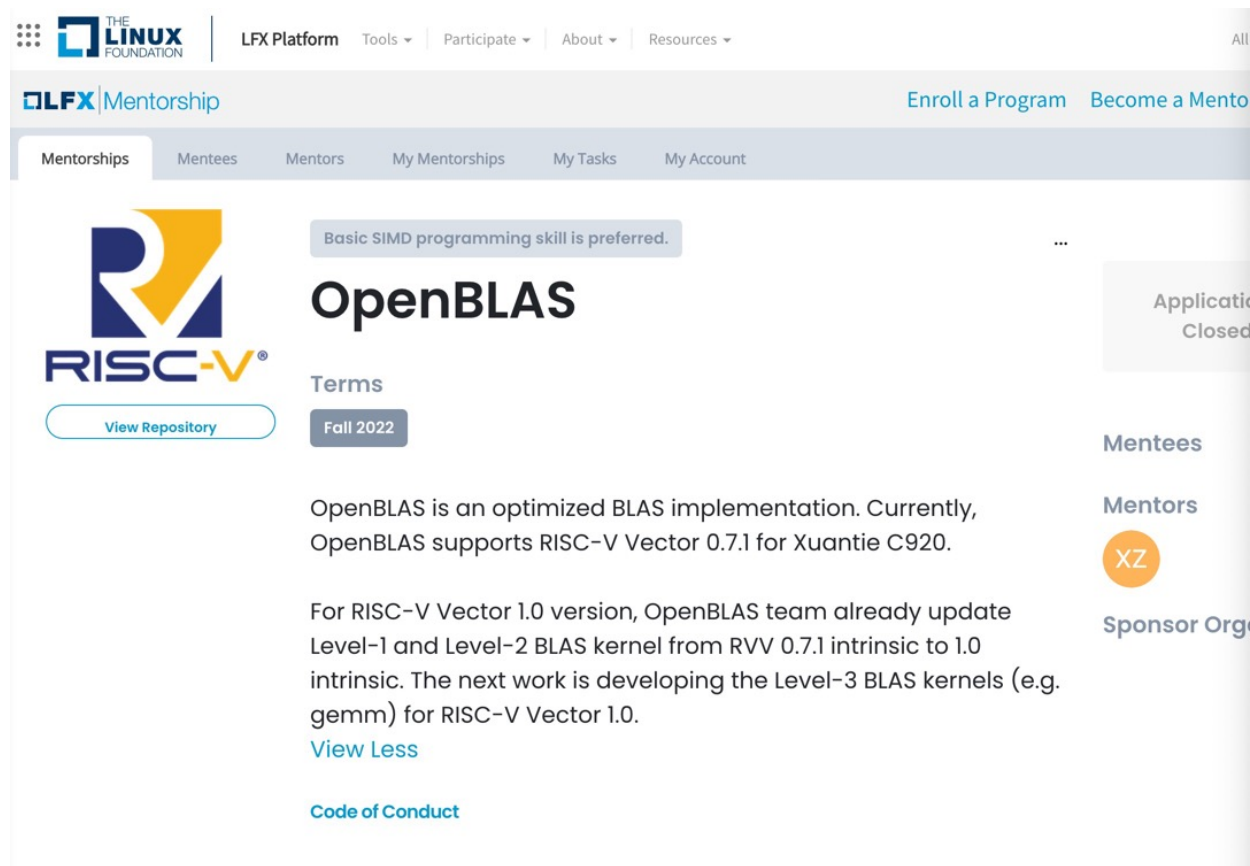
網易 NETEASE
www.163.com

LOONGSON 龙芯



OpenBLAS RVV 1.0支持进展

- BLAS 1级/2级
 - 向量化采用intrinsic实现
 - 已经支持RVV 1.0 intrinsic规范
- BLAS 3级
 - 当前向量化采用内嵌汇编实现，RVV 0.7.1
 - 9月份开始，进行RVV 1.0移植



The screenshot shows the LFX Mentorship program page for OpenBLAS. The page is part of The Linux Foundation's LFX Platform. It features a navigation bar with links for LFX Platform, Tools, Participate, About, and Resources. The main header includes the LFX Mentorship logo and links to 'Enroll a Program' and 'Become a Mentor'. The page is divided into sections for 'Mentorships', 'Mentees', 'Mentors', 'My Mentorships', 'My Tasks', and 'My Account'. The 'Mentorships' section is active, displaying the OpenBLAS logo, a 'View Repository' button, and a 'Terms' section. The 'Terms' section includes a 'Fall 2022' button and a description of the program: 'OpenBLAS is an optimized BLAS implementation. Currently, OpenBLAS supports RISC-V Vector 0.7.1 for Xuantie C920. For RISC-V Vector 1.0 version, OpenBLAS team already update Level-1 and Level-2 BLAS kernel from RVV 0.7.1 intrinsic to 1.0 intrinsic. The next work is developing the Level-3 BLAS kernels (e.g. gemm) for RISC-V Vector 1.0.' There is also a 'View Less' link and a 'Code of Conduct' link. On the right side, there is a sidebar with a 'Mentees' section showing a list of mentees, including one with the initials 'XZ'.

THE LINUX FOUNDATION

LFX Platform Tools Participate About Resources

LFX Mentorship

Enroll a Program Become a Mentor

Mentorships Mentees Mentors My Mentorships My Tasks My Account

Basic SIMD programming skill is preferred.

OpenBLAS

Terms

Fall 2022

OpenBLAS is an optimized BLAS implementation. Currently, OpenBLAS supports RISC-V Vector 0.7.1 for Xuantie C920.

For RISC-V Vector 1.0 version, OpenBLAS team already update Level-1 and Level-2 BLAS kernel from RVV 0.7.1 intrinsic to 1.0 intrinsic. The next work is developing the Level-3 BLAS kernels (e.g. gemm) for RISC-V Vector 1.0.

[View Less](#)

[Code of Conduct](#)

Application Closed

Mentees

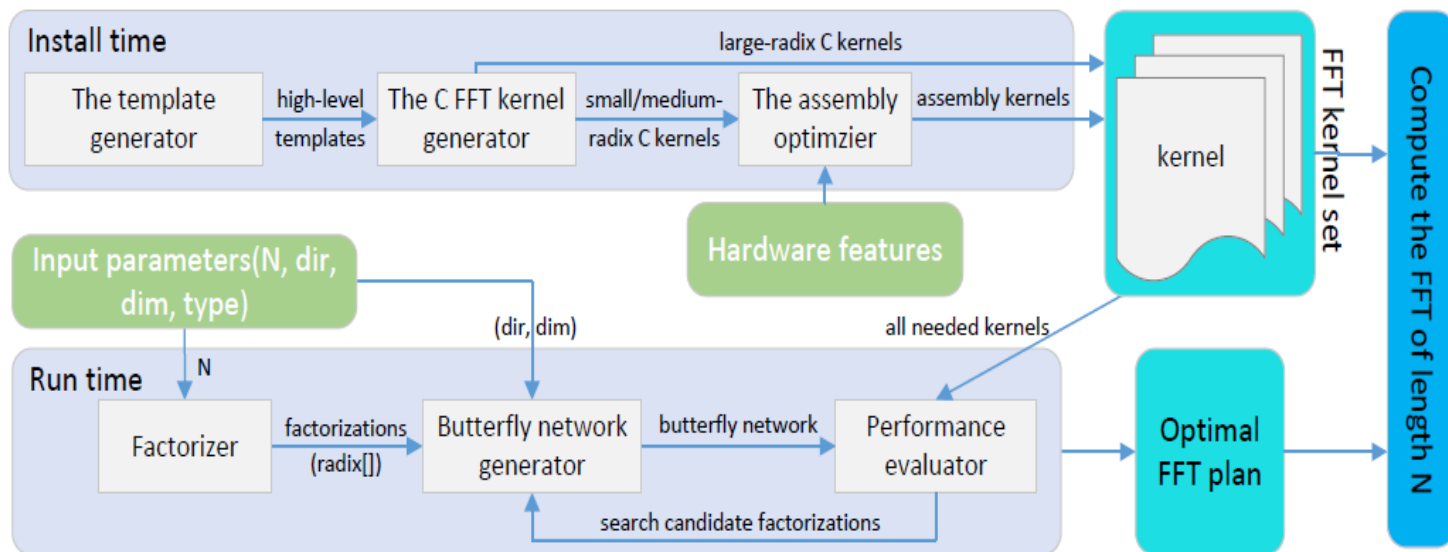
Mentors

XZ

Sponsor Org

高性能FFT库OpenFFT/PerfFFT

- 自主研发(不基于FFTW), 版权协议可控
 - 采用Cooley-Tukey FFT算法
 - 核心代码纯汇编编写, 高效实现了 $2^m * 3^n * 5^k * 7^j * 11^i * 13^l$
 - 支持多种精度、多种维度、包括复数和实数在内的多种变换类型
- 支持主流处理器, 性能优秀
 - X86/ARM/RISC-V



典型应用

- 信号与信息处理使用傅里叶变换应用非常广泛。所以, FFT运算能力已经成为处理器的性能标尺之一。
- 例如图像领域中的压缩,

OpenFFT相对其他FFT库性能提升

	Intel Xeon E5-2670 V3	ARMV8
Intel MKL 2019	60%	NA
ARM PL	NA	72%
FFTW	138%	61%

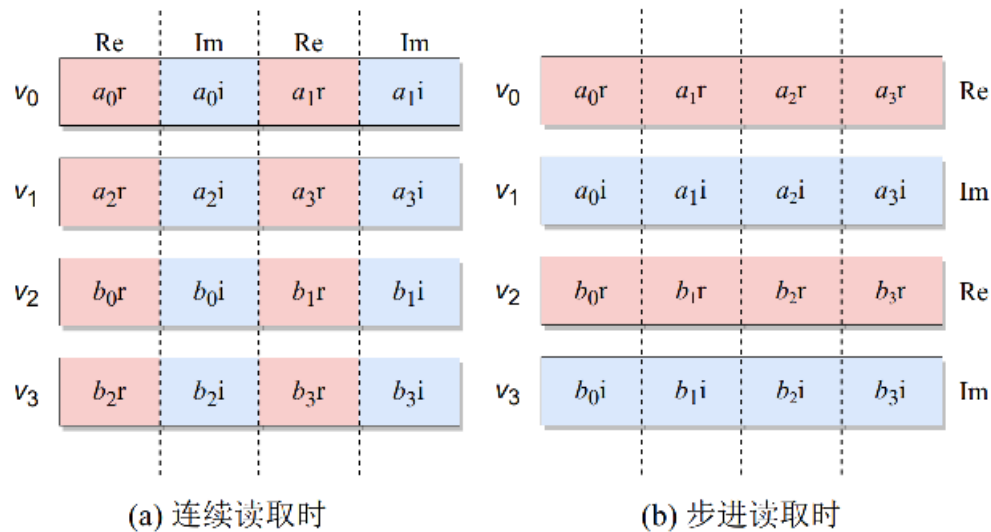
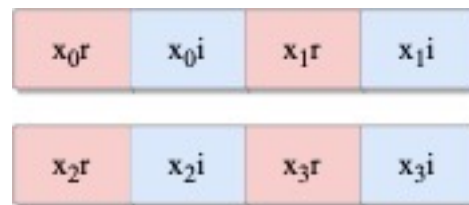
高性能FFT库OpenFFT/PerfFFT

• 数据访存——读取方便

- 实部、虚部交替存储
- 二维数组列优先存储
 - 实部虚部之间的计算——冗余指令
- 行优先读取

```
vlseg<nf>e.v vd, (rs1)
# offset = sew/8
# for (k = 0; k < nf; k++)
#   V_(d+k)[i] = mem[rs1 + k*offset + i*nf*offset]
```

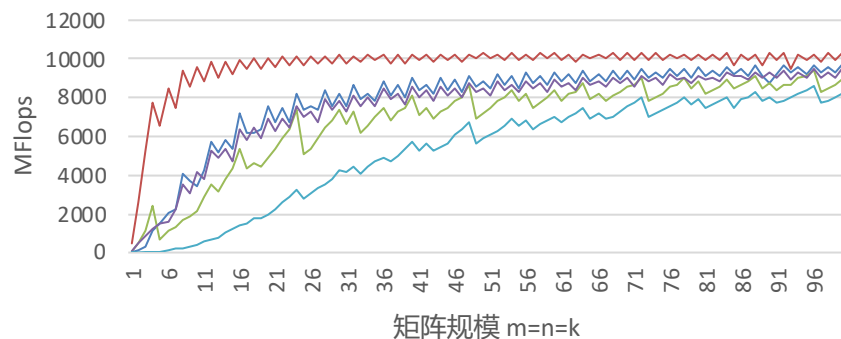
```
v0[0]=mem[x5],      v0[1]=mem[x5+4],
v0[2]=mem[x5+8],    v0[3]=mem[x5+12]
v1[0]=mem[x5+16],   v1[1]=mem[x5+16+4],
v1[2]=mem[x5+16+8], v1[3]=mem[x5+16+12]
```



高性能数学库 (PerfMPL) on ARM CPU

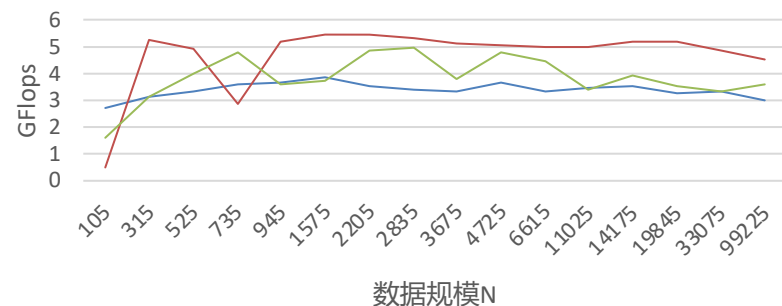
优势：1) 性能对标国际；2) 一致精度；3) 安全、可靠、稳定。

双精度实数 GEMM NN单线程性能 (越高越好)



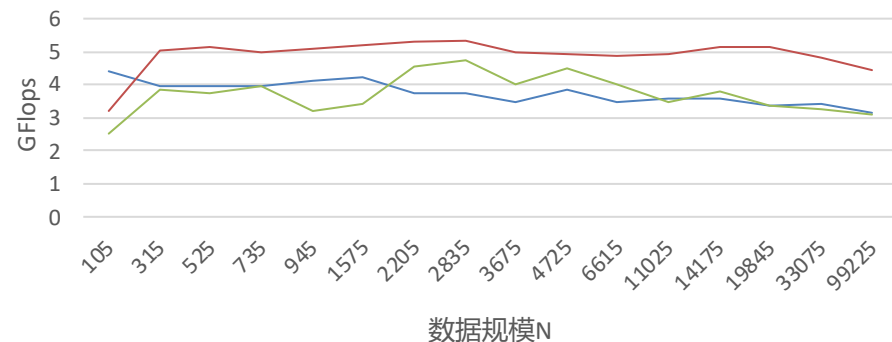
华为鲲鹏920

1D FFT C2C InPlace单线程性能 (越高越好)

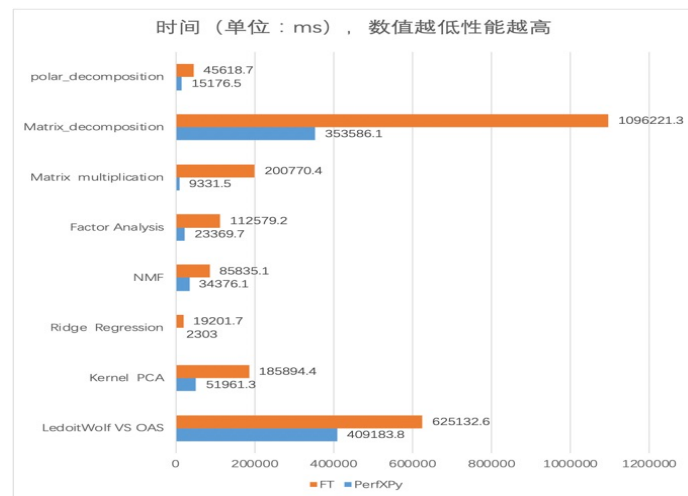


FT-2000 (ARM)

1D FFT C2C Out-of-Place单线程性能 (越高越好)



FT-2000 (ARM)

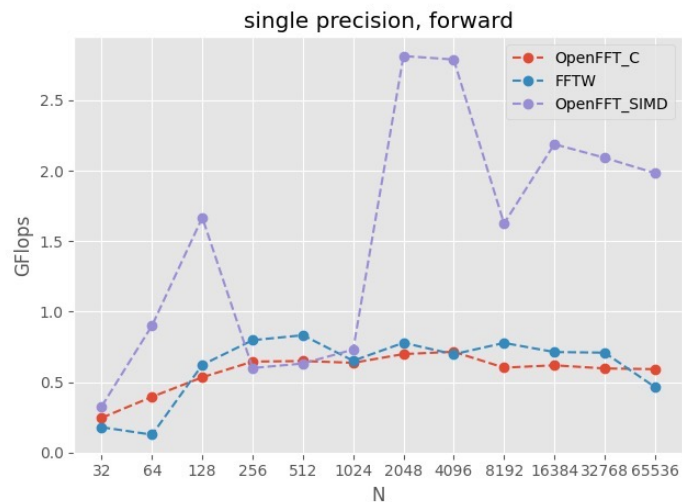


FT-2000/4

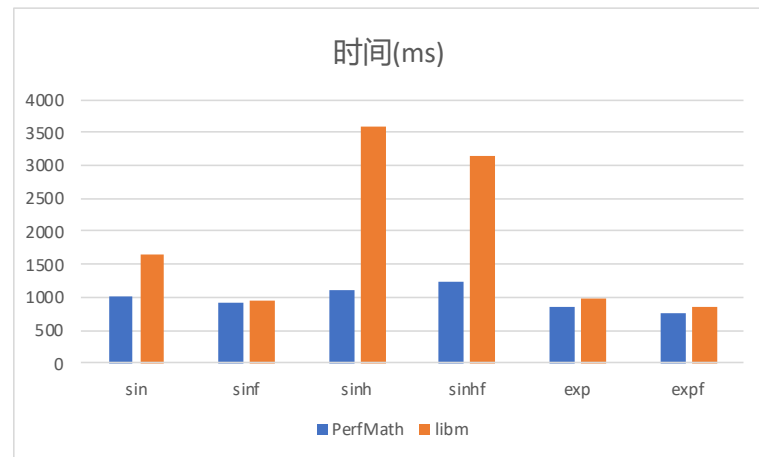
PerfMPL应用在PerfXPy科学计算软件上，与原生Python比较，带来性能大幅提升。

高性能数学库 (PerfMPL) on RISC-V

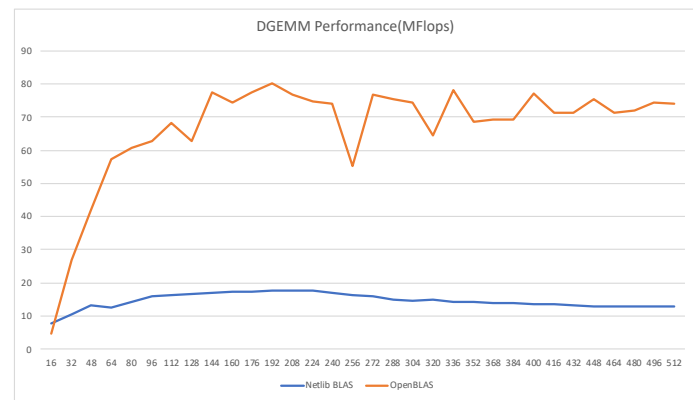
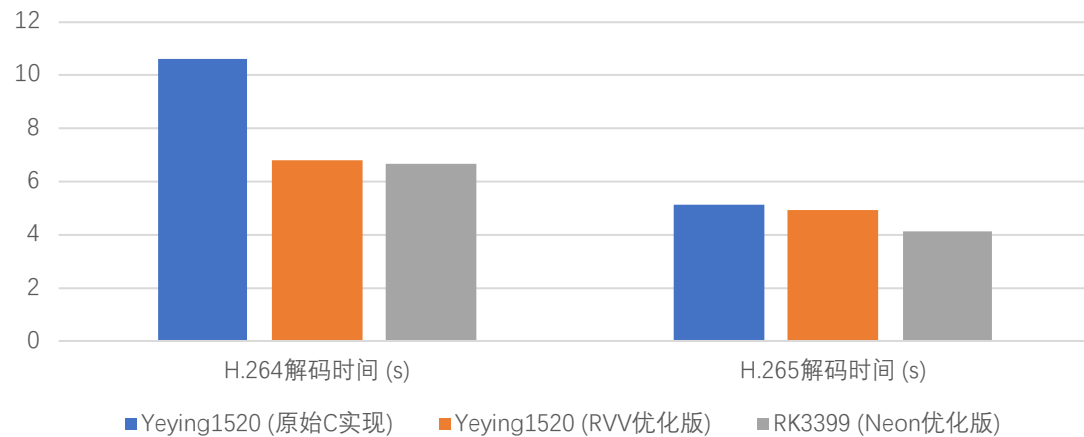
优势：1) 性能对标国际；2) 一致精度；3) 安全、可靠、稳定。



RISC-V C920

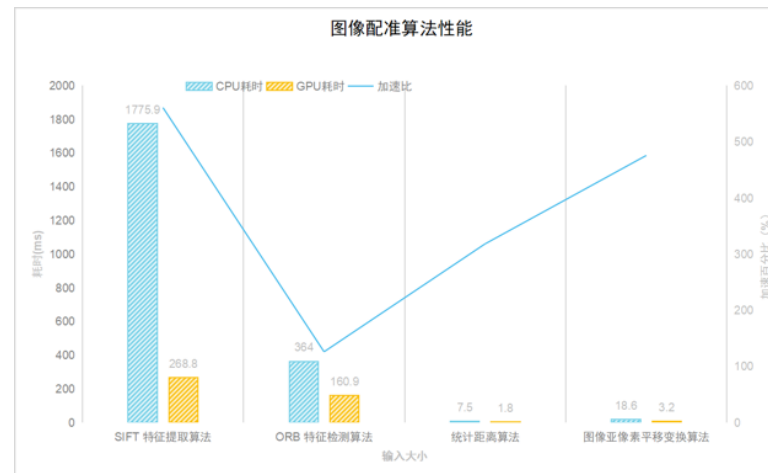
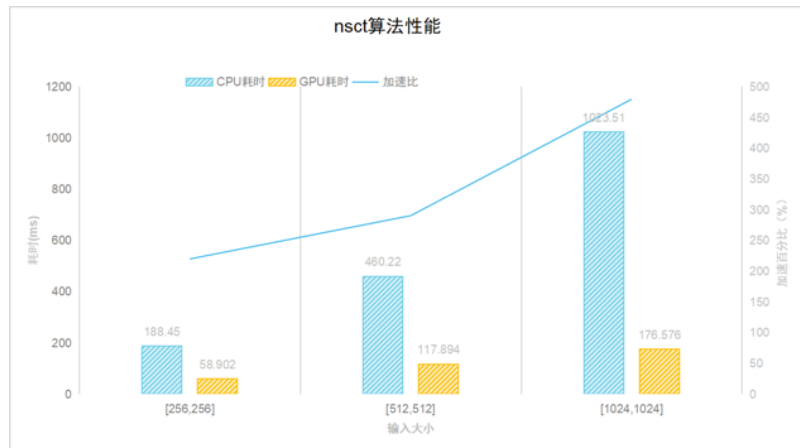
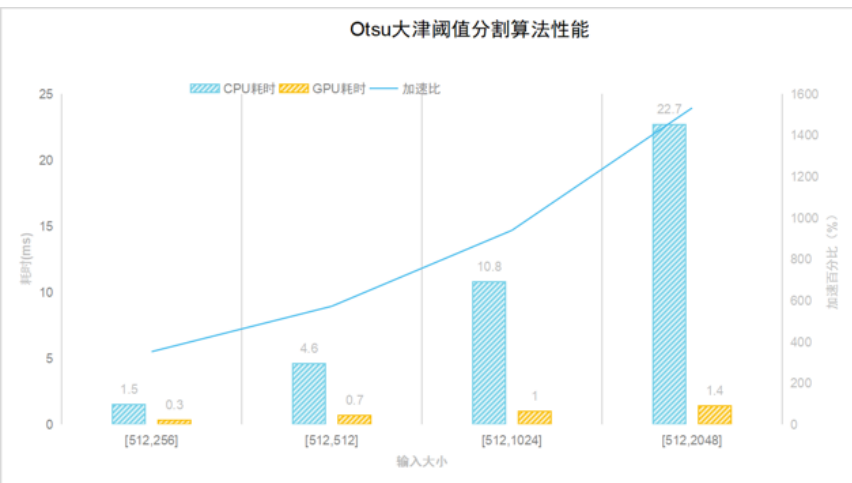
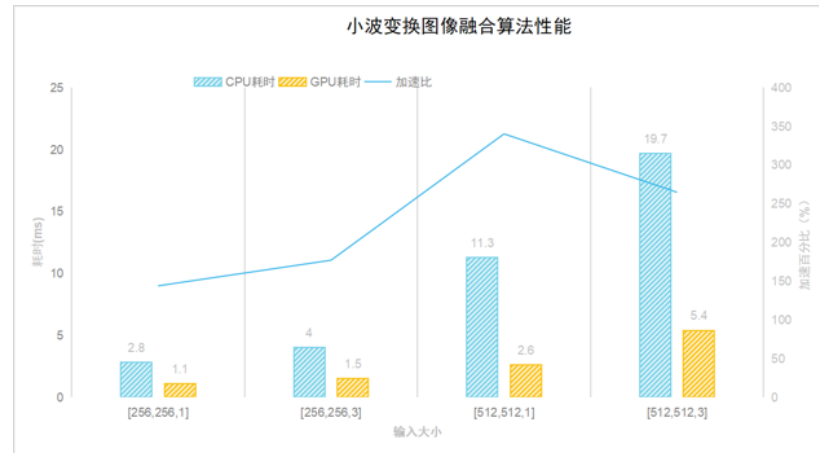
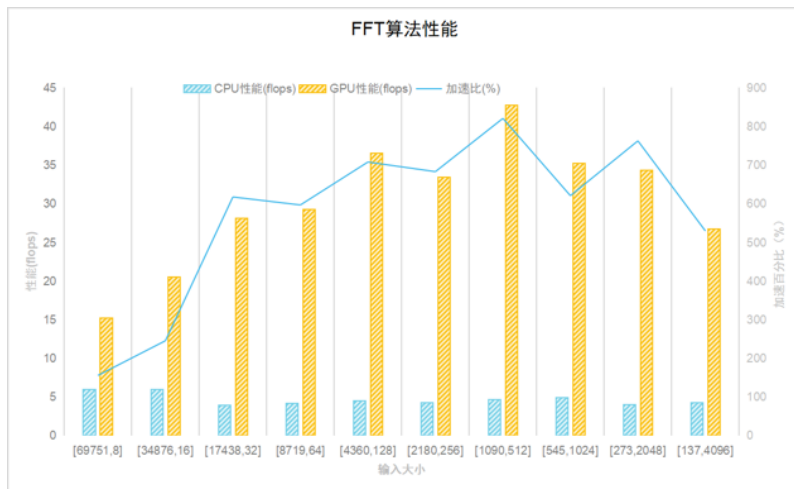


FFMpeg 解码计算时间



高性能数学库 (PerfMPL) on OpenCL GPU

- GPU型号：景嘉微9230
 - 支持OpenCL 3.0
 - FP32峰值性能：1.2T FLOPS



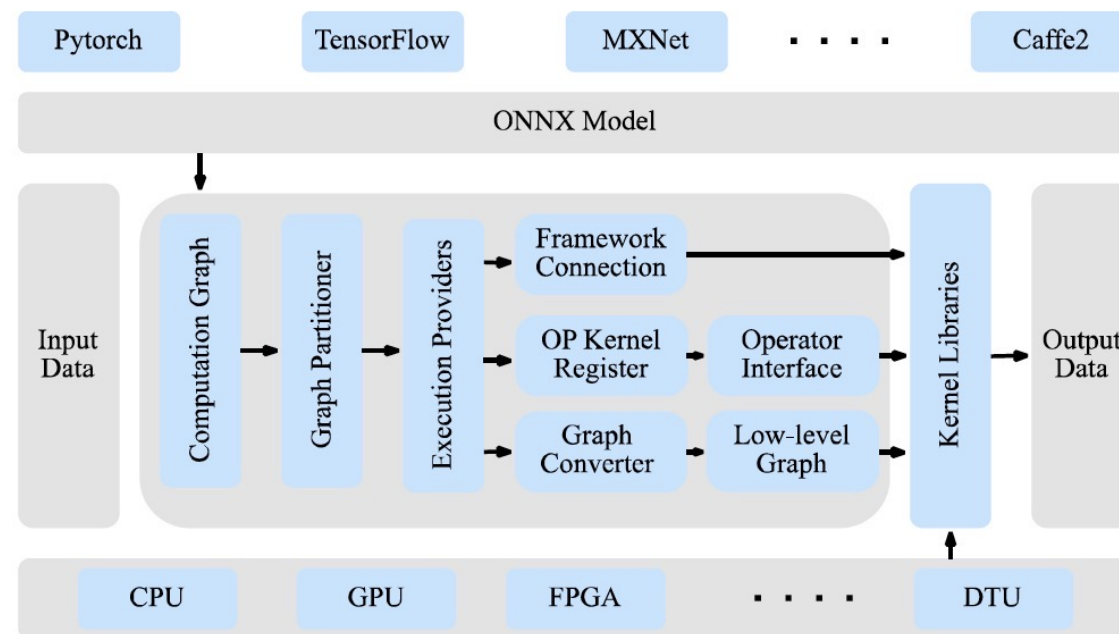
AI推理框架 – InferXLite 基于ONNXRuntime

- NPU硬件平台多样
 - 支持的算子种类不同
 - Onnx opset ?
 - 精度区别 , int8? Fp16? Fp32?
- 人工参与过多
 - 模型转化工作大,学习成本高
 - 各个芯片厂商都提出一套自己的工具链
 - 手工拆分图
 - 不同子图运行到不同的NPU上
 - 支持不同的训练框架
 - PyTorch, Tensorflow, PaddlePaddle,

ONNXRuntime介绍

- 微软领导开源项目
 - <https://github.com/microsoft/onnxruntime>
- 跨平台
- ONNX格式
 - Opset 版本更新最快
- 已支持多种后端
 - CPU (x86, ARM)
 - GPU (NVIDIA, AMD)
 - FPGA (Xilinx Vitis AI)
 - ...
- 支持新算子注册模式

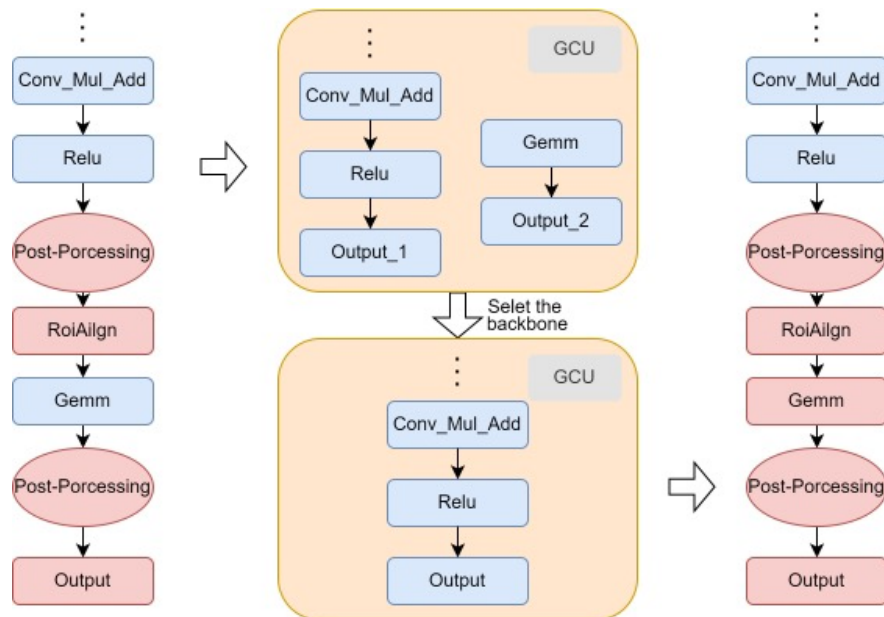
ORT架构图



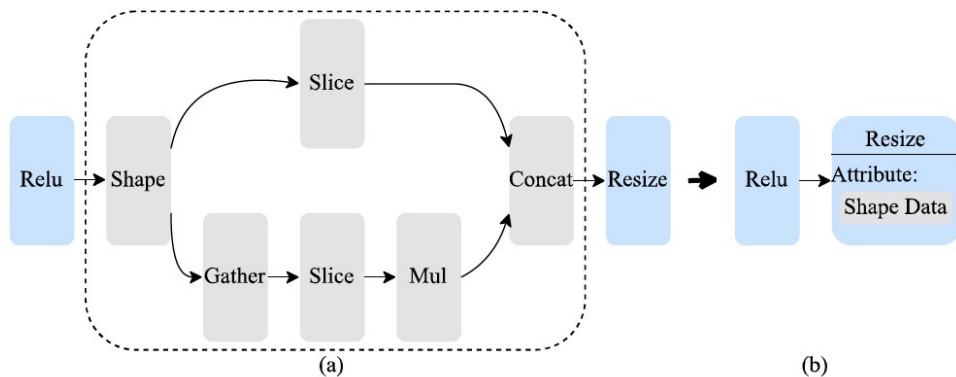
AI推理框架 – InferX Lite工作

- 支持更多国产异构硬件
 - 燧原I20 (已完成)
 - 华为Atlas (已完成)
 - OpenCL GPU (初步完成)
 - 寒武纪 (进行中)
- 增强易用性和性能
 - 面向异构设备的自动切图
 - 设备相关图优化等

PerfXLab



常量折叠



智能计算案例 – AI推理框架OpenCL加速



AMD V1605B APU + InferXLite 2.0 **VS** NVIDIA TX2

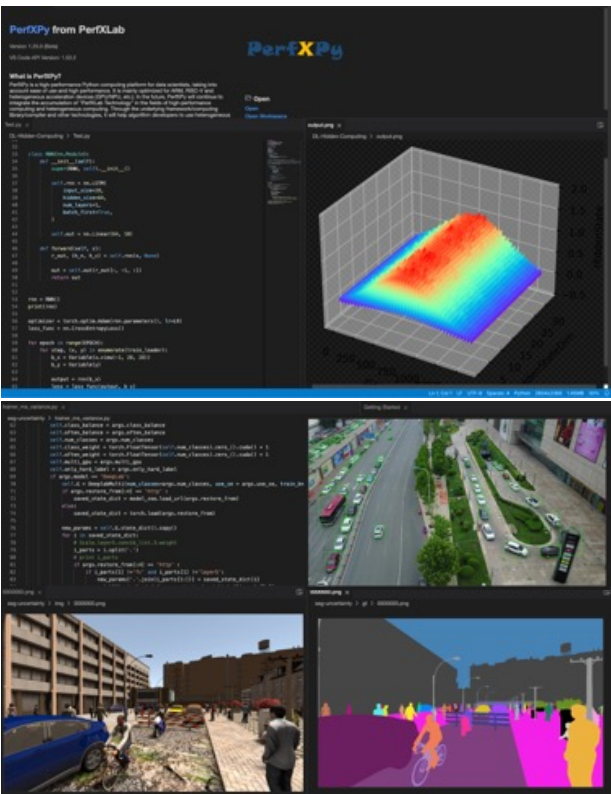
• YOLO网络性能对比

网络类型	规模	V1605B FP16 (单位: FPS)	V1605B FP32 (单位: FPS)	NVIDIA TX2 (单位: FPS)
YOLOv3	320*320	12.9	7.6	4.7
	416*416	9.2	6	2.9
	608*608	4.8	3.1	1.5
YOLOv2	320*320	28	18.5	10.8
	416*416	16.6	13.8	6.7
	608*608	8.7	7.7	3.6
YOLOv3-tiny	320*320	70	43	24
	416*416	58.7	27	19
	608*608	25.4	16.7	10
YOLOv2-tiny	320*320	84.2	58.6	25
	416*416	62.5	46.5	20
	608*608	28	23.3	11.1

• 应用效果（视频演示）



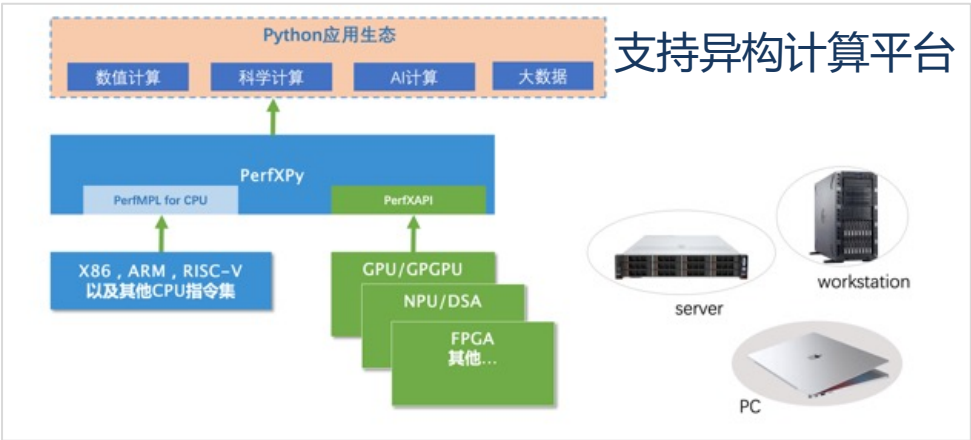
PerfXPy 新一代科学计算软件



科学计算

AI 计算

PerfXPy 是一个面向**科学家**和**算法工程师**的高性能**Python**计算平台，兼顾易用性与高性能，支持ARM、RISC-V 通用算力平台和异构加速平台。多个版本：单机版（免费），SaaS版本，集群版本。



支持异构计算平台



在更多高校展开实施

产学研新范式

过去



- 支持国产硬件
- 一键迁移，发挥极致计算性能。
- 传统科学计算工具对国产硬件基本不支持。

总结

- 计算软件栈
 - 芯片与上层应用（AI、工业软件）桥梁
 - 国外投入巨大，生态成熟
- 发展思路：开源 + 商业
 - 参与开源，共建生态
 - 商业发行版等
- 澎峰的工作
 - 高性能数学库PerfMPL，解决计算性能
 - 异构计算框架PerfXAPI，解决跨平台/异构
 - 高性能Python计算平台PerfXPy，解决易用性

感谢各位

期待与各位交流，为计算软件生态发展贡献力量

欢迎全职/兼职/实习生

