基于国产加速卡的科学大模型研究与实践

曙光信息产业股份有限公司

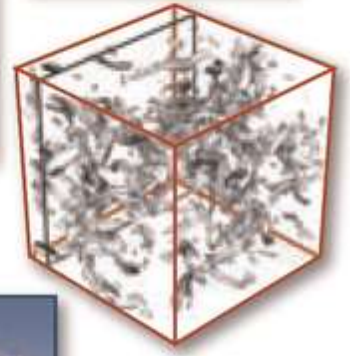
2024.08.06 长春

科学研究的四个范式

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

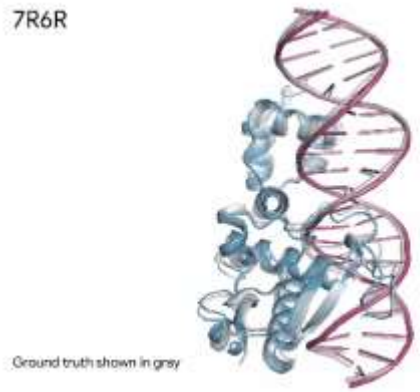
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



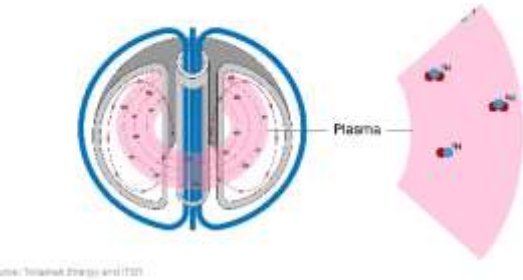
Jim Gray on E-Science (2007)

AI for Science 新范式

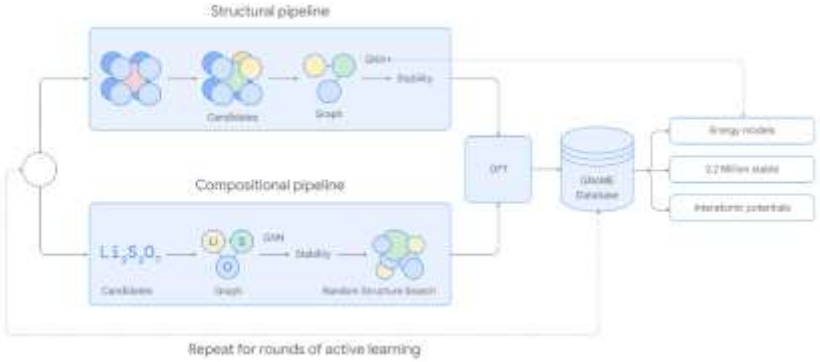
HPC和AI的结合方式，加速科学的研究与发现：将以往科学研究中的“经验”，实验观测数据与理论计算数据计算机化，通过智能计算提升研究成果的产出率



继AlphaFold2后，AlphaFold3公布，前所未有的精度预测所有生命分子的结构和相互作用。



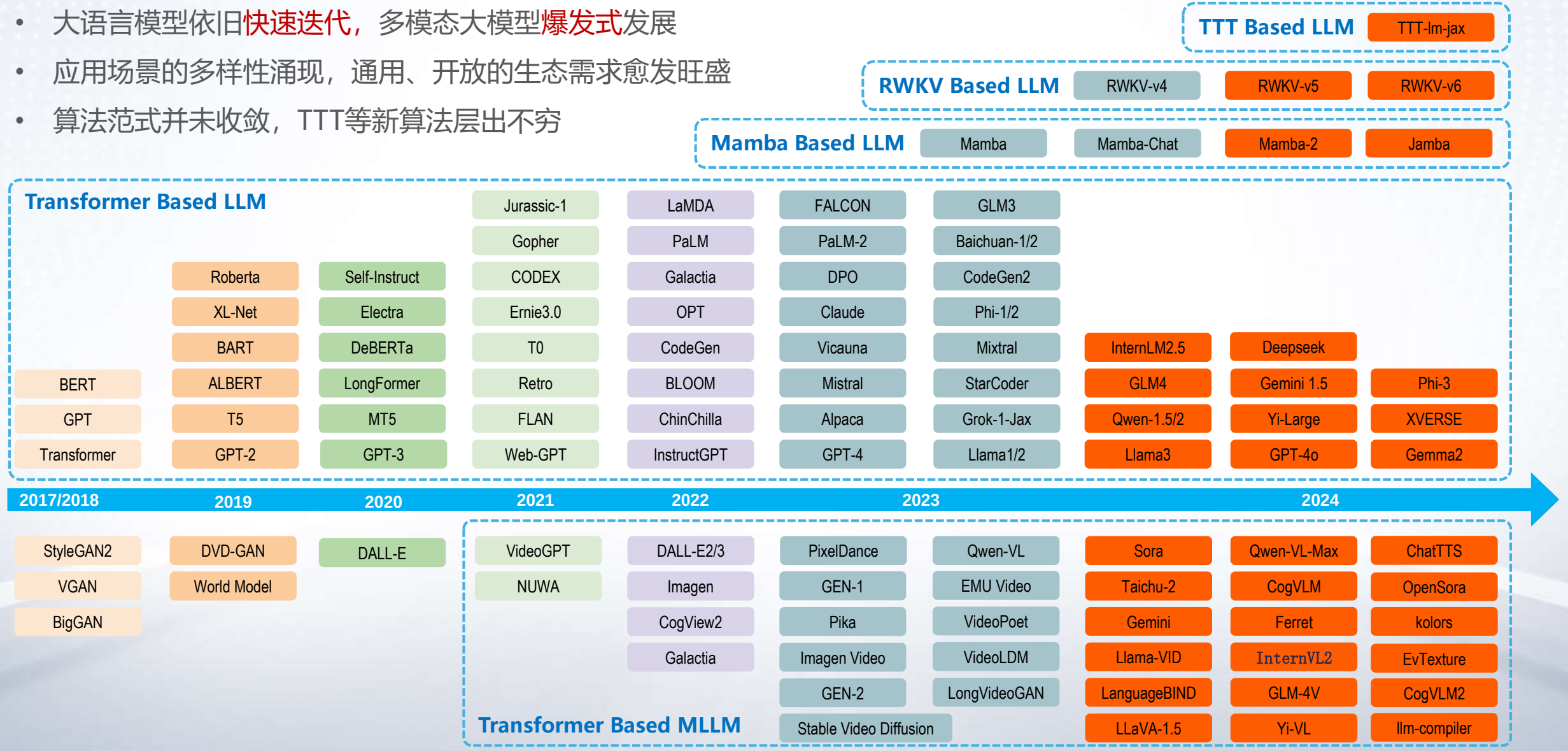
AI控制系统能够提前300毫秒预测核聚变等离子体不稳定性，助力于托卡马克的等离子体维持，该报告已经登上Nature。



DeepMind发布GNoME发现220 万种新晶体，相当于人类科学家 800 年的实验产出，其中 38 万种新晶体可以成为未来高新技术的稳定材料。

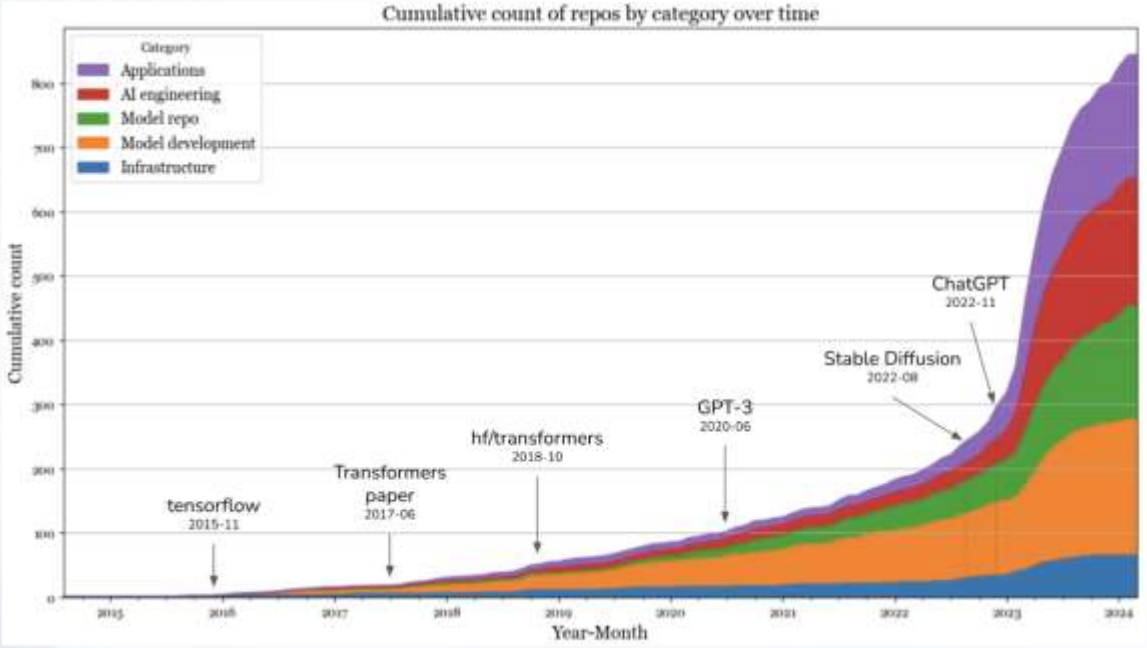
大模型发展趋势

- 大语言模型依旧快速迭代，多模态大模型爆发式发展
- 应用场景的多样性涌现，通用、开放的生态需求愈发旺盛
- 算法范式并未收敛，TTT等新算法层出不穷



AI开源生态发展趋势

AI生态爆发式发展，AI芯片生态通用性需求日益增长



在2023年SD和ChatGPT引爆大模型之后，大模型工具库的数量呈爆炸式增长。

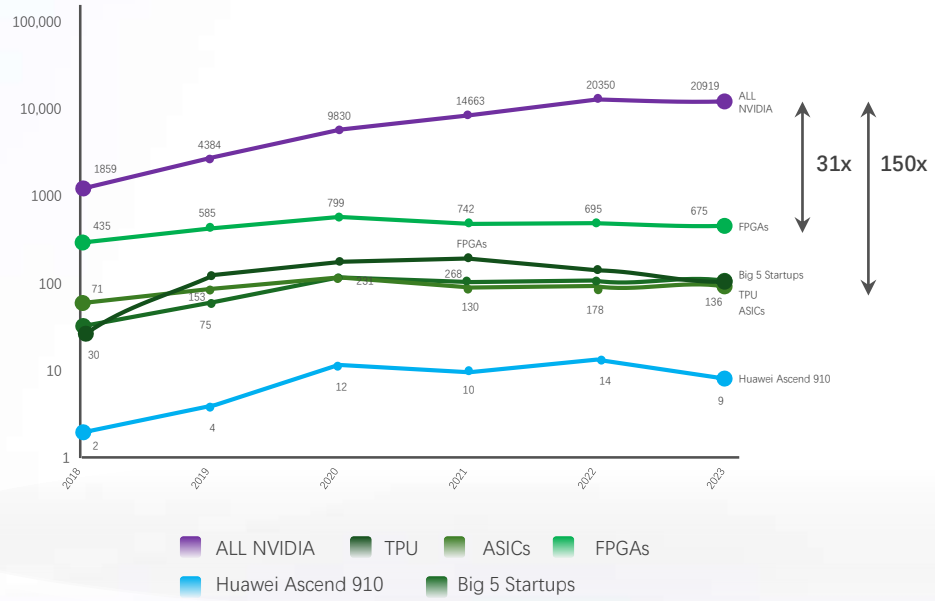
模型适配：不仅仅是基模适配

复合AI系统	由多个模型（通用+专用）和工具构建的AI应用系统正在成为可靠的大模型实施策略，AI芯片要适配众多的专用模型和工具
训练策略	预训练/增量预训练、监督微调、SFT、PPO、DPO、KTO、ORPO、 etc
训练算法	LoRA、QLoRA、GaLore、Badam、DoRA、LongLoRA、LLaMA Pro、LoRA+、LoftQ 、Agent tuning、etc
精度格式	训练/推理精度：FP32、TF32、**BF16**、FP16、INT8、INT4
性能优化	FlashAttention-2/3、PageAttention、Unsloth、RoPE scaling、NEFTune 、 rsLoRA、Prefill/Decoding分离、etc
工具	框架：Pytorch、JAX、TF、Paddle、Jittor、etc AI编译器：Triton、XLA、TVM、BladeDISC、etc 推理及服务：vLLM、LMDeploy、Fastertransformer、TGI、etc 集成化工具：Llamafactory、Ollama、Xtuner、Swift、etc

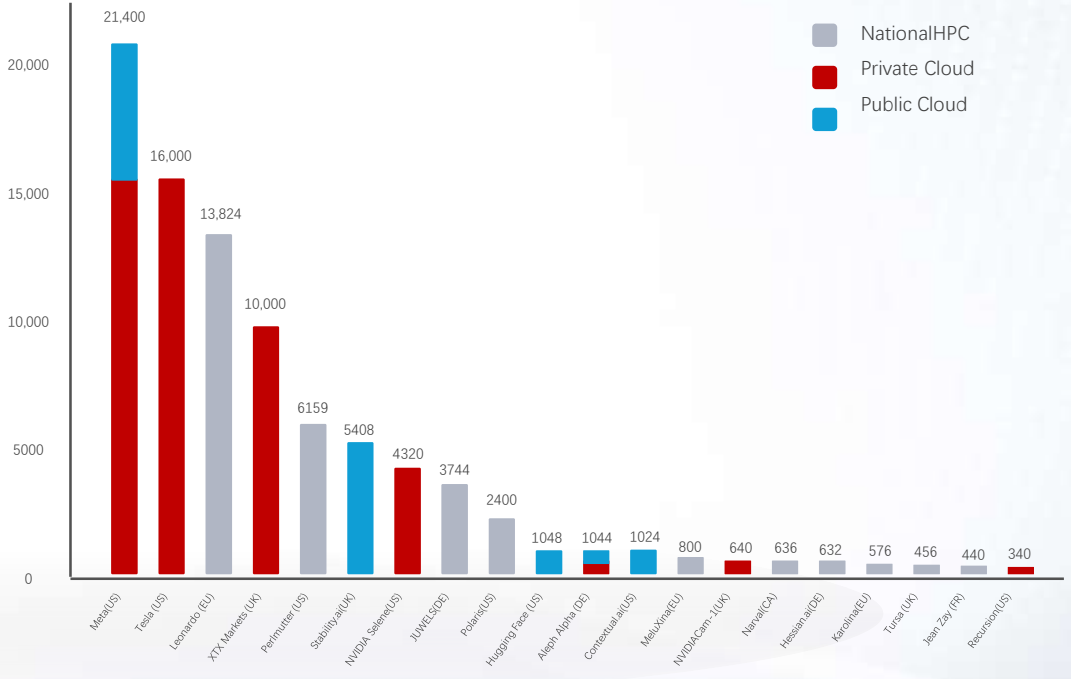
Ref：https://huyenchip.com

NVIDIA占据主流生态

2023年人工智能研究论文中特定半导体的利用情况
统计发现NVIDIA芯片被引用的次数远远多于其他：
比FPGAs多31倍，比TPUs多150倍



使用NVIDIA A100 的 GPU集群数量（2023）



Source: State of AI Report Compute Index

NVIDIA凭借硬件能力和CUDA生态覆盖，在学术界和工业界仍占据绝对优势，
如果完全绕过NVIDIA,必然与国外前沿模型算法和应用脱节。

先进算力芯片全面封锁

买不到

通过行政命令限制高性能计算和AI训练产品

- 2022/8/26 停止所有销售、停止对中国研发支持
- 2022/10/21 扩大先进算力芯片范围，不再局限NVIDIA和AMD
- 2022/11/7 停止实体名单在途订单产品向中国交付
- 2023/3/1 停止对在中国的美国公司研发支持
- 2023/9/1 停止在中国的所有制造

造不出

升级法律限制，限制为国产AI芯片生产制造

- 2022/10/7 管控中国境内先进制程半导体制造设备、软件及技术销售
- 2022/1/30 美日荷达成协议将限制向中国出口部分尖端设备

停止研发支持

- 2022/10/12 管控美国主体为中国从事研发、生产及相关活动
- 2022/12/15 管控AI芯片的研发、制造和销售

2023年10月 更新制裁

2023/10/17

更新制裁从“**传得慢**”→“**算得慢**”

3A090a: 针对最高性能芯片

- TPP > 4800
- TPP>1600，且PD超过5.92

3A090b: 针对次高端芯片

- TPP: [2400,4800），且PD[1.6, 5.92)
- TPP: [1600, ∞），且PD{3.2, 5.92)

国产加速卡技术路线

- GPGPU(General-purpose computing on graphics processing units)，通用图形处理器，利用现代图形处理器强大的并行处理能力和可编程特性。来处理非图形数据。
- DSA(Domain Specific Architecture)，领域专用架构
- AI算法不断创新，多模态模型模型快速发展，人工智能应用更加广泛。
- 通用计算架构是唯一被广泛采用开发新AI算法的软硬件平台，对多模态支持更友好，满足AI场景扩大的需求。

加速卡软件生态

数百个活跃的开源AI软件工具

自研AI工具链为主

Cuda

CUDA语言

Nvidia系列产品

DTK

函数级无缝迁移

ROCm

HIP语言

AMD系列产品

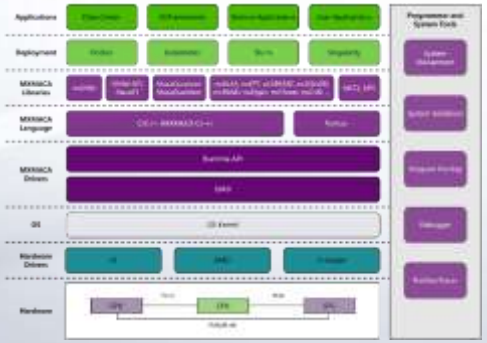
其他

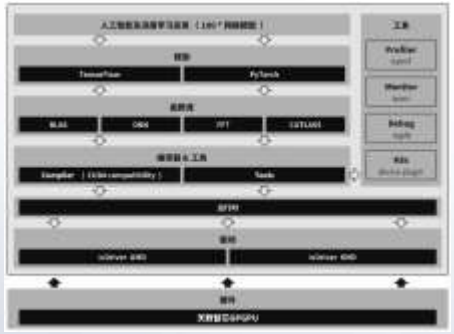
厂家自研
AI异构计算平台

各种NPU、 ARM+NPU芯片

GPGPU架构

DSA（NPU、ARM+NPU…）



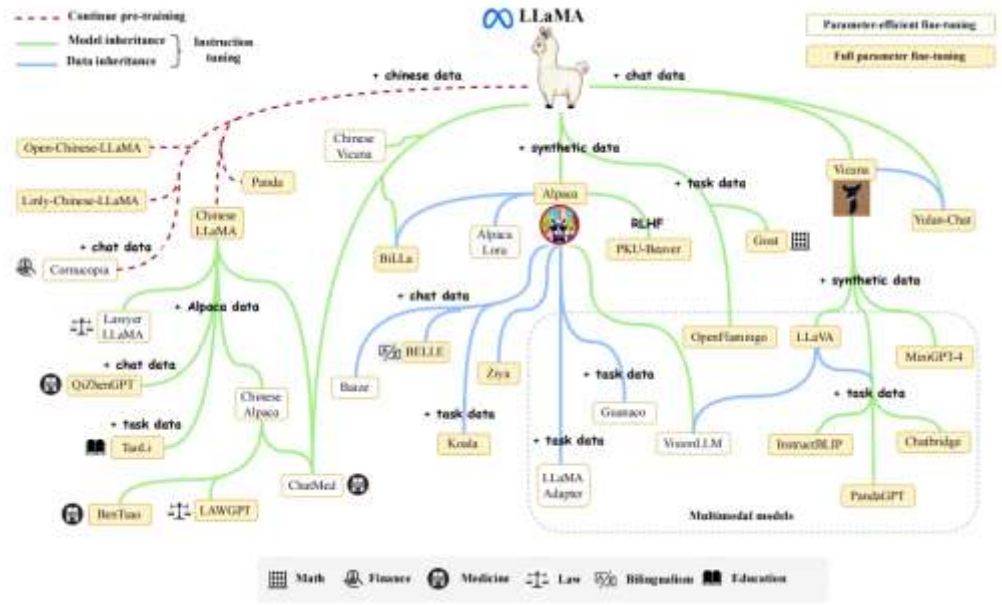


国产替代的生态挑战

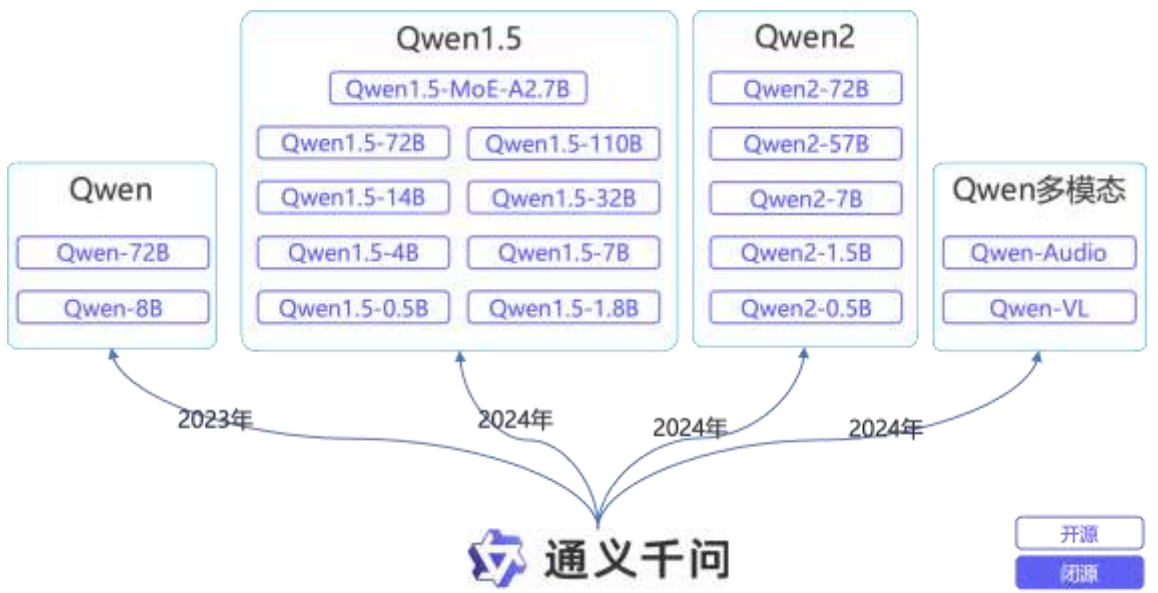
1. AI场景

- 数以万计的小模型
- 数量爆炸式增长的LLMs和LLMs生态工具，大模型适配，不仅仅是适配常见的SOTA基座模型
- 大模型应用通常是由大模型+多个工具/AI模型协同实现，其中大模型选型需要随着开源生态发展而快速更新。

Llama生态



Qwen模型家族



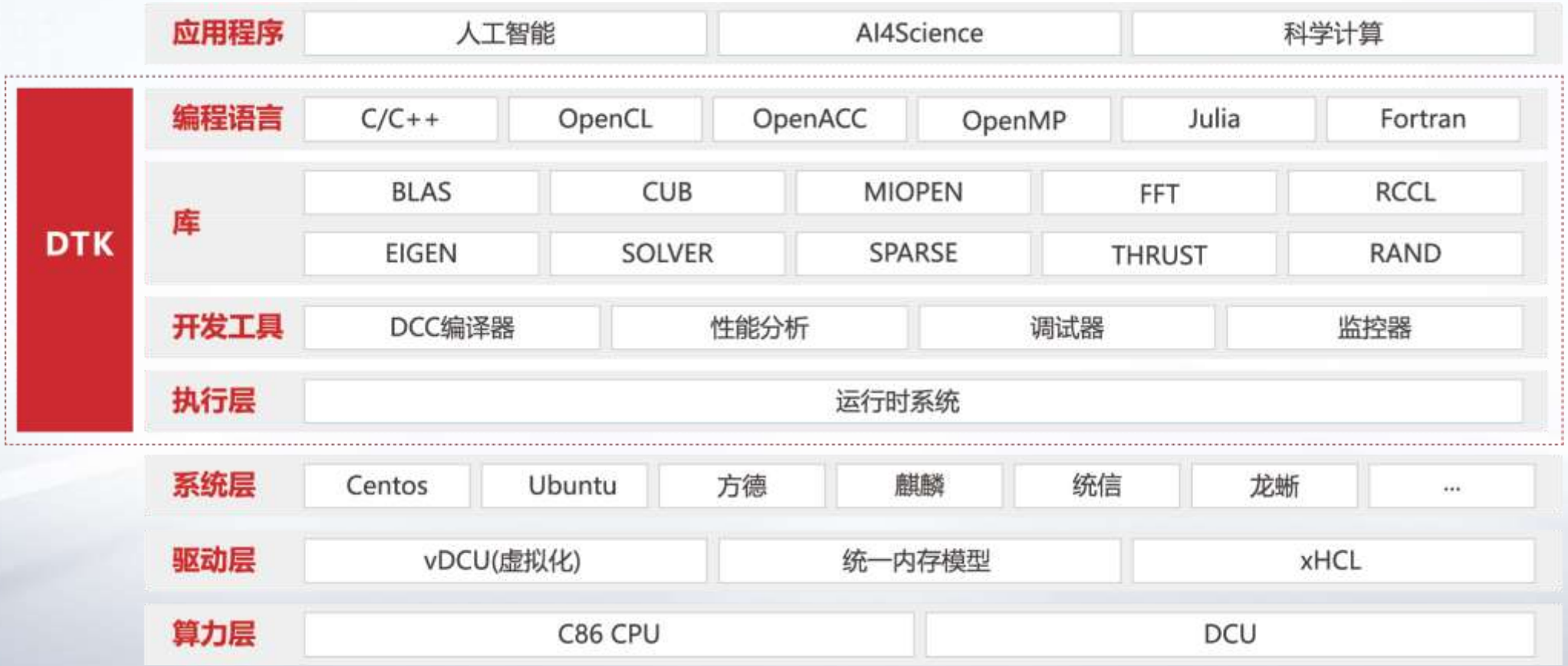
2. Science场景

- 各学科众多经典科学计算工具，和无数科学计算包
- 各领域都有维护一套GPU C/CPP代码的诉求，并与论文发表和与开源生态对接
- AI4Science快速发展，AI4Science workflow的Infra往往是AI基础工具 + 科学计算工具

DTK

DTK
DCU TOOLKIT

致力于让客户只维护一套代码，让客户国产化替代工作“软着陆”。



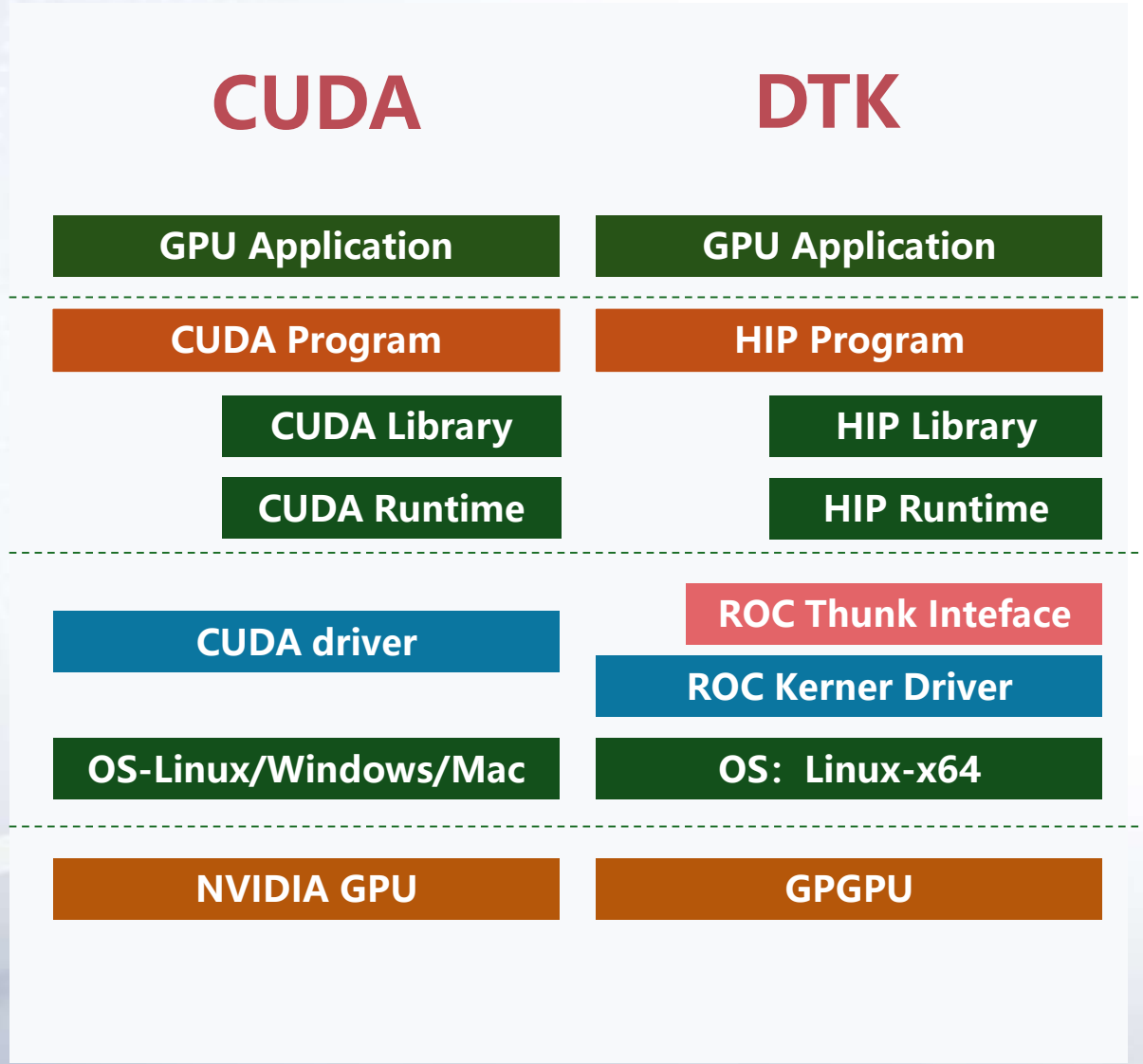
全栈自研

自主迭代

数学库完善

全场景覆盖

DTK vs CUDA：数学库对比



数学库功能	CUDA数学库	DTK数学库
深度学习基础数学库	cudnn	miopen
基础矩阵运算数学库	cublas	rocblas/hipblas
通信库	nccl	rccl
随机数数学库	curand	hiprand
稀疏矩阵数学库	cusparse	hipsparse
快速傅立叶变换数学库	cufft	rocfft/hipfft
基础算法库	cub	hipcub
并行库	thrust	rocthrust
计算密集型求解器	AMG-X	rocALUTION

开放优化——DAS

Coverage
全面覆盖

Performance
高效性能

Agility
快速迭代

OpenDAS
开源开放

光源

开发者社区
（AI版块讨论区/ FAQ版块）

OpenDAS扩展组件层

通用训练组件
apex | mmcv | ...

大模型训练组件
bnb | fa | ...

大模型推理组件
lmdeploy | vllm

fd | xform | ...

框架工具层

快速适配工具
FastPT

通用推理框架
ORT/MIGraphx/AIT/...

AI框架
PyTorch/TF/Jax/Paddle/Jittor/OneFlow/...

基础算子层

融合算子库
LightOP

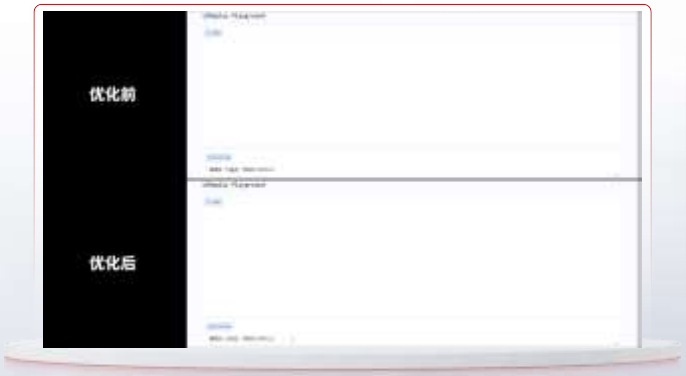
AI编译
Triton/XLA/TVM

开发工具栈

基础硬件系统

Pytorch视频检测优化
1.5X

Migraphx OCR优化
1.4X



Lmdeploy 推理优化
1.5X

Stable Diffusion 优化

Layout转换

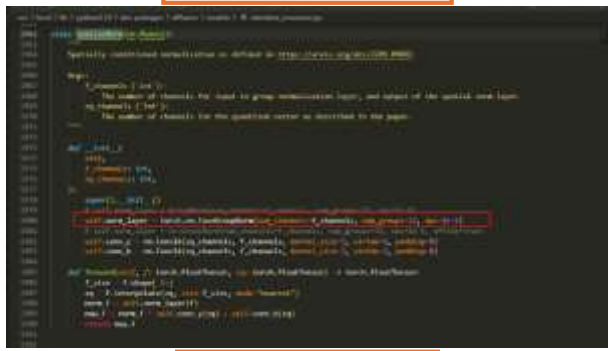
TN →NN

NCHW → NHWC

Lightop算子应用优化

gn+silu融合

gn优化替换



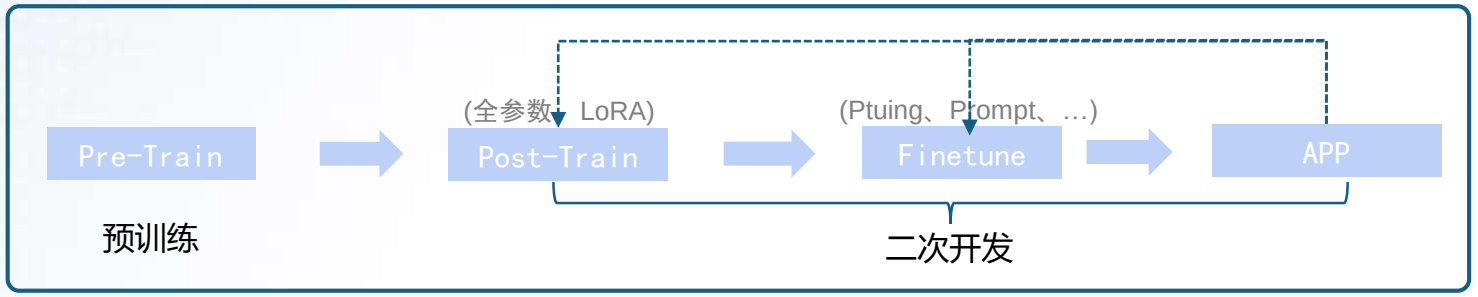
conv优化

gemm优化

Triton优化

	等待时间(s)	加速比
基线模型	4.52	1
优化后(非dc)	2.33	1.94
优化后(dc)	1.83	2.47

大模型全栈能力



大模型开发全流程覆盖

GPT Llama BLOOM GLM baichuan GROK
ERNIE Qwen RWKV T5 Stable Diffusionn GEMMA
VIT CLIP Conformer CPM SAM OpenSora …

主流开源模型全覆盖

Transformer → Grok-JAX
0.08B 312B

DeepSpeed Megatron FasterTransformer MLC-LLM vLLM
Oneflow ColossalAI Apex xformer Fairscale
libai FastLLM Lmdeploy FastMoe BMTrain
HF-Transformer HF-Accelerate FairSeq Jittor-LLM LightSeq
Triton TVM ONNXRuntime MIGraphx BladeDISC …

Pytorch JAX Tensorflow PaddlePaddle Jittor …

主流开源工具全覆盖

DTK

基础计算库全覆盖

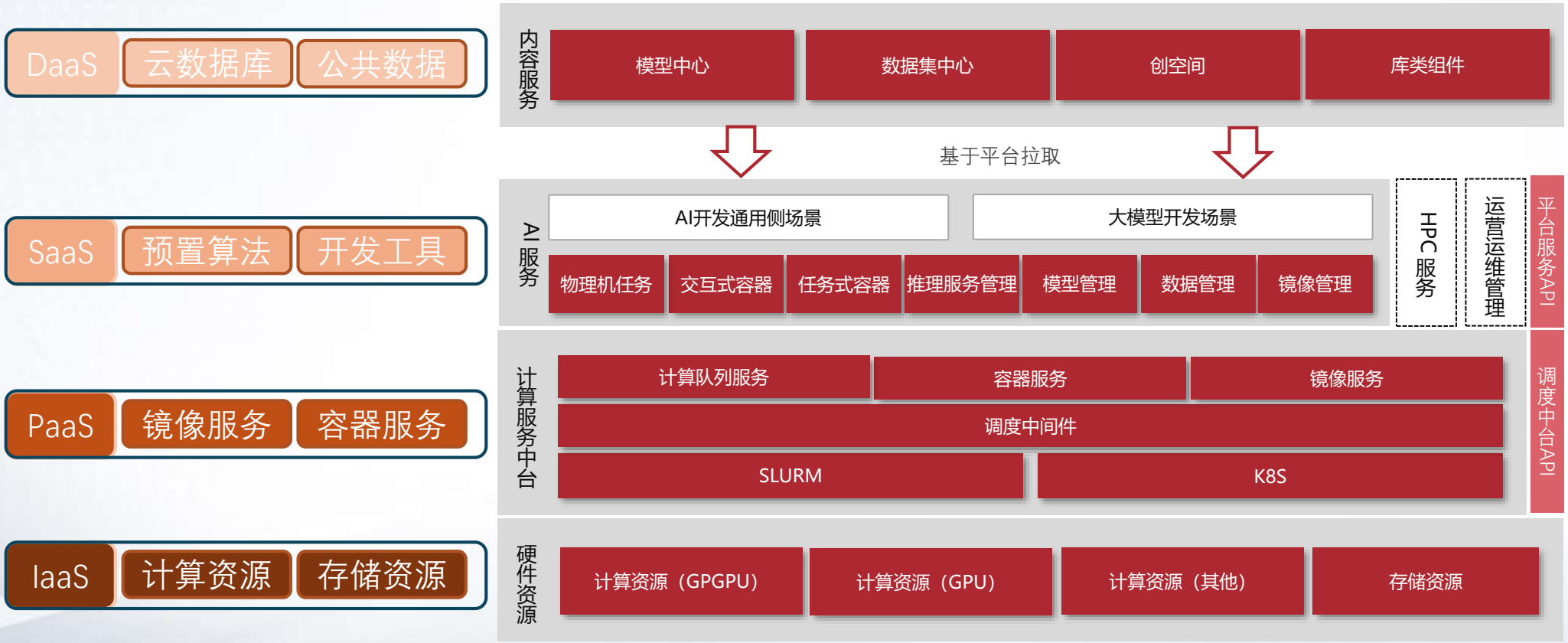
GPGPU

通用硬件架构,全精度覆盖

Science 生态版图

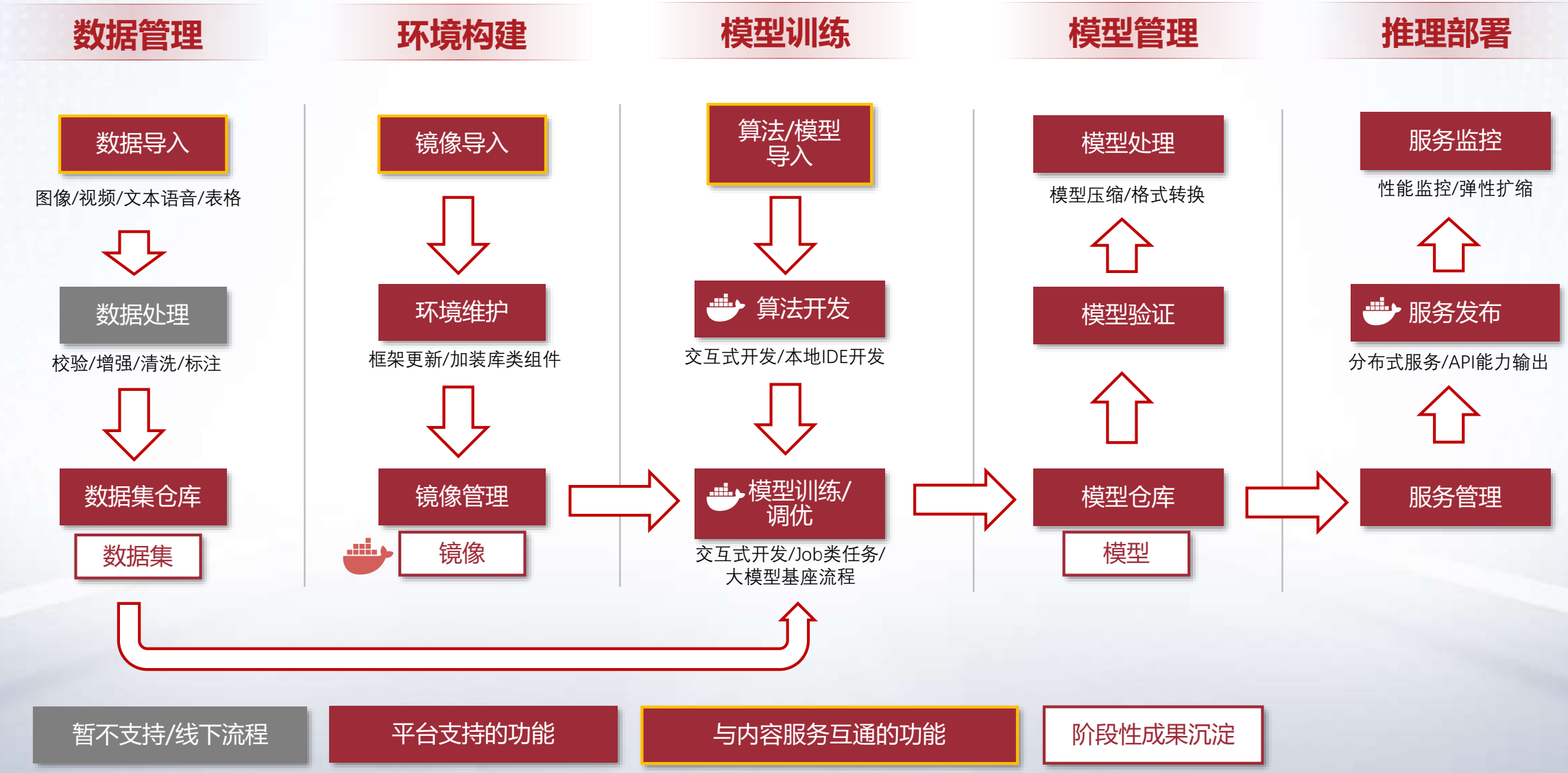
分子动力学	计算化学	工业仿真（CAE/CFD/电磁）		气象环境	计算物理	天体物理	Other	
GROMACS	CP2K	OpenFOAM	PYFR	SD3	QUDA	PHoToNs	hashcat	COSMA
NAMD	VASP	OpenCFD	xflow	CALPUFF	GWU-code	Gadget-2	SWsnn	Tiled-MM
LAMMPS	NWChem	RapidCFD	MxSim-Explicit	NAQPMS	Chroma	HSPM	Gridtools	SLATE
HOOMD-blue	PWMat	GESTS	SAPTIS	LICOM3	PIConGPU	GluoNs	TADOC	Ginkgo
OpenMM	FHI-aims	CCFD-V3.0	LBPM	IAP-AGCM	TOPS	OSKAR	Aries	PSEPS
Amber	BigDFT	Sunflow	pmlfma-ternary	RRTMG_SW	ANT-MOC	Cholla	Magma	PETSc
misa-md	PWDFT	Aries	cuFEM-DDM		LPAM	HACC	SIRIUS	Trilinos
Qbio	LR-TDDFT	HydroMap	SCLETD-PF	生信基因	Grid		Kokkos	SuperLU
GALAMOST	DGDFT	IPELBM	DGTD	BarraCUDA	Kripke	地球物理	RAJA	AMReX
eMD	AIMD-HF	PiFlow	LaspcemMoM	Blast	Laghos	SPECFEM3D	HPCPortfolioOpt	xsolver
CovalentMD	RT-TDDFT	Nekbone	COMBUSTION	GenomeWorks	PEPS	KarstSim	hipSYCL	HPSSE
Sponge	BDF	Xyce	SwallowSound	DeepVariant	Quicksilver	GeoEast	stdgpu	HYPRE
	WESP	Pflows		BWA	SymPIC	LSRTM	cupy	ParTI
	LS3DF			RELION3		SES3D	DMTCP	VkFFT
	QMCPACK			AutoDock		SOFI3D	DBCSR	Spfft
	TeraChem			HiBFlowFSI				

基于云计算的大模型应用落地



- **内容支持：**降低门槛、加速DCU开发进程，平台一键拉取
- **功能覆盖：**AI开发全流程支持，兼大模型及通用侧
- **环境交付：**容器化弹性灵活的环境部署，简单易用
- **资源管理：**支持多种调度，面相AI+HPC，完善的运营运维功能，提高资源利用率
- **生态支持：**兼容多种加速器，原生支持DCU

大模型业务流程覆盖


数据管理
环境构建
模型训练
模型管理
推理部署
数据导入
图像/视频/文本语音/表格
数据处理
校验/增强/清洗/标注
数据集仓库
数据集
镜像导入
环境维护
框架更新/加装库类组件
镜像管理
镜像
算法/模型导入
算法开发
交互式开发/本地IDE开发
模型训练/调优
交互式开发/Job类任务/大模型基座流程
模型处理
模型压缩/格式转换
模型验证
模型仓库
模型
服务监控
性能监控/弹性扩缩
服务发布
分布式服务/API能力输出
服务管理
暂不支持/线下流程
平台支持的功能
与内容服务互通的功能
阶段性成果沉淀

AI能力化实践

紫东太初 | 中国科学院自动化研究所

AlphaFold2 | DeepMind

ChatGLM

文心一言

悟道 WUDAO

LLAMA | Meta

FastChat |

OpenAI GPT-3

BLOOM

… …

RAW数据/爬取数据 → 预处理/清洗 → 生成标准pretrain/finetune数据 → 生成可喂入网络的数据 → Pretrain → 全参微调 → LoRA SFT → RLHF → 量化 → 推理 → Web Socket → 页面展示

大模型产研究实践

科研合作

中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

- 实现单/多模态理解与生成的多任务统一建模
- 针对平台特性开展多模态预训练模型架构设计与优化

北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

- 完成FastMoE等工具移植优化
- 悟道2.0达1.75万亿稠密参数，实现NLP等多个领域的世界第一。

产业合作

阿里	百度	科大讯飞	智谱
百川	易道博识	文因互联	360
天壤智能	青云科技	九章云极	博云科技
捷通华声	译图智讯	睿真科技	致宇科技
文通	硅心智能	思必驰	来也科技
上海弘玑	非十科技	实在智能	数巅科技

...

- 排名不分前后

高能所合作

中国科学院高能物理研究所是中国从事高能物理研究、先进加速器物理与技术研究及开发利用、先进射线技术与应用的综合性研究基地。

同高能所多个科研组展开全面合作，合作内容包括：

高能同步辐射光源HEPS

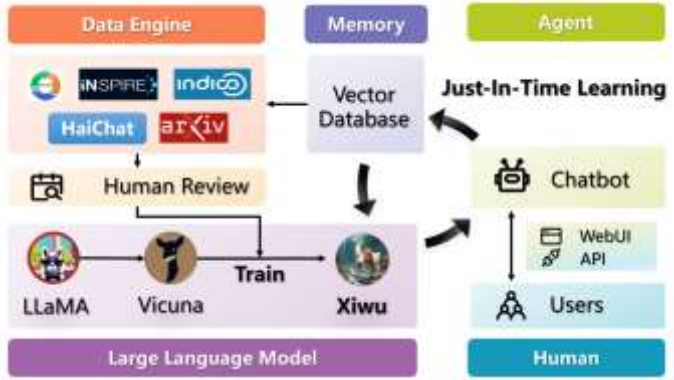
HEPS是国家重大科技基础设施建设“十三五”规划确定建设的十个重大科技基础设施之一

合作内容：

- 光源关键应用Hepsptycho在GPGPU上的移植
- 合作进行Hepsptycho性能优化
- 完成CUPY、Tomocupy、OpenCL-UFO、Astra ToolBox等科研工具的DCU移植和优化

溪悟高能物理大模型

溪悟大模型是高能所自研的、是第一个专用于高能物理领域知识挖掘和发现的L2级LLM

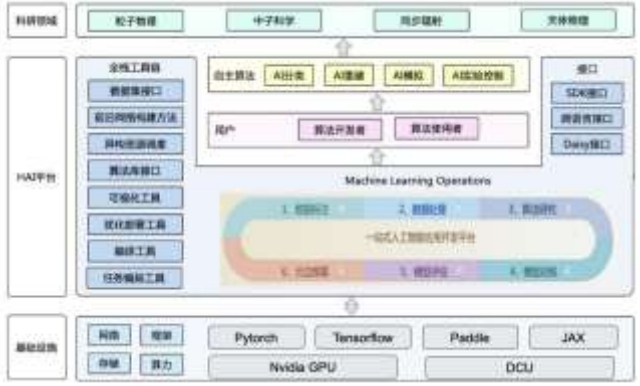


合作内容：

- LLAMA等溪悟基座模型适配、优化
- 协助溪悟workflow选型和适配优化工作,包括LlamaFactory、Xtuner、vLLM等

HEPAI高能物理AI平台

HEPAI高能物理人工智能平台用于加速多学科场景下的科学研究,基于溪悟构建,服务于粒子物理、天体物理、同步辐射等领域的科学研究.



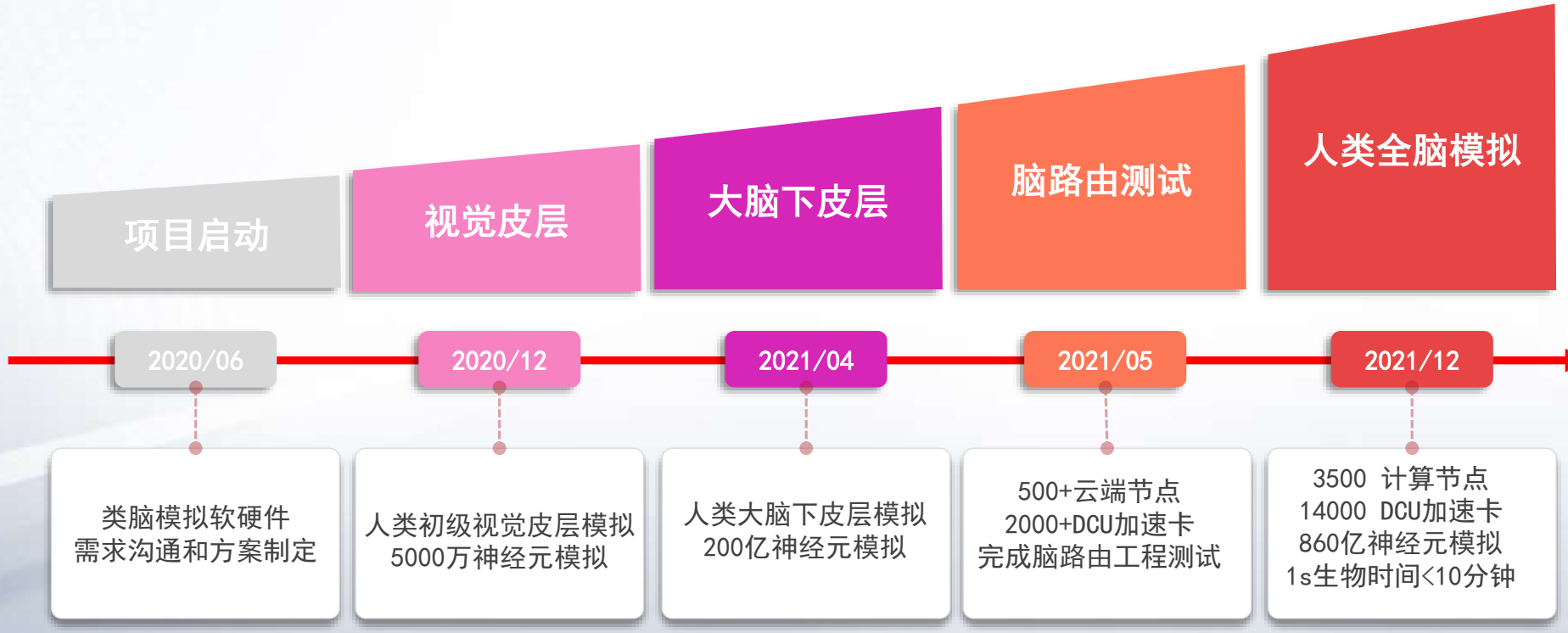
合作内容：

- GPGPU 接入 HEPAI平台
- 协助HEPAI调度工具选型(K8S、Ray)
- 协助基于GPGPU的溪悟模型服务接入HEPAI平台

脑与类脑智能

复旦大学类脑人工智能科学与技术研究院基于国产加速卡完成全球首次红毛猩猩、人类全脑模拟（860亿神经元、10^14神经突触），多次冲击世界最高水平，并构建多模态多尺度脑数据库，挖掘多动症和睡眠障碍的联系、大脑灰质和个性发育与醉酒频率的关系、推进AI在自动驾驶领域等课题研究及应用落地，极大的推进了类脑学科的进展。

截止2023年，为客户累计提供**5739万+**卡时，其中最大并行规模**14000**卡(**140P FP64**算力)，并协助客户完成超大规模类脑并行计算优化、网络优化等工作。



脑与类脑智能基础转化应用研究历程

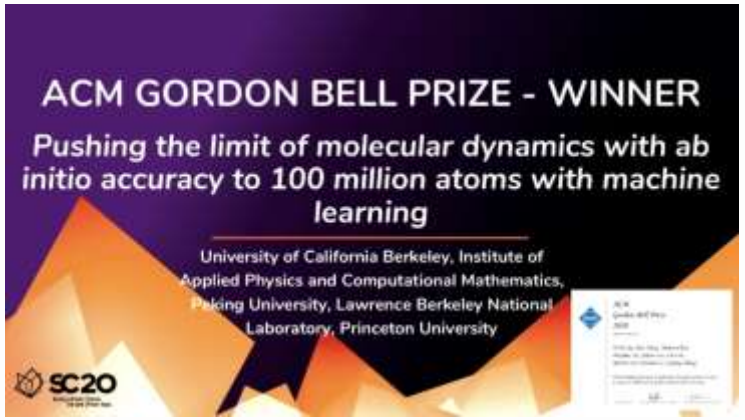
国际类脑科研进展概要

HBP+NEST	7.3亿神经元 生物时间30分钟
Spinnake	10亿神经元
Juqueen	17.3亿神经元 生物时间40分钟
复旦类脑	**860亿神经元 生物时间<10分钟**

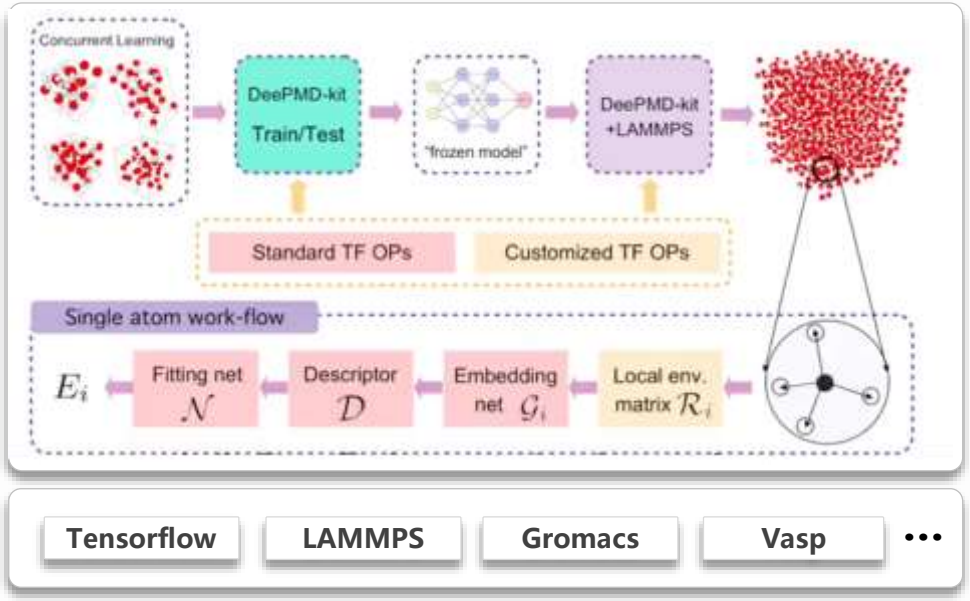
DeePMD

DeePMD-kit是深势科技开源的一个基于神经网络拟合的第一原理数据，用于多体势能表示和分子动力学的深度学习软件包。在保持量子力学的精确性和准确性的基础上，DeePMD-kit可以将分子动力学的计算速度提高几个数量级。

- 深势科技DeePMD曾获SC20 ACM 戈登·贝尔奖。
- DeePMD在8192卡将分子动力学模拟的原子体系规模扩展至十亿量级，原子体系规模超过 SC20 戈登贝尔奖的工作 。



DeepMD软件架构



实测对比：GPGPU vs A800

与客户进行优化合作，依托DeePMD进行动力学模拟，以大量的第一性原理计算数据来拟合原子之间的势能面，并基于阿里BladeDisc算法性能优化。

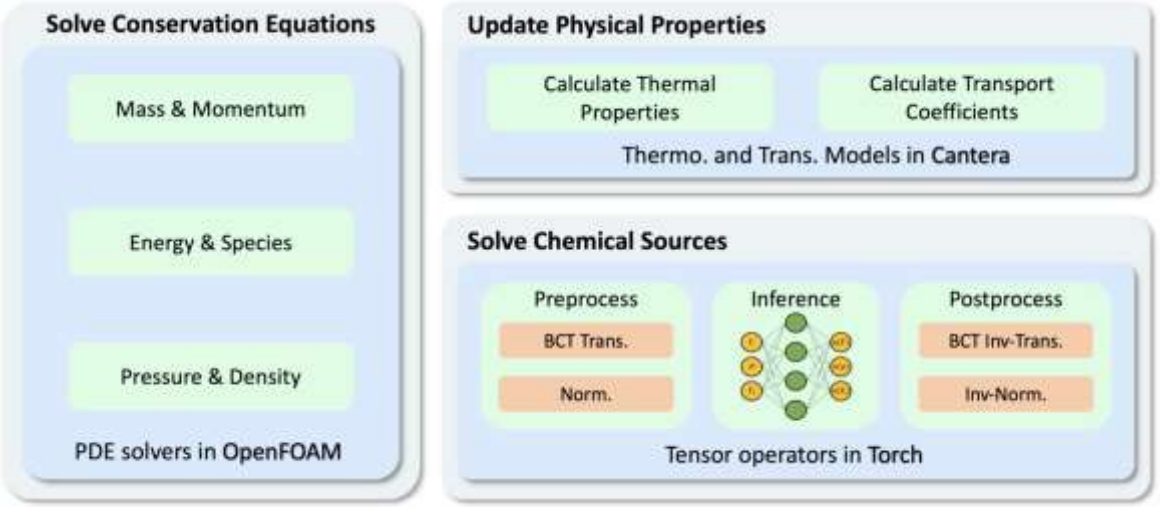
客户Case	GPGPU	A800	性能比 GPGPU /A800
case A	2.67 s	3.10 s	116.10%
case B	2.70 s	1.78 s	65.93%

注：时间越小越好

DeepFlame

DeepFlame AI for Science时代的燃烧反应流体计算平台

DeepFlame是一套具备机器学习能力的计算流体力学软件，适用于各种速度的单相或多相、层流或湍流的反应流。它旨在提供一个开源平台，汇集OpenFOAM、PyTorch库的各自优势，用于机器学习辅助反应流模拟。



DeepFlame 技术框架

DeepFlame战队获得2022先导杯大赛一等奖

先导杯期间，DeepFlame战队完成了开源代码在国产高性能异构计算平台的部署以及稳定运行，实现了多卡推理特性，完成了深度高性能优化，并进行了万核千卡规模的大体系算例验证，相关成果即将随DeepFlame首个大版本v1.0一同发布

Ref：https://mp.weixin.qq.com/s/zosUexZ_8hSGH8CPn_F2wg

共建产研生态，定智千行百业

总结

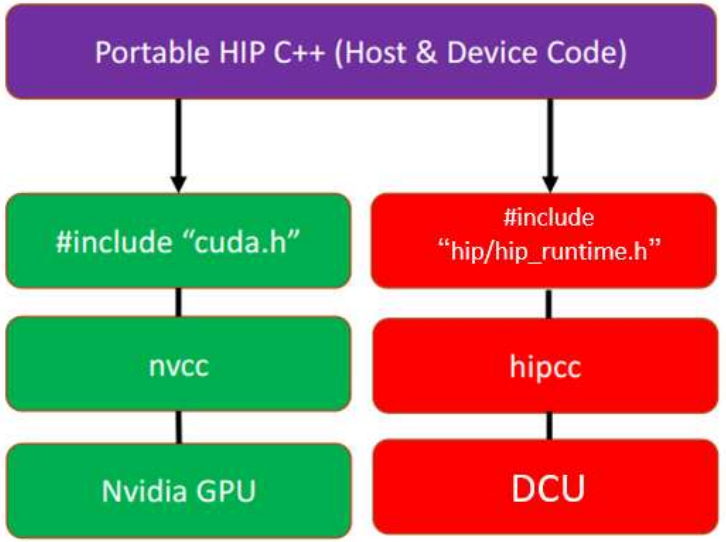
- 当下AI取得成就的本质原因是在**算力**和**数据**基础上**算法**对高维函数处理能力的大幅提升。
- 2023年科技部会同自然科学基金委启动“人工智能驱动的科学研究”(AI4S)专项部署工作。
- AI4S会是AI下一个主战场，将不断拓展Science 和 AI的边界。
- 当AI4S发展道路上也充满挑战，需要各行各业的人们，打破壁垒、凝聚共识、优势互补、不断迭代。
- 国产加速卡及生态已经成为必然趋势，将不断助力AI4S发展。

谢谢大家

DTK vs Cuda： 编程模型对比

HIP (Heterogeneous-compute Interface for Portability) 特性:

- 开源,以C/C++为基础的显式异构并行编程模型。 。
- 提供完善的异构编程模型和Runtime API，同时支持CUDA、ROCm、DTK。
- Runtime API与CUDA Runtime API兼容。



术语对照:

NVidia GPU	DCU	描述
Streaming Multiprocessor (SM)	Compute Unit(CU)	是由多组并行计算单元组成的计算结构，在一个block内的所有thread会被分配到同一个CU中。
Kernel	Kernel	核函数，可以并发执行的异构函数。
Warp	Wavefront	线程束/波前是硬件执行基本线程单元。DCU线程束的大小为64。
Thread	Work-item / Thread	线程，执行核函数的基本单元。
Block	Work-group / Block	线程块，由数个线程构成的集合，会由一个CU执行。

DTK vs CUDA：代码示例

CUDA与HIP核函数对比：

CUDA DAXPY

```
__global__ void add(int n, double *x, double *y)
{
  int index = blockIdx.x * blockDim.x + threadIdx.x;
  int stride = blockDim.x * gridDim.x;
  for (int i = index; i < n; i += stride)
    {
    y[i] = x[i] + y[i];
    }
}
```

HIP DAXPY

```
__global__ void add(int n, double *x, double *y)
{
  int index = blockIdx.x * blockDim.x + threadIdx.x;
  int stride = blockDim.x * gridDim.x;
  for (int i = index; i < n; i += stride)
    {
    y[i] = x[i] + y[i];
    }
}
```

KERNELS ARE SYNTACTICALLY THE SAME

RuntimeAPI对比示例：

CUDA	HIP
cudaMalloc(&d_x, N*sizeof(double));	hipMalloc(&d_x, N*sizeof(double));
cudaMemcpy(d_x, x, N*sizeof(double), cudaMemcpyHostToDevice);	hipMemcpy(d_x, x, N*sizeof(double), hipMemcpyHostToDevice);
cudaDeviceSynchronize();	hipDeviceSynchronize();

Kernel发起语法对比

CUDA KERNEL LAUNCH SYNTAX

```
some_kernel<<<gridsize, blocksize,
shared_mem_size, stream>>>(arg0, arg1, ...);
```

HIP KERNEL LAUNCH SYNTAX

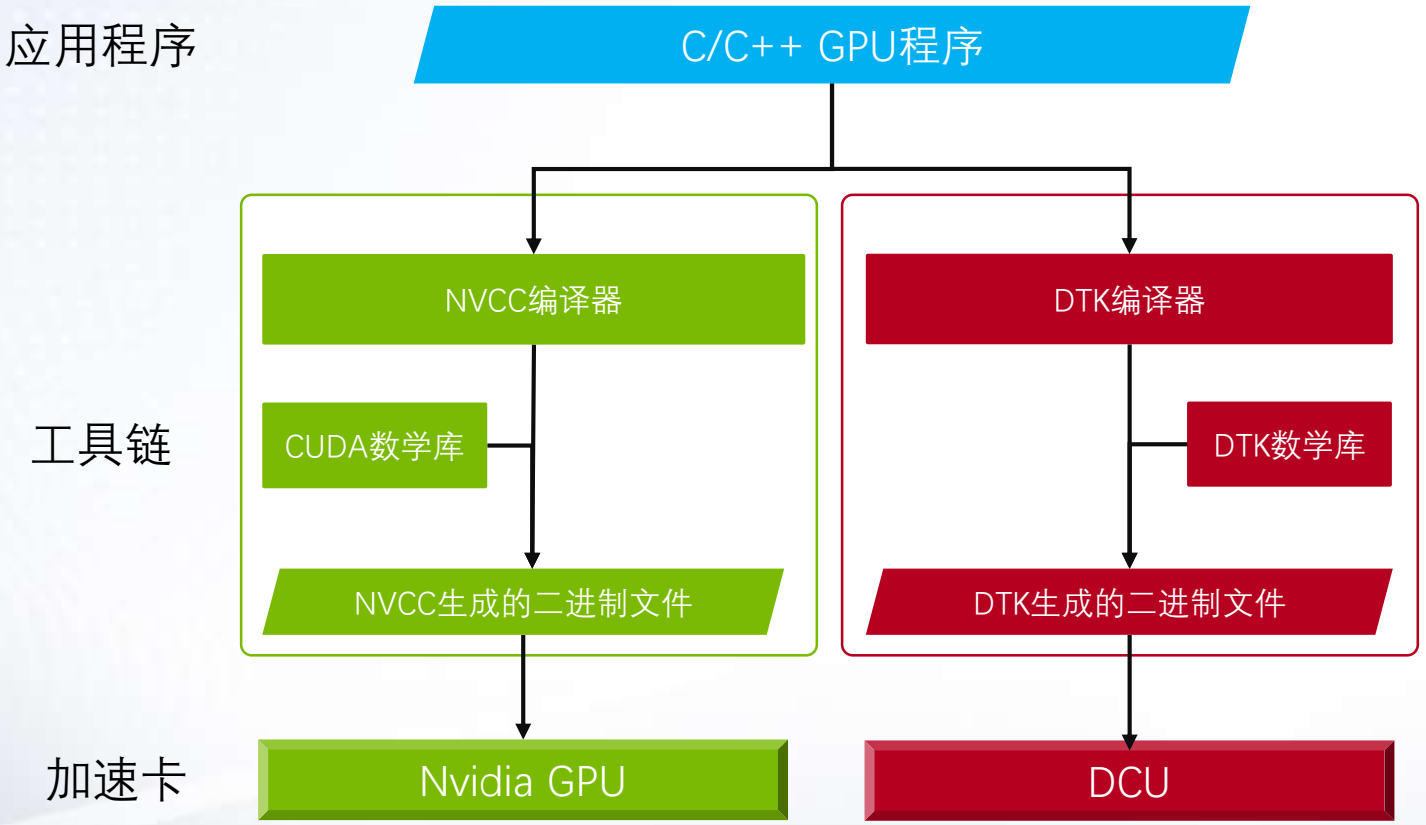
```
hipLaunchKernelGGL(some_kernel, dim3(gridsize),
dim3(blocksize), shared_mem_size, stream,
        arg0, arg1, ...);
```

DCU数学库

CUBLAS	ROCBLAS	Basic Linear Algebra Subroutines
CUFFT	ROCFFT	Fast Fourier Transforms
CUDNN	MIOPEN	Deep Learning Library
CUB	ROCPRIM	Optimized Parallel Primitives
EIGEN	EIGEN	C++ Template Library for Linear Algebra

MORE INFO AT: GITHUB.COM/ROCM-DEVELOPER-TOOLS/HIP → HIP_PORTING_GUIDE.MD

DTK GPUFusion



应用测试

应用	编译	测试
Amber	通过	通过
NAMD-2	通过	通过
NAMD-3	通过	通过
bwa	通过	通过
blast	通过	通过
libxc	通过	通过
HooMD-blue	通过	通过
Eigen	通过	通过
Magma	通过	通过
风雷软件	通过	通过
东方晶源	通过	通过
物探研究院	通过	通过
中石化胜利油田	通过	通过
Gromacs	通过	通过
MMDeploy	通过	通过
PPL.CV	通过	通过
Transfomer(C++版)	通过	通过
Xgboost	通过	通过

注:

- DTK目前兼容CUDA 10.2 和11.8的CUDA API且可以直接编译
- DTK支持Cmake构建系统,用户可以不改变GPU应用程序的构建工程