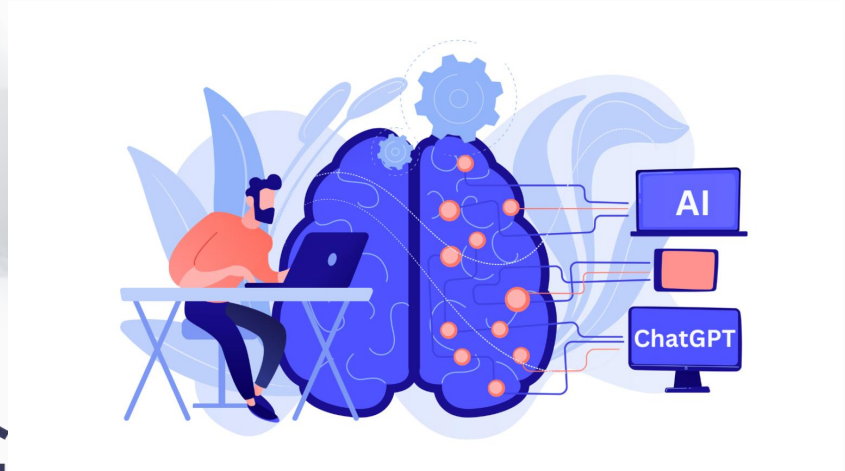

GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「AI is Everywhere」专场
本期议题：旷视算法量产与 MegEngine 生态建设
陈其友 2023年6月1日

AI is Everywhere

- 越来越多的领域，如视觉、语音、机器翻译、游戏甚至数据库领域都在使用 AI 算法
- AI 的优势：智能，固定运行时间和内存占用、可移植、容易固化成硬件、响应快速等等

AI 已经不可替代
AI 需求爆炸式增长

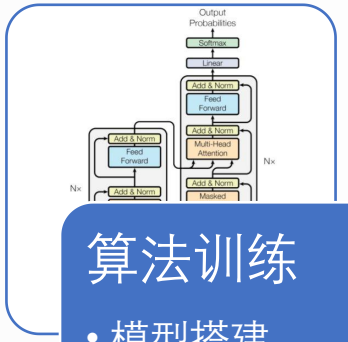


一般算法开发流程

算法研究员深度参与

数据生产

- 数据采集
- 数据标注
- 数据清洗



算法训练

- 模型搭建
- 模型优化

算法部署

- 模型量化/转换
- 模型推理

算法测试

- 精度/性能测试
- Bad Case 收集

迭代提升模型精度

一般算法开发

- 周期长
- 效率低
- 维护成本高
- 高度定制化

算法量产

算法量产核心思想：
- 流程标准化
- 工作自动化
- 技术低门槛化

人工对接
人工实操
研究员实操

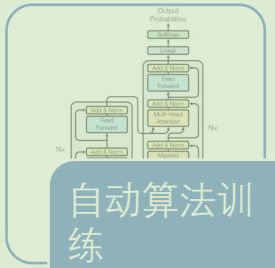
算法量产

标准化接口
自动化
算法量产工程师实操

算法量产平台
算法量产工程师

数据生产
- 数据采集
- 人工标注
- 数据清洗

自动化数据标注



自动算法训练
- 自动选择模型
- 自动进行模型优化

自动算法部署
- 自动模型量化/转换
- 自动模型推理

自动化测试

人工测试
- 精度/性能测试
- Bad Case收集

迭代提升模型精度

算法量产平台

AIS 旷视 MLOps 算法生产平台

GOTC

业务端

数据管理

- 数据清洗
- 数据集管理
- 数据标注
- 智能标注
- 训验对管理
- 数据质检

模型训练

- 自动学习
- 实验优化
- 实验对比
- 实验管理
- workspace
- 第三方训练

模型管理

- 模型评测
- 模型转换

模型部署

- 在线部署
- 服务器算法部署
- 端设备算法部署

团队管理

- 权限管理
- 事件管理
- 用户管理
- 项目管理

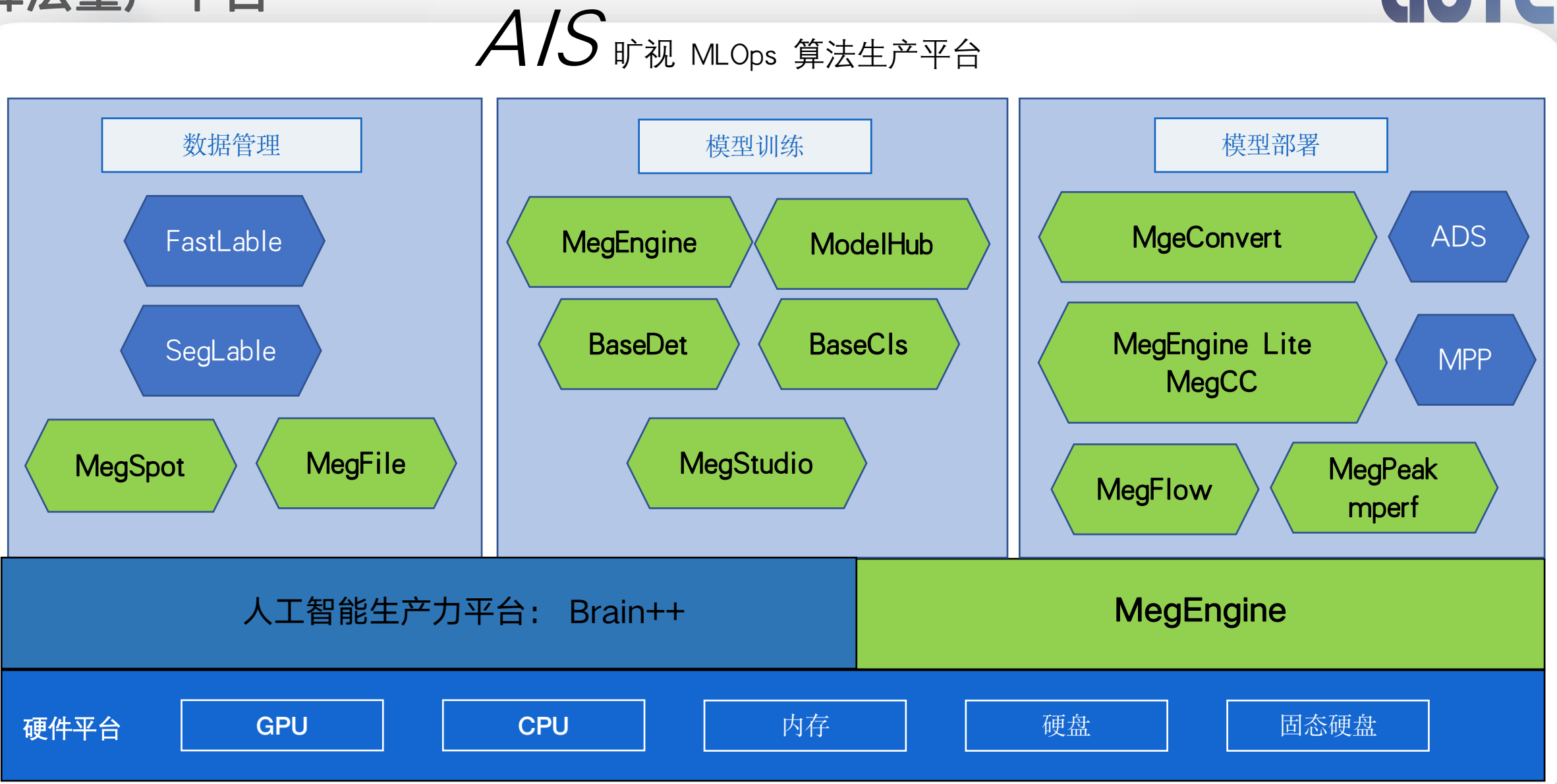
基础服务

- K8S
- Nori
- 数据库
- 对象存储
- 镜像服务
- 日志服务

运营端

- 全局用户管理
- 全局事件管理
- 资源管理
- 机器管理
- 数据统计/审计

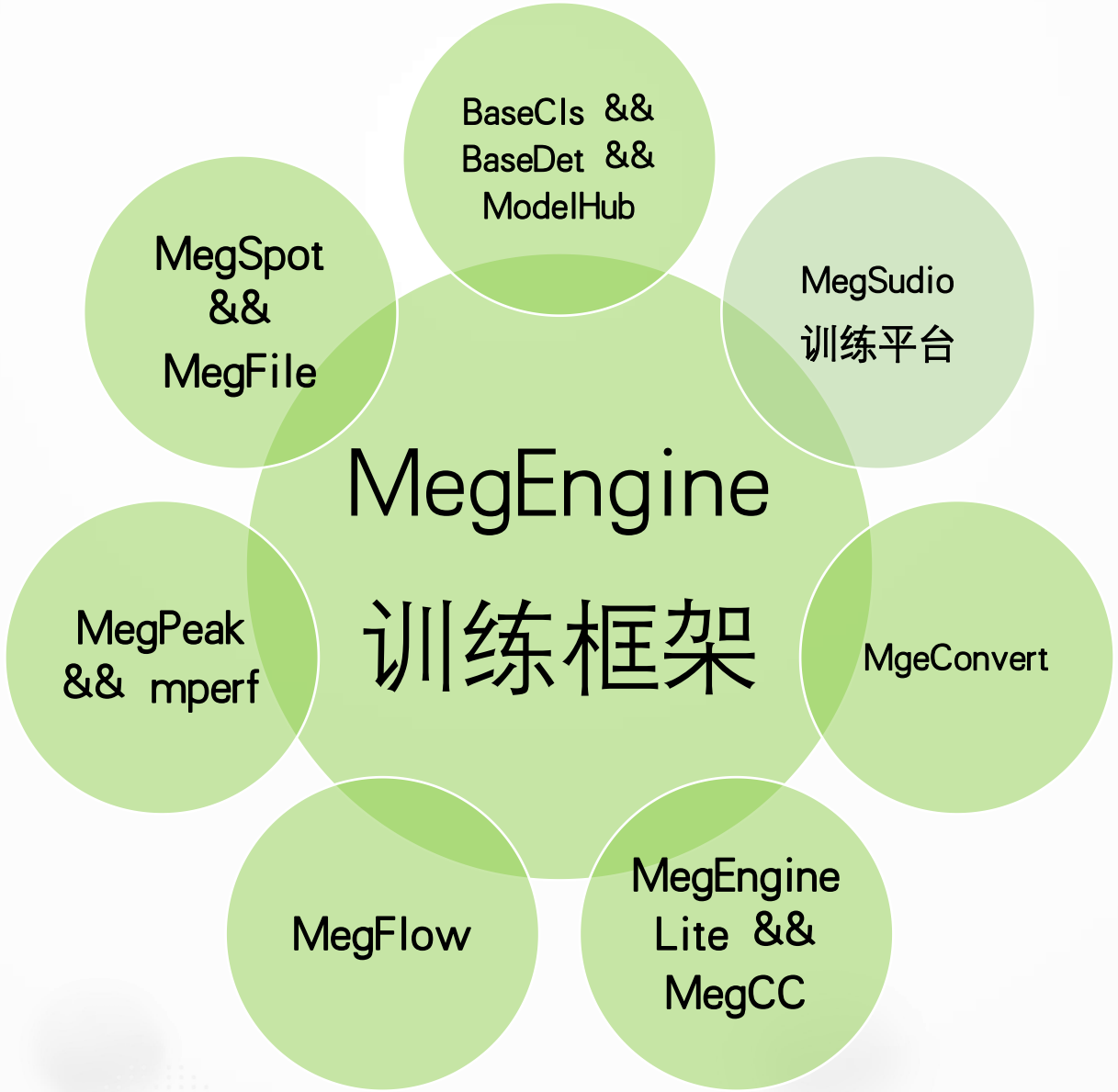
算法量产平台


AIS 旷视 MLOps 算法生产平台
数据管理
FastLable
SegLable
MegSpot
MegFile
模型训练
MegEngine
ModelHub
BaseDet
BaseCls
MegStudio
模型部署
MgeConvert
ADS
MegEngine Lite
MegCC
MPP
MegFlow
MegPeak
mperf
人工智能生产力平台： Brain++
MegEngine
硬件平台
GPU
CPU
内存
硬盘
固态硬盘

MegEngine 开源生态

覆盖功能

- 数据预处理
- 模型训练
- 模型转换
- 模型推理
- 算法 pipeline 搭建
- 硬件性能测评



算法量产生态 MegEngine

从 Alpha 到 V1.13， 不断优化，提升产品竞争力

- 用户使用体验的优化
- 训练性能，推理性能优化
- 以及生态建设

持续迭代，打造自主可控的国产开源框架

算法量产生态 MegEngine

GOTC

训练推理一体

一套内核支撑训练到推理，无需模型转换，精度损失最小化。

前后处理可放入计算图推理，训练推理精确对齐，python C++ 不用写两遍。

traced module → megengine lite → megflow，从模型到高并发视频流处理服务只需几行python代码。

超低硬件门槛

DTR 算法减少 75% 显存占用，1080 也能训 transformer。

自研 pushdown 内存分配算法，带来最低的内存/显存占用。

自动代码裁剪可使部署文件 binary size 下降 10 倍，有效降低推理硬件成本。

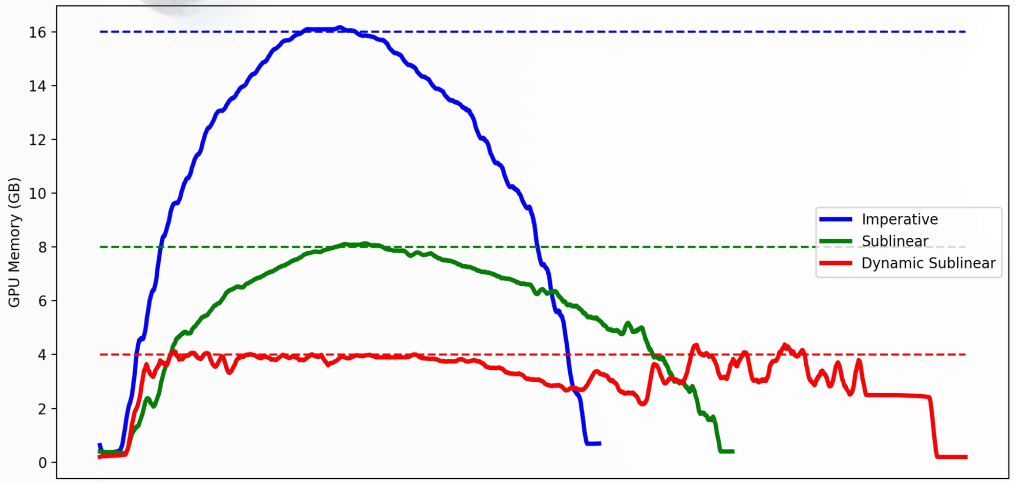
全平台高效推理

高效的推理性能，在各类 CPU、GPU 上均可享受到极致性能体验。

自动 layout、kernel 算法选择机制，轻松达到最优推理性能。

跨平台模型精度对齐，一套代码走天下。

首个支持动态图显存优化的开源框架




```
megengine.dtr.eviction_threshold = "4GB"
megengine.dtr.enable()
```

训大模型神器

显存占用降至 1/4

动静均支持

一键开启DTR[1]

无需阈值设置

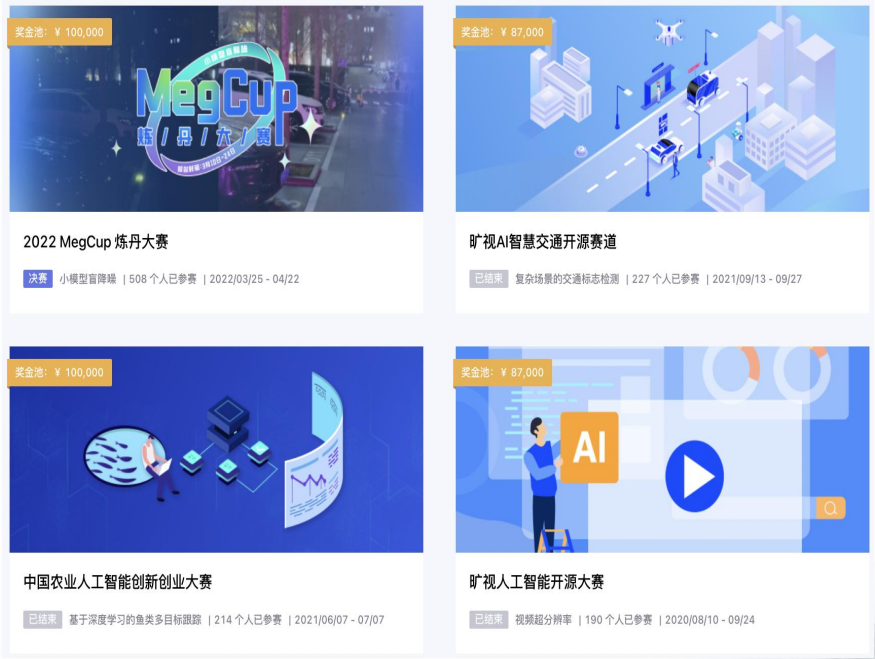
可帮助企业和高校大幅节省研发硬件投入成本

[1] Kirisame M, Lyubomirsky S, Haan A, et al. Dynamic tensor rematerialization[J]. arXiv preprint arXiv:2006.09616, 2020.

算法量产生态

GOTC

MegStudio 一站式 “深度学习”模型开发平台



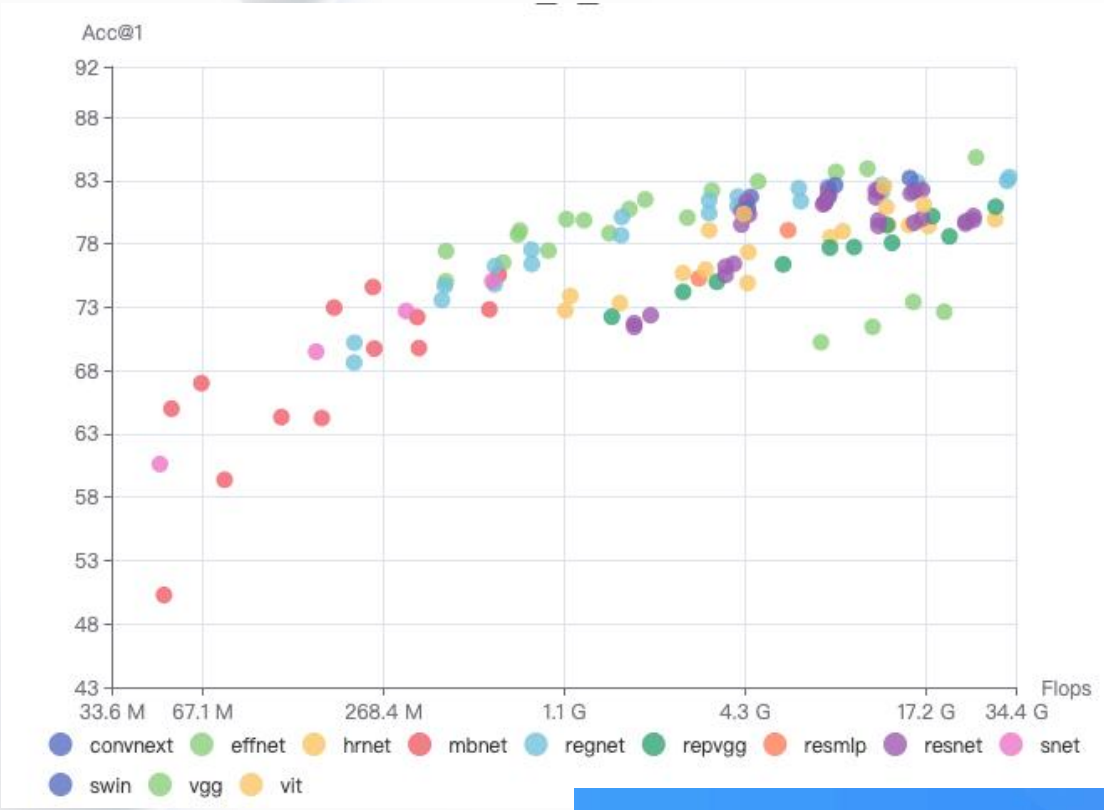
MegStudio 是面向 C 端海量 AI 入门级选手的在线一站式 AI 模型学习与实训平台。

为初学者提供了课程、项目、数据集、模型、计算框架、超级算力、持久存储等多种免费资源，

大大降低了 AI 技术的准入门槛。

算法量产生态

BaseCls



BaseDet

▼ model zoo

Model name	input size	lr sched	box mAP	weights
Faster R-CNN R50-FPN	800	1x(12e)	37.7	github
RetinaNet R50-FPN	800	1x(12e)	36.2	github
FreeAnchor R50-FPN	800	1x(12e)	38.4	github
FCOS R50-FPN	800	1x(12e)	39.0	github
ATSS R50-FPN	800	1x(12e)	39.5	github
OTA R50-FPN	800	1x(12e)	41.0	github
DETR R50	800	150e	39.9	github

模型中心

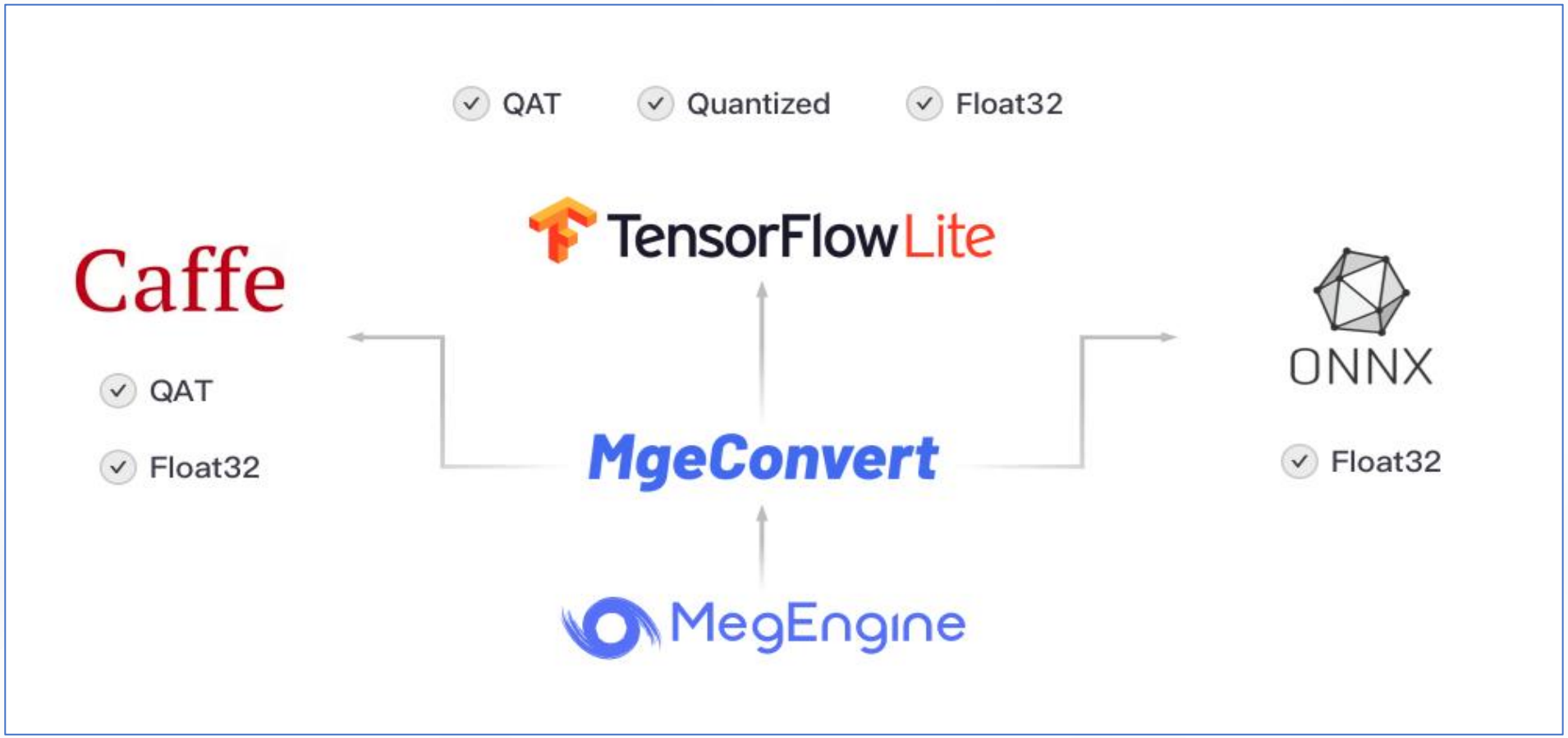
GitHub »»

基于旷视研究院领先的深度学习算法，提供满足多业务场景的预训练模型

算法量产生态

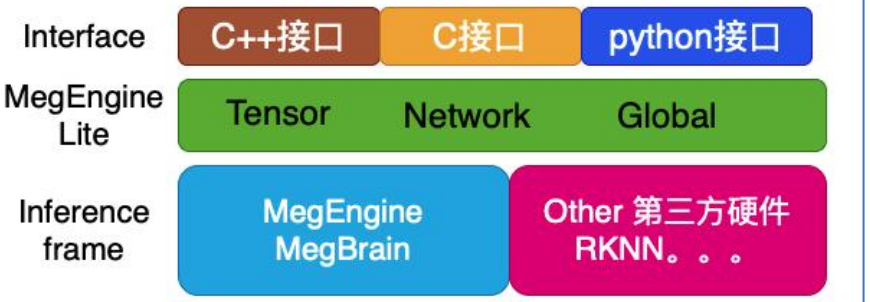
MegConvert

MgeConvert 实现多平台模型互转



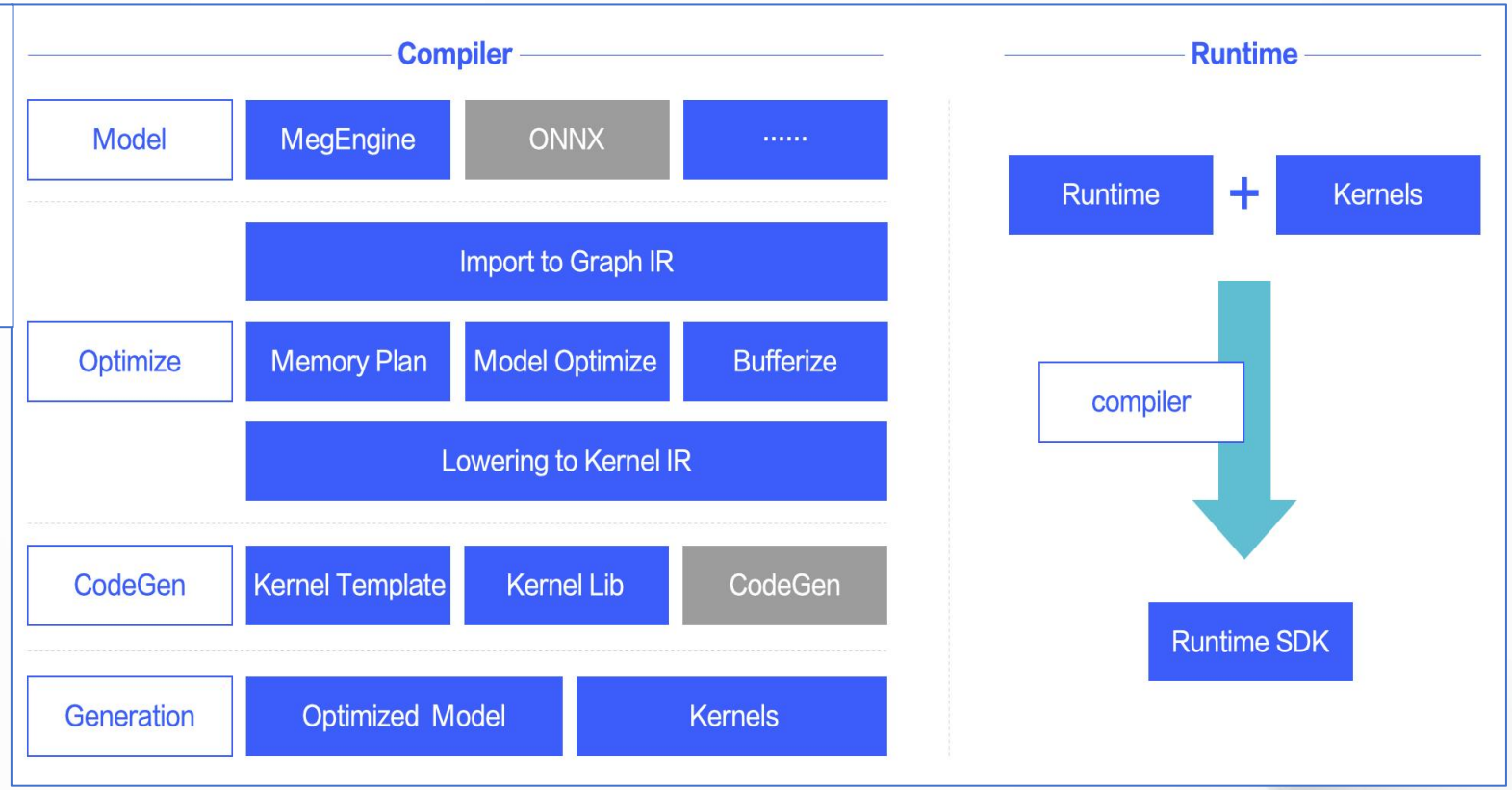
算法量产生态

MegEngine Lite/MegCC



覆盖

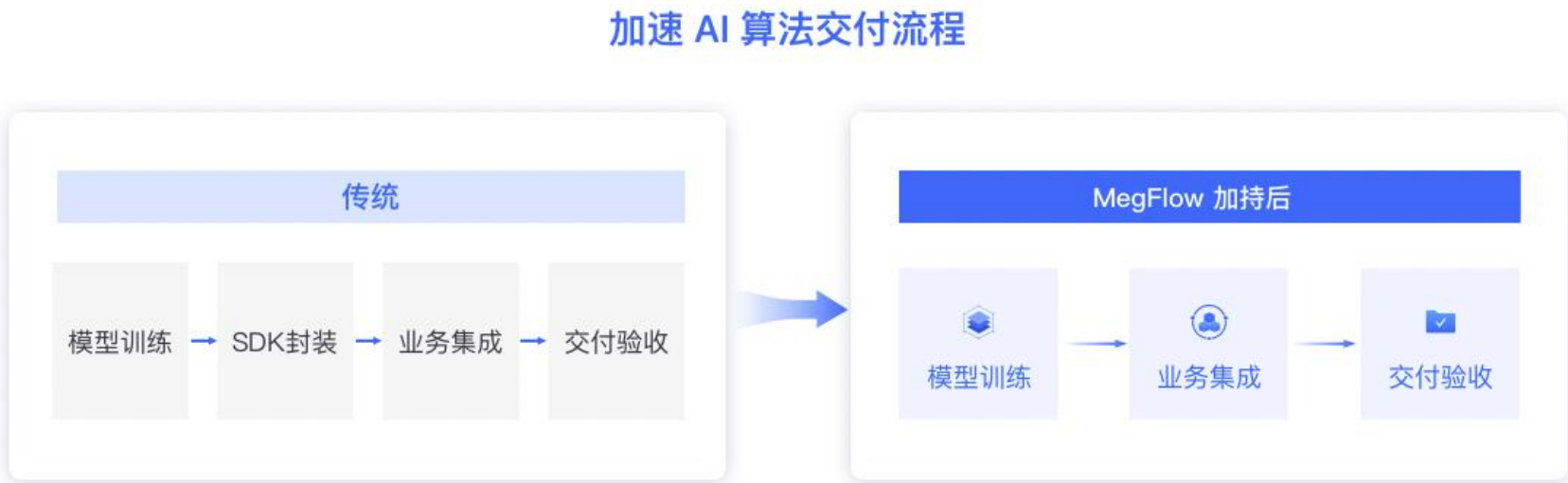
- 服务器推理
- 移动端推理
- 边缘推理
- 特殊情况推理(MegCC)



算法量产生态

MegFlow

MegFlow 提供了一种更为简洁的视觉应用落地流程，用户可以直接用 Python 搭建计算图，不必关心 C++、图优化相关问题，省去了 SDK 封装流程，可快速实现算法交付。



MegEngine 生态全景图

- GitHub：https://github.com/MegEngine/MegEngine
- 官网：https://www.megengine.org.cn/
- 论坛：https://discuss.megengine.org.cn/

群号：1029741705

技术交流 QQ 群

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE