沐曦股份 MXMACA-3.3.0.X 简要技术报告

一、 前言

1. 版本概述与核心定位

近期，沐曦股份发布了 MXMACA 软件栈（简称“MACA”）的 3.3.0.X 版本，MACA 套件是面向沐曦曦云 C 系列、曦思 N 系列 GPU 研发的异构计算软件栈核心计算平台、引擎、运维工具和规范化操作范本，作为沐曦“自主 GPGPU 硬件+全栈软件体系”的关键协同载体，如图 1 所示，MACA 承担着连接硬件算力单元与上层应用生态的核心纽带作用，覆盖底层驱动、用户态接口、编译器、算子适配、训练框架、推理框架、行业场景优化等全链路能力，是支撑国产 GPU 生态落地与行业赋能的算力基座。

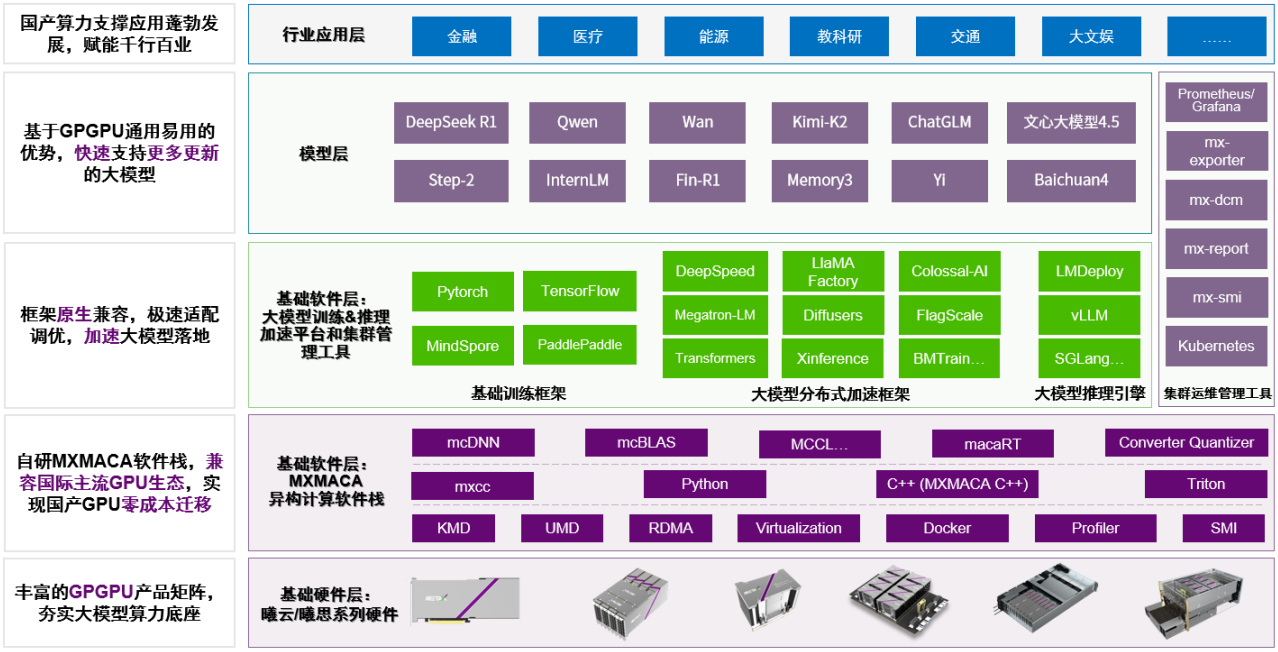

图 1 MACA 套件的定位和作用

本次 MACA 套件版本发布聚焦“生态强化与场景深度适配”，涵盖底层基础能力迭代与主流 AI 框架、大模型训推、搜广推、科学计算等多维度生态适配，但本报告不针对技术细节做全面罗列，而是聚焦版本对行业应用的实际赋能价值，选取核心场景进行深度解析。

本报告围绕 AI 领域行业核心场景，系统呈现 MACA-3.3.0.X 版本的场景适配成果、效能表现及生态价值，向开发者与合作伙伴清晰传递沐曦软硬件协同的行业赋能能力，为相关方的技术选型与产业落地提供专业参考。

2. 版本测试保障

为确保 MACA 版本作为核心协同载体的稳定性、功能完整性与性能优越性，切实支撑国产 GPU 生态落地与行业赋能，每个 MACA 版本正式发布前均经过多维度、大规模的严格测试验证，构建起覆盖软硬件协同优化和行业应用支撑底座的全流程质量管控体系。

测试体系以行业场景的全面匹配度和覆盖度为目标，共包含超过 1 万 5 千个 MACA 软件栈测试用例和超过 1 万个行业相关场景应用测试用例，这些用例在整个测试周期中反复迭代执行，从功能正确性、性能达标率、长期稳定性等多维度验证产品质量，确保满足商业落地的严苛要求。测试覆盖近 30 种国际主流及国产操作系统及内核（如表 1 所示），仅曦云 C 系列 GPU 产品测试相对应的测试就占用超过 60,000 个 GPU 小时，以大规模资源投入保障测试的全面性与有效性。

表 1 MACA-3.3.0.X 版本适配 CPU、操作系统和内核对照部分列表

CPU 架构	操作系统	操作系统内核版本
x86_64	Ubuntu18.04	5.4.0-42-generic
x86_64	Ubuntu18.04	5.4.0-131-generic
x86_64	Ubuntu20.04	5.15.0-58-generic
x86_64	Ubuntu22.04	5.15.0-72-generic
x86_64	Ubuntu22.04	5.15.0-88-generic
x86_64	Ubuntu24.04	6.14.0-27-generic
x86_64	CentOS 7	5.14.0
x86_64	CentOS8	4.18.0-240.el8.x86_64
x86_64	RHEL 9.6	5.14.0-570.12.1.el.9.6.x86_64
x86_64	Kylin V10 SP2	4.19.90-24.4.v2101.ky10.x86_64
x86_64	KylinV10	4.19.90-24.4.v2101.ky10.x86_64
x86_64	KylinV11	6.6.0-32.7.v2505.ky11.x86_64
x86_64	TencentOS 3.1	5.4.119-19.0009.44
x86_64	TencentOS 3.1	5.4.119-19.0009.54
x86_64	TencentOS 3.3	Kernel:5.4.241-24.0017.23
x86_64	TencentOS 4.4	6.6.92-34.1.tl4.x86_64

CPU 架构	操作系统	操作系统内核版本
x86_64	Anolis OS 8	5.10.134-18.an8.x86_64
x86_64	Anolis OS 23	6.6.25-2.an23.x86_64
x86_64	Alinux 3	5.10.134-13.1.al8.x86_64
x86_64	BCLinux R8 U2	4.19.0-240.23.11.el8_2.bclinux.x86_64
x86_64	CCLinux 22.09.2	5.15.131-2.cl9.x86_64
x86_64	CTYunOS 23.01	5.10.0-136.12.0.86.ctl3.x86_64
x86_64	CULinux 3.0	5.10.0-60.67.0.116.ule3.x86_64
x86_64	CULinux 4.0	6.6.0-72.0.0.95.ule4.x86_64
x86_64	Debian 10	5.10.0-0.deb10.28-amd64
x86_64	KeyarchOS 5.8	4.18.0-477.27.1.3.kos5.x86_64
x86_64	RockyOS 9.2	5.14.0-284.11.1.el9_2.x86_64
x86_64 Hygon C86-4G	Ubuntu22.04	5.4.0-144-generic
x86_64 Hygon C86-4G	Ubuntu22.04	5.15.0-119-generic
aarch64 FT5000C	Kylin V10	5.15.0-1.10.6.v2307.ky10h.aarch64
aarch64 Kunpeng 920	Kylin V10	5.15.0-1.10.6.v2307.ky10h.aarch64

在核心测试模块覆盖上，MACA 软件栈测试精准对标全链路技术能力。其中，开发效率引擎层的测试能力包括 CTS（兼容性测试套件）、编译器、数学库（算子库）、通讯库、工具链、虚拟化、视频编解码等基础模块，确保底层基础能力的稳定可靠；涵盖算子库广度以及算子压力测试（算子库范围和数量如表 2 所示）、用户态接口的完备性和性能优化测试、集合通信能力验证等核心模块。在垂直场景赋能层中，生态适配体系层面则涵盖 PyTorch、TensorFlow、PaddlePaddle 为代表的主流 AI 框架兼容、Megatron-LM、DeepSpeed 等大模型训练框架和 vLLM、SGLang 等推理框架及加速库支持以及科学计算场景适配等，针对生态适配体系中的 AI 框架与模型兼容需求，测试环节专门覆盖 40 余种主流 AI 框架（如表 3 所示）及接近 500 个模型类别，每个模型会在多种参数组合和多种 GPU 配置环境下完成性能验证，最终每个细分场景下都会输出超过 5,000 条性能数据，全面保障模型运行效率与兼容性，为开发者提供低成本、高性能运行的坚实保障。

表 2 部分常用算子库及其测试用例数

kernel 类型	算子级别测试用例数
mcblas	450221
mccub	82
mcdnn	849
mcgeqrf	1
mcfft	2162
mcrand	33
mcsolver	32
mcsparse	101
mcthrust	213
mctlass	100
mcapex	53
mcflash_attn	372608
mcflash_mla	8
mcflash_infer	51
mcpytorch	2410
mcsage_attn	4
mcspconv	26
mctriton	8481
mcxformers	175

表 3 MACA-3.3.0.X 版本支持 AI 框架部分列表

名称	框架类别	名称	框架类别
OnnxRuntime/MACART	推理	ColossalAI	训练
Triton-Inference-Server	推理	DeepSpeed	训练

名称	框架类别	名称	框架类别
vLLM	推理	Llama-Factory	训练
SGLang	推理	Megatron-DeepSpeed	训练
PPL.LLM.Serving	推理	Diffusers.Training	训练
LMDeploy	推理	InternLM	训练
KTransformer	推理	Megatron-LM	训练
Transformers	推理	PaddlePaddle	训练
FastDeploy	推理	PyTorch	训练
LightX2V	推理	TensorFlow	训练
ComfyUI	推理	Swan Labs	训练
xDiT	推理	XTuner	训练
tvm	推理	VeRL	训练
ChiTu	推理	MS-Swift	训练
SiliconLLM	推理	siiRL	训练
Diffusers	推理	JAX	训练
InfiniLM	推理	InfiniTrain	训练
Mooncake	推理	FlagScale	训练
AIBrix	推理	BMTrain	训练

3. 核心发布信息汇总

	核心信息
适配硬件	沐曦曦云 C 系列 GPU、曦思 N 系列（基于全自研 GPGPU 核心 IP 及架构，原生支持全精度计算、MetaXLink 高速互连及硬件级虚拟化、软切分能力）
版本定位	生态强化版：聚焦算子全量覆盖、主流框架深度兼容、多场景性能优化，全面提升商业落地适配性与开发者使用体验
核心升级方向	1. PyTorch 2.8 版本算子全量覆盖； 2.大模型训推性能对标国际旗舰产品；

	3.搜广推场景多技术栈深度适配； 4.垂直领域（AI4S、传统小模型）专项优化
正式发布时间	● MACA-3.3.0.X 软件栈于 12 月 8 日发布 ● AI 框架适配版本于 12 月 15 日发布
版本迭代核心亮点	**MACA SDK：** ● 单机多卡环境下，支持任意数量 GPU 动态锁定/解锁同一主机内存，实现多卡 H2D 传输 ● 优化 Stream 优先级，即保证 Graph 额外创建的 Stream 和 Graph Launch API 使用的 Stream 优先级一致 **通讯库：** ● 适配 MIXL 库 ● DeepEP 适配 Hidden Size 和专家数等更多参数规格，以支持更多 MoE 大模型 ● 分层算法支持多机 Reduce Scatter、All Gather 通信功能 **数学库：** ● XFormers 将 attention backend 所用的 flashAttn2.5.3 升级到 2.6.3，并支持全部 memory efficient forward API 功能 ● mctlassEx 新增 w8a16 contiguous group gemm 接口功能 ● mcDNN 新增 int8/fp16 fwd conv+gelu 融合功能 ● FlashMLA 支持 DeepSeek v3.2 所需的 sparse prefill 和 decode 功能 **MACA PyTorch:** ● 支持 torchcodec-0.6.0 ● 发布 PyTorch2.8 版本 **MACA JAX:** ● 正式发布 mcJAX-0.4.34 AI 训推框架 **PaddlePaddle:** ● 适配 Paddle3.0 版本 ● 支持 Customer Kernel 注册 ● 支持大模型训推一体 ● 支持科学计算高阶微分 **vLLM：** ● 适配 0.11.0 版本发布 **SGLang：** ● 适配 0.5.4 版本发布，并优化性能

二、 生态适配度详情

1. 全栈基础能力：生态全域覆盖与技术底座革新

1.1 PyTorch 框架适配与算子全量覆盖

沐曦 MACA-3.3.0.X 版本完成了对 PyTorch 2.8 版本的深度适配工作，实现了对原生算子配置文件 native_functions.yaml 中定义算子体系的全面兼容。本次适配覆盖全部 2650 个核心算子（其中 GPU 算子 2410 个），涵盖基本算术运算、线性代数操作、卷积/池化类算子、规约操作、随机采样、索引与切片快速傅里叶变换（FFT）、Attention 等关键算子类别；在张量运算维度，同时支持稠密张量与稀疏张量的完整运算逻辑，数据类型层面则覆盖整数、浮点、布尔、复数及量化类型等多类数据形态，保障了算子能力的完整性与场景适配性。

在此基础上，沐曦基于该完备的算子系统，进一步完成分布式训练、torch.compile 等高级特性的适配与落地，实现了从基础算子层到高阶训练与编译优化能力的全栈式兼容，为基于 PyTorch 2.8 的各类深度学习训练与推理场景提供了稳定、全面的底层算子支撑。该适配方案基于沐曦全栈软件体系打造，向上兼容 PyTorch 原生接口与核心模块，向下深度契合自研 GPU 硬件特性，无需调整工程构建逻辑即可实现现有模型无缝使用。

为保障生态兼容性，MACA 套件通过生态适配工具链实现构建系统平滑切换，支持 C++扩展功能及 Megatron-LM、DeepSpeed 等主流大模型训练框架，支持 vLLM、SGLang、Transformers 和 KTransformer 等主流大模型推理框架，兼容 Ubuntu、CentOS、RHEL、openEuler、Anolis OS 、银河麒麟等主流 Linux 发行版。同时完整支持混合精度训练、分布式训练、torch.compile 编译优化与图模式任务下发的深度集成等关键特性，搭配性能分析与优化工具链，核心场景性能对标主流 GPU 水平。此外，通过内存分配优化、数据布局适配等底层调优，进一步释放硬件算力，结合轻量化部署方案与丰富示例程序，大幅降低使用门槛。不仅夯实了国产 GPU 的生态基础，更为深度学习开发者提供了开箱即用、高效稳定的技术支撑，加速 AI 训练与推理场景的产业化落地。

1.2 第三方开源仓库资产复用测试

CUDA 是 GPGPU 领域的行业标准，能便捷实现 GPU 并行编程，支撑各类软件与框架运行，GitHub 上相关项目近 3 万个，覆盖并行计算、科学计算等关键场景，影响力远超同类技术。对于已有资产的适配意义重大：一方面，其适配后可快速接入成熟开源生态，拓展 AI、数据处理、气象预报、计算化学等多元应用场景；另一方面，能满足 HPC 软件、PyTorch 等主流框架需求，降低用户学习成本，提升平台竞争力，填补国产异构计算平台的 GPU 加速生态空白。

质量保证团队以 GitHub 为核心数据源，按“含 CUDA 关键字且 star 数量大于 1 且具有活跃度”的规则筛选代码仓，本版本测试选取 4490 个仓库进入正式测试；这些代码按依赖库集中于 MPI、BLAS 等高频库，按应用领域可划分为 AI 模型和应用、高性能并行计算、气象模拟、计算化学等场景，按编程语言 C/C++原生语言为主。通过“双环境验证+自动化流水线”的方式推进，最终 4490 个开源项目适配测试结果如图 2 所示：57 个暂不可用（编译失败 11 +运行失败 46）、260 个修改后可用（含结果不

一致 45)、4173 个直接可用（结果一致），直接适配成功率 92.94%。

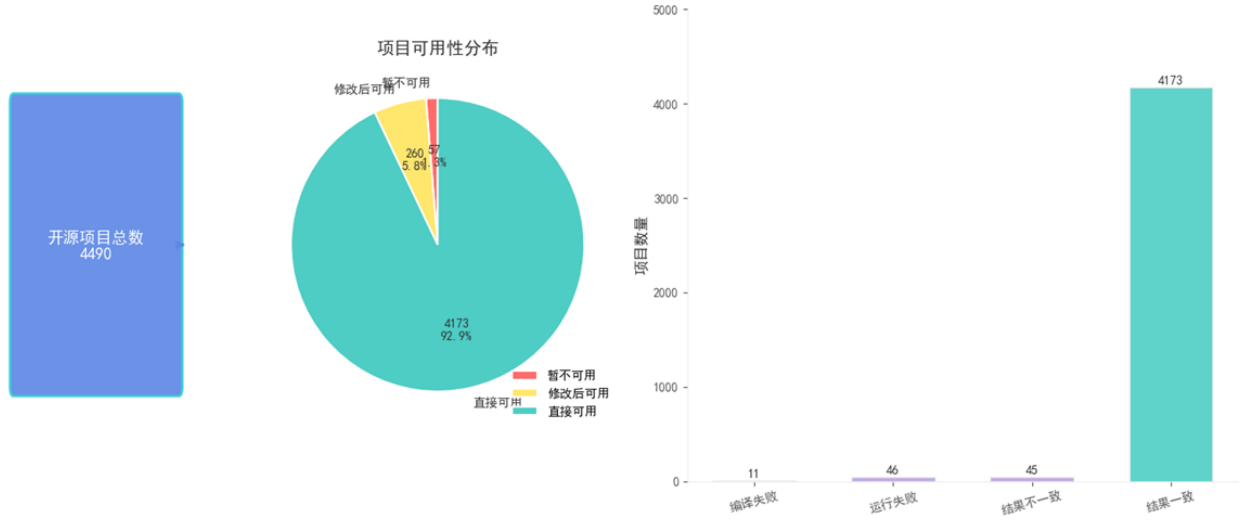

图 2 MACA 套件中开源项目适配测试结果

这些直接通过的项目无需额外改动代码，从 GitHub 筛选后拿过来即可在 MACA 平台稳定运行，覆盖 MPI、BLAS 等核心依赖库及气象模拟、计算化学等主流应用场景，充分体现了 MACA 对现有生态的兼容深度。仅小部分项目需手动微调，这类项目共 260 个、占比 5.79%（不足 6%），且修改量极小。主要集中在 cmake 配置优化、少量头文件适配或编译器脚本调整，无需改动核心业务逻辑，平均每个项目手动修改耗时不超过半天。结合自动化流水线的批量验证能力，整体适配效率与可用性处于行业较好水平，为用户快速使用并行加速应用提供了可靠支撑。

1.3 全栈工具链优化与多场景适配主要特性一览

在开发效率引擎层，MACA 套件通过高性能算子库、智能编译工具链、专业性能分析工具及配套工具库，构建起降低异构开发门槛的技术体系。其中，六大核心高性能算子库（mcBLAS、mcDNN、mcFlashAttention 等）针对多 GPU 拓扑优化内存访问与并行逻辑，如 mcBLAS 支持按 GPU 数量动态切分矩阵，mcFlashAttention 通过三级存储体系减少跨 GPU 通信；编译器工具支持 MACA C/C++、Fortran 等多语言，结合指令重排、内存合并、任务自动切分等多 GPU 优化策略，将高级语言转化为高效可执行程序；性能分析工具则通过系统级追踪与核函数级指标采集，助力定位计算瓶颈，搭配 mcPytorch、mcTriton 等工具库，进一步简化异构开发全流程，相关技术细节如图 3 所示。

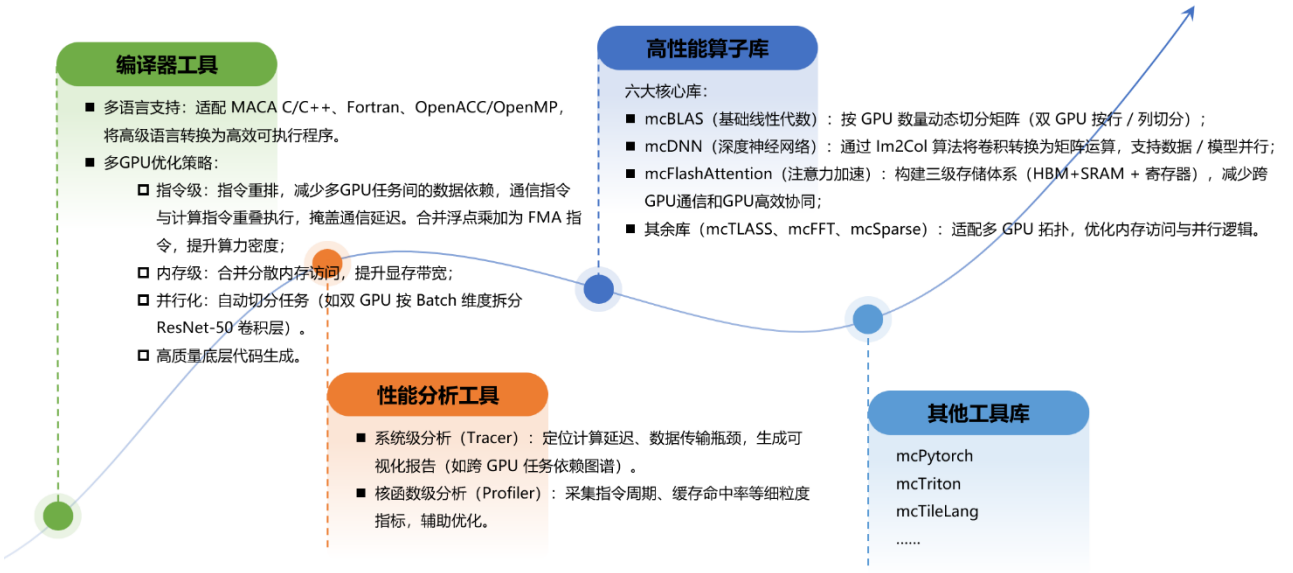

图 3 MACA 套件中开发效率引擎——降低异构开发门槛

在垂直场景赋能层，MACA 套件围绕 AI 与科学计算两大领域，通过针对性的优化策略与框架适配，实现算力与行业需求的精准融合。AI 领域中，训练优化兼容 PyTorch、BMTrain 等框架，依托硬件流水线并行实现通信与计算重叠，优化分布式并行策略；推理优化则适配 ONNX Runtime、vLLM、SGLang 等框架，采用 INT8 量化、KVCache 跨卡管理提升长序列处理效率。科学计算领域通过重构 MPI、BLAS 库提升内存带宽，并定向移植 OpenFOAM、GROMACS 等专业科学计算框架，结合容器化部署方案，确保算力能高效支撑流体仿真、分子动力学等垂直场景，完成从算力供给到行业价值转化的关键衔接，具体实施方案如图 4 所示。

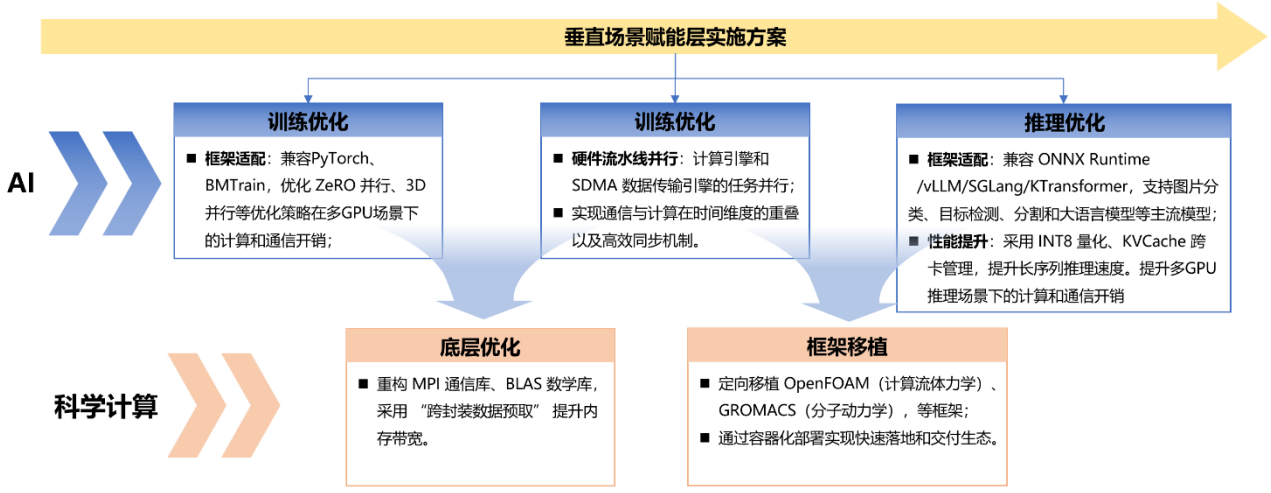

图 4 MACA 套件中垂直场景赋能层——算力与行业需求的融合

2. 大模型训推一体化：大模型算力支撑底座和效能突破

沐曦 MACA-3.3.0.X 版本构建起覆盖大模型训练与推理全流程的一体化算力支撑底座，通过软硬件深度协同、核心算子优化、分布式架构升级，破解大模型超大规模参数训练的通信瓶颈、高算力需求、长周期部署等核心痛点，实现训推效能的跨越式突破。

2.1 训推一体化算力底座核心架构

2.1.1. 硬件算力基座支撑

依托沐曦自研 GPGPU 的高算力密度、高内存带宽与高速互连优势，底座提供从单卡到万卡级集群的弹性算力供给。单卡原生支持多精度混合计算，内存容量与带宽适配千亿参数模型的存储需求；跨节点通过 MetaXLink 自研高速互连技术，构建低时延、高带宽的分布式通信网络，为大规模集群训推奠定硬件基础。

2.1.2. 全栈软件协同赋能

以 MACA 异构计算软件栈为核心，构建起端到端协同体系，实现软硬件能力的深度耦合与效能最大化。该体系全面兼容 PyTorch、PaddlePaddle、TensorFlow、JAX、Megatron-LM、DeepSpeed、XTuner 等主流大模型训练框架，全面兼容 vLLM、SGLang、LMDeploy 等大模型推理框架，图 5 展示了 MACA 套件在大模型推理场景下的优化技术汇总。总体特征是无需大幅修改代码即可支持现有模型，降低开发者使用门槛；依托 MetaXLink 自研高速互连技术与 MCCL 高性能通信库，构建低时延、高带宽的分布式通信架构，有效破解分布式训练中的通信瓶颈；集成拓扑感知 MCCL 分布式通信库，能够动态识别集群拓扑结构并适配最优通信策略，为多机多卡训推提供高效数据协同支撑；同时内置自研编译器优化模块，通过算子自动融合、循环展开等编译级智能优化，充分挖掘硬件底层算力潜力，实现计算资源的高效利用，为大模型训推全流程提供稳定、高效的软件底层支撑。

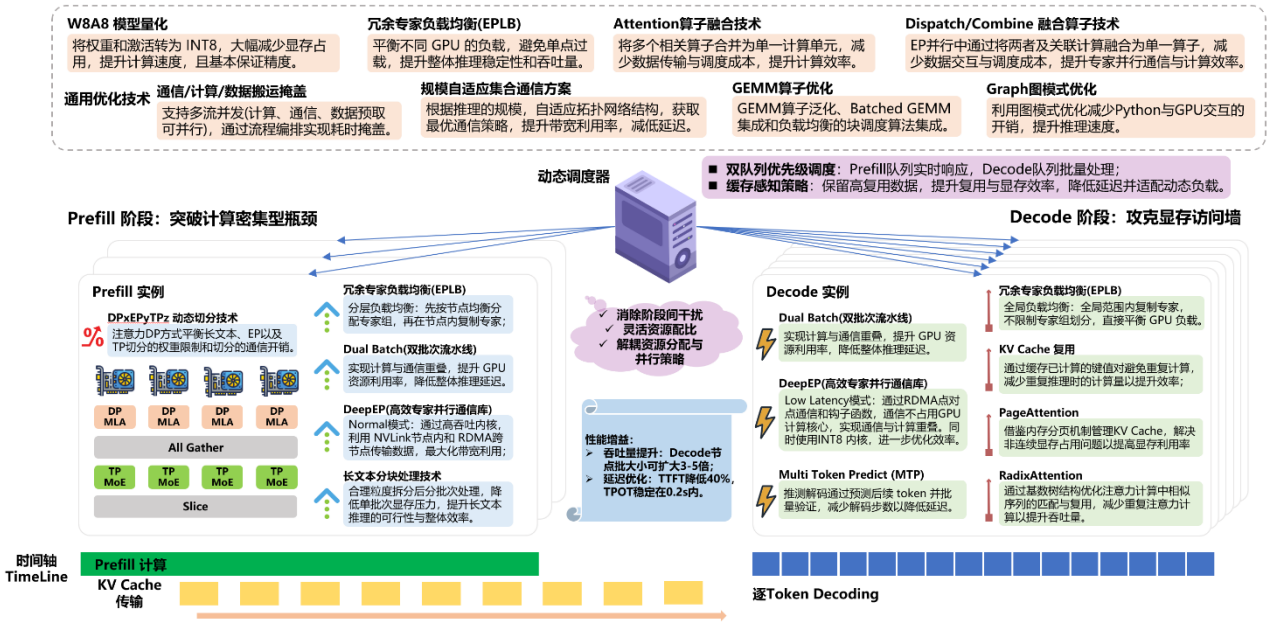

图 5 MACA 套件大模型推理优化技术汇总

2.1.3. 训推无缝切换能力

MACA 软件栈底座打破训练与推理的场景壁垒，支持模型训练后的轻量化转换与直接部署，无需二次适配。通过统一的模型格式与接口规范，实现“训练-微调-推理-部署”全流程链路打通，大幅降低大模型从研发到落地的周期成本。

2.2 核心效能优化技术突破

2.2.1. 关键算子深度调优

针对大模型训推核心算子开展硬件亲和性优化：

FlashAttention 算子：优化数据布局与访存流水线设计，融合计算与数据搬运操作，适配大模型长上下文生成需求。按 GPU 片上高速缓存大小拆分 Q/K/V 数据块，让计算全程在高速缓存内完成，不用反复读写外部 HBM 高速内存。同时整合矩阵相乘、Softmax 归一化等多步操作，中间结果不落地，大幅减少 HBM 数据传输开销。支持 FP16/BF16 多精度与超长序列，长序列场景吞吐量提升，内存带宽占用降低，模型精度完全不受影响，高效缓解访存瓶颈。

分布式集合通信库：作为分布式训推的“数据协同中枢”，负责多机多卡间高效数据同步与交换，是大规模集群发挥算力的核心支撑。针对 AllReduce、All2All、AllGather 等高频算子开展全维度优化：AllReduce（聚合核心）采用算法自适应策略，根据数据量动态切换 Ring/Tree/Recursive Doubling 算法，结合节点内预聚合+跨节点拓扑感知路由，减少 20%跨节点通信延迟；All2All（MoE 专家并行关键）通过动态分组通信、流量均衡调度优化，避免专家数据交换时的网络拥堵，如图 6 所示在 EP144 的实践中，使用了优化后的 All2All 通信库，专家并行效率提升 15%，；AllGather（数据汇聚）采用分块流水线传输+异构网络适配，提升数据分片聚合速率。同时叠加通信压缩（梯度量化/稀疏化）、预通信调度等技术，千卡集群线性度稳定在 95%以上，保障大模型分布式训推的高吞吐与低延迟。

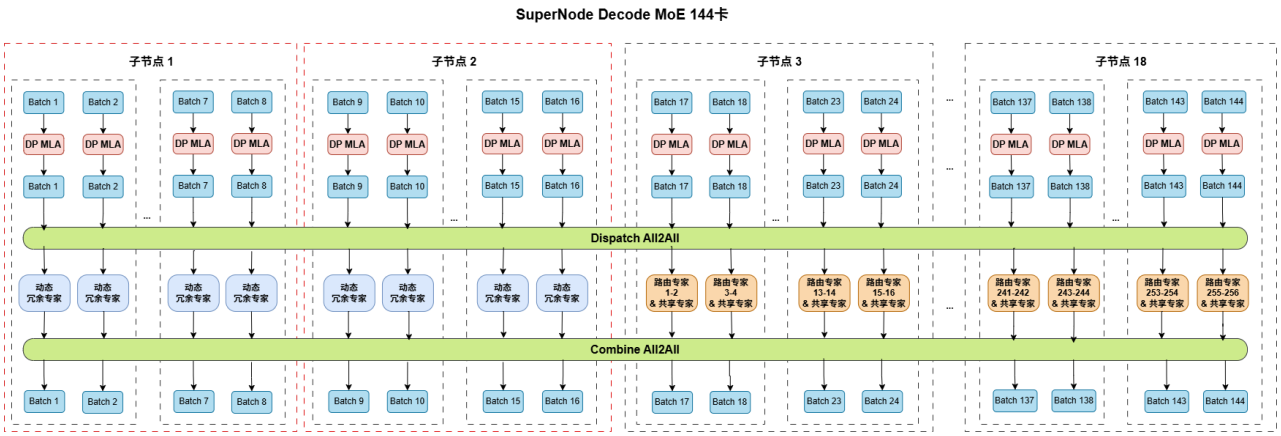

图 6 分布式集合通信库在大 EP 并行中的使用

通信-计算重叠优化：通信-计算重叠优化是突破 GPU 训推性能瓶颈的核心技术，旨在解决数据通信与计算任务串行执行导致的资源闲置问题。核心通过异步通信机制实现：依托 MACA 自研 MCCL 集合通信库的非阻塞接口，将数据传输任务与 GPU 计算任务解耦；结合任务调度引擎预加载远程数据、拆分通信粒度，利用 GPU 空闲周期并行处理数据传输；部分架构通过硬件级专有通信单元卸载，进一步降低 CPU 干预开销。图 7 所示，在大模型分布式训练场景的实践中，通过计算和通信并行可显著缩短端到端延迟，提升 GPU 利用率 15%-30%，在分布式训练、大模型推理等场景中，有效缓解跨节点/跨卡通信瓶颈，支撑更大批量数据处理与更复杂模型的高效运行。

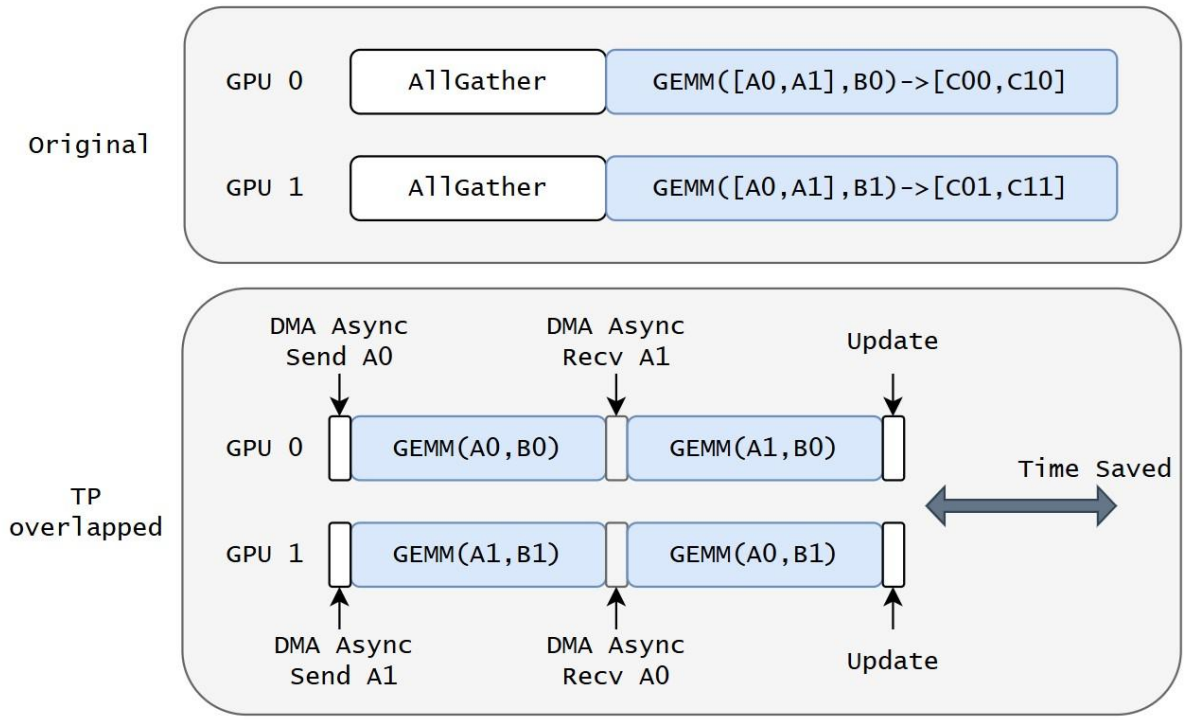

图 7 大模型分布式训练场景的计算和通信并行

2.2.2. 编译与部署优化

编译级效能提升：深度支持 torch.compile 动态图编译优化，通过算子自动融合、循环展开、指令调度优化等手段，最大化硬件算力利用率，模型训练迭代速度提升；

推理引擎轻量化适配：针对大模型推理场景打造专用轻量化引擎，优化算子调度与批处理策略，覆盖长短序列差异化需求，短序列推理延迟降低，长序列推理吞吐量提升；

企业级部署适配：兼容容器化部署与云原生调度架构，支持集群快速扩容与弹性伸缩，简化环境配置与运维流程，适配企业级大规模落地需求，降低部署与运维成本。

2.3 效能突破核心表现

1、训练效能：针对大规模大模型训练场景，显著缩短训练周期，在大规模集群分布式训练中展现优异线性度，可支持长周期无故障稳定运行，保障训练任务高效推进；

2、推理效能：对主流大模型推理性能进行深度优化，显著降低推理延迟、大幅提升吞吐量，在长上下文推理场景下仍保持高效稳定的运行表现，适配复杂业务需求；

3、兼容性：全面兼容当前主流大模型生态体系，覆盖全系列主流模型，无需进行代码修改即可直接开展训练与推理工作，降低模型优化与适配成本；

4、扩展性：具备从小规模调试到大规模训推的全场景平滑扩展能力，可灵活适配不同规模企业的技术研发与生产部署需求，提供高效、可扩展的算力支撑方案。

2.3.1 大模型训练性能数据

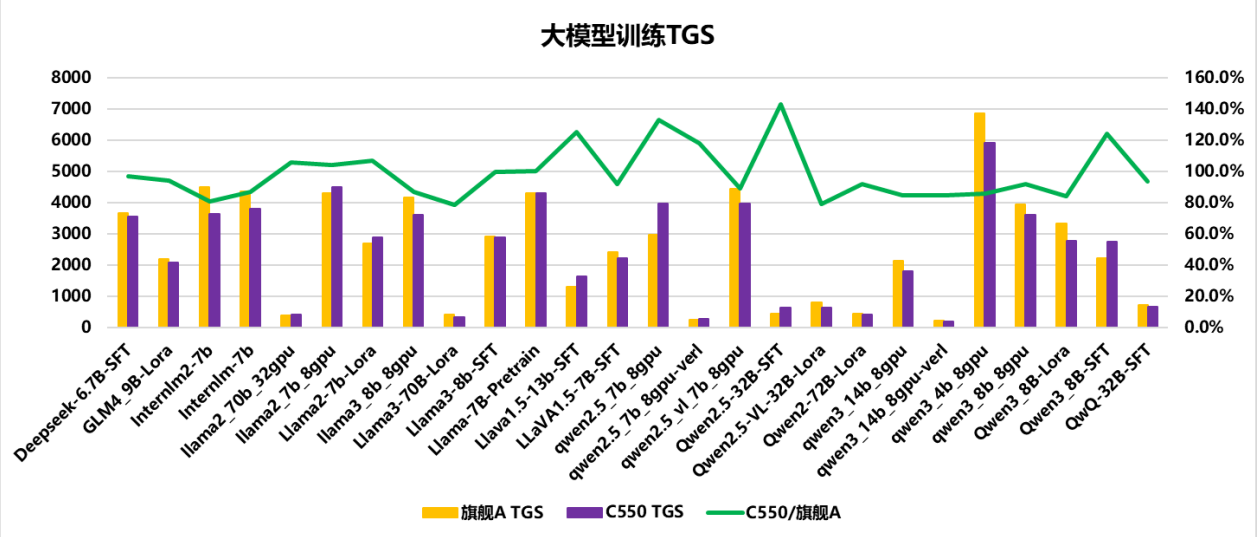

图 8 不同大模型训练任务的 TGS 对比

图 8 展示了 DeepSeek、GLM、InternLM、Llama、Qwen 等多系列大模型，在不同参数规模（如 7B、13B）及任务类型（SFT、Pretrain）下的训练 TGS 数据，包含“旗舰 A TGS”（黄色柱）、“C550 TGS”（紫色柱）及两者效率比值（绿色折线）。

2.3.2 大模型推理性能数据

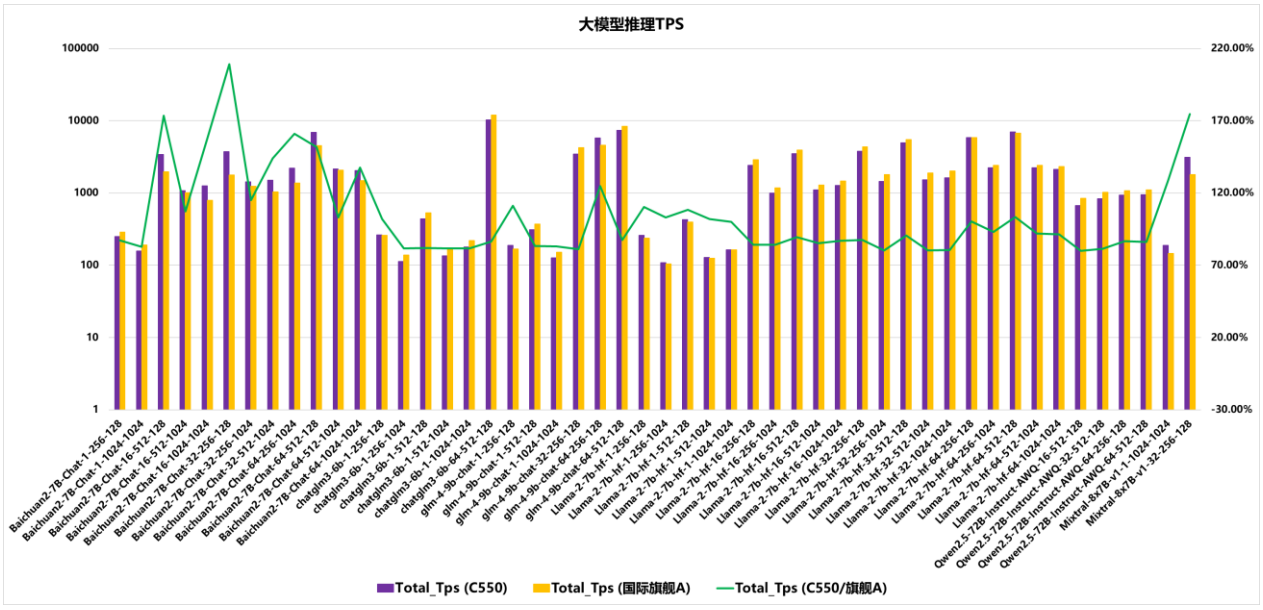

图 9 不同大模型推理任务的 Total TPS 对比

图 9 呈现大模型推理阶段的 Total_Tps 指标对比，横轴为组合型 Model-Name（格式：模型名 - 并发数 - Input size-Output size），涵盖 Baichuan2、chatglm3、glm4、Llama2、Owen2.5、Mixtral 等模型及不同并发、输入输出尺寸的配置。纵轴左侧为 Total_Tps 数值，右侧为 Total_Tps（C550 / 国际旗舰 A）的比值；紫色柱代表 Total_Tps（C550），黄色柱代表 Total_Tps（国际旗舰

A)，绿色折线表征两者的比值，展示了不同模型及配置下的 TPS 表现与相对比值波动。

3. 搜广推业务全场景：多技术栈协同的全链路训推适配升级

搜广推（搜索、广告、推荐）是互联网核心流量变现与用户体验优化场景，其核心诉求是在海量数据中精准匹配用户需求，并支持高并发、低延迟的实时决策。随着数据规模爆发式增长和模型复杂度提升，GPU 凭借大规模并行计算能力、高内存带宽、专用计算核心，成为搜广推场景的核心算力支撑。本版本技术报告重点讨论 TensorFlow/JAX 与 XLA 技术的深度融合，暂不展示 TorchRec 体系。

3.1 训练适配：TensorFlow/JAX + XLA 深度协同，打造高效训练新范式

3.1.1. 技术栈支持：全链路覆盖与深度融合

全面完成 TensorFlow、JAX 双框架与 XLA 技术栈的深度协同适配，打通从数据输入、模型构建、编译优化到分布式执行的全链路流程。其中，TensorFlow 依托成熟的模型开发体系、工业级分布式训练框架及搜广推场景生态优势，提供低门槛的开发体验；JAX 则凭借函数式编程特性、原生 XLA 深度集成能力及灵活的自动微分机制，适配高性能、定制化的训练需求；XLA 编译器作为统一优化层，通过算子融合、内存智能调度、静态编译优化等核心能力，解决传统训练中内核调用频繁、内存开销大的痛点。三者形成“TensorFlow 生态便捷性+JAX 高性能灵活性+XLA 编译高效性”的三重优势，覆盖搜广推场景模型训练全流程，无论是基于 TensorFlow 的工业化落地，还是基于 JAX 的极致性能调优，均无需额外适配即可实现高效接入。

3.1.2. 核心优化：多卡训练与精度优化双轮驱动

单机多卡高效适配：以 TensorFlow 单机多卡训练框架为核心，完成其数据并行模式与 XLA 的深度协同适配，打通 TensorFlow 多卡调度接口与 XLA 跨设备编译链路；同时针对 JAX 的分布式特性，适配其分布式接口与 XLA 跨设备编译逻辑，解决 JAX 多卡场景下数据分片、设备通信的效率瓶颈。借助 XLA 对 TensorFlow 静态计算图、JAX 函数式计算图的统一优化能力，解决单机场景下多卡间数据同步、算子调度协调等关键问题，实现大规模批次样本的高效并行训练；依托 TensorFlow 成熟的多卡资源管理能力、JAX 轻量化的分布式调度特性与 XLA 的编译优化联动，让单机多卡资源利用率提升。

混合精度计算原生兼容：深度适配 TensorFlow 混合精度训练接口，基于其对 FP16/BF16/TF32/FP32 数据类型的原生支持，结合 XLA 编译器的精度自适应优化逻辑；同时充分利用 JAX 原生对 BF16/FP16 的轻量化支持，针对 JAX 函数式计算图的精度传播特性，优化 XLA 的精度适配规则。XLA 可统一解析 TensorFlow、JAX 计算图中各模块的精度需求，自动识别模型核心计算模块与精度敏感模块，对 TensorFlow 场景侧重“生态兼容下的精度平衡”，对 JAX 场景侧重“极致性能下的精度可控”，实现高精度与高性能的动态平衡。相较于传统 CPU 训练方案，依托 TensorFlow/JAX 与 XLA 的协同优化，训练周期缩短，有效支撑大规模稀疏特征与复杂模型的长时间稳定训练。

编译优化深度迭代：针对 TensorFlow 定义的搜广推高维稀疏算子及特征处理流程，优化 XLA 的编译适配逻辑；同时针对 JAX 生态下的稀疏计算需求，定制 XLA 对 JAX 稀疏算子的编译规则，解决 JAX 稀疏特征处理中编译开销高、算子碎片化的问题。XLA 基于 TensorFlow 静态计算图、JAX 函数式计算

图的结构特点，精准识别“特征查找-交叉-激活”等关键子图，通过子图聚类、算子自动融合等技术，将多步 TensorFlow 操作或 JAX 函数调用融合为单一编译单元，减少数据在 TensorFlow/JAX 算子与 XLA 编译单元间的搬运开销，让核心计算模块训练效率提升。

3.1.3. 适配模型：覆盖核心场景与复杂架构

深度兼容搜广推领域全量核心模型，无论是基于 TensorFlow 开发的传统机器学习模型（LR、GBDT）、深度学习基础模型（DeepFM、Wide&Deep、DCN、NFM）、复杂序列模型（DIN、DIEN），还是基于 JAX（结合 Flax/Haiku 高层神经网络库）实现的同类型模型，均无需大幅修改代码即可接入适配体系。针对 Transformer 类模型的注意力机制、DeepFM 的特征交互模块等复杂计算单元，分别定制 XLA 编译优化规则：对 TensorFlow 版本侧重“生态兼容下的算子稳定性优化”，对 JAX 版本侧重“函数式计算图的编译效率优化”，既保障 TensorFlow 模型工业化训练的稳定性，也提升 JAX 模型在 GPU 集群下的计算效率，全面覆盖 CTR/CVR 预估、推荐排序、搜索召回等核心训练场景。

3.1.4. 效果展示：TensorFlow/JAX +XLA 的深度融合

图 10 为部分搜广推模型在 XLA 技术的深度融合下的，与国际旗舰产品 A 的推理时间对比效果。横坐标为选取的各种主要模型，柱状图的主纵轴表示训练单个 step 的平均耗时，折线图的次纵轴表示国际旗舰产品 A 与沐曦 GPU 产品的比值。

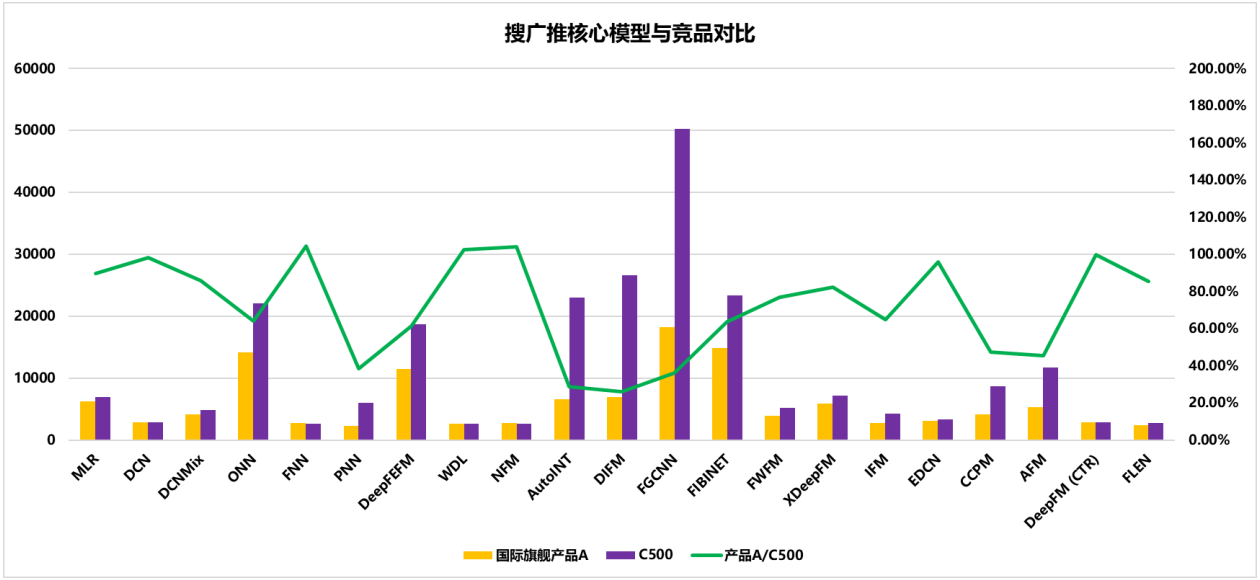

图 10 搜广推模型使能 XLA 与国际旗舰产品 A 的对比效果

3.2 推理适配：TVM + XLA 协同，构建低延迟推理体系

3.2.1. 技术栈支持：编译与部署一体化支撑

完成 TVM + XLA 推理技术栈全流程适配，构建“模型转换-图优化-图切分-图编译-算子编译优化-部署落地”的一体化支撑体系。TVM 提供跨硬件平台的模型优化与部署能力，支持 TensorFlow 等多框架模型的统一转换；XLA 则作为核心编译引擎，承接模型的算子优化、子图编译等关键环节，两者协同实现“模型无需改造即可编译，编译结果可直接部署”的高效流程，适配 GPU、CPU 等加速器等多硬件

环境，降低跨平台部署成本。

3.2.2. 核心优化：多维度技术降低推理开销

算子融合与编译优化：通过 XLA 的算子融合能力与 TVM 的图优化策略，实现推理计算图的深度优化，将 Concat、Transpose、Reduce、Split、Elementwise 等串行算子融合为复合算子，减少内核调用次数，推理计算效率提升。

W8A8 低比特量化落地：针对搜广推推理场景对延迟的严苛要求，实现 W8A8 低比特量化技术的全流程支持，在 XLA 编译器的量化感知编译与 TVM 的量化部署工具协同下，模型体积压缩，推理延迟大幅度优化，同时精度损失可控，满足业务指标要求。

动态批处理智能适配：结合搜广推业务流量波动特点，支持动态批处理技术，通过 TV 的批处理调度模块与 XLA 的动态 shape 编译能力，自动适配不同流量场景下的请求批次大小，在高并发场景下吞吐量提升，低并发场景下延迟降低。

3.2.3. 适配场景：全面覆盖核心业务全流程

全面覆盖搜广推业务核心推理场景，实现全场景高效支撑：

搜索场景：适配召回（向量召回、协同过滤召回）、粗排、精排全链路推理，通过低延迟优化保障搜索结果毫秒级返回，提升用户检索体验；推荐场景：支撑个性化推荐的精排、重排环节，动态批处理技术适配流量峰值波动，确保推荐列表实时更新与高效推送；广告场景：覆盖 CTR/CVR 预估、广告排序、出价决策等核心环节，低比特量化与算子融合技术保障广告投放的实时性与精准性，提升广告转化效率。

3.2.4. 效果展示：TensorFlow/JAX +XLA 的深度融合

在搜广推场景中，TVM 与 XLA 技术形成“双轮驱动”的推理生态，实现了高效且适配的技术效果：mcTVM 针对搜广推需求，支持稀疏算子、兼容 PyTorch/ONNX 等主流框架，提供端到端编译部署工具链，其适配的数十个搜广推开源模型（如 EasyRec/DeepCTR），平均推理性能超越国际旗舰 GPU 产品 A（121.04%）、产品 B（131.64%），上百个模型可开箱即用。mcXLA 则兼容 TensorFlow/JAX 框架，支持动态 shape 与 JIT 编译，能无缝对接 TF Serving 快速部署，同样适配上百个搜广推模型。二者结合既兼顾通用性编程，又充分优化硬件性能，最终让沐曦 GPU 具备支撑千亿级流量的搜广推全栈推理能力，精准适配这一 AI 落地最成熟的商业场景，高效释放硬件潜力。

图 11 为部分搜广推模型在 XLA 编译技术与 TVM 编译框架深度融合方案下，与国际旗舰产品 A 和 B 的性能对比效果。

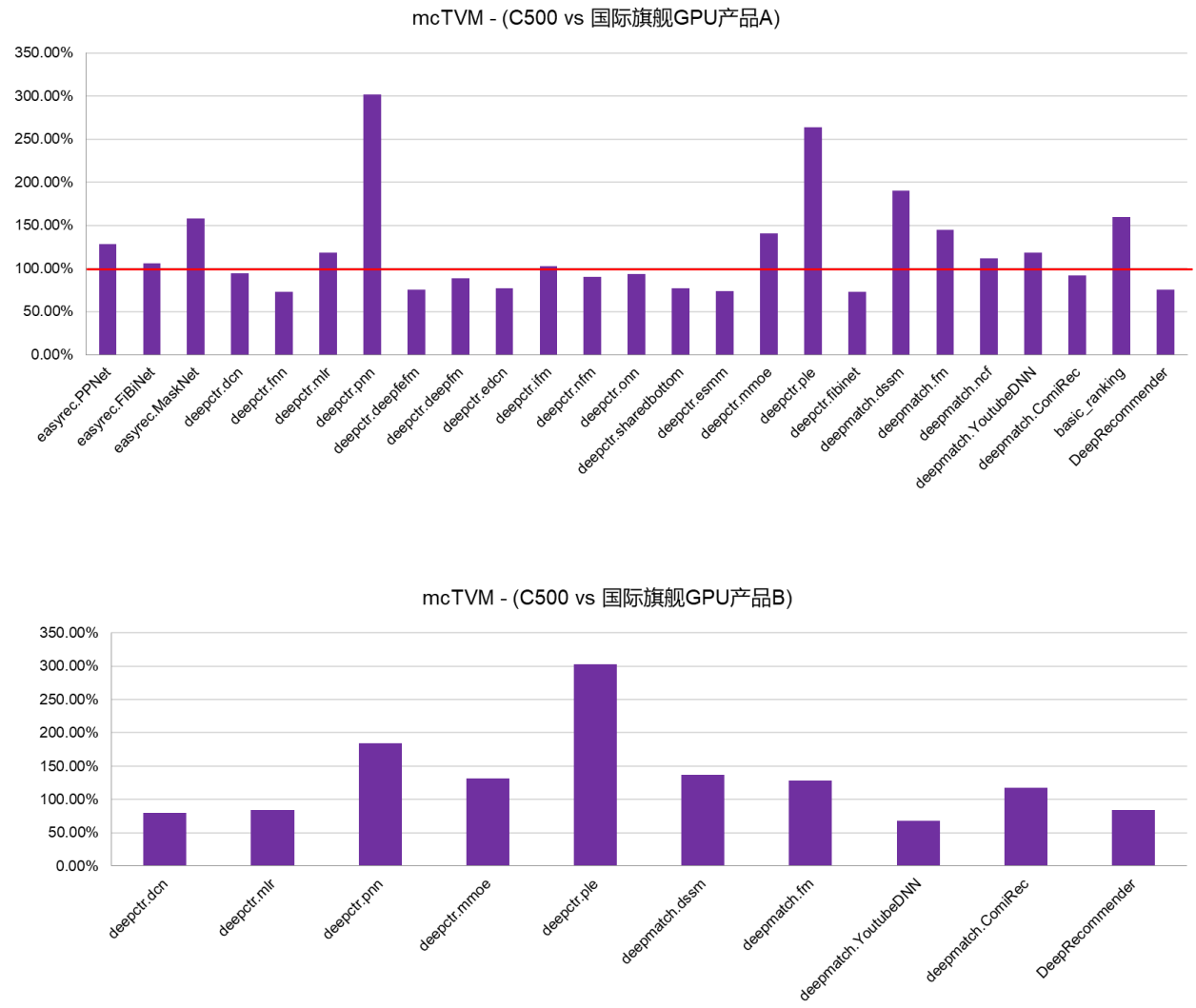

图 11 MACA 套件中主流搜推广开源模型与国际旗舰 A 和 B 的推理性能对比

图 12 呈现推荐系统和向量检索的性能对比测试结果，涉及三类核心指标：it/s（每秒迭代次数，数值越高代表模型运算速度越快）、s/Nits（完成指定迭代量的耗时，数值越低代表效率越高）、QPS（每秒查询数，数值越高代表向量检索吞吐量越强）。测试对象覆盖多目标排序 rechub 等推荐模型、Deep Image/SIFT 系列向量检索模型，对比 C500（紫柱）与国际旗舰 B（黄柱）的性能表现，绿色曲线为 C500 相对国际旗舰 B 的性能占比。

数据层面：推荐模型中，多目标排序 rechub 的 it/s 指标 C500（约 32）优于国际旗舰 B（约 18）；HugeCTR DIN 的 s/8000its 指标中国际旗舰 B 耗时更短。向量检索模型中，Deep Image-ivflat 的 QPS（C500 约 607、国际旗舰 B 约 205）C500 领先显著，仅 SIFT-ivfpq 的 QPS 中国际旗舰 B 略高。多数场景下 C500 性能占比超 100%，Deep Image-ivflat 占比近 300%，体现其在多数测试场景的性能优势。

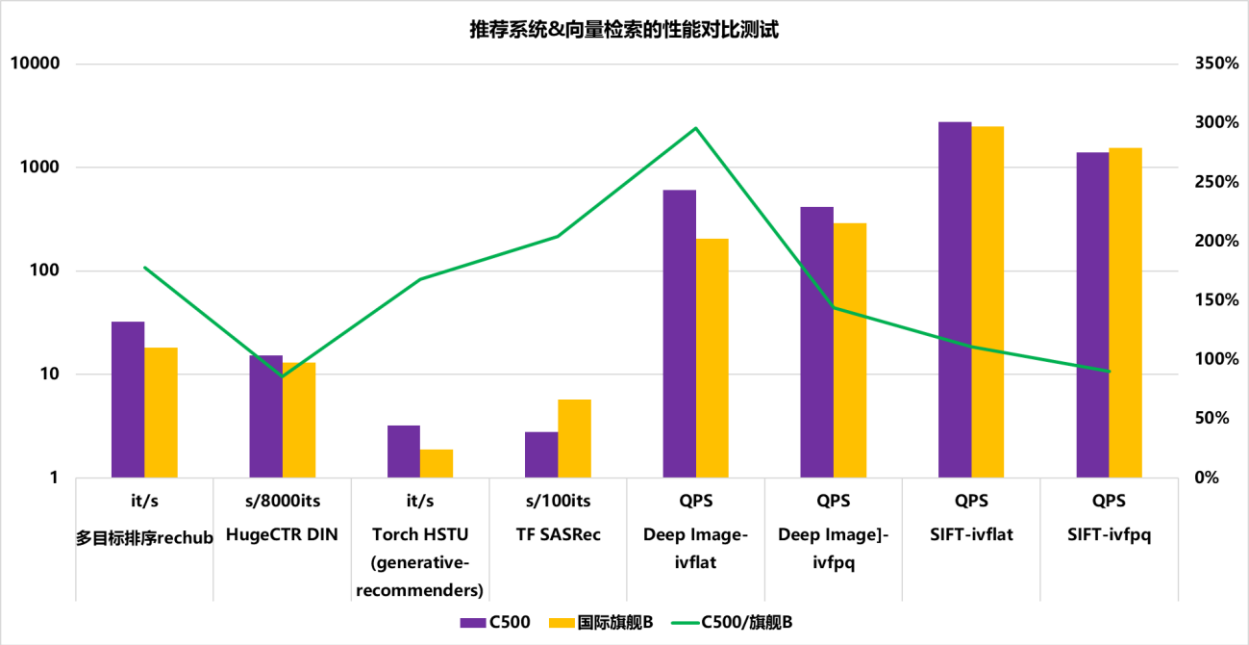

图 12 推荐系统和向量检索与国际旗舰 B 的 XLA 推理性能对比

4. 传统小模型支持：泛场景的低成本与高效落地赋能

针对产业级 AI 应用中传统小模型的部署需求，MACA 套件构建了一套支持多技术领域、低成本、高算力利用率的技术体系，聚焦计算机视觉、自然语言处理及传统机器学习等核心场景，通过兼容主流模型格式、优化底层计算逻辑及简化流程，实现小模型的高效落地与性能提升，为相关技术应用提供标准化技术支撑。方案全面覆盖传统小模型的核心应用场景，无需额外构建专属适配框架，其中计算机视觉场景支持图像分类、目标检测等基础任务，适配各种图像输入格式，可满足工业质检、智能监控、物流分拣等典型场景的轻量化推理需求，兼容主流轻量化视觉模型结构；自然语言处理场景适配文本分类、识别等高频任务，支持多语言文本及不同长度文本的处理需求，可应用于舆情分析、智能客服意图识别、金融信息抽取、法律文书处理等场景，兼容轻量化 NLP 模型的推理逻辑；传统机器学习场景则全面兼容线性回归、聚类分析、决策树、随机森林等经典机器学习算法，适配结构化数据建模需求，可应用于预测、评估、分类、聚类等场景。

MACA 套件具备多模型格式兼容、底层计算优化及优异性能表现等核心技术特性，支持 ONNX、TensorFlow Lite、PyTorch 等主流模型格式，搭配格式转换工具链，可实现模型的直接导入与运行，降低跨框架适配的技术成本；依托 MACA 基础计算库，对 BLAS（矩阵运算）、FFT（频域处理）、Sparse（稀疏数据计算）三大核心模块进行针对性优化，实现 GPU 硬件的算力精准调度，减少算力冗余消耗，提升计算资源利用率，可满足实时推理与高并发处理需求。

与国际旗舰产品 A 相比，部分典型模型的性能测试对比数据如图 13 所示：

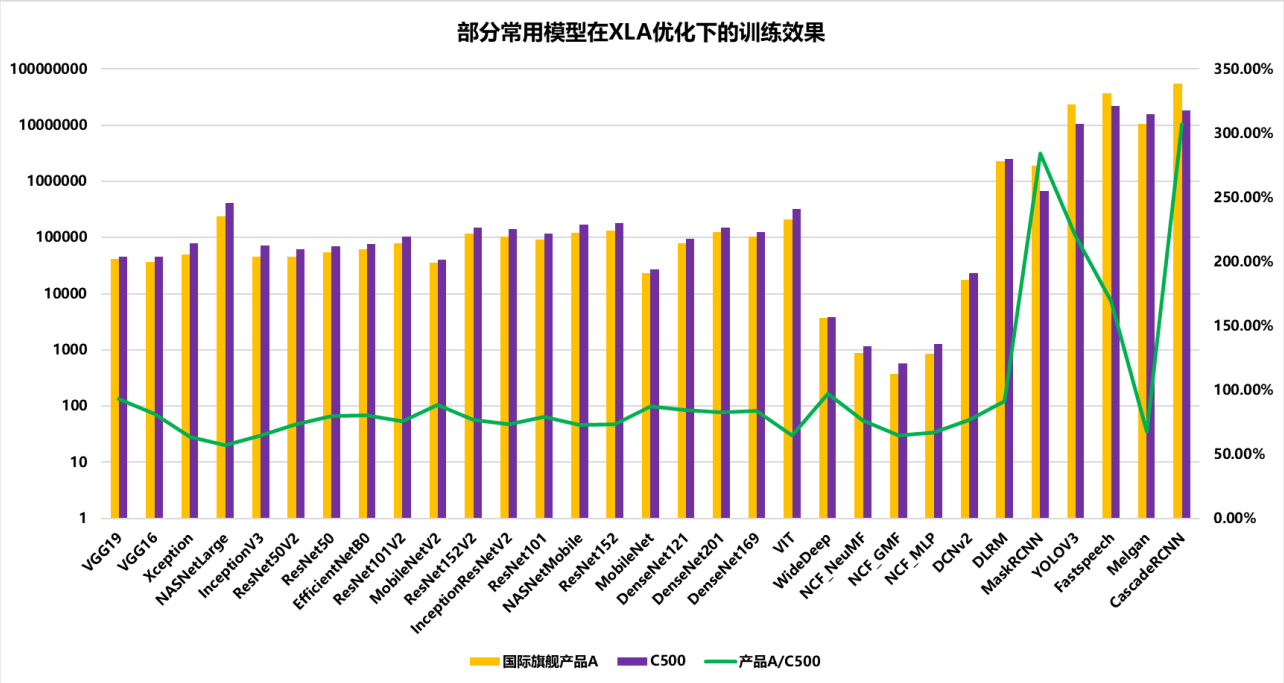

图 13 MACA 套件中 TensorFlow+XLA 的训练效果

图 13 针对 XLA 优化场景下的常用模型训练效果展开对比，横轴覆盖 VGG19、Xception、NASNetLarge 等多类典型模型，纵轴左侧为训练相关每个迭代步的平均耗时（由于不同模型的运行时间差异比较大，为了便于显示使用了对数纵轴坐标系），右侧为“国际旗舰产品 A”与“C550”的指标比值。图中以橙色柱形表示“国际旗舰产品 A”的指标值，紫色柱形表示“C550”的指标值，绿色折线表征两者的比值。该图呈现了多类模型在 XLA 优化下的训练指标差异，比值折线在多数模型区间呈小幅波动，仅在 DRM、YOLOV3 等少数模型处出现显著抬升。

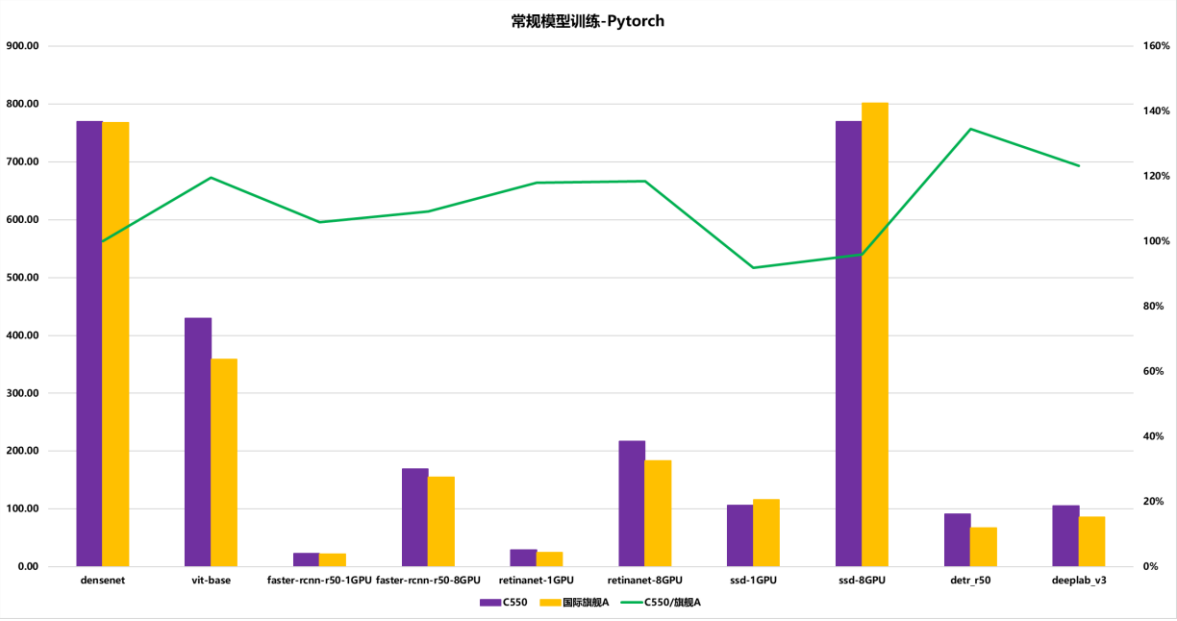

图 14 MACA 套件中常规小模型 PyTorch 框架的训练效果

图 14 展示 Pytorch 框架下常规模型的训练指标对比，横轴包含 densenet、vit-base 及不同

GPU 配置的模型（如 faster-rcnn-f50-1GPU/8GPU），纵轴左侧为训练阶段的吞吐量指标，右侧为“C550”相对“国际旗舰产品 A”的比值。可视化元素包括紫色柱形（C550）、黄色柱形（国际旗舰产品 A）、绿色折线（C550 / 旗舰 A）。图中覆盖了单 / 多 GPU 配置下的模型训练指标，折线反映了不同模型及硬件配置下 C550 相对旗舰产品的指标比值变化。

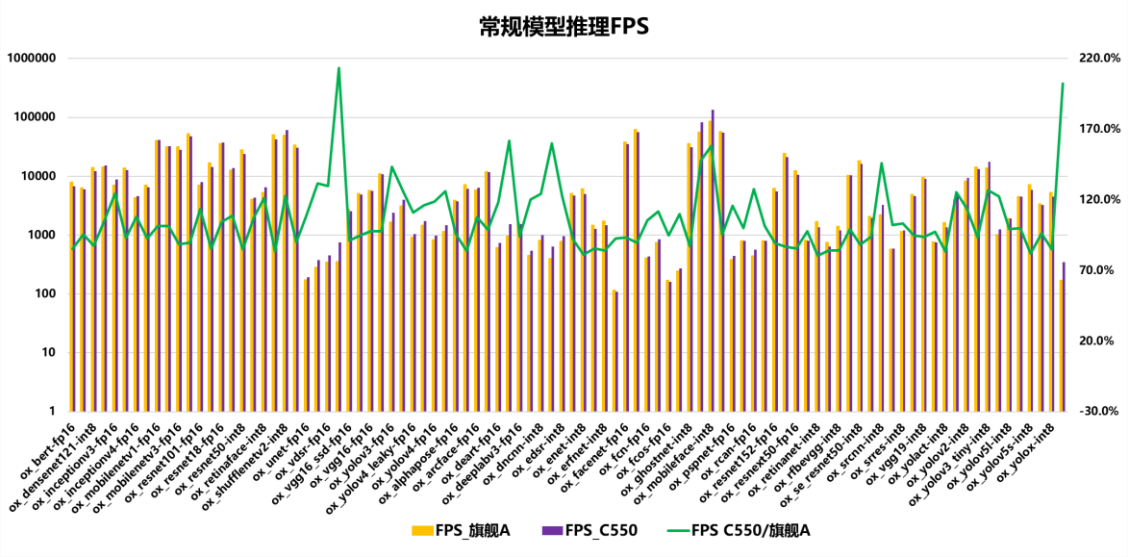

图 15 常规模型推理阶段的 FPS

图 15 聚焦常规模型推理阶段的 FPS（每秒帧率）指标对比，横轴为含精度配置的多类模型（如 ox_bert-fp16、ox_densenet121-int8），纵轴左侧为 FPS 数值（区间 0 至 160000），右侧为“C550 / 旗舰 A”的比值。图中黄色柱形代表“FPS_旗舰 A”，紫色柱形代表“FPS_C550”，绿色折线代表两者的比值。该图涵盖了 fp16、int8 等精度下的多模型推理 FPS，折线呈现 C550 相对旗舰产品的 FPS 比值在不同模型及精度配置间的波动特征。

5. AI4S 核心场景：沐曦推动第五范式科研革新的实践进展

AI4S（AI4Science）是继实验、理论、计算模拟、数据驱动后的第五代科学研究范式，2024 年诺贝尔物理/化学奖对 AI 在基础科学中贡献的表彰，标志其已成为科学创新的核心工具。依托自研 GPGPU 及 MACA 生态套件，沐曦目前对 AI4S 多领域核心场景均已实现覆盖，同时深化了与主流 AI 框架的生态协同，可推动科研与产业的智能化转型，如图 16 所示。

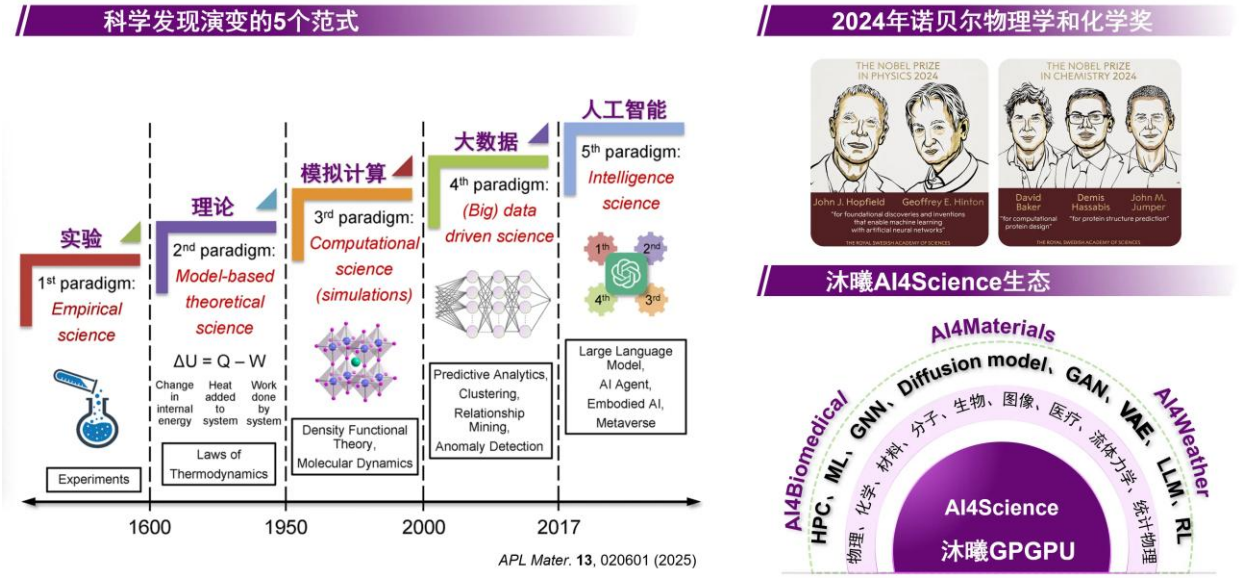

图 16 MACA 套件推动第五范式科研革新的实践进展

5.1 主流 AI 框架的生态适配

Paddle 框架支持：PaddleScience 是基于 PaddlePaddle 开发的科学计算套件，凭借深度学习能力与自动（高阶）微分机制，可解决物理、化学、气象等领域问题，支持物理机理驱动、数据驱动、数理融合三种求解方式，并提供基础 API 与详尽文档供二次开发。目前 PaddleScience 已完成与沐曦 AI 芯片的适配，双方展开深度合作，覆盖智能仿真、高性能计算、科学建模等方向，成功验证 50+科学计算模型全量训练的精度与性能，充分展现国产 AI 硬件在科学计算场景的潜力。后续沐曦将与飞桨继续在高性能科学计算、智能模拟等方向联合攻关，推动 AI for Science 从实验室走向产业落地。

JAX 框架支持：针对 JAX 在科学计算中"高效自动微分+大规模并行计算"的特性，沐曦平台已实现对 JAX 生态的兼容适配，借助其技术能力，可助力物理建模、数据驱动型科学研究的高效开展，进一步丰富了 AI4S 的技术工具链。

5.2 AI4Materials：破解材料研发低效痛点

针对传统“试错式”材料研发的高成本、长周期问题，AI4Materials 构建“第一性原理 + ML 势能+分子动力学+GNN +大模型”的一体化生态。目前，沐曦平台已兼容 ABACUS、DeepMD-kit、LAMMPS 等工具链，融合多物理场耦合与原子尺度生成模型，实现国产化材料模拟平台自主可控，推动产学研融合以提升新型功能材料的研发速度。

5.3 技术科学场景：流体仿真的国产化工具赋能

在技术科学的流体仿真与求解器耦合方向，沐曦平台适配了 PaddleScience 旗下的 **paddleCFD** 组件，可支持圆柱绕流、颅内动脉瘤、空气激波等典型流体问题的仿真计算，结合 CFD-GCN、NSF-Nets 等模型，进一步提升了流体仿真的效率与精度，为汽车控制臂、心脏仿真等工业级场景提供了国产化工具支持。

5.4 AI4Weather：支撑高精度气象风险管控

极端天气对多行业冲击显著，AI4Weather 以秒级响应的 AI 模型，弥补传统数值天气预报的高成本短板。沐曦平台适配 WRF 数值模式及 FourCastNet 等 AI 大模型，可支撑高精度极端天气预警，助力行业风险管理与决策优化。

5.5 AI4Biomedical：重塑生物医药创新格局

在药物研发领域，沐曦 AI4Drug discovery 平台覆盖分子表征、蛋白结构预测等全流程，集成 AlphaFold3、DiffDock 等工具，同时融入 PaddlePaddle 生态下的 **paddleHelix** 工具；该工具可支持分子生成、蛋白配体相互作用预测等关键环节，完善了药物研发的全流程国产化工具链，有效缩短研发周期、降低成本；在医学影像领域，沐曦提供图像重建、分割等工具集，支撑虚拟增强影像、冠脉血流模拟等临床科研方案。

AI4S 通过继承前四范式的优势，实现科研效率与精度的跨越式提升。沐曦的实践证明：基于自主可控的算力底座，结合 Paddle、JAX 等框架适配及 paddleCFD、paddleHelix 等专用组件，可为多领域提供软硬件协同赋能，推动科学研究的范式变革。

6. 版本迭代前瞻：软硬件生态的前瞻布局与能力升级预告

后续将推出曦云 C600 GPU Beta 版，聚焦硬件性能升级并对标国际高端 GPU 水准；软件层面将同步强化前沿技术适配、多模态模型全流程兼容、行业专用框架适配及边缘部署优化等核心能力，相关版本的正式发布时间后续补充。

三、 总结

沐曦 MACA-3.3.0.X 版本依托全自研 GPGPU 架构与 MACA 异构计算软件栈，构建了“1+6+X”战略生态体系：沐曦“1+6+X”战略是其算力生态商业化落地的核心布局，构建了“算力底座-行业赋能”的闭环体系。

“1”为数字算力底座：以沐曦 GPU 为依托，通过自主 GPGPU 硬件与全栈软件栈，支撑国家人工智能公共算力平台、互联网、运营商、智算中心等主体，提供自主可控的稳定算力基础。

“6”是 6 大核心行业赋能：聚焦金融、医疗健康、能源、教科研、交通、大文娱领域，针对各行业场景需求输出行业定制化算力方案，实现场景级效能提升。

“X”为泛行业拓展：基于标准化算力能力快速适配其他行业需求，扩大生态覆盖边界。

本次版本发布进一步验证了沐曦软硬件生态的“高性能、高兼容、高可用”核心特性，不仅实现了与国际旗舰产品的性能对标，更通过降低学习成本、全场景深度适配，为国产算力替代提供了成熟、可靠的解决方案，助力产业数字化转型与技术自主创新。