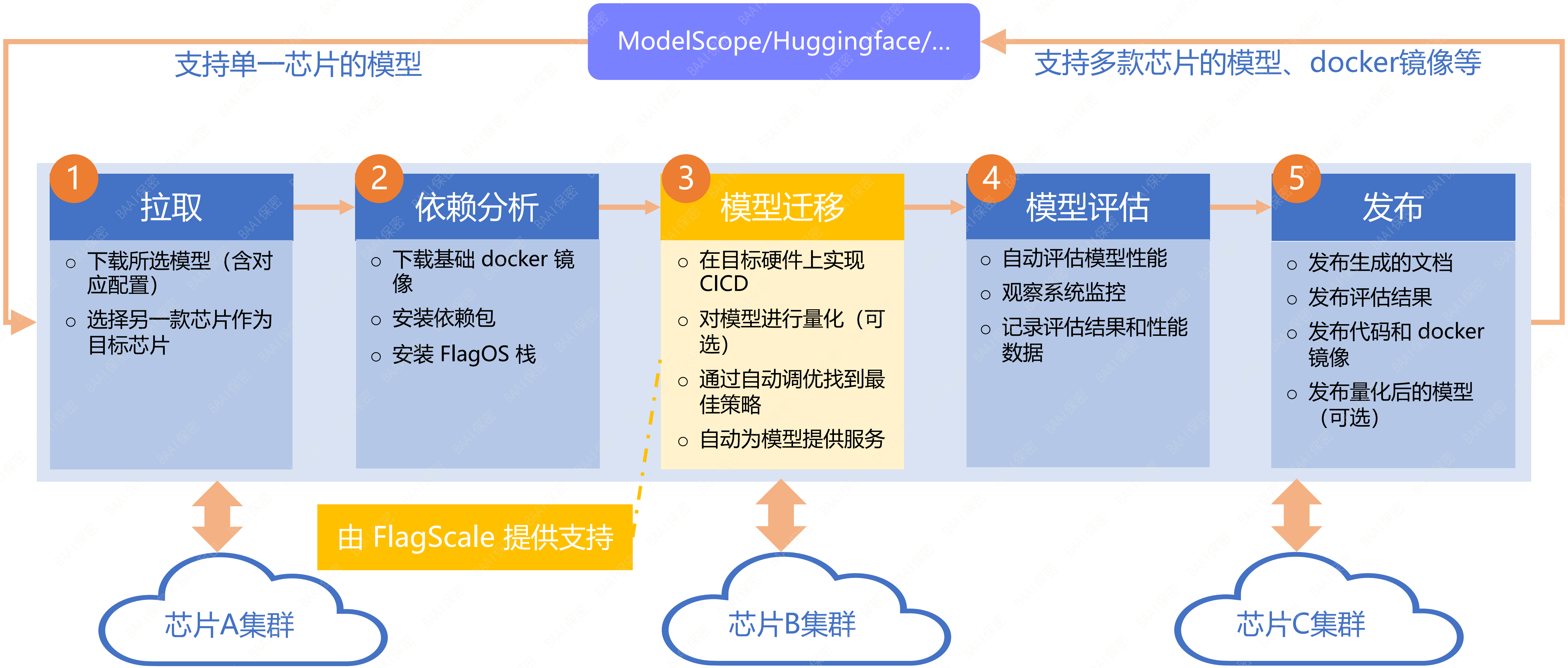
北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

基于 FlagRelease 实现模型到不同芯片上的推理部署

支持FlagRelease实现模型多芯片自动发版

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



ModelScope/Huggingface/...

支持单一芯片的模型

支持多款芯片的模型、docker镜像等

1 拉取
- 下载所选模型（含对应配置）
- 选择另一款芯片作为目标芯片

2 依赖分析
- 下载基础 docker 镜像
- 安装依赖包
- 安装 FlagOS 栈

3 模型迁移
- 在目标硬件上实现CICD
- 对模型进行量化（可选）
- 通过自动调优找到最佳策略
- 自动为模型提供服务

4 模型评估
- 自动评估模型性能
- 观察系统监控
- 记录评估结果和性能数据

5 发布
- 发布生成的文档
- 发布评估结果
- 发布代码和 docker 镜像
- 发布量化后的模型（可选）

由 FlagScale 提供支持

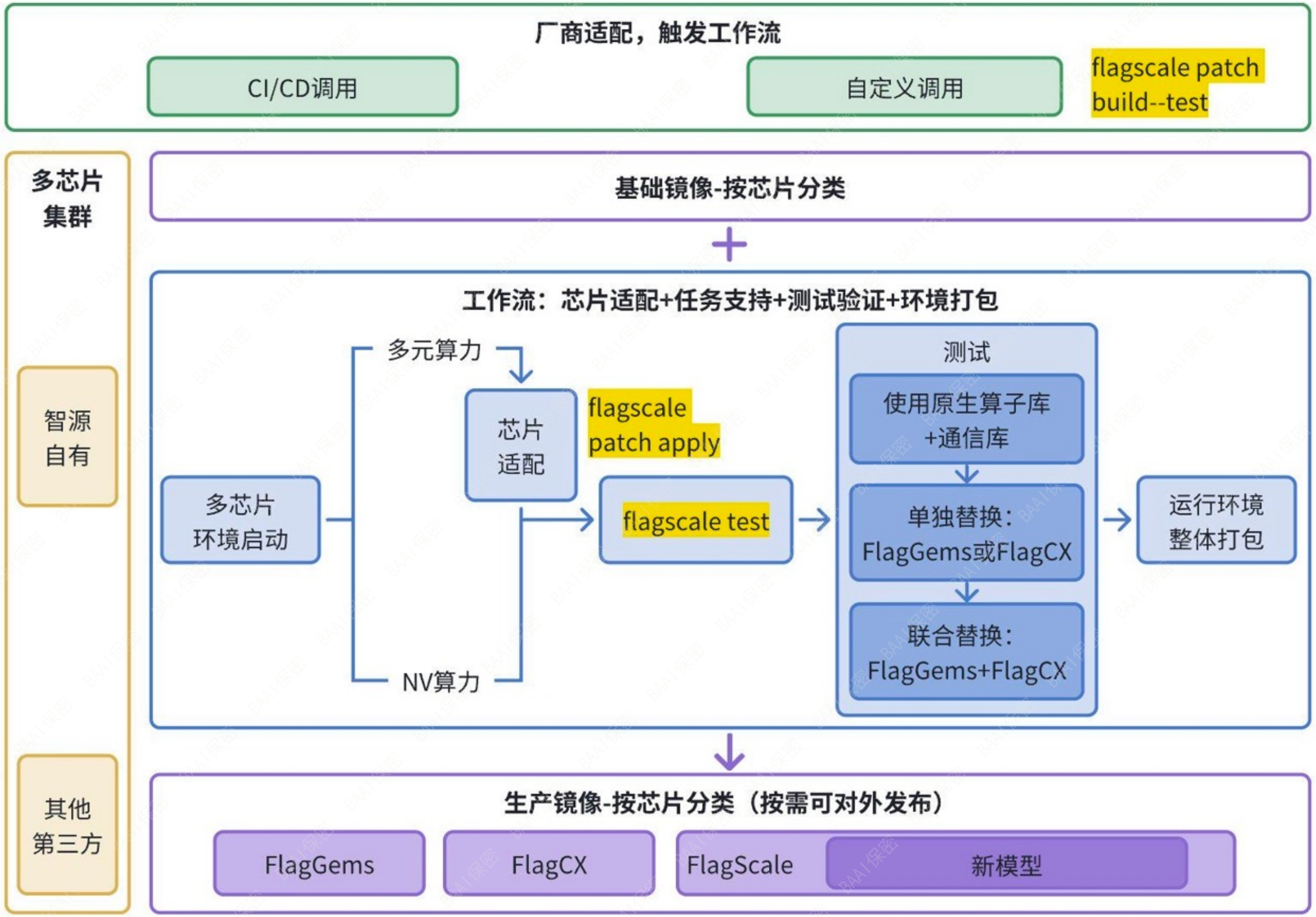
芯片A集群

芯片B集群

芯片C集群

自动化CI/CD持续升级完善，更好支持FlagRelease

- 提交PR时自动触发
- 本地可手动出发
- 集成高性能算子库FlagGems测试
- 未来将继承高性能统一通信库FlagCX测试
- 为FlagRelease新模型的发布提供支撑



厂商适配，触发工作流

CI/CD调用

自定义调用

flagscale patch
build--test

多芯片
集群

智源
自有

其他
第三方

基础镜像-按芯片分类

+

工作流：芯片适配+任务支持+测试验证+环境打包

多芯片
环境启动

多元算力

NV算力

芯片
适配

flagscale
patch apply

flagscale test

测试

使用原生算子库
+通信库

单独替换：
FlagGems或FlagCX

联合替换：
FlagGems+FlagCX

运行环境
整体打包

生产镜像-按芯片分类（按需可对外发布）

FlagGems

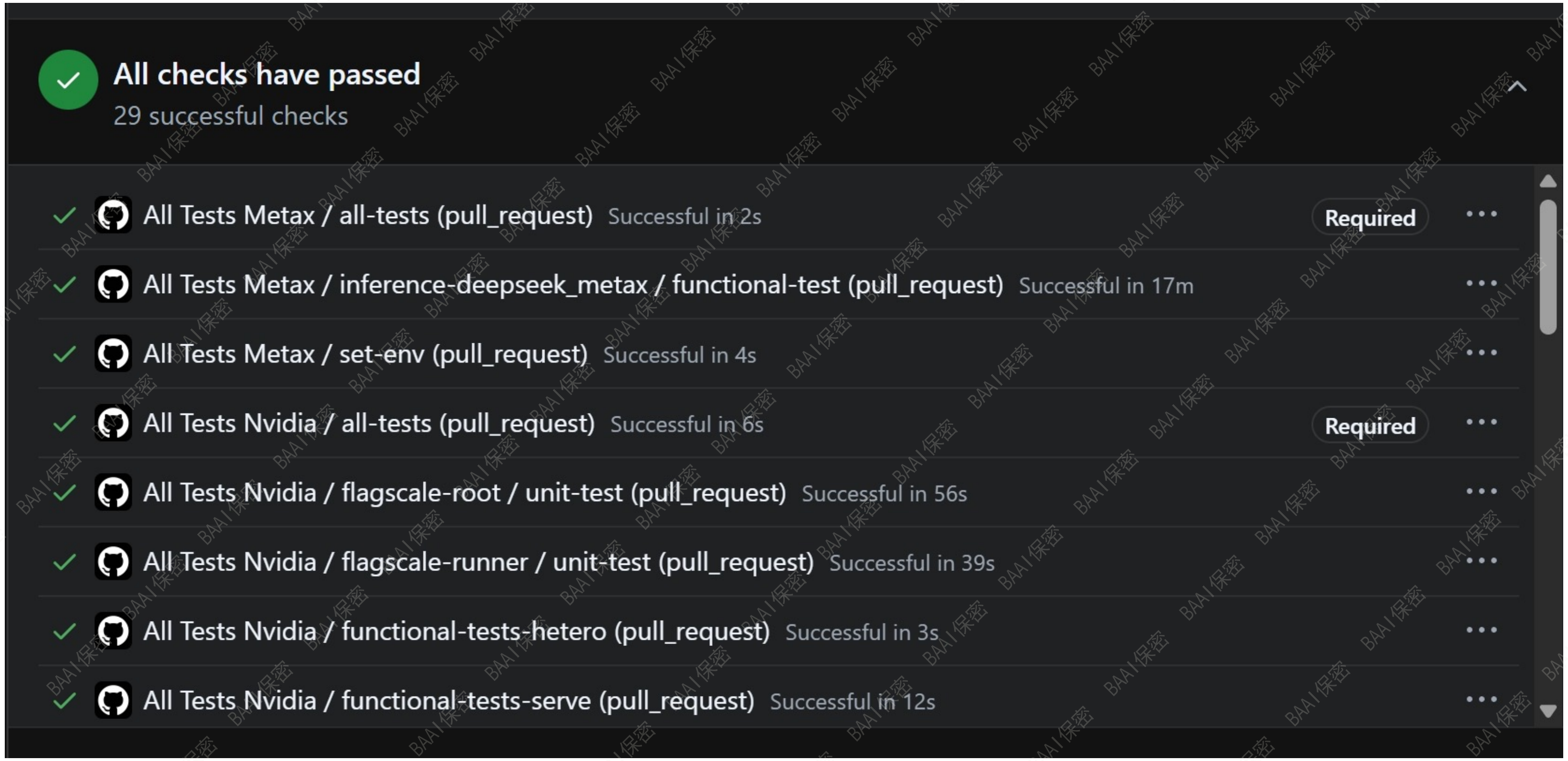
FlagCX

FlagScale

新模型

自动化CI/CD持续升级完善，更好支持FlagRelease

- 当前已上线沐曦CICD，更多国产芯片CI/CD未来持续上线中



All checks have passed
29 successful checks

✓ All Tests Metax / all-tests (pull_request) Successful in 2s Required

✓ All Tests Metax / inference-deepseek_metax / functional-test (pull_request) Successful in 17m

✓ All Tests Metax / set-env (pull_request) Successful in 4s

✓ All Tests Nvidia / all-tests (pull_request) Successful in 6s Required

✓ All Tests Nvidia / flagscale-root / unit-test (pull_request) Successful in 56s

✓ All Tests Nvidia / flagscale-runner / unit-test (pull_request) Successful in 39s

✓ All Tests Nvidia / functional-tests-hetero (pull_request) Successful in 3s

✓ All Tests Nvidia / functional-tests-serve (pull_request) Successful in 12s

实现模型到不同芯片上部署示例—英伟达芯片

- 下载镜像

```
docker pull flagrelease-registry.cn-beijing.cr.aliyuncs.com/flagrelease/flagrelease:nv_vllm085_gemscbbf39
```

- 启动容器

```
docker run --rm --init --detach \ --net=host --uts=host --ipc=host \ --security-opt=seccomp=unconfined \ --privileged=true \ --ulimit stack=67108864 \ --ulimit memlock=-1 \ --ulimit nofile=1048576:1048576 \ --shm-size=32G \ -v /nfs:/nfs \ --gpus all \ --name flagos \ flagrelease-registry.cn-beijing.cr.aliyuncs.com/flagrelease/flagrelease:nv_vllm085_gemscbbf39 \ sleep infinity docker exec -it flagos bash
```

- 安装modelscope并下载模型

```
pip install modelscope
modelscope download --model Qwen/Qwen3-4B --local_dir /nfs/Qwen3-4B
```

- 启动推理服务

```
flagscale serve qwen3
```

https://www.modelscope.cn/models/FlagRelease/Qwen3-4B-FlagOS-Nvidia

实现模型到不同芯片上部署示例—沐曦芯片

- 下载镜像

```
docker pull flagrelease-registry.cn-beijing.cr.aliyuncs.com/flagrelease/flagrelease:flagrelease_metax_qwen3
```

- 启动容器

```
docker run -it --device=/dev/dri --device=/dev/mxcd --group-add video --name test_flagrelease --device=/dev/mem --network=host --security-opt seccomp=unconfined --security-opt apparmor=unconfined --shm-size=100gb --ulimit memlock=-1 -v /usr/local/:/usr/local/ -v /nfs:/nfs flagrelease-registry.cn-beijing.cr.aliyuncs.com/flagrelease/flagrelease:flagrelease_metax_qwen3 /bin/bash
```

- 安装modelscope并下载模型

```
pip install modelscope
modelscope download --model Qwen/Qwen3-4B --local_dir /nfs/Qwen3-4B
```

- 启动推理服务

```
flagscale serve qwen3
```

https://www.modelscope.cn/models/FlagRelease/Qwen3-4B-FlagOS-Metax

谢谢聆听，
欢迎合作与建议。