AI Compiler in Alibaba

Wei Lin

weilin.lw@alibaba-inc.com

为什么做AI Compiler

- AI爆发期，模型创新加快，需要快速上线
- 多样的异构加速器件，特别是端侧，需要通过编译的技术来构建快速执行器
- 模型越来越大，从数据并行演化到模型并行，需要通过编译技术来构建分布式训练范式

使命（AI优化自动化）

WHAT → HOW

- 可用性
- 规模化
- 高性能
- 低成本

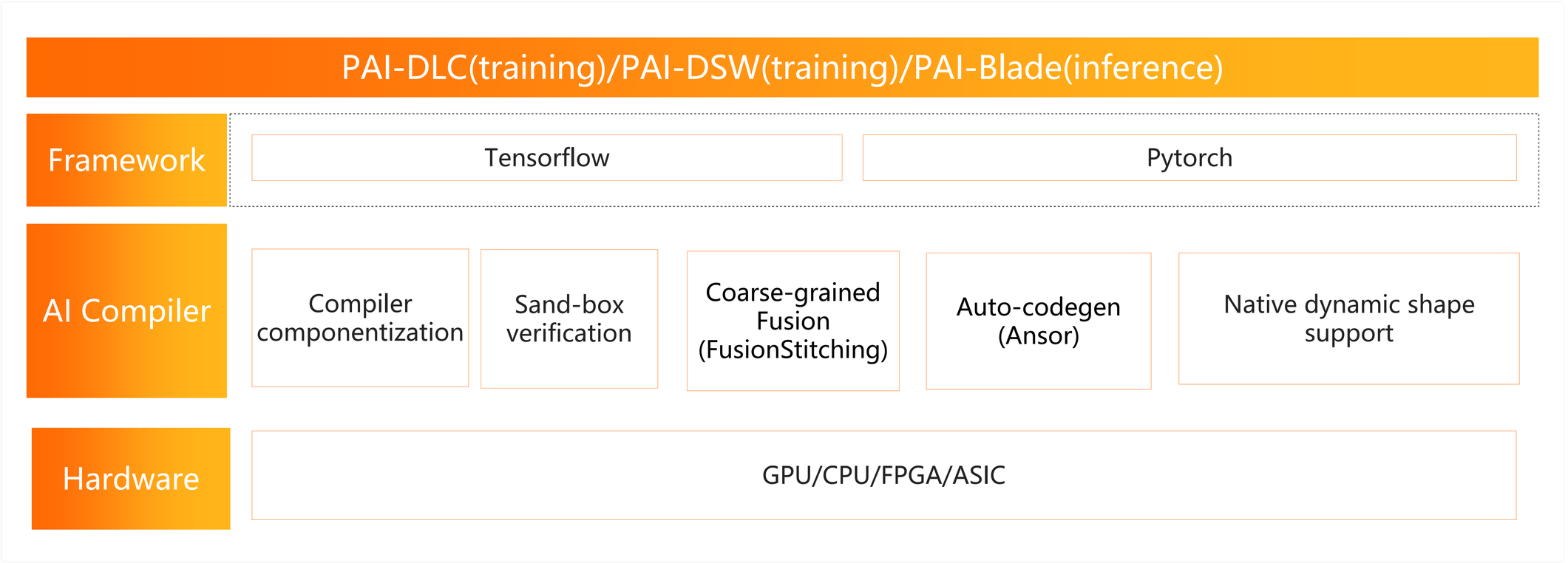
提纲

- 挑战
- AI Compiler的整体结构
- 编译优化流程
- 扩大kernel fusion范围的FusionStitching技术
- 密集运算kernel的TVM上自动代码生成技术（Ansor）
- Dynamic shape上编译优化技术
- AI Compiler的阿里云应用

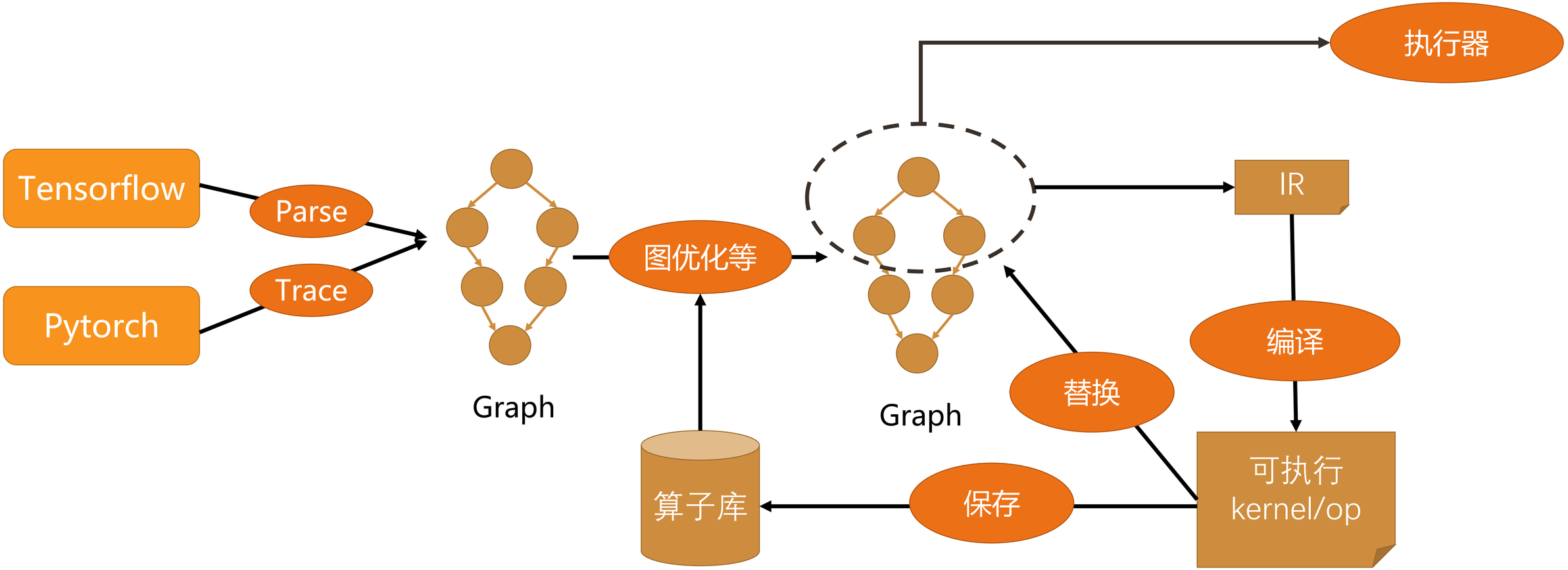
挑战

- 编译的正确性
- 编译的普适性，how to avoid worst case
- 易用性，如何能够对用户透明
- 优化目标多样性：推理，训练，多种设备，多种框架
- 编译本身的负载

AI Compiler@Alibaba整体结构

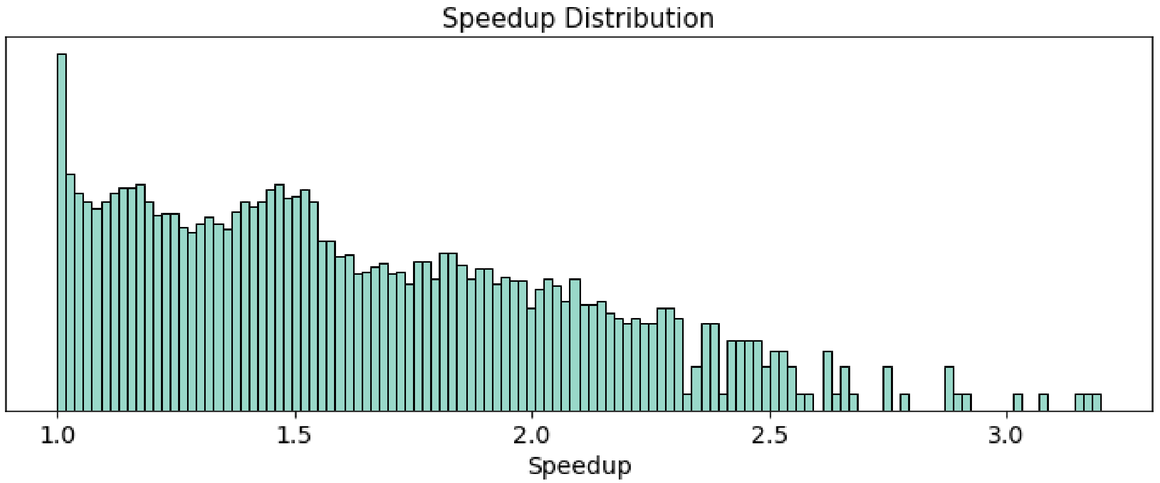


编译优化流程



一些结果

- 阿里巴巴内部集群
  - 覆盖数万任务
  - 30%任务得到 **> 1.1x** 加速



AICompiler end2end speed-up distribution

FusionStitching （扩大fusion范围）

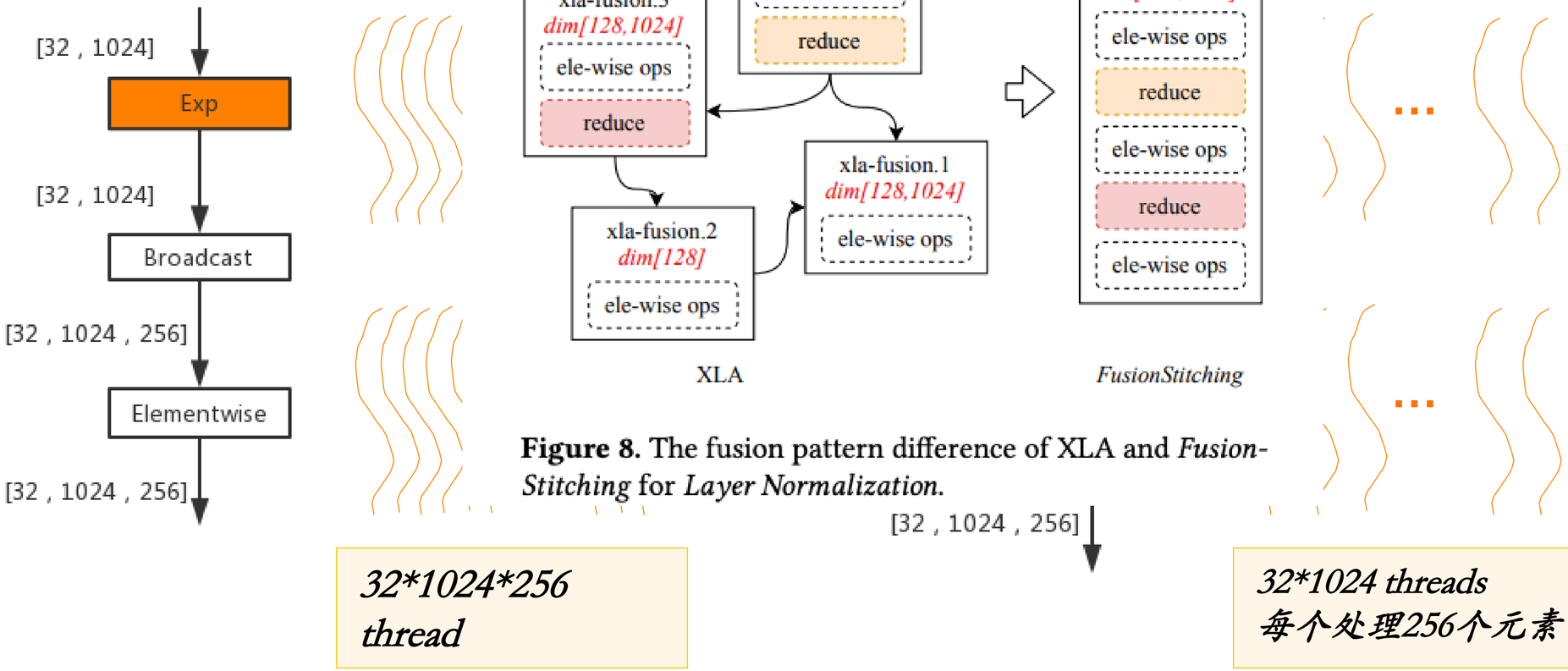
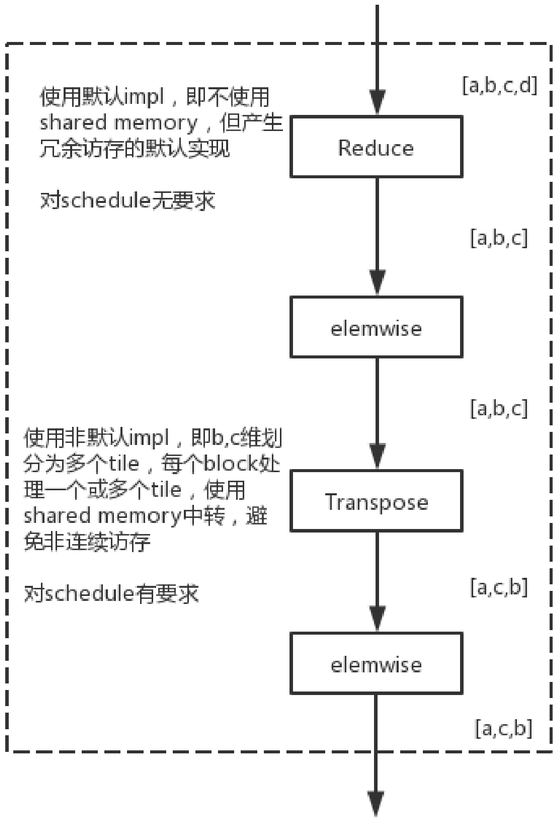
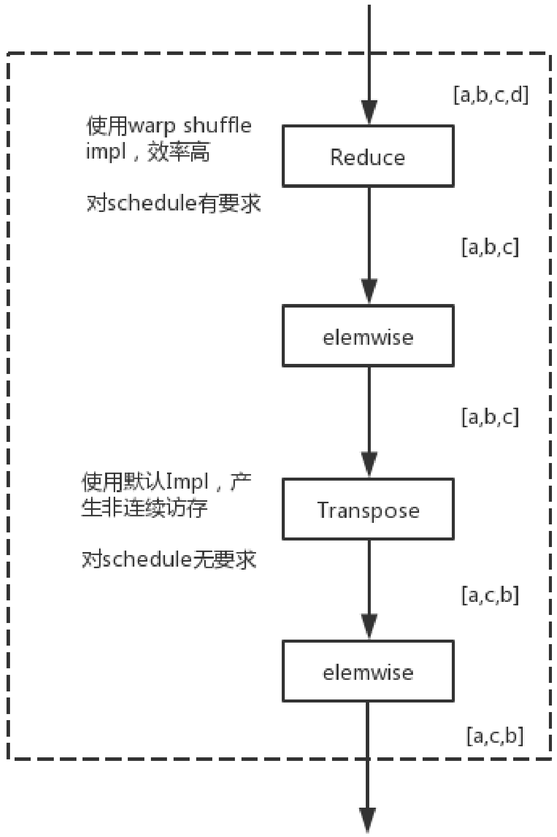
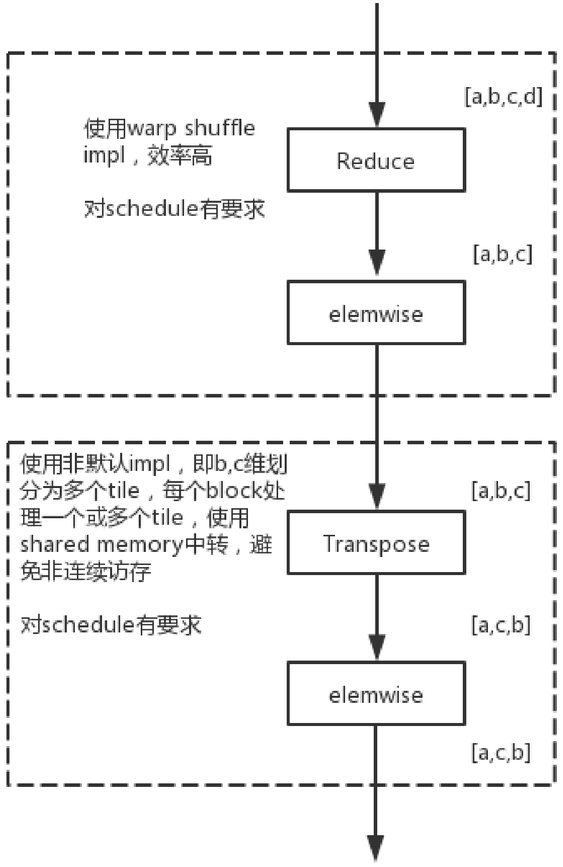


Figure 8. The fusion pattern difference of XLA and *Fusion-Stitching* for *Layer Normalization*.

多种选择







在不同具体size的情况下，上面三种CodeGen策略存在寻优空间，而Rule-based的策略则存在局限

FusionStitching整体设计

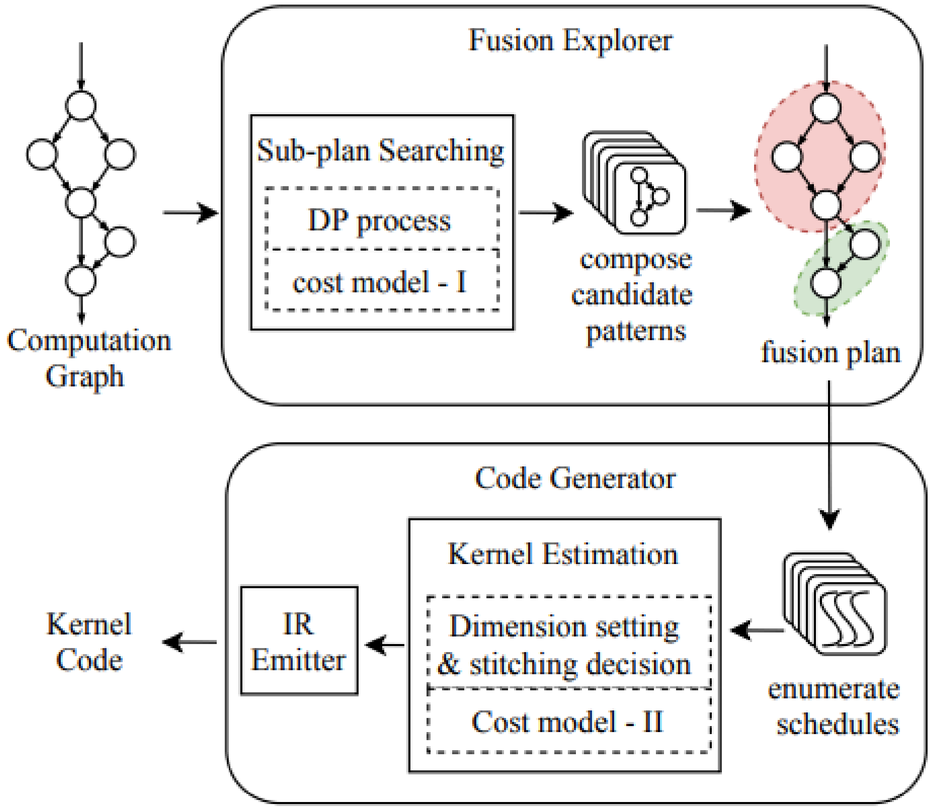


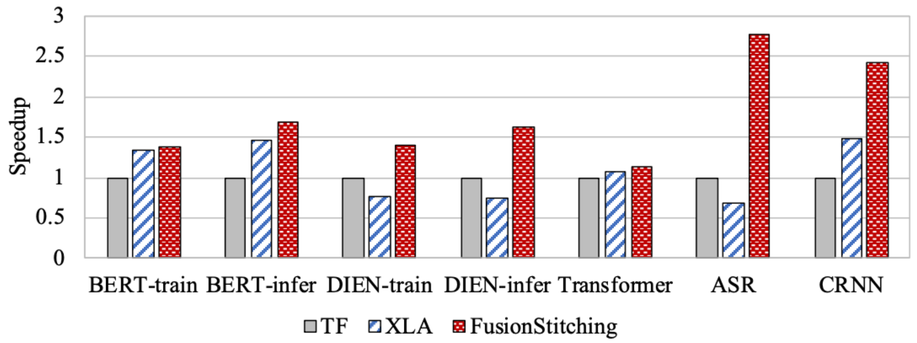
Figure 1. *FusionStitching* Overview.

- 基于代价评估来选择合理的计划
- 更为复杂的融合给codegen带来的挑战和机会

FusionStitching的结果

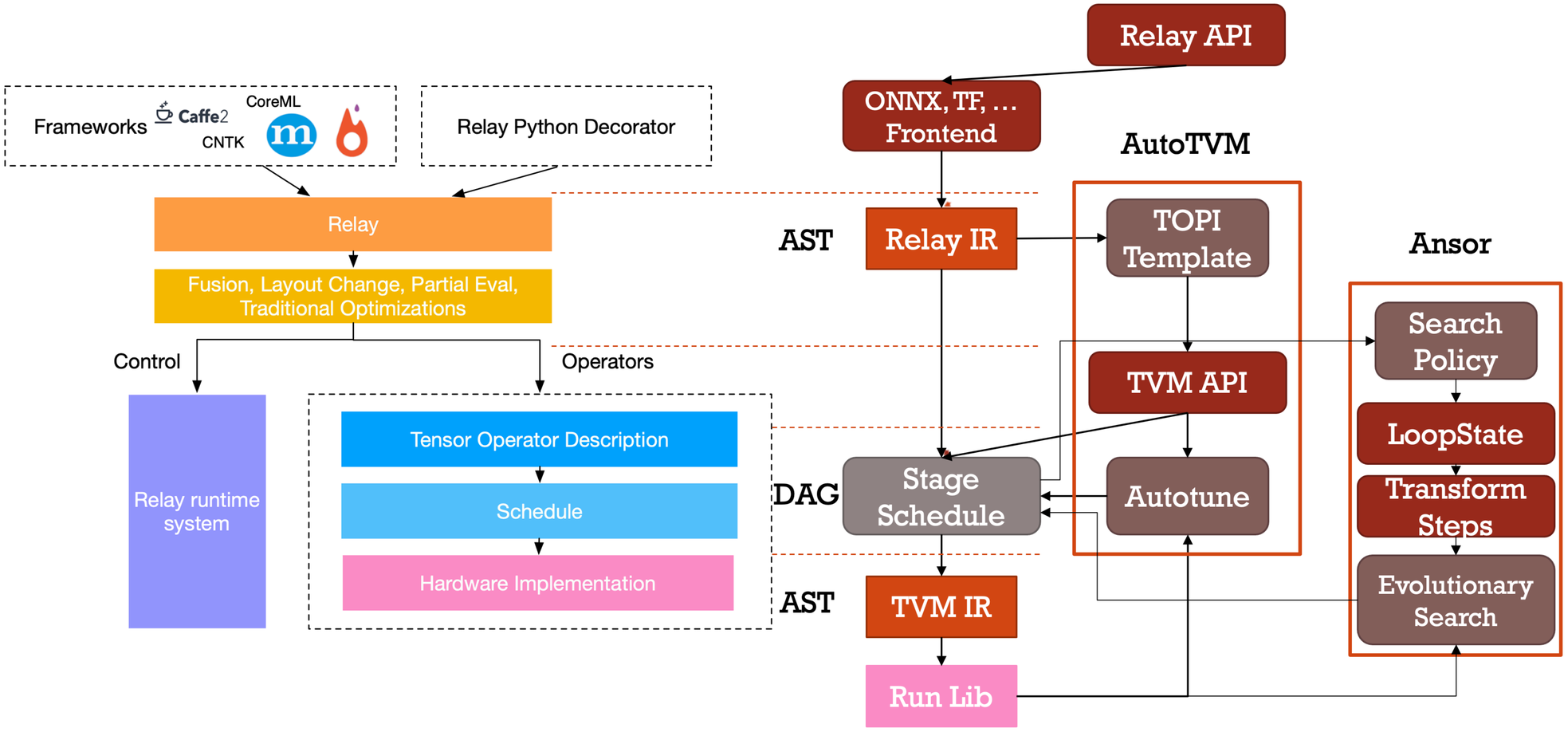
Table 1. Workloads for evaluation.

Model	Field	Mode	Batch Size
BERT	NLP	Both	32
DIEN	Recommendation	Both	256
Transformer	NLP	Training	4096
ASR	Speech Recognition	Inference	8
CRNN	OCR	Inference	8

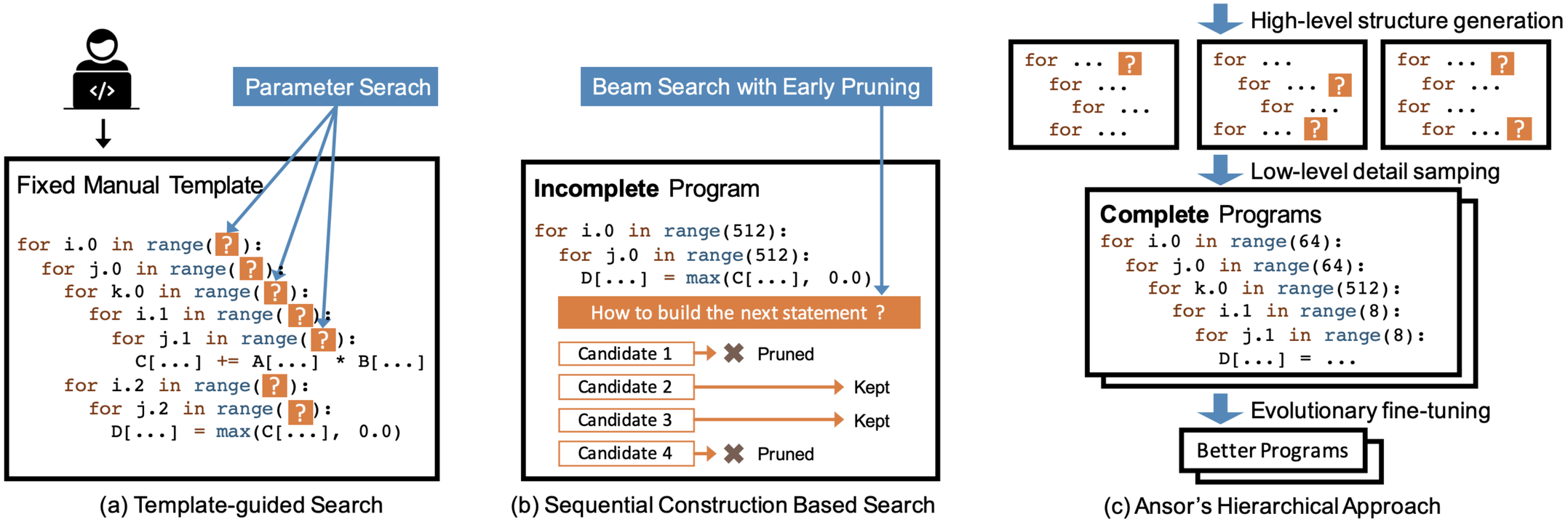


Model	Tech	T/#	CPU	Math	Mem	Cpy	E2E
BERT -train	TF	T	1.55	41.69	28.45	0.15	71.84
		#	-	98	561	102	761
	XLA	T	2.3	41.89	9.56	0.15	53.9
		#	-	98	200	97	395
	FS	T	2.8	42.11	7.02	0.03	51.96
		#	-	98	98	20	216
BERT -infer	TF	T	3.24	1.65	0.83	0.14	5.86
		#	-	70	365	106	541
	XLA	T	0.78	2.50	0.60	0.13	4.02
		#	-	98	277	94	469
	FS	T	0.59	2.46	0.40	0.04	3.49
		#	-	98	77	30	205
DIEN -train	TF	T	90.13	7.77	32.54	7.12	137.56
		#	-	1218	10406	1391	13015
	XLA	T	124.04	9.06	37.50	6.56	177.16
		#	-	1215	6842	1996	10053
	FS	T	48.42	7.91	35.84	5.55	97.72
		#	-	1215	2109	1395	4719
DIEN -infer	TF	T	27.36	2.58	7.55	1.99	39.48
		#	-	406	3680	225	4311
	XLA	T	44.21	2.24	6.12	0.94	53.51
		#	-	405	2585	627	3617
	FS	T	17.54	2.45	3.51	0.7	24.20
		#	-	405	815	422	1642
Trans former	TF	T	7.99	109.13	69.53	1.63	188.28
		#	-	309	3860	724	4893
	XLA	T	23.63	107.48	40.20	4.24	175.55
		#	-	309	1923	2065	4297
	FS	T	8.21	110.70	42.57	3.05	164.53
		#	-	243	1384	1765	3392
ASR	TF	T	21.02	2.14	3.63	0.78	27.57
		#	-	116	1292	534	1942
	XLA	T	17.51	1.66	1.81	19.76	40.74
		#	-	84	496	376	956
	FS	T	6.00	1.92	1.63	0.36	9.92
		#	-	108	212	199	519
CRNN	TF	T	23.31	6.05	6.14	1.60	37.10
		#	-	256	3674	890	4820
	XLA	T	12.17	0.30	11.37	1.04	24.88
		#	-	7	993	406	1406
	FS	T	6.35	0.31	7.69	1.01	15.36
		#	-	8	311	388	707

密集算子的优化技术（Ansor）



Ansor

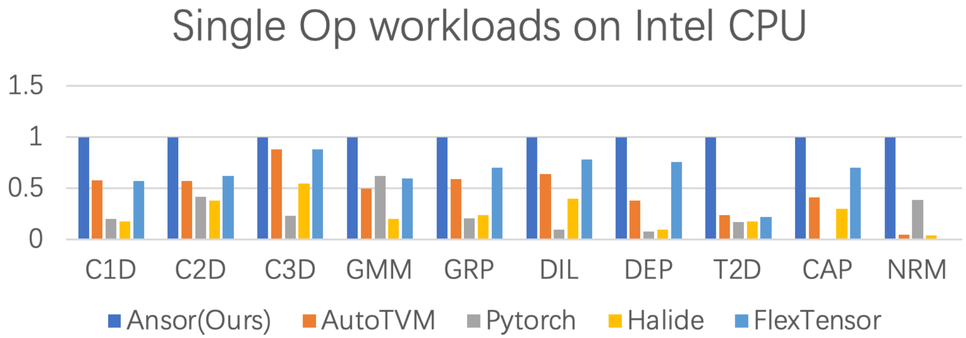


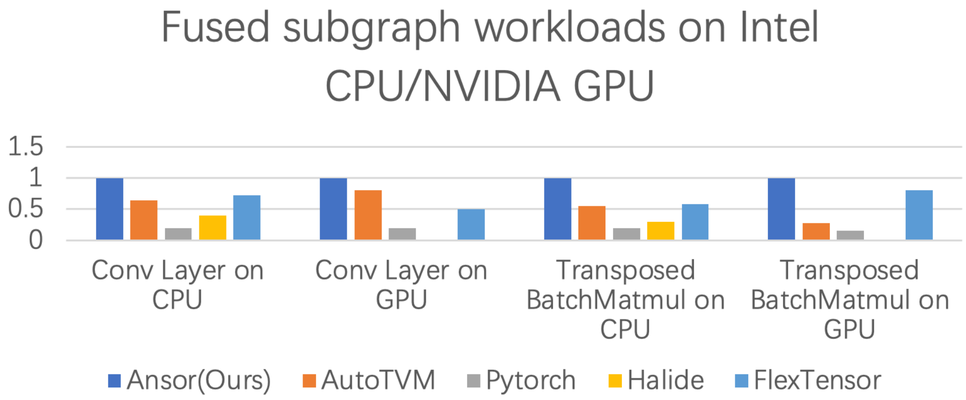
(a) Template-guided Search (b) Sequential Construction Based Search (c) Ansor's Hierarchical Approach

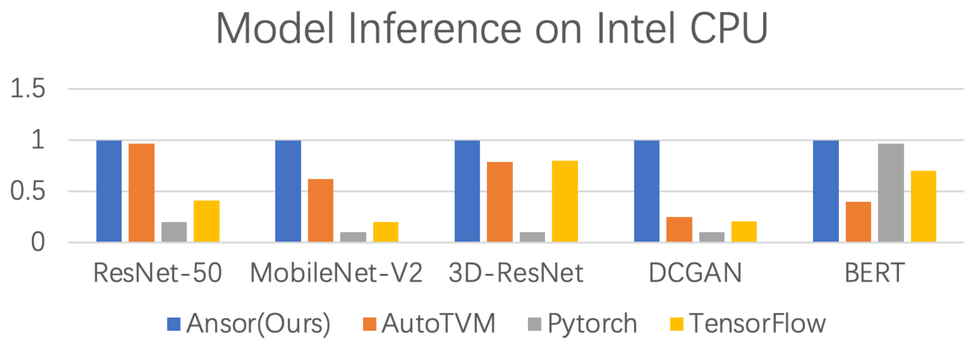
- 自动构建搜索方式，避免人工书写搜索模板
- 利用采样在完整的程序上进行性能验证，提高搜索效率和质量
- [RFC] Ansor: An Auto-scheduler for TVM (AutoTVM v2.0)

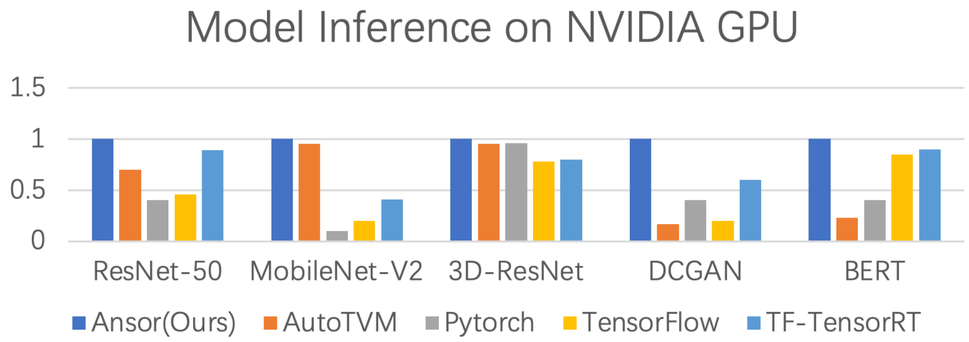
Ansor结果

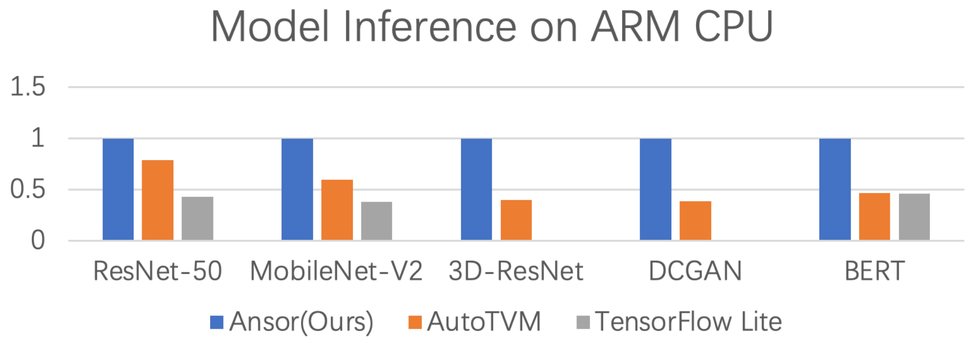
- 被OSDI 2020接收
  - https://arxiv.org/abs/2006.06762





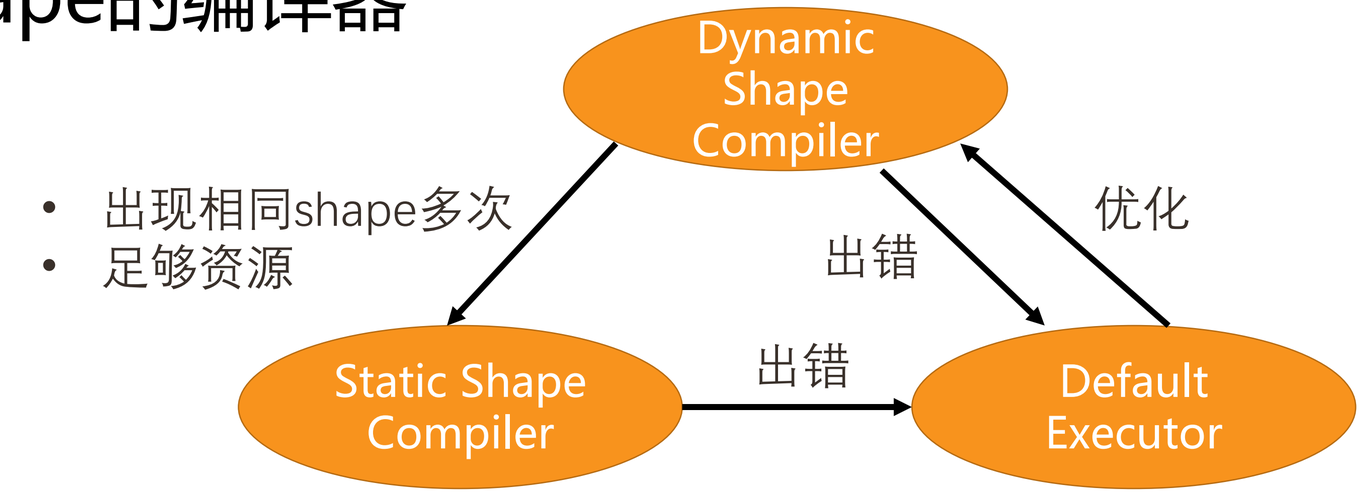






Dynamic Shape的支持

- 现有AI编译框架更多在固定shape条件下进行优化
  - Shape的变化会造成优化中的计算和IO之间关系剧烈变化
  - 某些AI模型具有Dynamic Shape的特性，在固定shape的优化方式下性能影响比较大，有较大的优化空间
    - Seq2Seq模型
    - Sparse模型中的unique op
    - 模型训练中的不同batch size等等
- MLIR基础上支持Dynamic Shape的编译器
  - MLIR ODM
  - 和现有其他部分有机结合



PAI-Blade 通用推理优化框架

阿里云 | 奥运会全球指定云服务商

What’ s ne

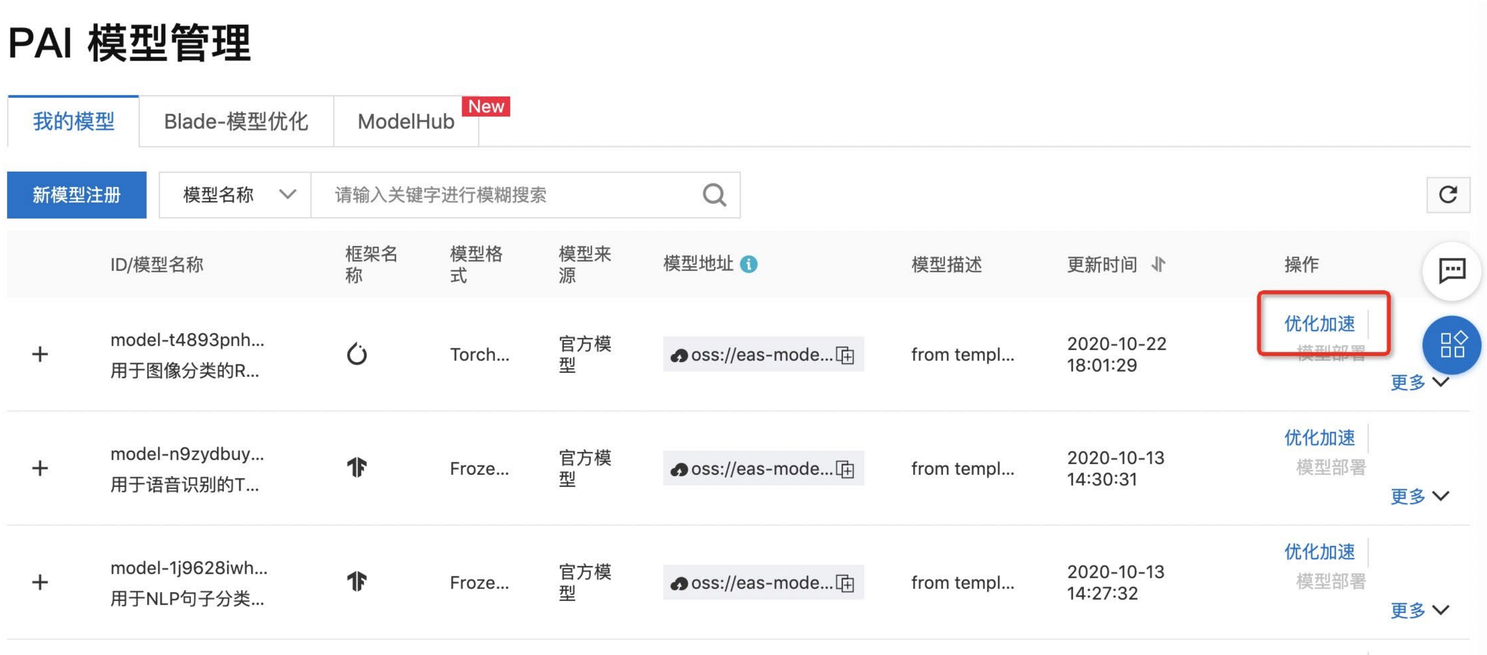
PyTorch支持

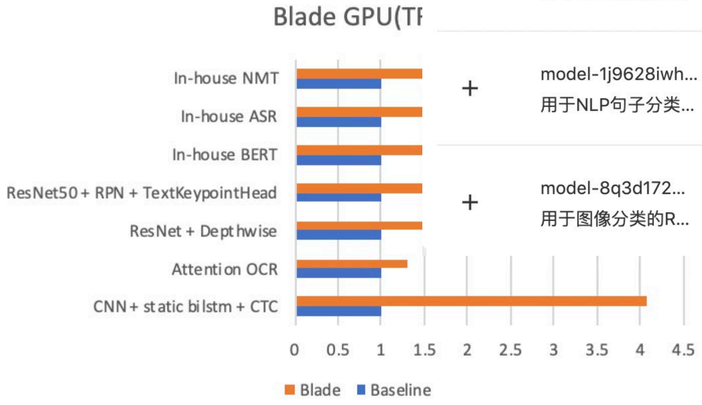
端侧MNN

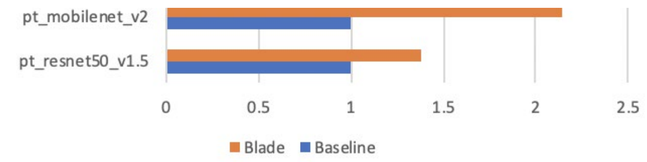
CPU

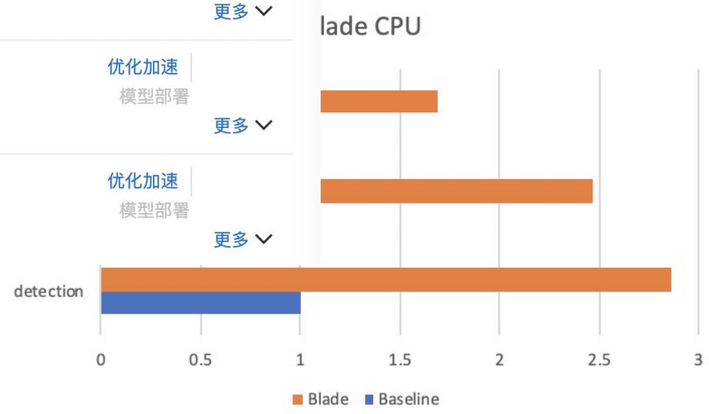
大量用户体验完善

AI Compiler集成









联合优化

(MNN共建)

联合优化(MNN共

ession API

充分利用集成了Intel® DL Boost加速指令集(INT8/BF16)的Intel Xeon CPU

Thanks