深度学习发展趋势和百度PaddlePaddle

黄硕

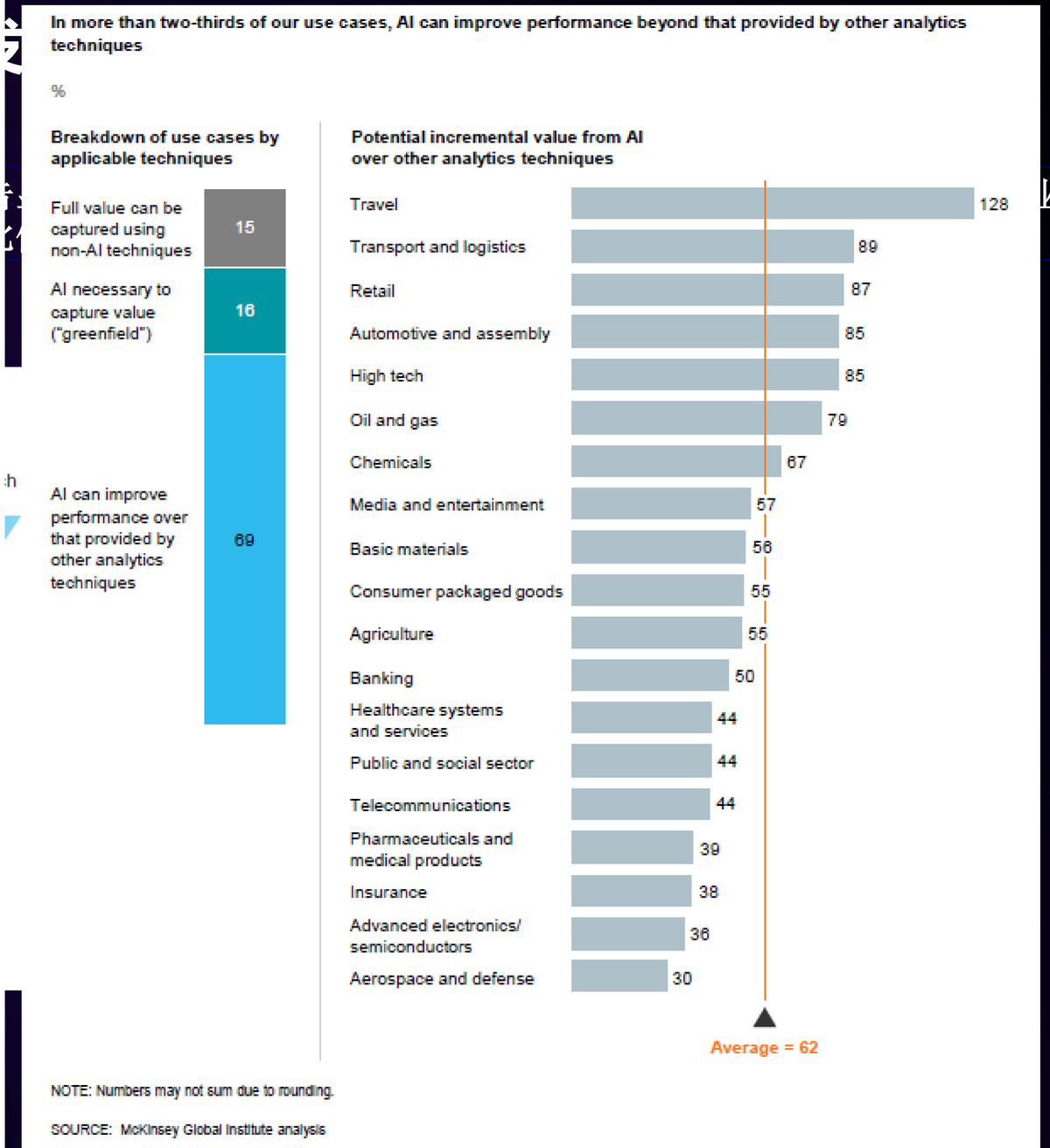
百度深度学习技术平台部

2018.11.16

深度学习技术与框架发展新趋势

深度学习技术经过多年发展，已经开始大规模应用

✓麦肯锡研究表明，深度学习可帮助2/3的企业应用提升经营表现，各行业产值平均增长62%

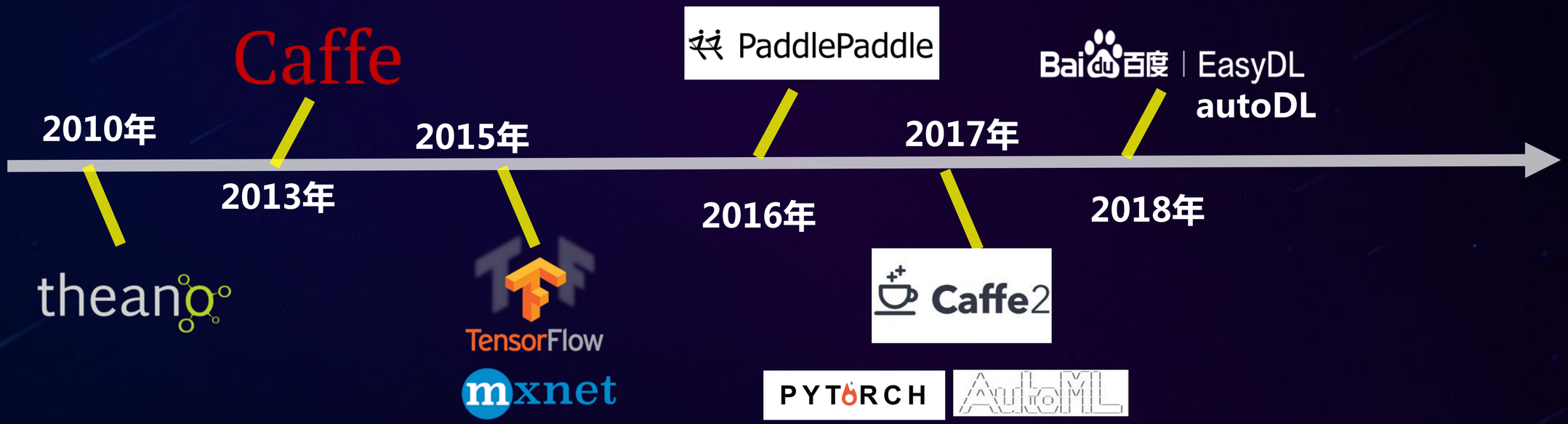


NOTE: Numbers may not sum due to rounding.

SOURCE: McKinsey Global Institute analysis

深度学习框架演变历程

- 学术界起源，逐渐演化为巨头竞争：Theano（2010，U Montreal）、Caffe（2013，Berkeley）、Tensorflow（2015，Google）、 MXNET（2015，AWS支持）、 PaddlePaddle（2016，百度）、PyTorch（2017，Facebook）、 Caffe2（2017，Facebook）



趋势：深度学习框架门槛持续降低，Everyone Can AI

低

门槛高

理论知识学习成本高，实现成本也很高，从业人员少

理论知识

PYTORCH

PaddlePaddle

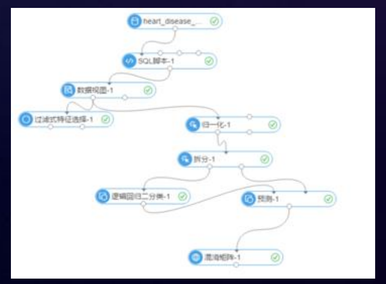
Caffe

theano

TensorFlow

深度学习框架陆续开源，极大降低了入门成本，但开发者仍然需要特定的知识背景的硬件资源支持

框架开源



AutoDL

AutoML

网络结构自动化设计工具
开发者无需经过特殊训练

AutoML/AutoDL

AutoDL网络结构

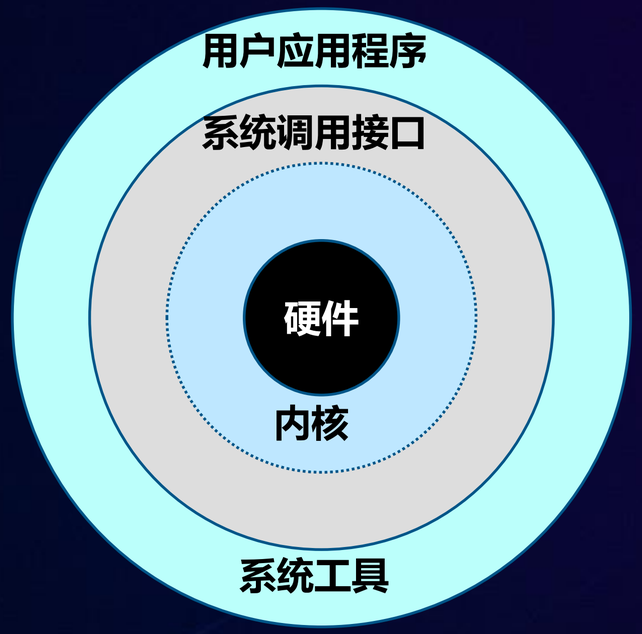
Baidu百度 | EasyDL

零算法基础的快速应用平台，无需任何专业背景即可轻松定制模型，与云端结合，使用户无任何后顾之忧

EasyDL

Baidu 大脑 | IT大咖说 知识共享平台

趋势：以深度学习框架为核心的“操作系统生态”



AI操作系统

AI应用程序
(如语音、NLP、视觉、推荐等)

各类深度学习业务模型
（如Resnet、Transformer等）

使用API构建业务模型，形成粘性

系统安装工具 | 辅助开发工具（VisualDL等） | 快速部署工具

数据

组网API | 训练API

核心训练框架

基础训练功能实现与执行、资源管理 | 并行训练技术

异构硬件支持和基于硬件特性性能优化

模型

预测调用API

核心预测引擎

预测功能实现 | Edge模型压缩

异构硬件支持和基于硬件特性性能优化

适配已有硬件优化性能，或研发专用处理器

CPU/GPU（集群） | FPGA | AI芯片 | ARM | 各类终端芯片

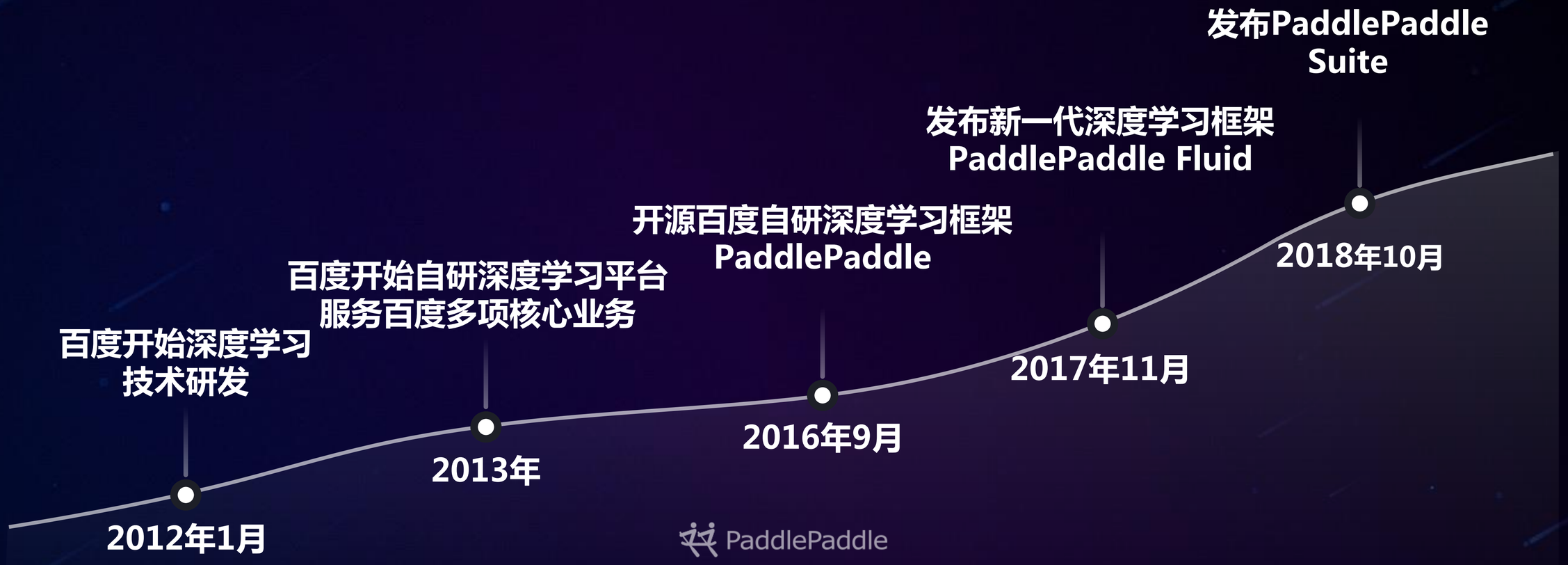
操作系统

介于用户程序和硬件资源之间
通过内核为用户程序提供资源调度
通过接口为用户程序提供开发便利

PaddlePaddle

最适合企业和开发者的深度学习框架

PaddlePaddle的发展历程



http://www.paddlepaddle.org

PaddlePaddle Suite

技术全面领先的深度学习全功能套件

服务平台

- **EasyDL**
  零基础定制化训练和服务平台
- **AI Studio**
  一站式开发平台
- **AutoDL**
  网络结构自动化设计

核心框架

- **PaddleRec**
  智能推荐
- **PaddleCV**
  智能视觉
- **PaddleNLP**
  智能文本处理
- 训练框架
- Paddle Serving
- Paddle Mobile

模块及组件

- **VisualDL**
  训练可视化工具
- **PARL**
  深度强化学习框架
- **EDL**
  弹性深度学习计算

技术领先的核心框架

官方支持最全面的业务模型

超大规模深度学习并行技术

全面领先的高速推理引擎

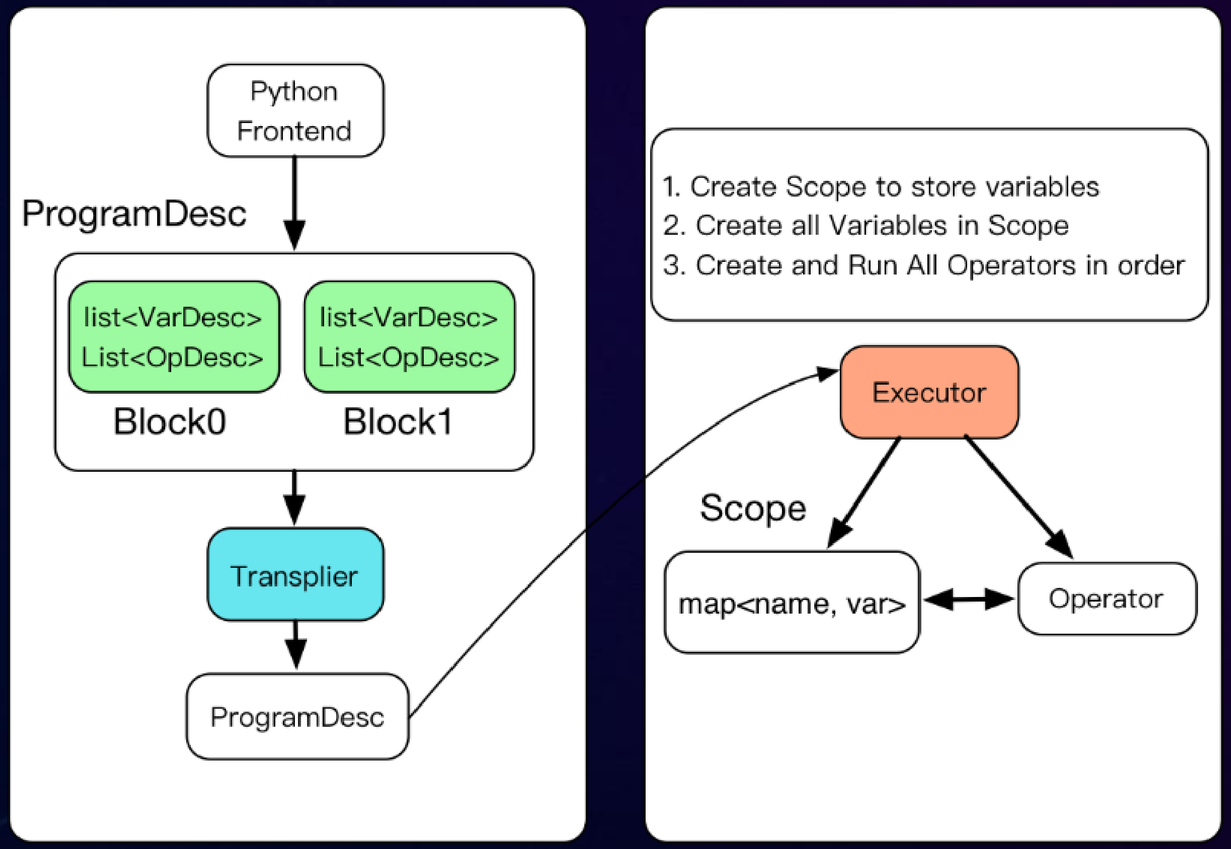
框架整体架构


组网模块
Python API
模型表达与优化
模型Program
统一中间表达
统一中间表达优化库
服务器预测
服务器预测API
服务器预测RunTime
训练模块
训练RunTime
CPU/GPU集群
移动端预测
移动端预测API
移动端预测RunTime

框架整体架构

Sequence of Layer → Graph of Operators → Program

CompileTime RunTime



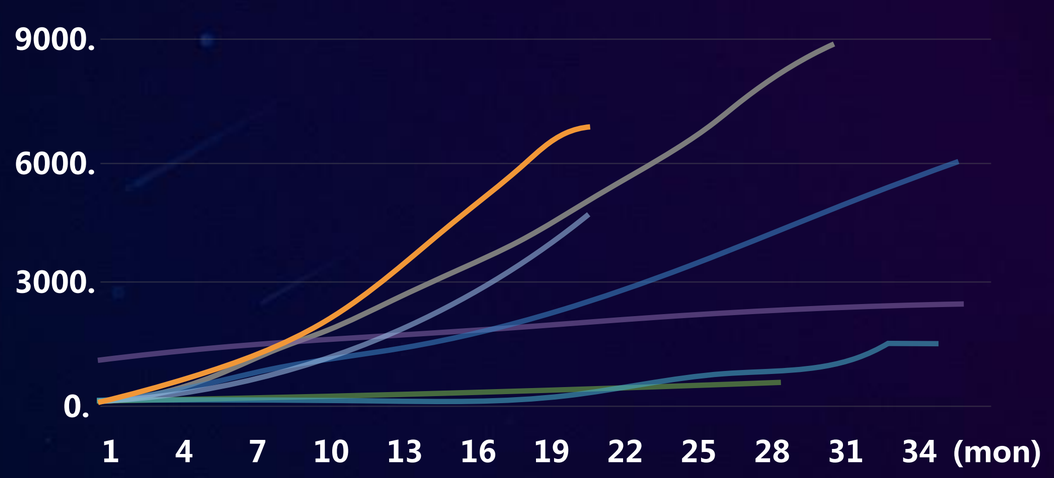

```
limit = fluid.layers.fill_constant_batch_size_like(
    input=label, dtype='int64', shape=[1], value=5.0)
cond = fluid.layers.less_than(x=label, y=limit)

ie = fluid.layers.IfElse(cond)
with ie.true_block(): # block 1
    true_image = ie.input(image)
    hidden = fluid.layers.fc(input=true_image, size=100, act='tanh')
    prob = fluid.layers.fc(input=hidden, size=10, act='softmax')
    ie.output(prob)
```

Tensor, Operator, Program (Blocks)
Control Flow
Transpiler, Executor

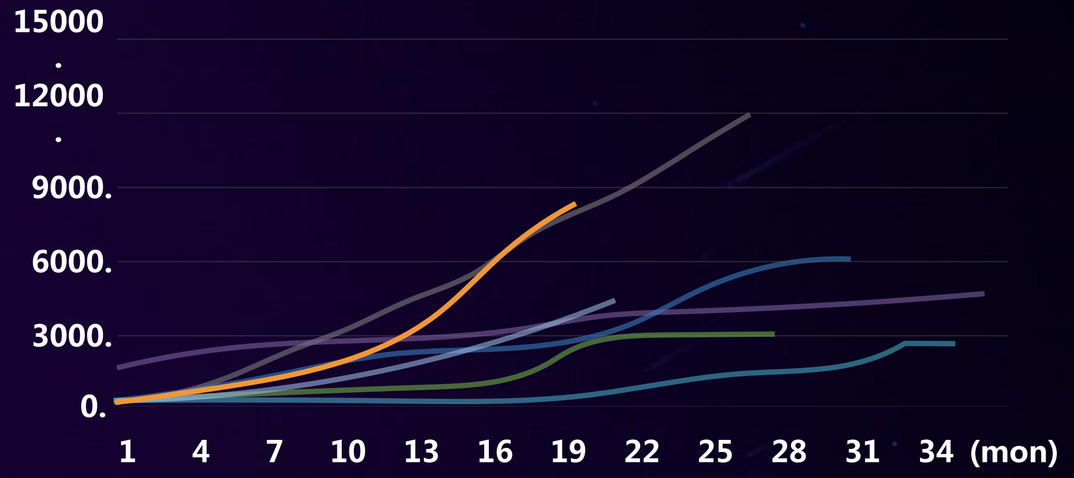
PaddlePaddle 社区快速发展

在GitHub开源社区中的活跃度较高，甚至高于其他框架同期水平

Pull requests 同期对比



Issues 同期对比



PaddlePaddle Tensorflow MxNet Caffe Caffe2 CNTK Pytorch

支持广泛的产业应用场景

官方模型支持和维护—最完善的模型集合

使用场景	PaddleRec – 智能推荐	PaddleCV – 智能视觉	PaddleNLP – 智能文本处理
	信息流　智能营销	视频分析　医学影像　智能驾驶　工业质检	舆情分析　搜索引擎　机器翻译　智能对话
模型集合	提供多种推荐场景下的召回、排序经典算法	机器视觉应用场景全覆盖	全方位满足主流 NLP 任务
	DeepCTR　GRU4Rec　文本标签模型	图像分类　目标检测　人脸检测	中文词法分析　语义匹配　阅读理解
	序列语义召回　Multi-view Simnet	OCR　图像语义分割　生成对抗网络	机器翻译　中文情感分析
		度量学习　视频分类	
应用案例	百度Feed　好看视频	百度地图　百度OCR　百度Feed　百度糯米	百度搜索　百度翻译

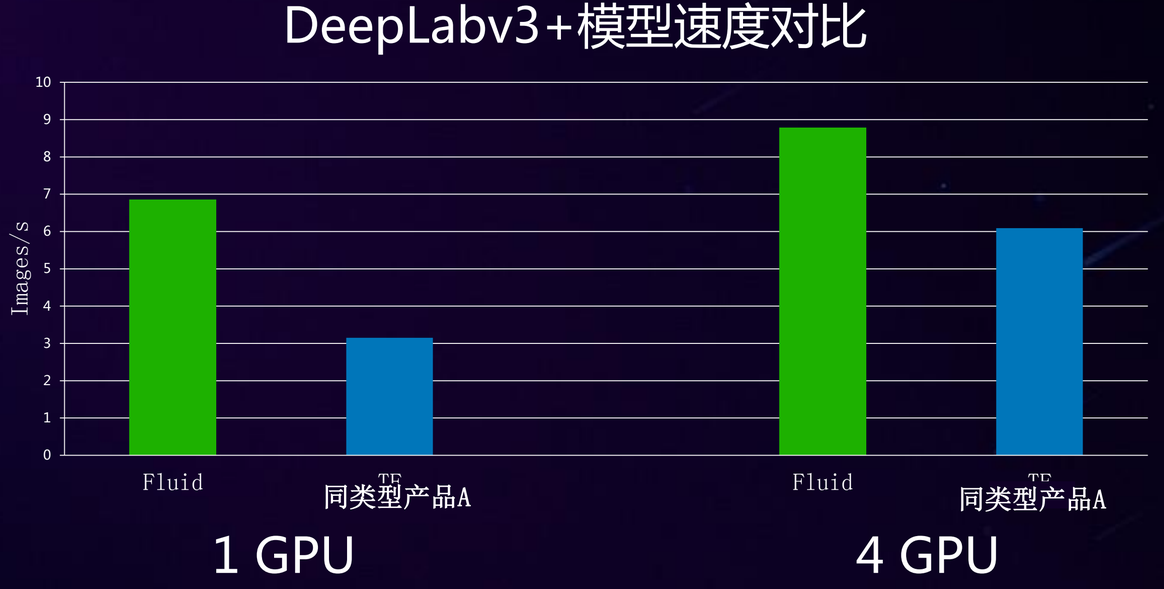
CV模型

分类	ResNet, SE-ResNeXt, GoogleNet, VGG
检测	SSD, Faster-RCNN, Mask-RCNN, Yolo v3，Fast RCNN
分割	DeepLab v3+, ICNet
关键点	OpenPose
视频分类	TSN
OCR识别	CRNN CTC, seq2seq+attention
OCR检测	East检测
人脸识别	Metric Learning、大规模softmax方法
图像生成	CycleGAN, CGAN, DCGAN

语义分割模型DeepLabv3+

Encode-Decoder架构
Xception + Depthwise separable convolution提高精度和计算效率
在多项数据集上取得state-of-art的mIoU
高度优化了 + Depthwise separable convolution

训练速度	GPU个数	Images/S
Paddle Fluid	1	6.856
同类型产品A	1	3.15
Paddle Fluid	4	8.787
同类型产品A	4	6.0691



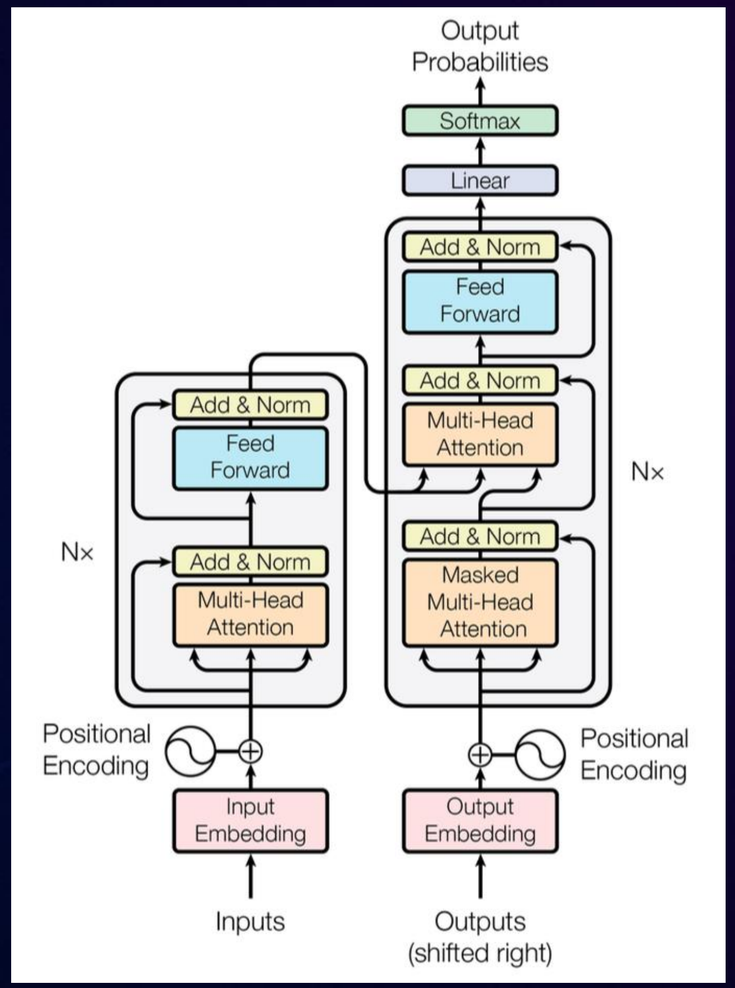
图像模型线上应用

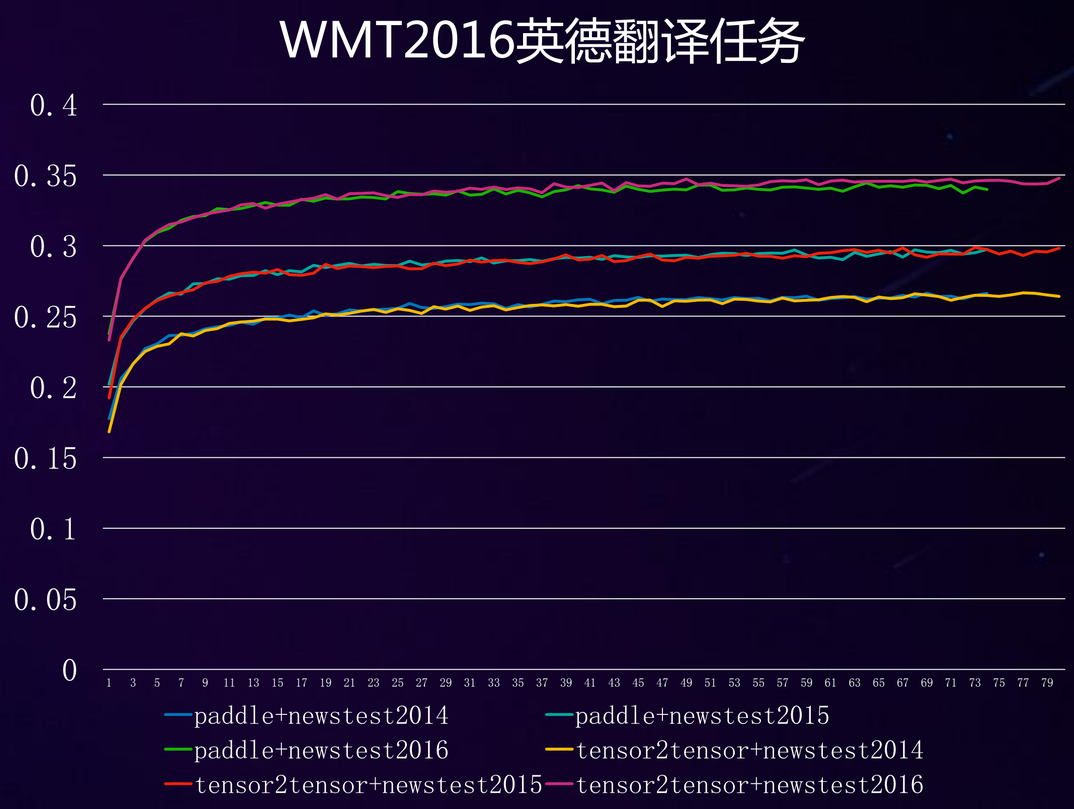
服务名称	模型	业务线	收益
OCR英文识别	CRNN-CTC模型	多模、翻译等	精度提升1%，QPS提升7%
OCR车牌识别	CRNN-CTC模型	AIP云平台，百度地图	统一CPU、GPU预测库版本，简化了服务对不同预测库的依赖。
148类通用物体检测	MobileNet-SSD	河图开放服务	效果和原先打平
23类手势识别	VGG-SSD	河图开放服务	手势识别使用Fluid首次在河图上线。
车辆检测	VGG-SSD	多模	整体速度提升约20%，相同资源下QPS提升约100%，多模端日PV 920W，AIPE日活80+
车辆识别	ResNeXt	多模	和上面服务同时上线，收益同上。
车辆REID	ResNet	通过AIPE给智慧城市调用	初次上线。
动植物相似识别	GoogleNet, ResNet 101	多模	效果和原先打平
主体目标检测	GoogleNet	河图开放服务	性能由40 qps（K1200）提升至300 qps （P4）

NLP模型

词法分析	分词、词性标注、专名识别
语义匹配	MM_DNN, DAM, PyramidDNN
分类	情感分析、黄反识别
语言模型	GRU, LSTM
问题生成	Seq2seq + Point Generator
机器翻译	Transformer
阅读理解	BiDAF

Transformer模型





机器翻译任务上，翻译质量跟竞品持平，代码量远少于同类型产品

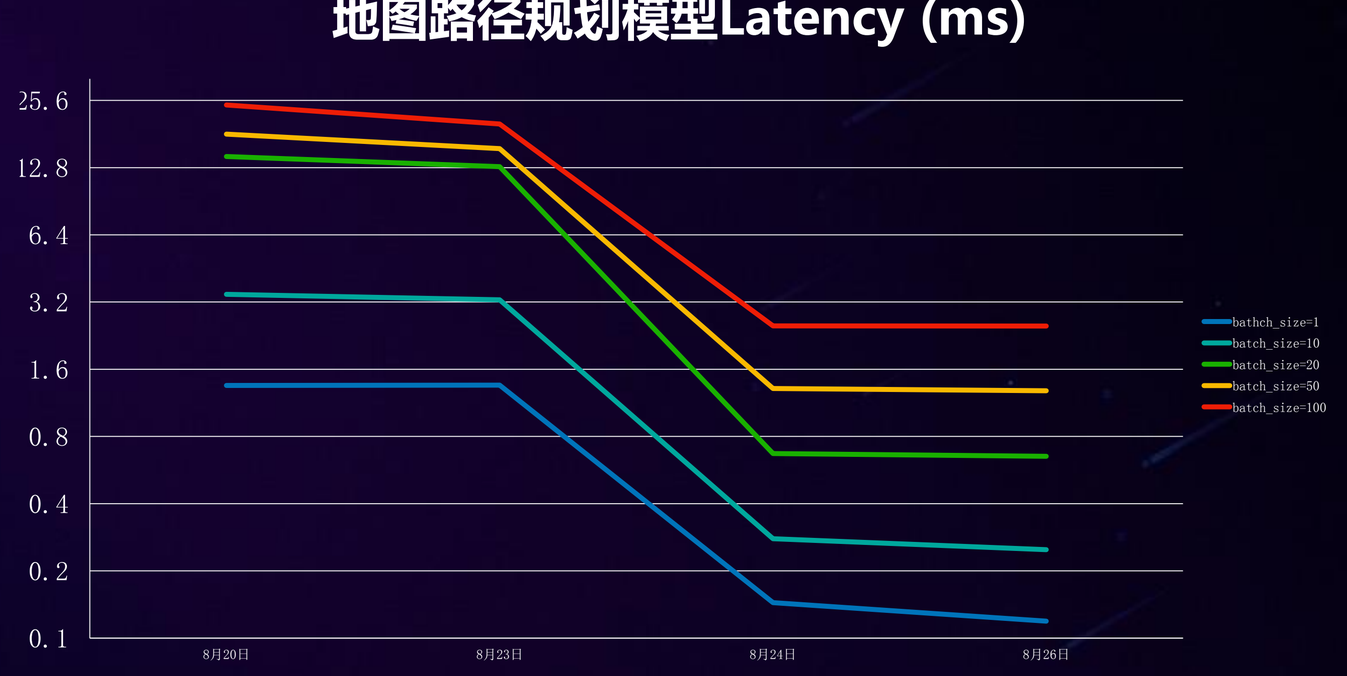
RNN地图路径规划

2个RD一周时间，提高地图路径规划模型在线预测性能10~30倍

基于可视化工具的模型结构和性能分析(Profile)

基于统一图中间表达(IR)和扩展机制(Pass)，注册定制化的图优化策略(Plug-in)

基于operator的扩展机制，注册fused算子(Plug-in)



Paddle Mobile多平台预测部署

多平台预测部署

Paddle 服务器预测

预测服务部署

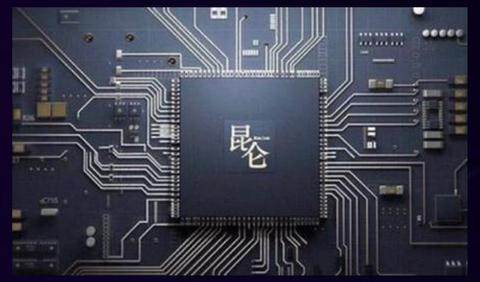
CPU、GPU深度性能优化

集成TensorRT和nGraph等顶级引擎

Paddle-Mobile

多硬件支持：
ARM CPU，Mali GPU，高通DSP，FPGA，定点化量化计算

Paddle Anywhere



Paddle Mobile 架构

选择性编译，轻量化表达

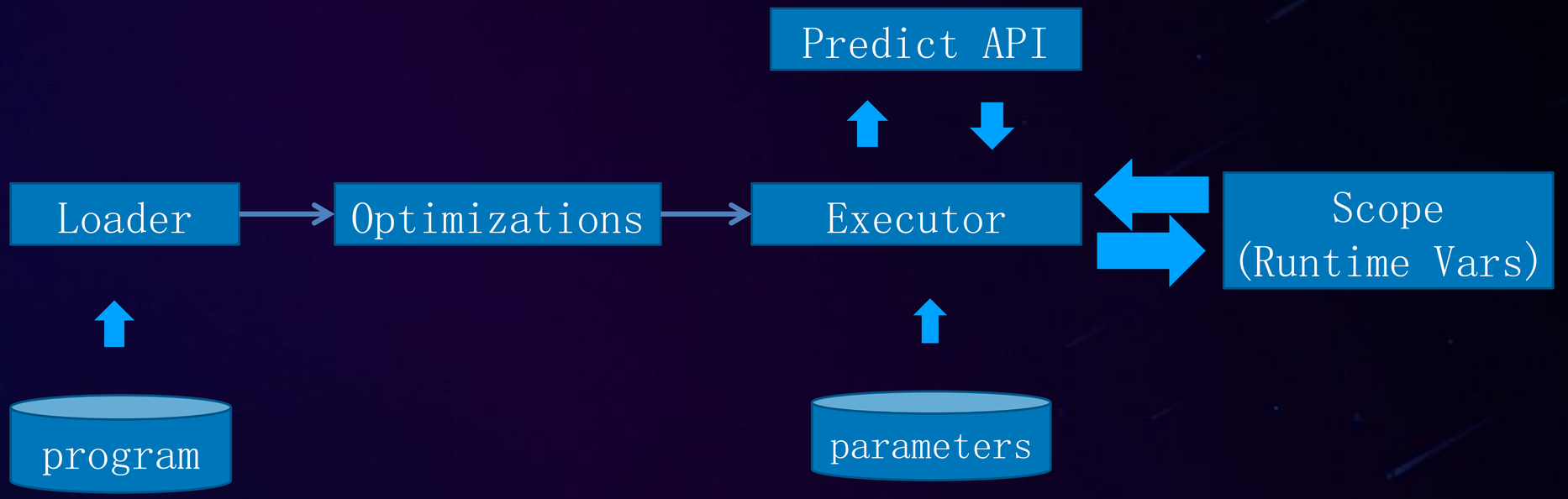
模型压缩，量化，剪枝

iOS, Android多平台支持

Arm CPU, Mali GPU, Adreno GPU, FPGA多设备支持

融合，汇编，自动化模型设计，极致性能优化。

服务器，移动端统一设计，共享中间表达和优化库



Performance

Mobilenet-ssd Unit:ms	1 Thread				2 Threads				4 Threads			
	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C
Kirin960	212.14	226.46	376.58	221.27	125.16	125.15	278.03	122.19	74.57	78.83	188.24	77.46
Qualcomm835	215.58	229.04	372.14	21.571	126.4	130.52	265.72	4.1081	76.41	85	183.53	76.97

Mobilenetv1 Unit:ms	1 Thread				2 Threads				4 Threads			
	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C	Paddle Mobile	同类型产品A	同类型产品B	同类型产品C
Kirin960	106.29	116.67	173.40	221.27	61.53	69.16	139.43	122.19	36.48	41.05	88.89	39.34
Qualcomm835	107.65	127.88	174.56	21.571	64.15	67.10	131.10	4.1081	37.66	42.51	88.85	41.57

More on: https://github.com/PaddlePaddle/paddle-mobile

超大规模并行深度学习

高效并行训练

和框架一体的并行设计
统一的Operator级操作
对用户透明的分布式部署

大规模文本类任务
高度异步化并行策略

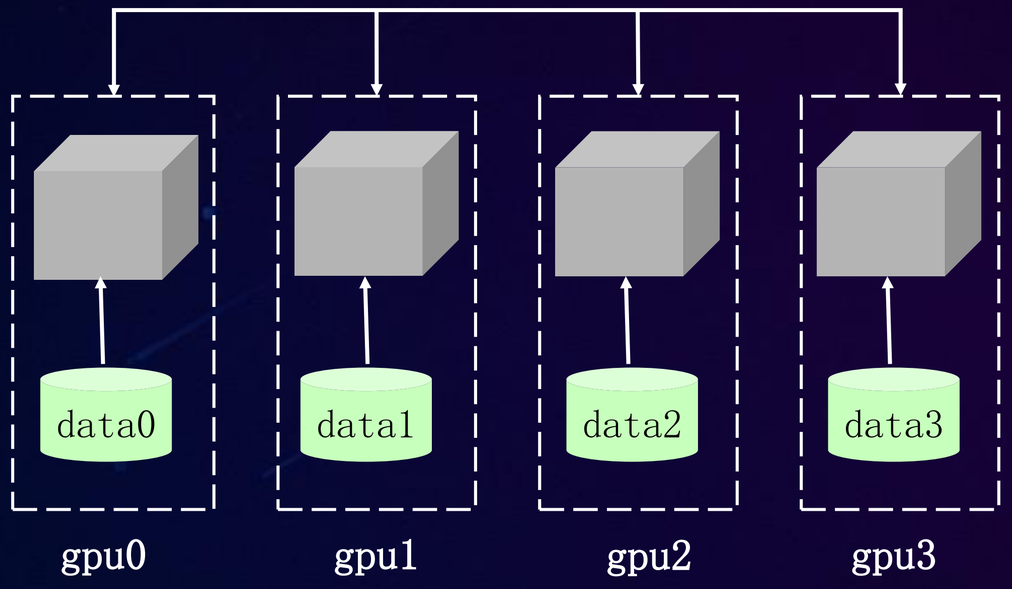
CTR预估类
超大规模稀疏特征任务
参数分布式存储

GPU多机多卡、CPU多机多线程支持

同步、异步支持

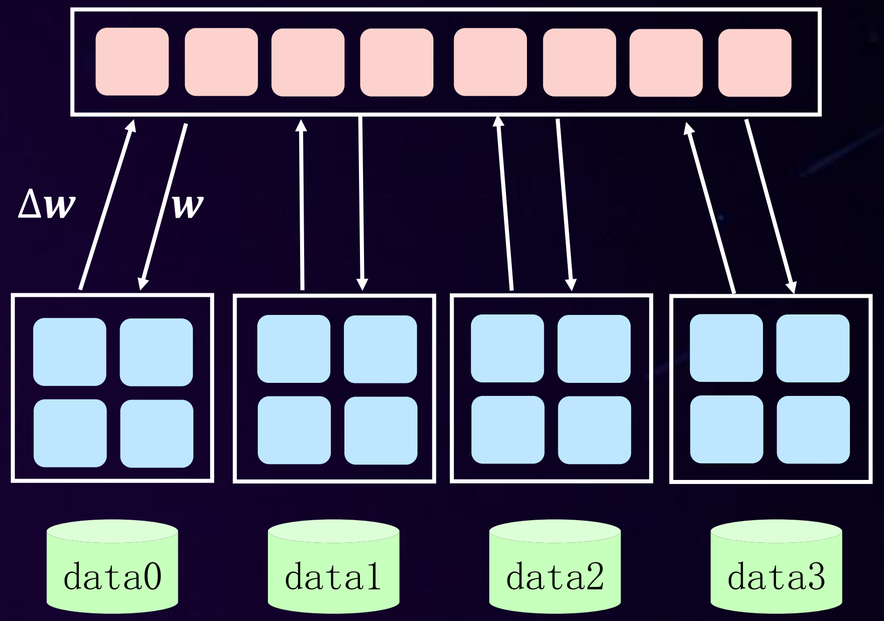
多种通信方式支持

PaddlePaddle支持的分布式训练场景举例



图像分类、机器翻译等计算密集度高的任务

参数同步方式：同步Collective操作



CTR预估、语义匹配等数据吞吐量大的任务

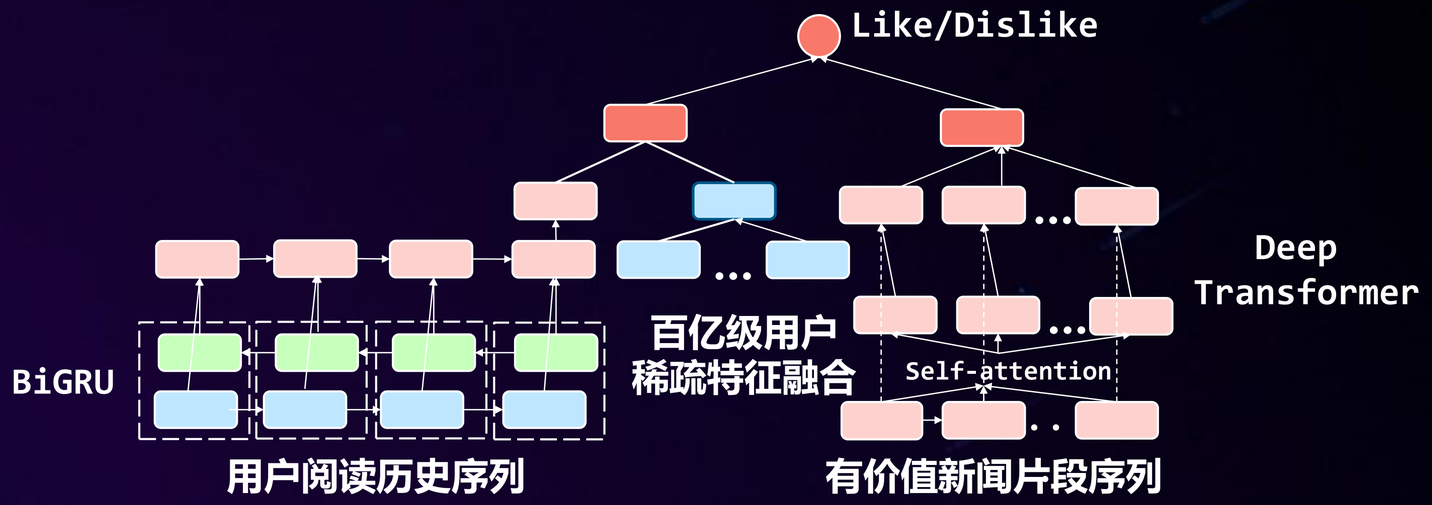
参数同步方式：异步大规模稀疏参数服务器

并行方式：以数据并行为主的任务，覆盖NLP、CV、搜索、推荐、广告等场景

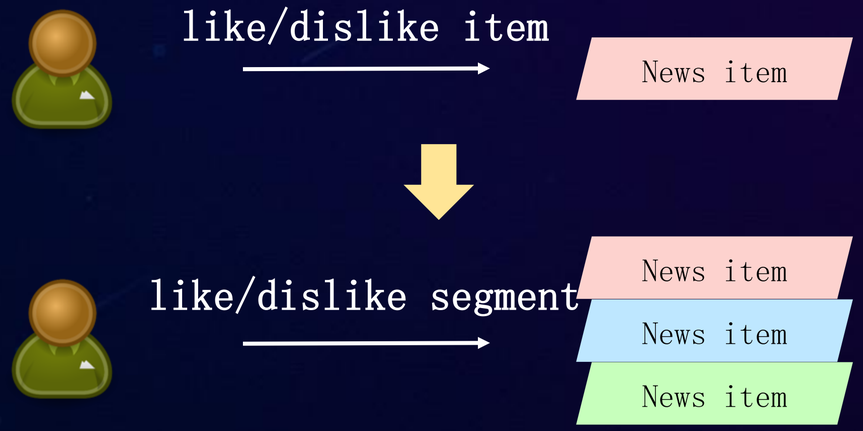
参数同步方式：GPU场景下以基于NCCL2为主的同步训练；CPU场景下基于RPC，支持经典实用的异步优化算法

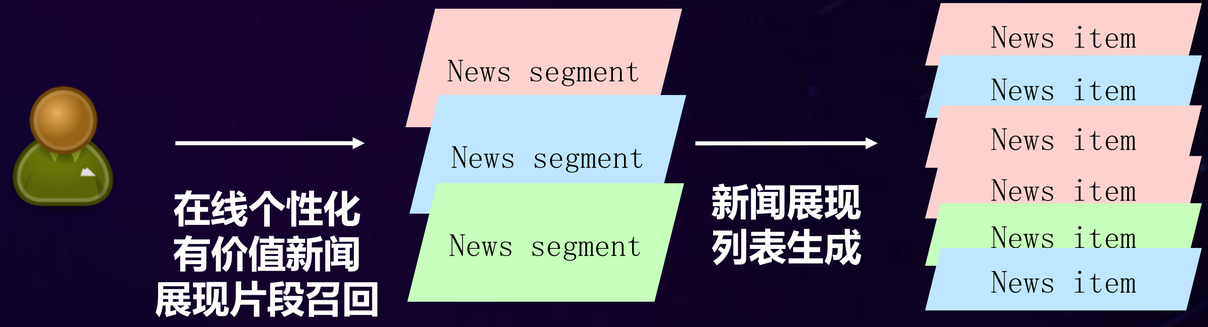
应用案例：Feed List Generation

Title	P(B|A)/P(B)	备注
A: 饿了么9月15日中午在部分地区遭遇28分钟故障，官方致歉	470.03	互联网热点
B: 拼多多回应售卖黑作坊散装纸尿裤：将先行在整个电商行业全量下架		
A: GIF精选：喝酒技术人才培训中心，进门喜洋洋，出门扶着墙！	438.04	搞笑段子
B: GIF精选：军训调整我的顺拐，我绝对有信心把教官带成顺拐！		



seq2seg model 亿级别数据小时级更新线上模型





- 展现新闻片段的结构具有价值，能够吸引用户点击
- 提出预估对象从news item 转换为 news segment，并利用 nt信息组成feed展现列表
- 小流量：Feed媒体总时长+1.27%，总分发+0.54%，总展现量+1.12%

PaddlePaddle Release 1.1

欢迎使用！

Github：https://github.com/paddlepaddle

官网：http://paddlepaddle.org/

微信公众号