中山大學
SUN YAT-SEN UNIVERSITY

面向泛在算力网络的边缘端协同智能计算系统

Collaborative Edge AI System in Ubiquitous Computing Network

叶盛源

中山大学计算机学院博士生（导师：陈旭教授）

http://www.brandonye.tech

算力与算力网络

- 随着各行各业数字化、智能化转型的推进，**算力**对数字经济发展的带动作用越来越突出

智能家居

智能自动化工厂

AI赋能智慧城市

- 我国**算力**仍存在**资源不充足**、**调度不协调**、**分布不均匀等**问题

2024
03/30
16:11:37
来源：新华网

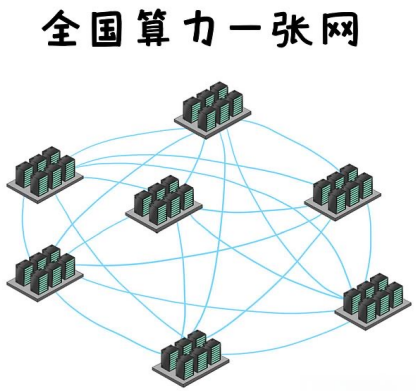
报告：我国算力资源分散、利用效率有待提高

报告显示，2022年我国以公共云形式提供服务的算力占比仅为28%，大部分服务器以私有化部署的形式存在。从使用效率看，公共云CPU利用效率可达40%，而专属云部署的CPU使用效率通常为5%-10%。我国算力产业呈碎片化分布，算力资源在规模、使用成本等方面难以满足人工智能的规模化应用和快速迭代创新的需要，建立适应“人工智能+”时代的高质量算力服务体系迫在眉睫。

然而，在数字经济的发展下，我们也看了一些问题，比如数据中心在地理上分布不均，带来资源不协调的矛盾。东部区域经济发达，对数据中心的要求很高，但在寸土寸金的东部区域，数据中心的建设带来了资源紧张，同时，数据中心对能耗的需求越来越高，电力成本激增，社会成本越来越高。而拥有丰富可再生资源、气候适宜、地广人稀的西部区域，其数据中心建设较少，但网络带宽低，跨省数据传输费用高。面对如此局面，我们思考，能否在 IT 基础

算力与算力网络

- 加快构建以**算力网络**为核心的新型基础设施建设，打通数字经济发展的“大动脉”



- 以“**东数西算**”、“**超算互联网**”为主题词的**算力网络**建设，正蓬勃发展中

东西部	枢纽节点	集群	起步区边界
西部	贵州枢纽	贵安数据中心集群	贵安新区贵安电子信息产业园
	内蒙古枢纽	和林格尔数据中心集群	和林格尔新区、集宁大数据中心产业园
	甘肃枢纽	庆阳数据中心集群	庆阳西峰数据信息产业聚集区
	宁夏枢纽	中卫数据中心集群	中卫工业园西部云基地

东西部	枢纽节点	集群	起步区边界
东部	京津冀枢纽	张家口数据中心集群	张家口市怀来县、张北县、宣化区
	长三角枢纽	长三角生态绿色一体化发展示范区数据中心集群	上海市青浦区、江苏省苏州市吴江区、浙江省嘉兴市嘉善县
		芜湖数据中心集群	芜湖市鸠江区、弋江区、无

国家超算互联网平台上线

国家超算互联网平台日前正式上线。国家超算互联网可将全国众多超算中心连接起来，构建一体化的超算算力网络和服务平台。

近年来，我国算力设施建设取得积极进展，但人工智能等技术的快速发展，对算力提出了更高要求，算力中心亟须突破现有单体运营模式。2023年4月，国家超算互联网正式启动建设。其目标是紧密连接供需方，通过市场化的运营和服务体系，实现算力资源统筹调度，有效支撑原始科学创新、重大工程突破、经济高质量发展等目标达成，成

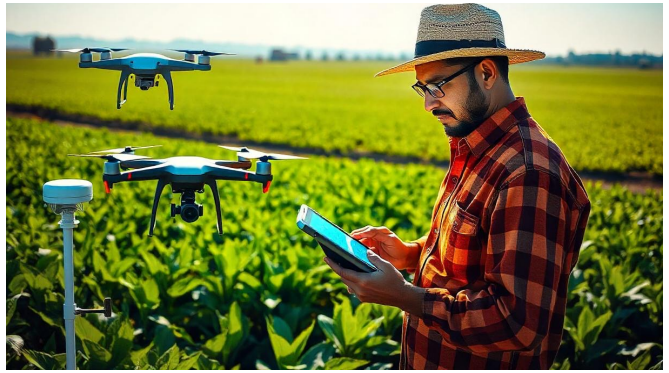
算力网络与“泛在边缘算力网络”

- **泛在边缘算力网络**是常规算力网络在终端侧的重要延展分支

智慧城市：无人车、智能路灯

智能家居：智能音箱、智能安防

智能工厂：机器人、机械臂



智慧农业：无人机、温湿监控器

端侧算力网络白皮书

（2022年）

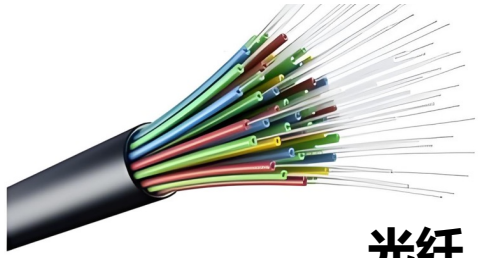
中国移动通信集团终端有限公司、北京邮电大学
中国信息通信研究院、中国通信学会

中国通信学会
https://www.china-cic.cn › upload PDF
端侧算力网络白皮书
2022年7月28日 — 随着端侧. 算力能力的提升，边缘计算技术和

中国移动、中国通信学会、北京邮电大学等联合发布《**端侧算力网络白皮书**》

泛在边缘算力网络的主要特征

- 不同于常规算力网络需要集中投资建设基础设施，泛在边缘算力网络往往是通过**现有的成熟的网络设施**，将一定空间内的**终端/边缘平台空闲算力**进行汇集和组织，聚沙成塔，成为常规算力网络中的一个**重要算力供给方**。



光纤

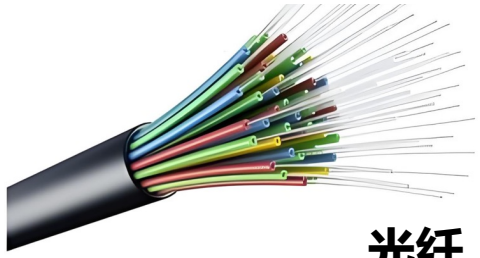
5G/6G

Wi-Fi

ZigBee® 3.0
Better Together IoT协议

泛在边缘算力网络的主要特征

- 不同于常规算力网络需要集中投资建设基础设施，泛在边缘算力网络往往是通过**现有的成熟的网络设施**，将一定空间内的**终端/边缘平台空闲算力**进行汇集和组织，聚沙成塔，成为常规算力网络中的一个**重要算力供给方**。



光纤

5G/6G

Wi-Fi

ZigBee® 3.0
Better Together IoT协议

- 区别于数据中心和超算互联平台严密的网络架构和强大的算力，泛在边缘算力网络一般具有**更松散灵活的网络架构**，且具有极强的**动态性**；同时，接入网络的单个边缘节点**算力资源**往往是**有限且异构**的。

手机、手表等移动物联网设备

个人PC/边缘服务器

英伟达 Jetson 系列同类边缘计算平台

面向泛在算力网络的边缘端协同智能计算系统

- **面向泛在算力网络的边缘端协同人工智能模型微调系统**

Asteroid (小行星) : Resource-Efficient Hybrid Pipeline Parallelism for Collaborative DNN Training on Heterogeneous Edge Devices. **ACM MOBICOM 2024**

Pluto and Charon (冥王星) : A Time and Memory Efficient Collaborative Edge AI Framework for Personal LLMs Fine-Tuning. **ACM ICPP 2024**

- **面向泛在算力网络的边缘端协同人工智能模型推理系统**

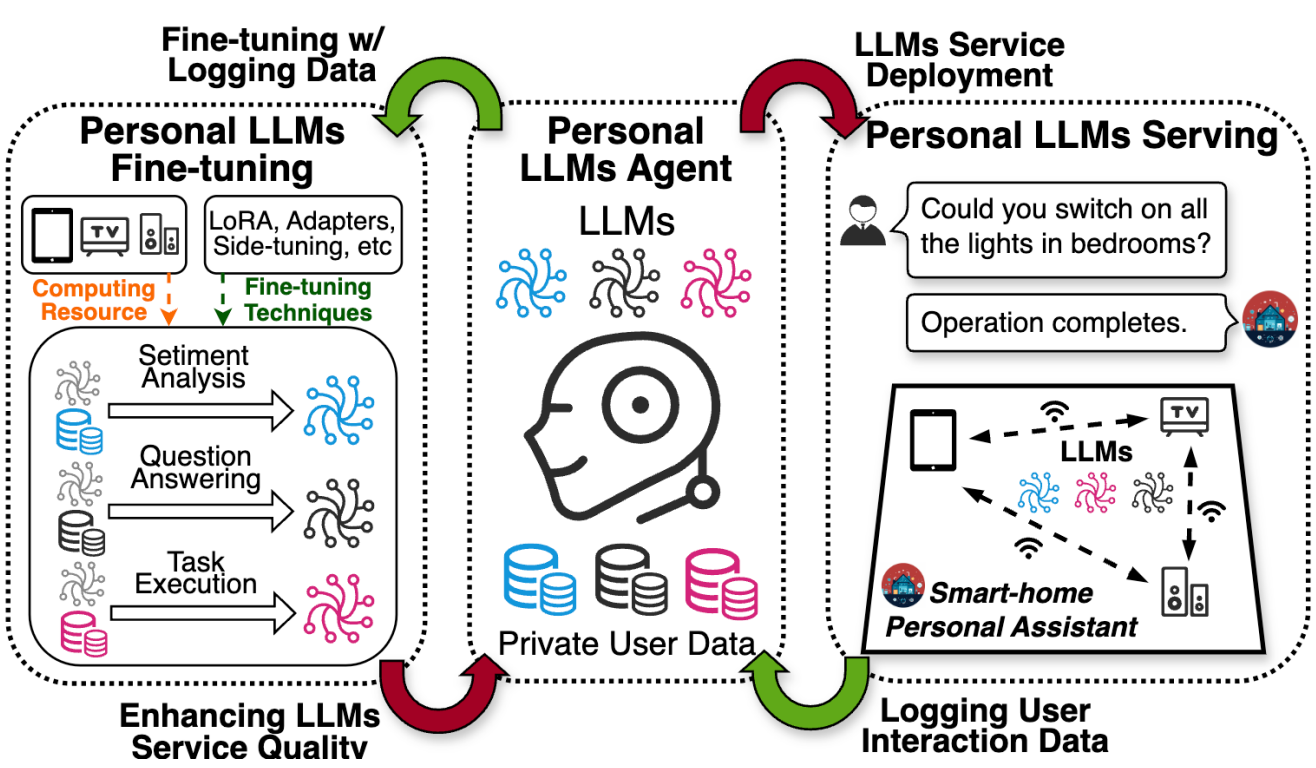
Jupiter (木星) : Fast and Resource-Efficient Collaborative Inference of Generative LLMs on Edge Devices. **IEEE INFOCOM 2025**

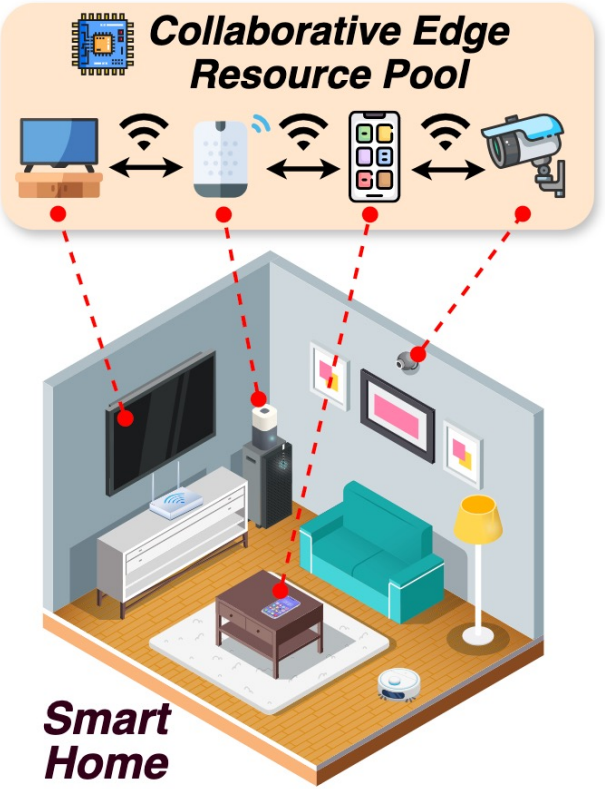
Galaxy (银河) : A Resource-Efficient Collaborative Edge AI System for In-situ Transformer Inference. **IEEE INFOCOM 2024**

研究目标：搭建高性能、低功耗、强隐私的边缘端协同智能计算系统，充分挖掘泛在边缘算力网络中的算力和通信潜力，使泛在边缘算力网络能成为常规算力网络中的一个重要算力供给方！

应用场景描述

- **本地部署的人工智能应用理想工作流程：**





上传用户隐私数据至商业公司运营的数据中心进行模型推理和微调，将**面临着严重的用户隐私安全问题。**

边缘端协同计算系统为本地人工智能应用提供**高性能、低功耗、强隐私的算力资源。**

边缘端协同人工智能模型微调计算系统：Asteroid

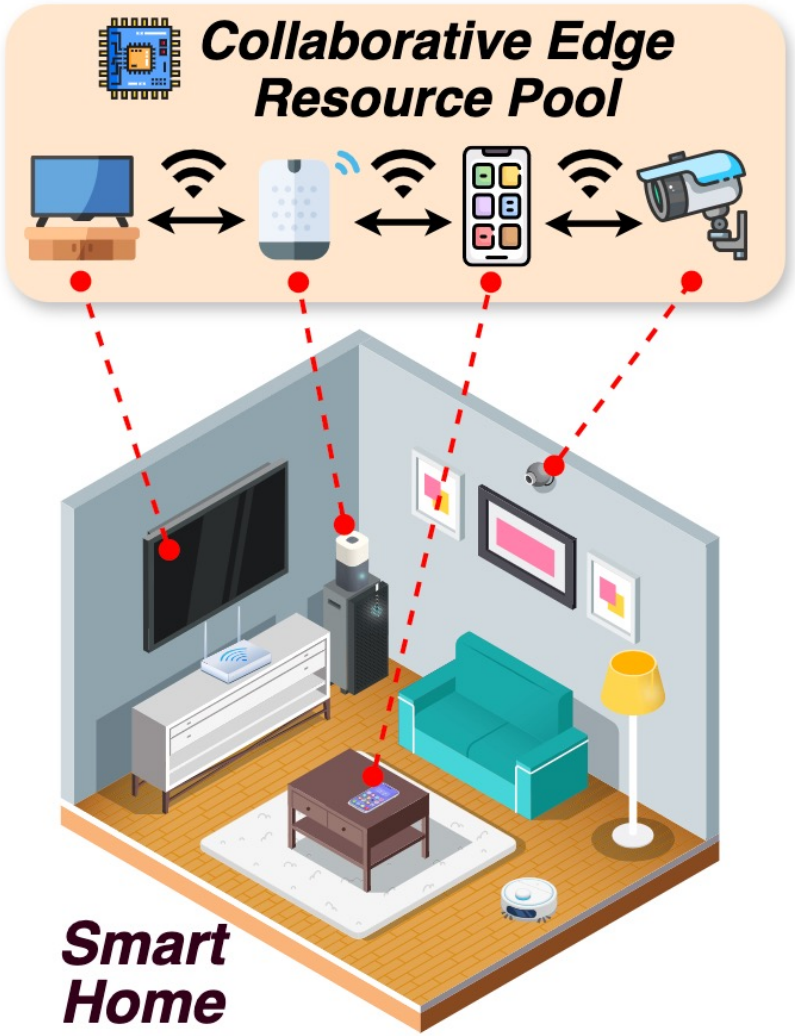
边缘端协同人工智能模型微调计算系统：Asteroid

边缘协同训练的研究问题

1. 如何选择**最适合边缘网络的并行策略**？是数据并行，流水线并行，还是张量并行？

2. 如何充分考虑多台**异构边缘设备的资源预算**，量身定制资源最高效的并行规划方案？包括模型切分，以及设备编排方式。

3. 如何在设备**高度动态的泛在边缘算力网络环境**下，实现稳定可靠的深度学习模型训练过程？



边缘端协同人工智能模型微调计算系统：Asteroid

边缘协同训练的研究问题

1. 如何选择**最适合边缘网络的并行策略**？是数据并行，流水线并行，还是张量并行？

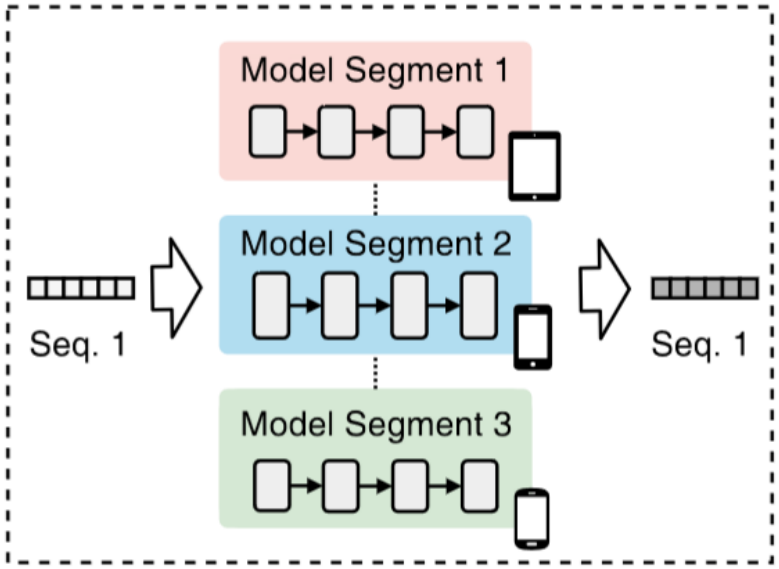
2. 如何充分考虑多台**异构边缘设备的资源预算**，量身定制资源最高效的并行规划方案？包括模型切分，以及设备编排方式。

3. 如何在设备**高度动态的泛在边缘算力网络环境**下，实现稳定可靠的深度学习模型训练过程？

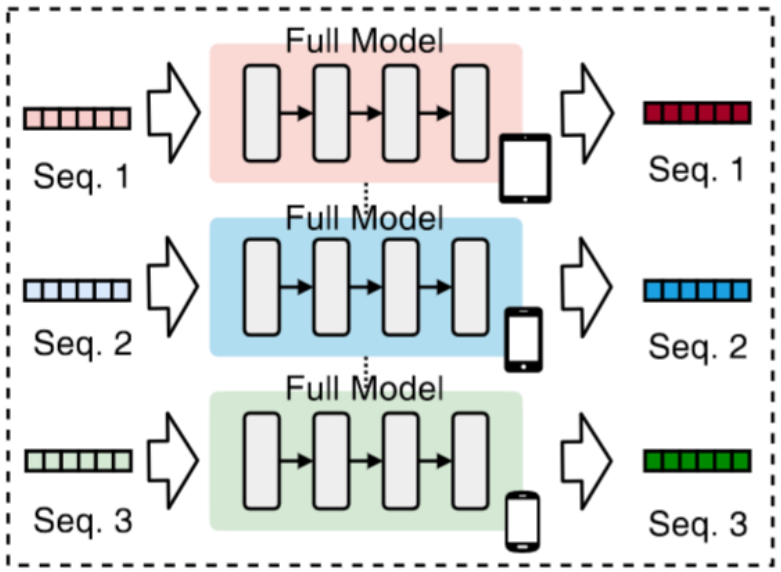
边缘端协同人工智能模型微调计算系统：Asteroid

- **选择何种并行训练架构?**

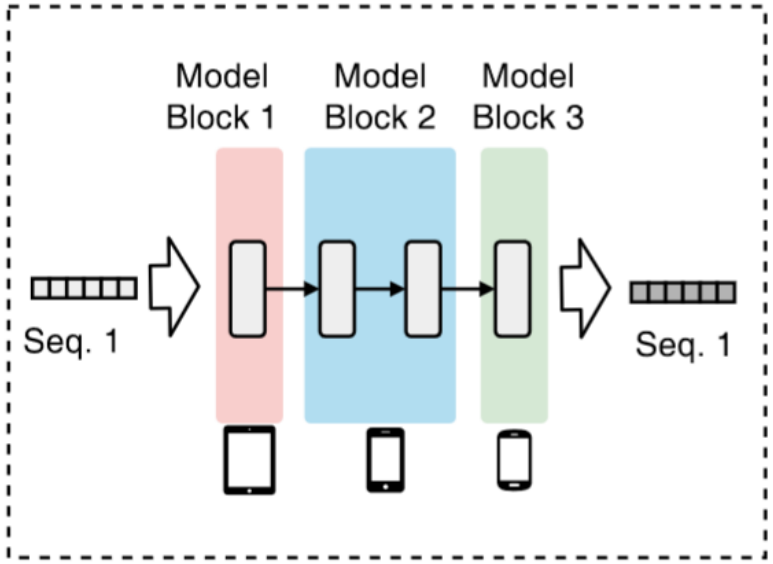
张量并行



数据并行



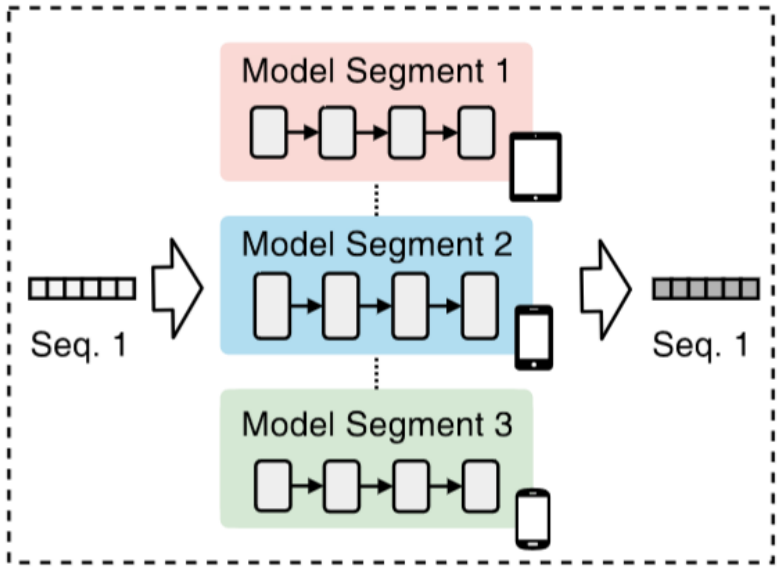
流水线并行



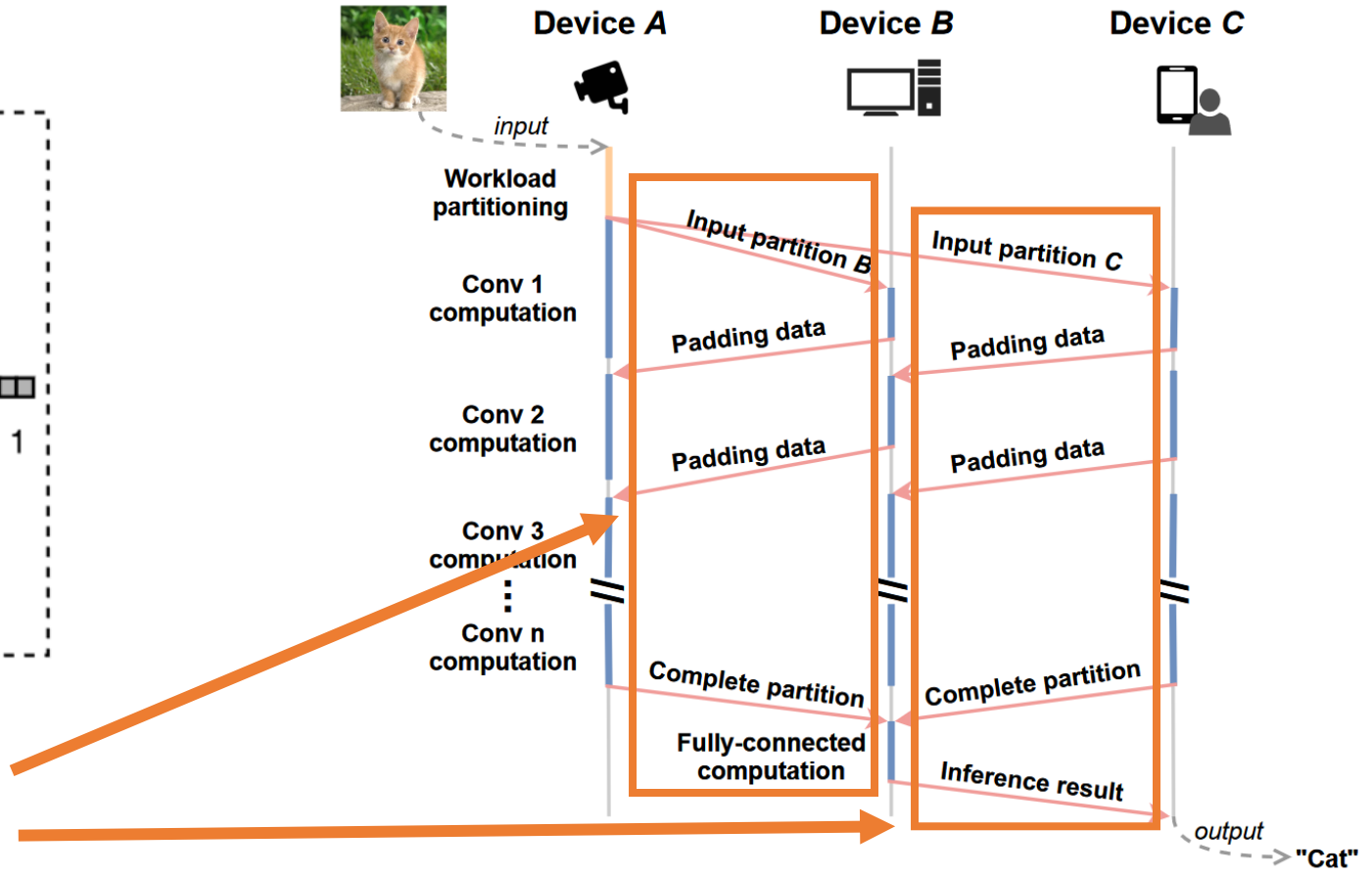
边缘端协同人工智能模型微调计算系统：Asteroid

- **选择何种并行训练架构?**

张量并行



在每一层神经网络的计算中都会产生大量的、无法忍受的张量通信同步开销。

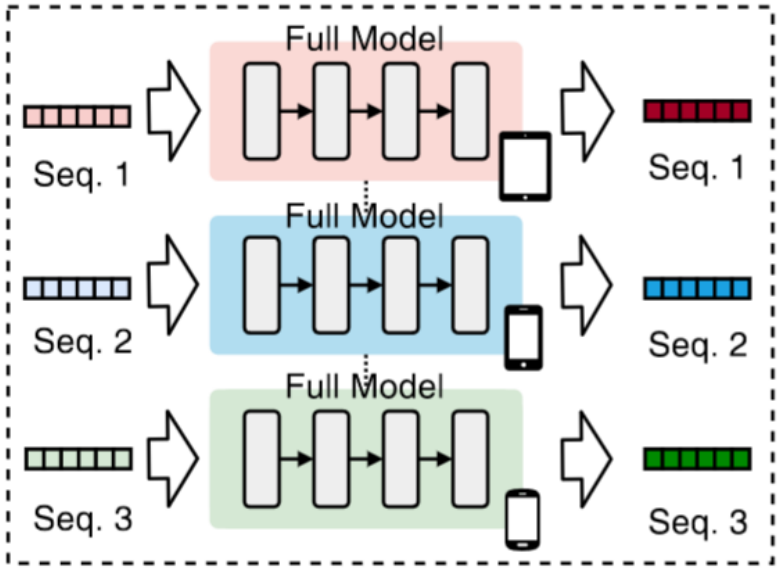


图片引用自: CoEdge: Cooperative DNN Inference with Adaptive Workload Partitioning over Heterogeneous Edge Devices

边缘端协同人工智能模型微调计算系统：Asteroid

- **选择何种并行训练架构?**

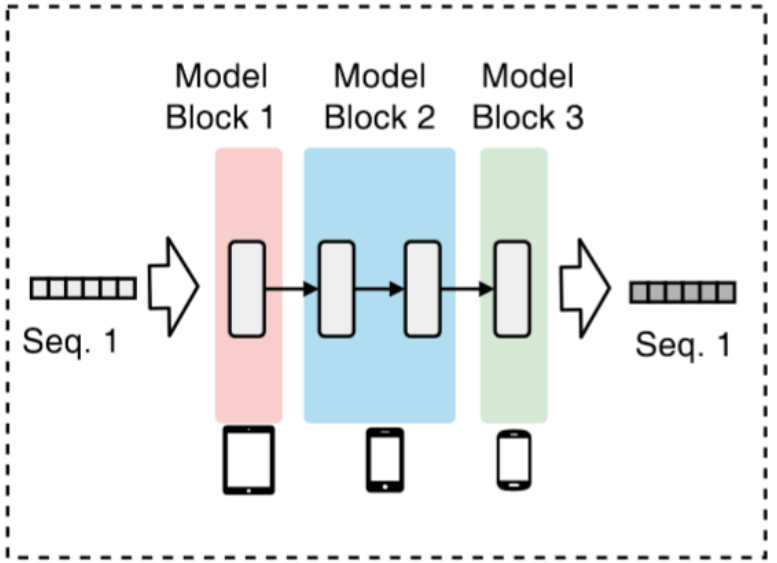
数据并行



更高效的资源可拓展性和资源利用率

需要每个设备保存完整模型，单机内存负担非常大

流水线并行



可以将模型切分到多个设备上，单机内存负担小

随着设备增加，资源拓展效率较差。且对不均匀的负载划分更敏感，优化难度大

边缘端协同人工智能模型微调计算系统：Asteroid

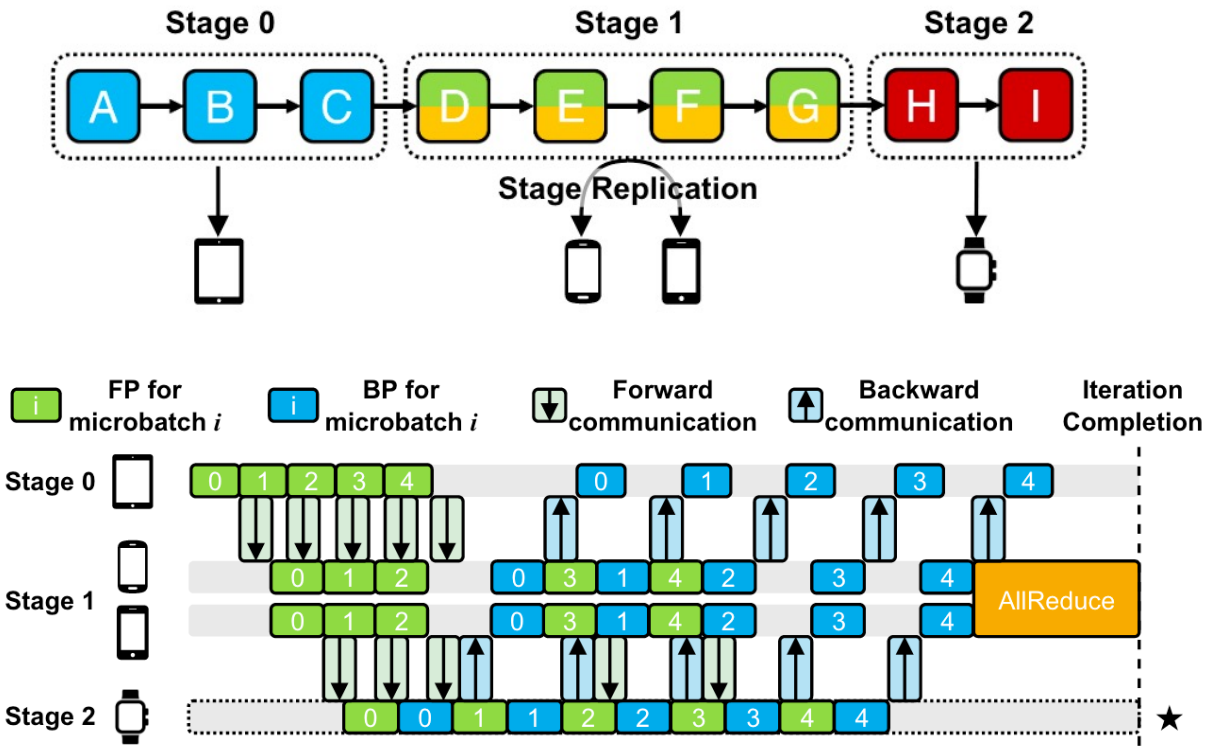
Asteroid采用了混合流水线和数据并行架构来编排边缘设备

训练工作流程：

1. 步骤一：**将深度神经网络切割**为多个流水线阶段，每个阶段包含一个阶段子模型。

2. 步骤二：**将边缘设备分组**，并将每一个设备组关联到其中一个的流水线阶段上。

3. 步骤三：将一个mini-batch的微调数据，切分成很多个更小的micro-batch，并将它们同时注入流水线。Asteroid训练系统将在**设备组间进行流水线并行训练，在设备组内进行数据并行训练**。



将一个mini-batch的数据切分为5个micro-batch，并同时注入流水线后的训练示意图

边缘端协同人工智能模型微调计算系统：Asteroid

边缘协同训练的研究问题

1. 如何选择**最适合边缘网络的并行策略**？是数据并行，流水线并行，还是张量并行？

2. 如何充分考虑多台异构边缘设备的资源预算，量身定制资源最高效的并行规划方案？包括模型切分，以及设备编排方式。

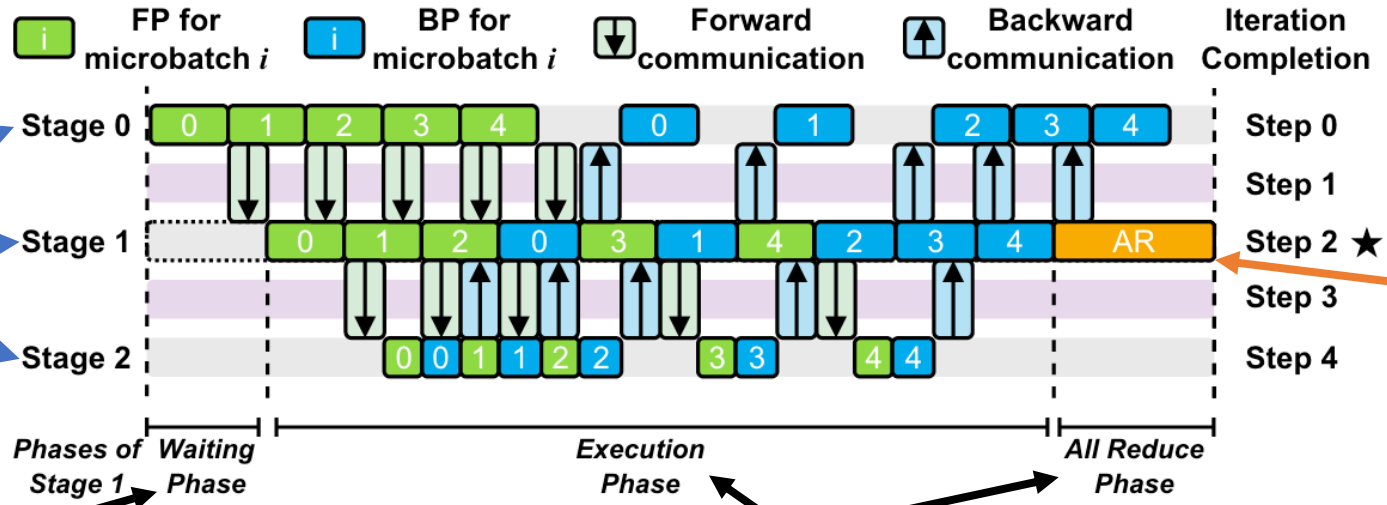
3. 如何在设备**高度动态的泛在边缘算力网络环境**下，实现稳定可靠的深度学习模型训练过程？

边缘端协同人工智能模型微调计算系统：Asteroid

- **优化目标：最小化一个mini-batch训练所需的时间(HPP-Round Latency)：**

$$\text{HPP-Round Latency} = \max_{s\in\{0,1,\ldots,S-1\}} \left(T_w^s + T_e^s + T_a^s\right)$$，需要满足内存预算约束，避免内存溢出！

使用1F1B调度策略优化设备的内存开销



设备数量大于1的阶段，在结束时将需要进行All-Reduce权重同步

$$T_w^s = \sum_{i=0}^{s-1} E_f^i,$$

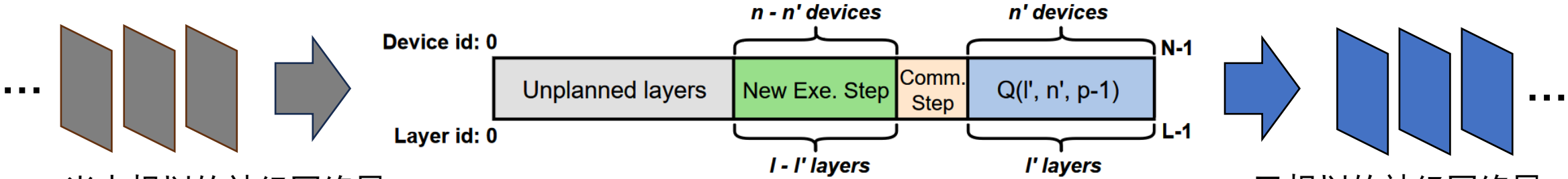
$$T_a^s = \frac{2\left(|\mathcal{G}_s| - 1\right) \cdot \sum_{l\in\mathcal{D}_s} w_l}{|\mathcal{G}_s| \cdot \min_{d,d'\in\mathcal{G}_s} b_{d,d'}}.$$

$$T_e^s = M \times \left(E_f^{dm} + E_b^{dm}\right) + \begin{cases} \sum_{i=s}^{dm-1} \left(E_f^i + E_b^i\right), & s < dm, \\ -\sum_{i=dm}^{s-1} \left(E_f^i + E_b^i\right), & s \geq dm. \end{cases}$$

边缘端协同人工智能模型微调计算系统：Asteroid

- **资源预算感知的并行配置搜索算法：输出模型切分和设备分组配置方案**

采用了动态规划思想来加速最优并行规划配置的搜索过程



尚未规划的神经网络层

已规划的神经网络层

Algorithm 2: Dynamic Programming HPP Planning

```
for p from 1 to min(L, N) do
    for n from 1 to N do
        for l from 1 to L do
            for n' from 0 to n do
                for l' from 0 to l do
                    Get E_f^s and E_b^s with Alg. 1 and Eq. (8);
                    Update Dominant Step with Eq. (11);
                    Get T_w^s, T_e^s and T_a^s with Eq. (5) and (6);
                    Get HPP-Round Latency with Eq. (4);
            Update Q(l, n, p) with Eq. (10);
```

Properties	PipeDream	Dapple	Alpa	HetPipe	Asteroid
Combining DP with PP?	✓	✓	✓	✓	✓
Resource Heterogeneous Awareness?				✓	✓
Memory Constraint Awareness?			✓		✓
Communication Modeling & Optimization?		✓			✓

边缘端协同人工智能模型微调计算系统：Asteroid

边缘协同训练的研究问题

1. 如何选择**最适合边缘网络的并行策略**？是数据并行，流水线并行，还是张量并行？

2. 如何充分考虑多台**异构边缘设备的资源预算**，量身定制资源最高效的并行规划方案？包括模型切分，以及设备编排方式。

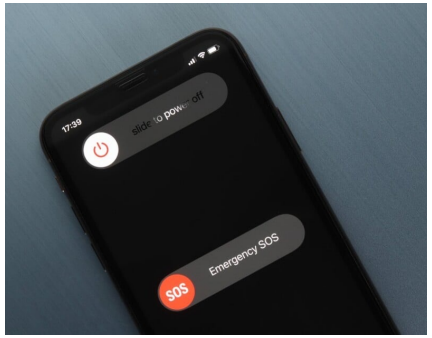
3. 如何在设备**高度动态的泛在边缘算力网络环境**下，实现稳定可靠的深度学习模型训练过程？

边缘端协同人工智能模型微调计算系统：Asteroid

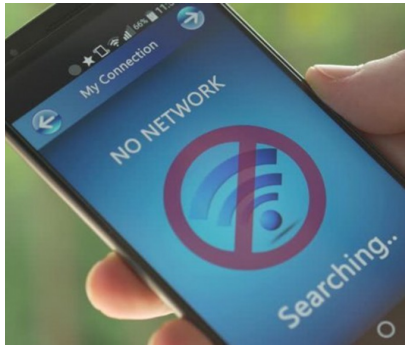
- **边缘流水线训练的容灾恢复机制**

为什么需要容灾恢复？

- 设备的退出或宕机可能导致训练权重丢失。
- 流水线中存在任何异常设备都会导致流水线阻塞进而影响全局训练进行。



Energy Depletion



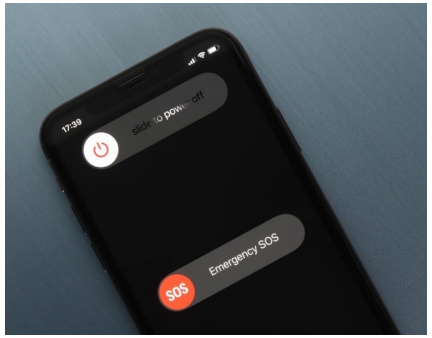
Network Anomalies

边缘端协同人工智能模型微调计算系统：Asteroid

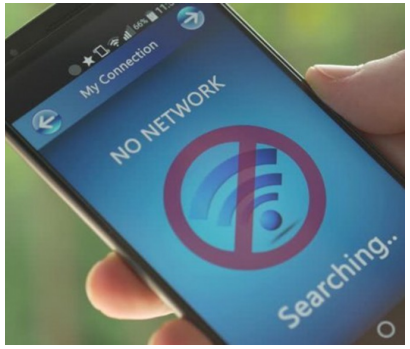
- **边缘流水线训练的容灾恢复机制**

为什么需要容灾恢复？

- 设备的退出或宕机可能导致训练权重丢失。
- 流水线中存在任何异常设备都会导致流水线阻塞进而影响全局训练进行。



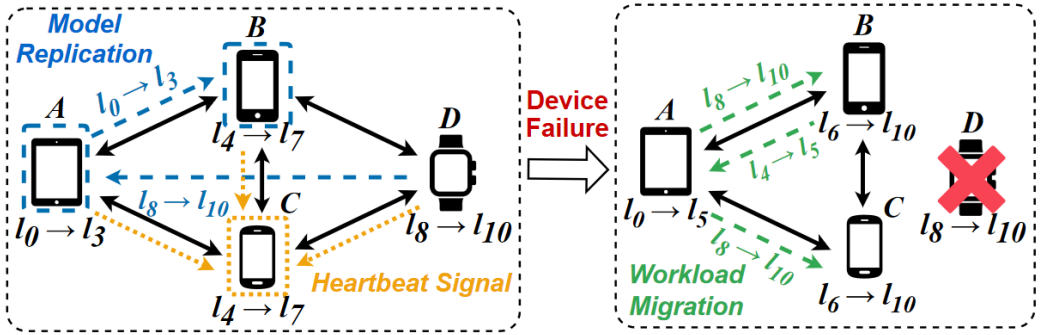
Energy Depletion

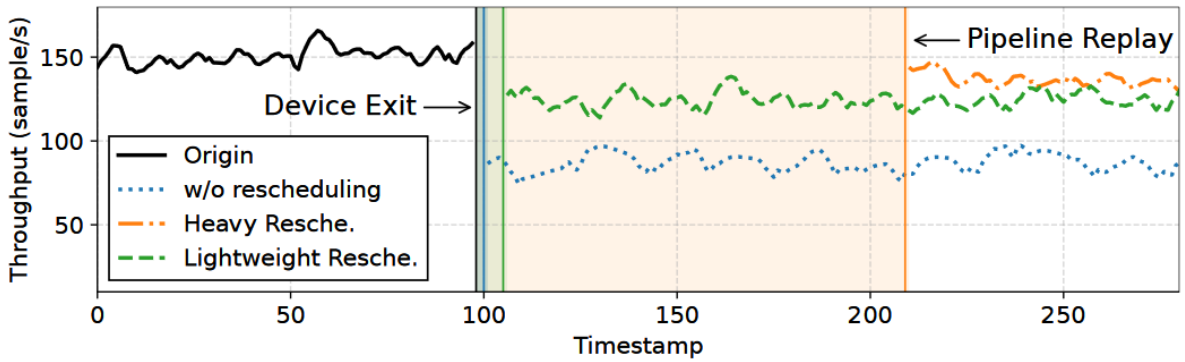


Network Anomalies

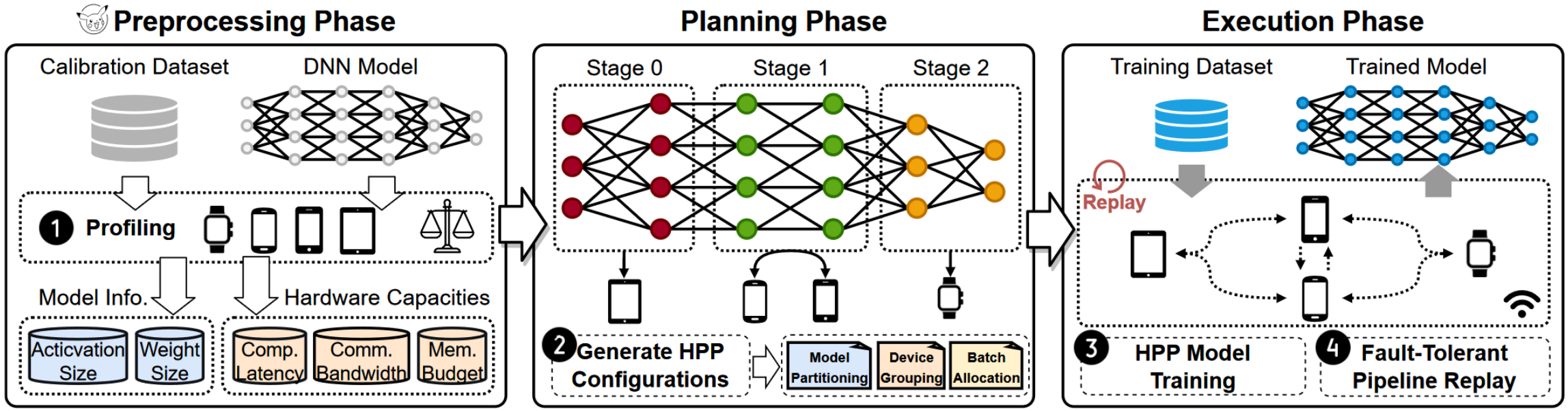
如何实现高效的容灾恢复机制

- 心跳感知的故障检测
- 定期模型权重备份
- 层粒度的权重迁移和流水线恢复





边缘端协同人工智能模型微调计算系统：Asteroid



◆ **预处理阶段：**

Asteroid Profiler 模块将在真实的边缘设备上进行测量，并跟踪记录规划阶段所需要的设备和模型信息。

◆ **并行规划阶段：**

Asteroid Planner 模块将从Profiler模块中获取设备和模型的信息，并执行动态规划算法，生成模型切分和设备分组配置方案。

◆ **并行规划阶段：**

Asteroid Runtime 模块将规划配置应用于目标模型和边缘设备上，并进行高效的边缘协同训练。同时，容灾恢复模块将在后台持续监控系统是否存在异常设备。

边缘端协同人工智能模型微调计算系统：Asteroid

- **实验设置**
  - 实验设备及边缘环境：使用3种异构的边缘AI平台：Jetson Nano, TX2 和 NX，模拟了四种同构或异构的边缘集群。在100Mbps(低速)和1000Mbps(高速)的网络环境下进行测试。

Table 5: Specifications of edge devices in experiments.

Edge Device	GPU Processor	Memory
Jetson Nano [2]	128-core NVIDIA Maxwell	4GB
Jetson TX2 [1]	256-core NVIDIA Pascal	8GB
Jetson NX [3]	384-core NVIDIA Volta	8GB

Table 6: Heterogeneous edge env. used in experiments.

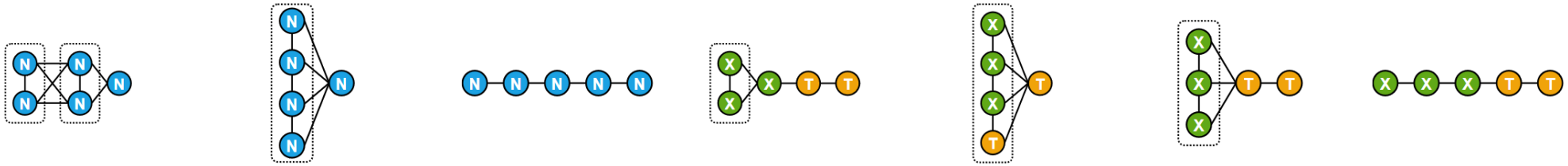
ID	Devices	ID	Devices
A	5 × Nano	C	1 × NX, 2 × TX2, 3 × Nano
B	3 × NX, 2 × TX2	D	1 × TX2, 3 × Nano

  - 模型及数据集：
    - 模型：4种被广泛应用于计算机视觉和自然语言处理的深度神经网络：EfficientNet, MobileNet, ResNet and BERT.
    - 数据集：CIFAR-10，Mini-ImageNet 和 GLUE数据集

边缘端协同人工智能模型微调计算系统：Asteroid

Asteroid在各种边缘环境和网络条件下保持高性能，与经典的数据并行和流水线并行相比，训练加速最高可达 12.2 倍。

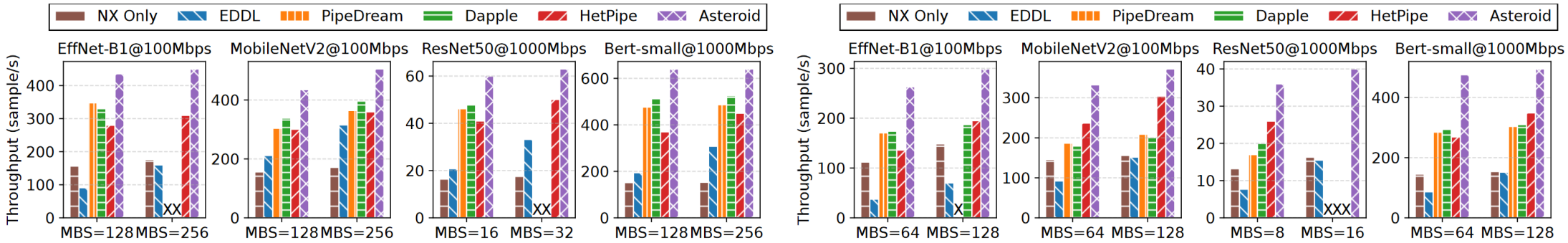
Task	Model	Dataset	Input Size	Edge Environment	Asteroid Config.	Speedup over Device	Speedup over DP	Speedup over PP
Image Classi-fication	EfficientNet-B1 [49]	Cifar-10 [15]	3 × 32 × 32	A (100Mbps)	❶	4.4×	2.1×	2.8×
				B (100Mbps)	❹	3.0×	4.8×	9.7×
				B (1000Mbps)	❹	3.7×	2.1×	1.4×
	MobileNetV2 [45]	Cifar-10 [15]	3 × 32 × 32	A (100Mbps)	❷	4.5×	1.5×	3.5×
				B (100Mbps)	❺	3.2×	2.3×	11.2×
				B (1000Mbps)	❻	3.8×	1.2×	1.3×
	ResNet50 [20]	Mini-ImageNet [52]	3 × 224 × 224	A (100Mbps)	❷	3.4×	3.6×	5.8×
				B (100Mbps)	❻	1.5×	6.1×	12.2×
				B (1000Mbps)	❹	3.7×	2.9×	3.1×
Language Model	Bert-small [14]	Synthetic Data	32 × 512	A (100Mbps)	❸	3.5×	6.4×	1×
				B (100Mbps)	❼	1.3×	6.8×	1×
				B (1000Mbps)	❼	3.9×	4.2×	1.3×



(a) Configuration ❶. (b) Configuration ❷. (c) Configuration ❸. (d) Configuration ❹. (e) Configuration ❺. (f) Configuration ❻. (g) Configuration ❼.

边缘端协同人工智能模型微调计算系统：Asteroid

与云数据中心并行训练算法相比，Asteroid 可将延迟降低最多 86%。并且在训练中达到目标准确度的速度比对比方法快最多 6.1 倍。



(a) Training throughput compared with existing approaches on Env. B.

(b) Training throughput compared with existing approaches on Env. C.

Figure 13: Training throughput comparison under various settings. × means out-of-memory error.

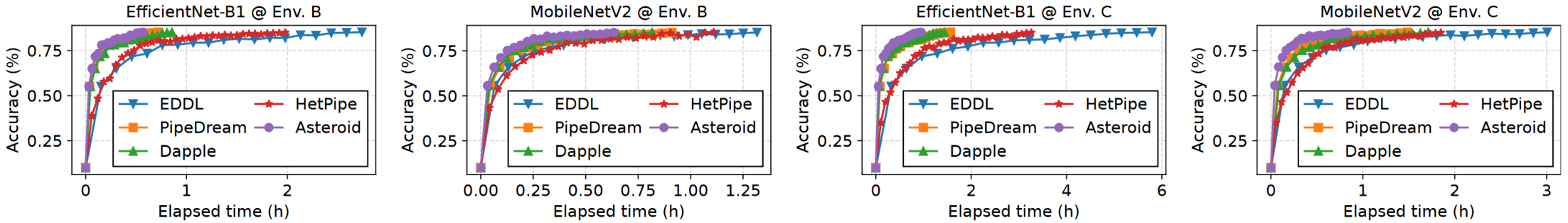
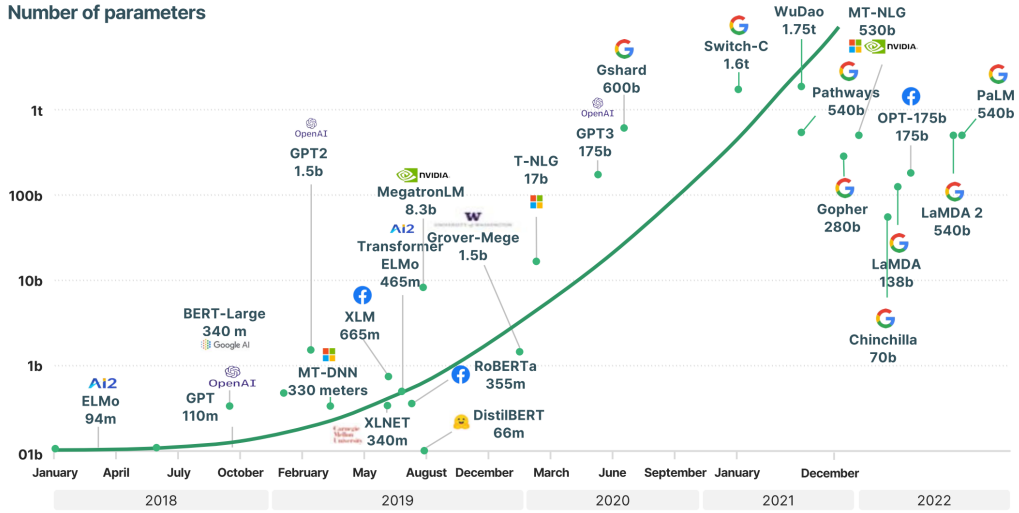


Figure 14: Training convergence of EfficientNet-B1 and MobileNetV2 on Env. B and C compared with baselines.

从小规模模型走向大规模预训练模型

- **以Transformer架构为代表的深度学习模型参数量呈指数级增长，全参数训练/微调算力需求极高，难以被泛在边缘算力网络有限的算力所承载。如何降低算力需求成为了关键难题。**



揭秘ChatGPT背后天价超算！上万颗英伟达A100，烧光微软数亿美元

作者：新智元　　2023-03-14 13:06:54

新聞

ChatGPT日燒70萬美金，OpenAI傳面臨破產

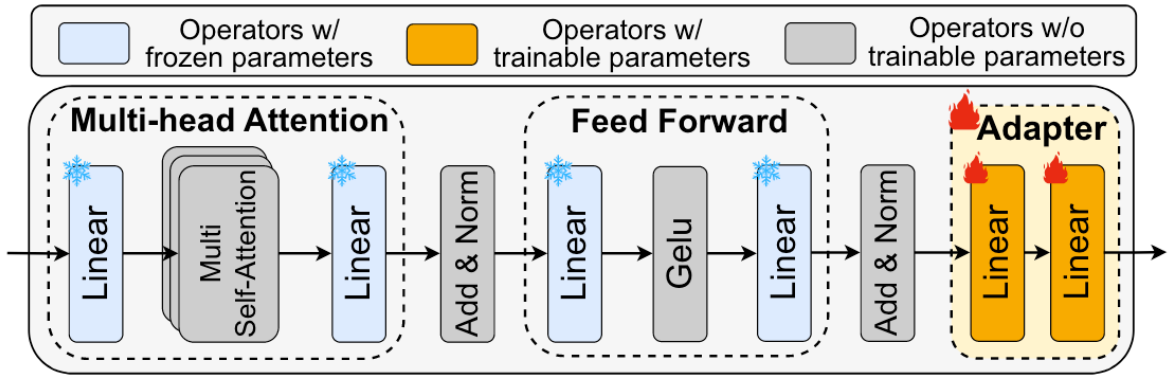
繼去年媒體指出OpenAI因開發ChatGPT，導致虧損擴大到5.4億美元，如今更有報導宣稱，OpenAI即便獲得微軟金援，但營收表現不理想仍讓這家新創浮現破產危機

51CTO　内容精选　视频　话题　技术期刊　活动

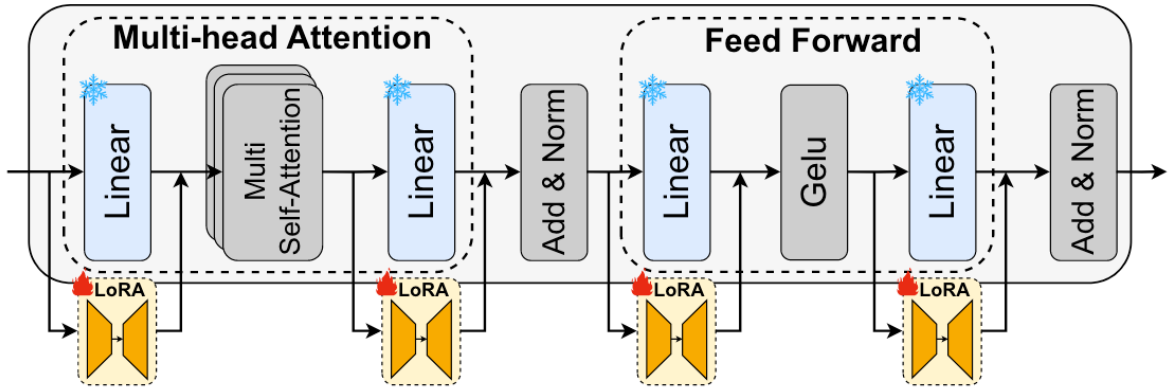
- **Asteroid 主要针对小规模通用神经网络架构，对模型微调采用了全参数微调技术**
  - 因此难以被直接应用在大规模预训练语言模型的边缘微调任务上。
  - 亟需降低大规模语言模型在泛在算力网络上微调所需的算力资源。

边缘端协同大语言模型微调计算系统：Pluto and Charon

- 参数高效的微调技术（例如：LoRA 和 Adapters），**只需要修改模型1%的参数**，就可以取得和全参数微调一样的效果。



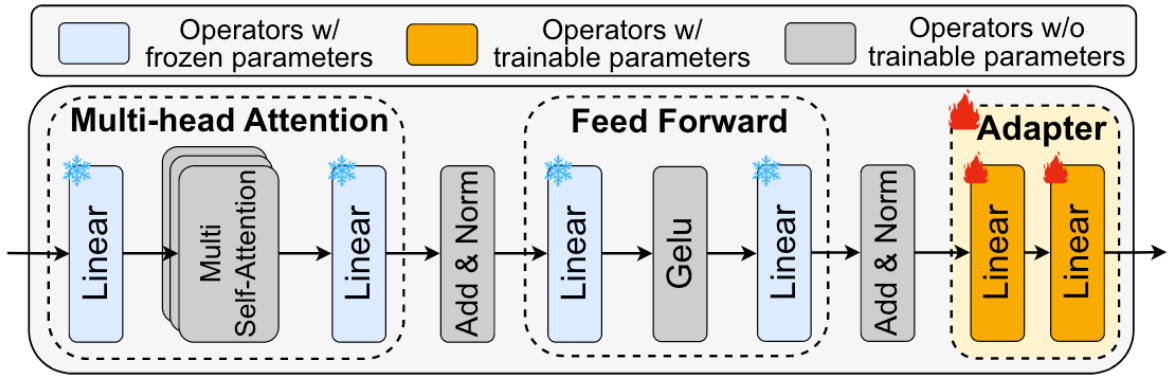
(a) The transformer layer structure of Adapters.



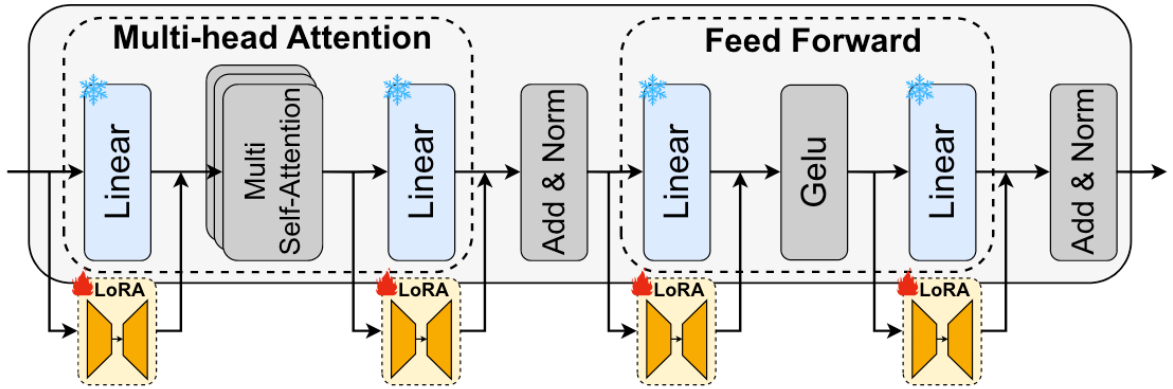
(b) The transformer layer structure of LoRA.

边缘端协同大语言模型微调计算系统：Pluto and Charon

- 目前最主流的参数高效微调技术（例如：LoRA, Adapters），对于资源受限的边缘设备**还并不足够**的算力与内存资源高效。



(a) The transformer layer structure of Adapters.



(b) The transformer layer structure of LoRA.

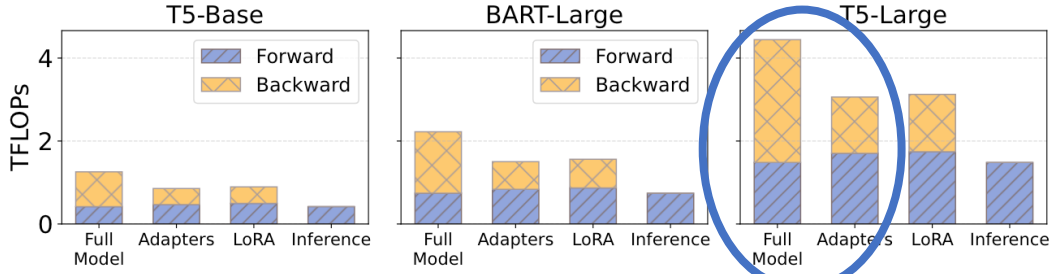


Figure 3: The comparison of floating point of operations

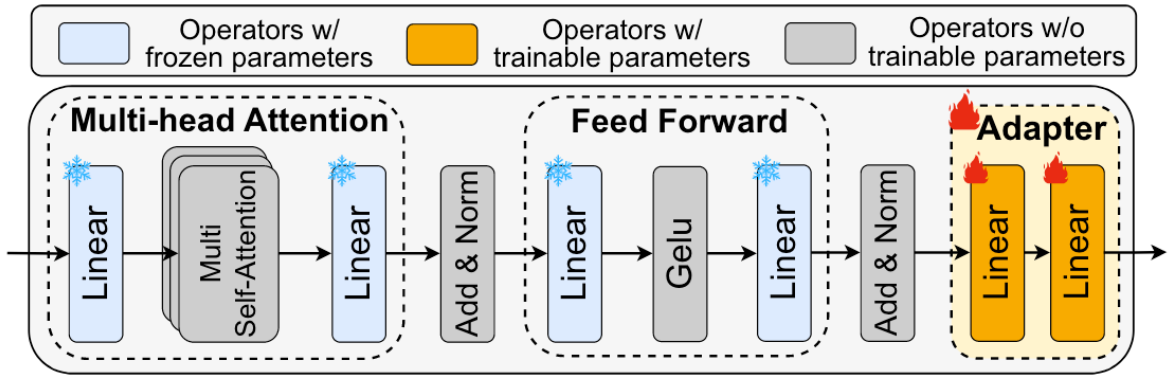
Techniques	Trainable Parameters	Memory Footprint (GB)			
		Weights	Activations	Gradients	Total
Full	737M (100%)	2.75	5.33	2.75	10.83
Adapters	12M (1.70 %)	2.80	4.04	0.05	6.89
LoRA	9M (1.26%)	2.78	4.31	0.04	7.13
Inference	/	2.75	/	/	2.75

Table 1: The breakdown of memory footprint. "Activations" contain the intermediate results and optimizer states. Model

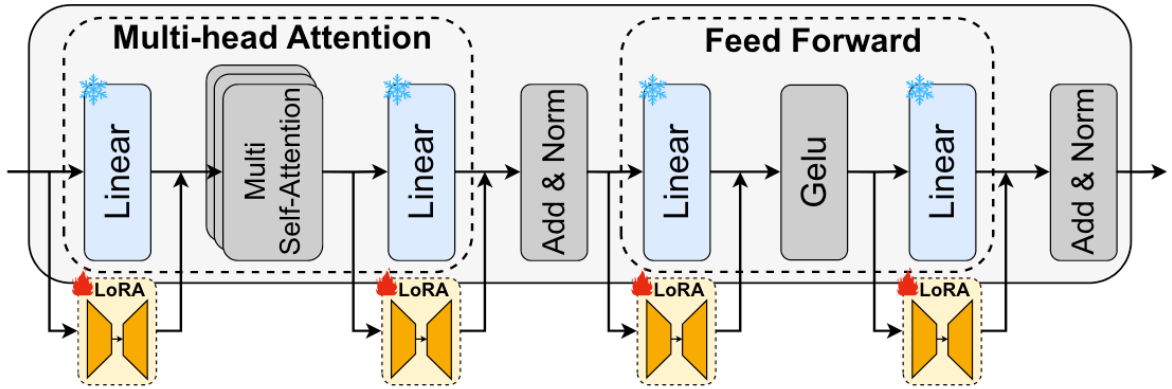
需要微调的**参数量**降低了**99%**，但所需的**算力**和**内存**资源仅仅减少了大约1/3！

边缘端协同大语言模型微调计算系统：Pluto and Charon

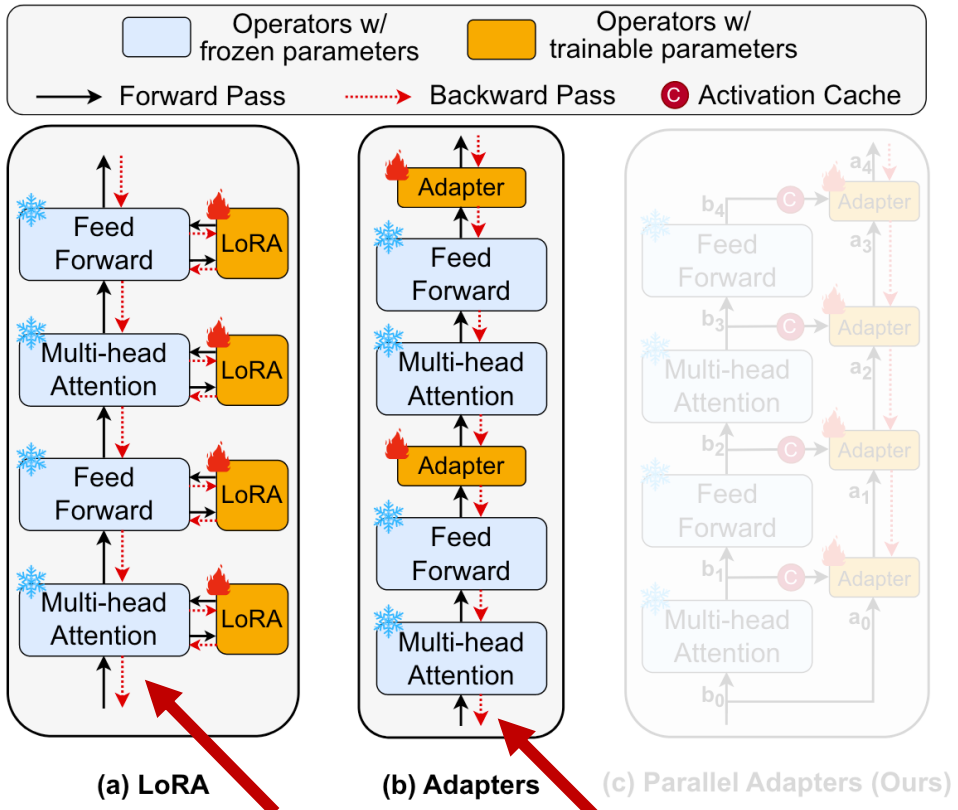
- 目前最主流的参数高效微调技术（例如：LoRA, Adapters），对于资源受限的边缘设备**还并不足够**的算力与内存资源高效。



(a) The transformer layer structure of Adapters.



(b) The transformer layer structure of LoRA.

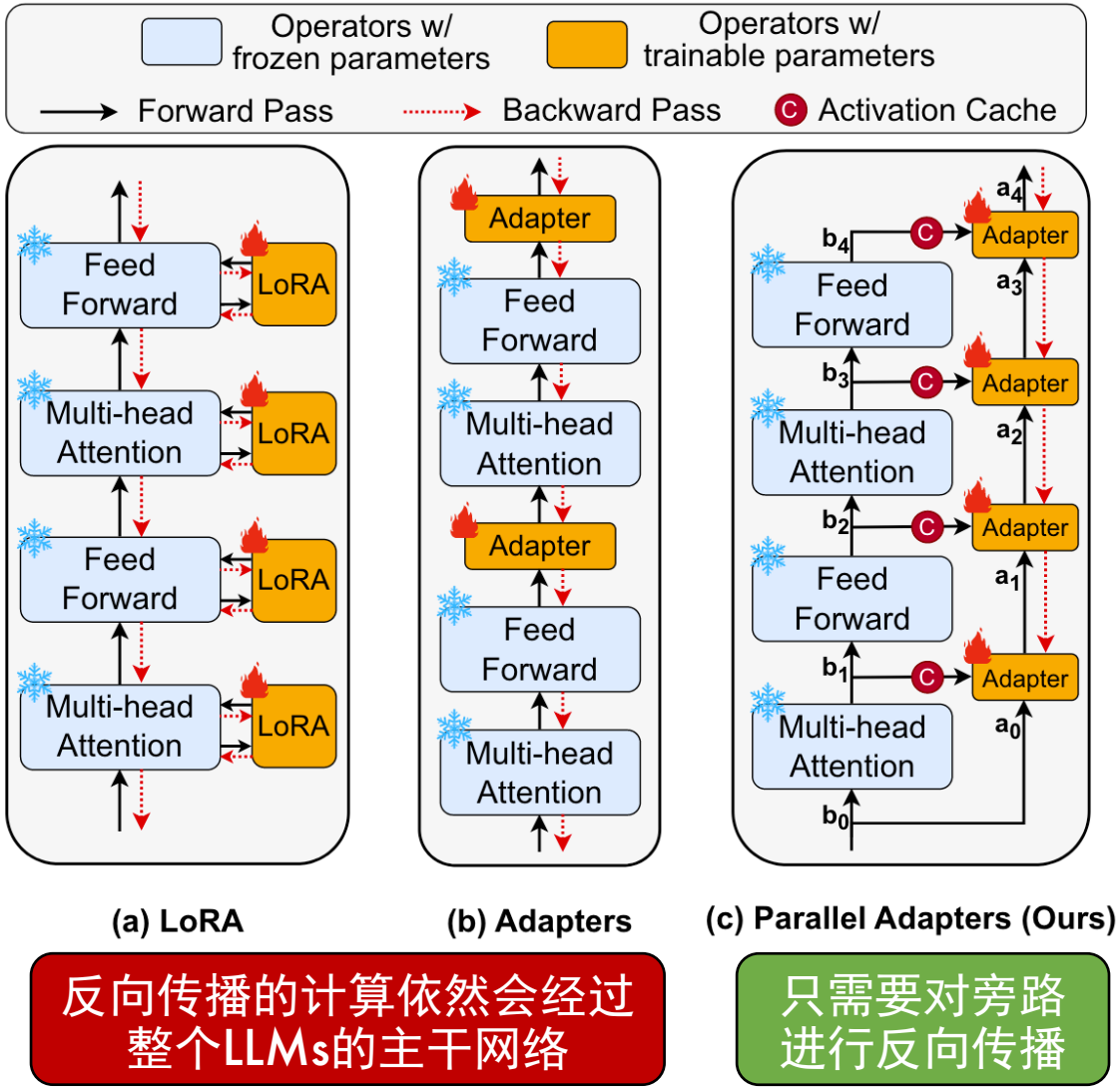


(a) LoRA (b) Adapters (c) Parallel Adapters (Ours)

反向传播的计算依然会经过完整的LLMs主干网络

边缘端协同大语言模型微调计算系统：Pluto and Charon

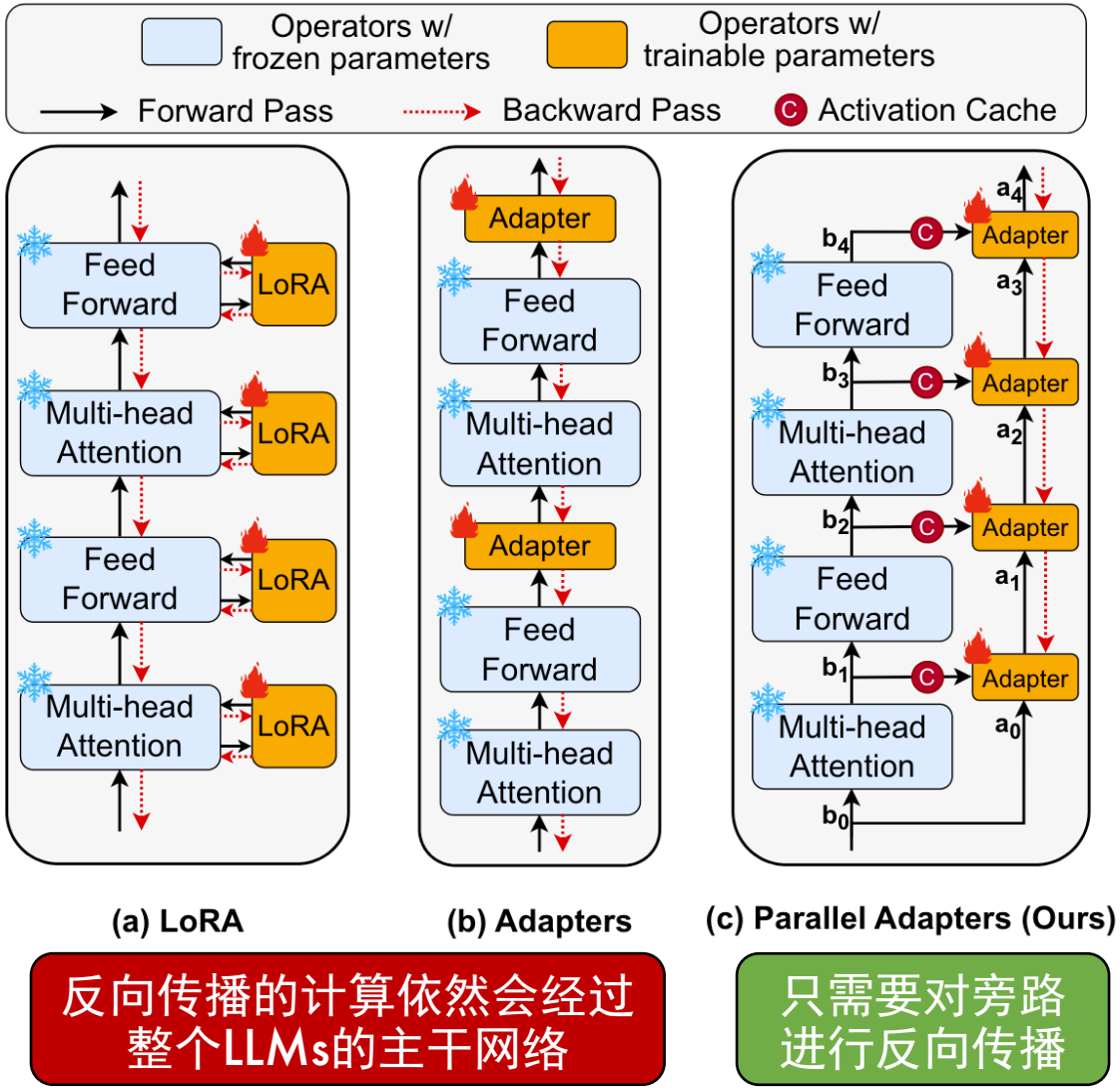
- **Parallel Adapters LLMs微调算法**
  - 大语言模型主干网络参数被冻结（如右图中蓝色模块所示）。
  - 在主干网络旁构建一个并行的轻量级子网络，用于参数高效的微调**（可以避免对主干网络进行反向传播！）**



边缘端协同大语言模型微调计算系统：Pluto and Charon

- **Parallel Adapters LLMs微调算法**
  - 大语言模型主干网络参数被冻结（如右图中蓝色模块所示）。
  - 在主干网络旁构建一个并行的轻量级子网络，用于参数高效的微调**（可以避免对主干网络进行反向传播！）**
  - 主干网络推理产生的激活结果可以被缓存下来以重用**（可以避免对主干网络进行前向传播！）**

避免了对LLMs主干模型进行前向传播和反向传播，显著降低了计算资源的需求！



反向传播的计算依然会经过整个LLMs的主干网络

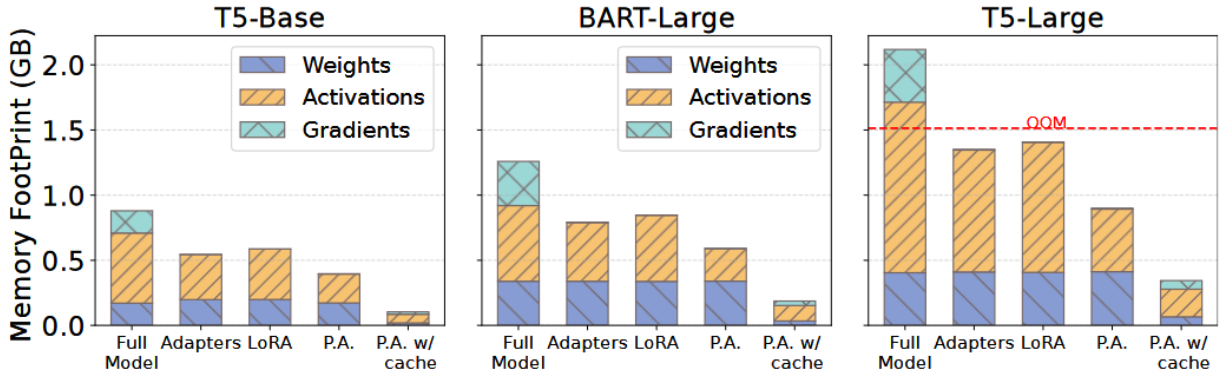
只需要对旁路进行反向传播

边缘端协同大语言模型微调计算系统：Pluto and Charon

- 端到端框架性能对比基线方法有最高**8.64倍**的微调加速

Full Model	Standalone	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.45	0.71	2.74	4.32	2.41	3.78	14.56	22.98	OOM	OOM	OOM	OOM
	EDDL	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Adapters	Standalone	1.21	1.9	7.29	11.51	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.39	0.61	2.35	3.71	0.54	0.85	3.27	5.16	2.75	4.31	16.59	26.19
	EDDL	0.34	0.53	2.06	3.25	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
LoRA	Standalone	1.21	1.89	7.28	11.49	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.41	0.64	2.45	3.87	0.55	0.87	3.33	5.26	2.73	4.28	16.48	26.02
	EDDL	0.31	0.48	1.86	2.94	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Parallel Adapters	PAC (Ours)	**0.14**	**0.22**	**1.34**	**2.12**	**0.29**	**0.45**	**2.69**	**4.25**	**0.69**	**1.09**	**8.88**	**14.02**

- 微调过程中的单机内存开销相比于全参数微调，最高可降低**88.2%**



边缘端协同大语言模型微调计算系统：Pluto and Charon

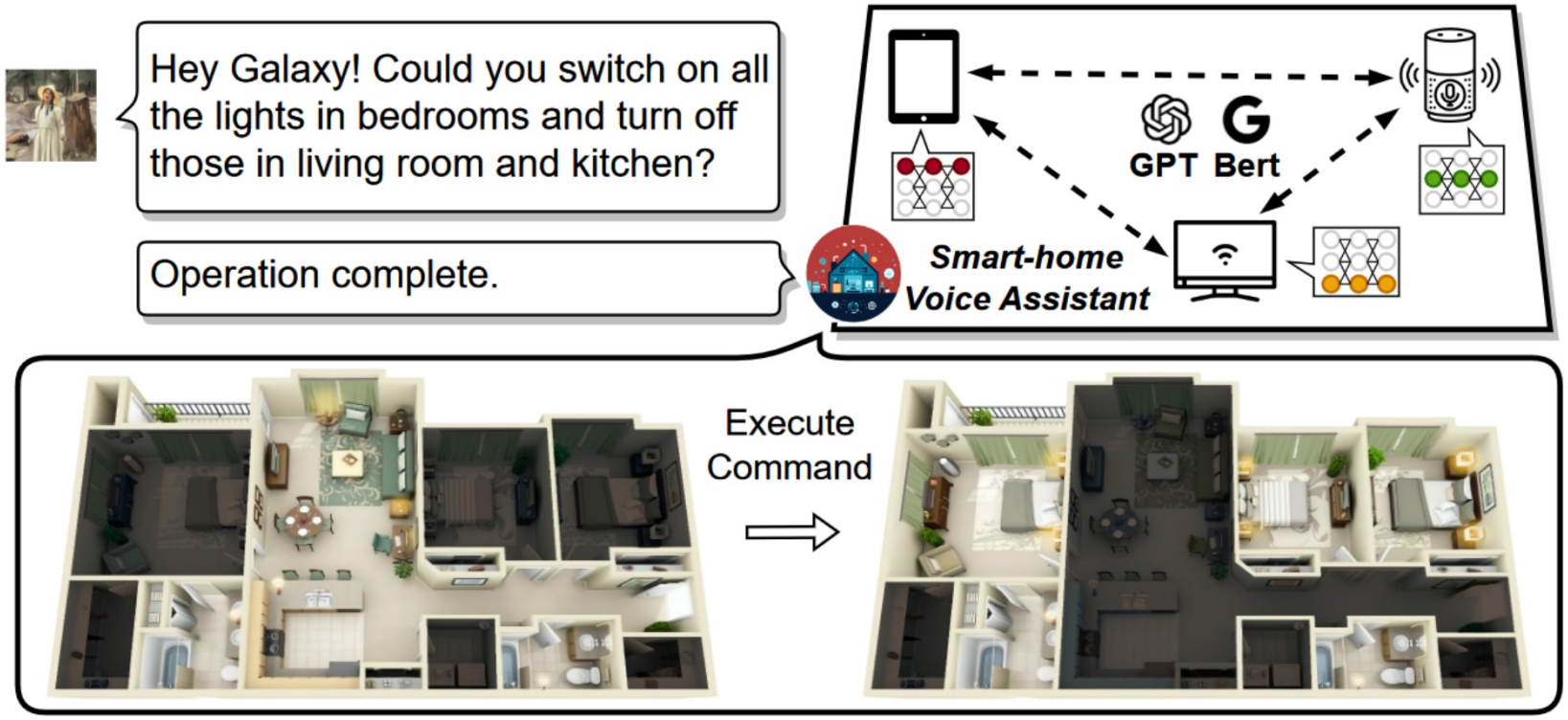
- 对比目前主流的LLMs微调算法（Full Model、Adapters、LoRA），可以取得**非常相近甚至更优**的微调后模型性能。

Table 3: Comparison of final performance between different fine-tuning techniques across four datasets. We report the average of F1 score and accuracy for MRPC. We use Pearson-Spearman Correlation as the metric for STS-B. For SST-2 and QNLI, we report accuracy. The mean value is the average performance of Full Model, Adapters and LoRA.

Fine-tuning Techniques	T5-Base				BART-Large				T5-Large			
	MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI
Full Model	89.71	90.94	94.03	93.08	88.16	91.10	95.64	94.40	92.78	91.08	95.30	93.30
Adapters	88.73	90.51	93.58	93.04	86.63	90.24	94.93	93.27	91.86	90.58	96.10	94.07
LoRA	86.27	90.73	93.69	93.30	87.46	90.36	95.23	94.48	90.27	92.08	95.53	94.18
Mean Value	88.24	90.73	93.77	93.14	87.42	90.57	95.27	94.05	91.64	91.25	95.64	93.85
Parallel Adapters (Ours)	88.24	90.43	93.46	93.25	87.71	90.54	95.25	93.68	91.7	91.57	95.76	93.7
Difference from Mean	+0.00	-0.30	-0.31	+0.11	+0.29	-0.03	-0.02	-0.37	+0.06	+0.32	+0.12	-0.15

边缘端协同大语言模型推理计算系统

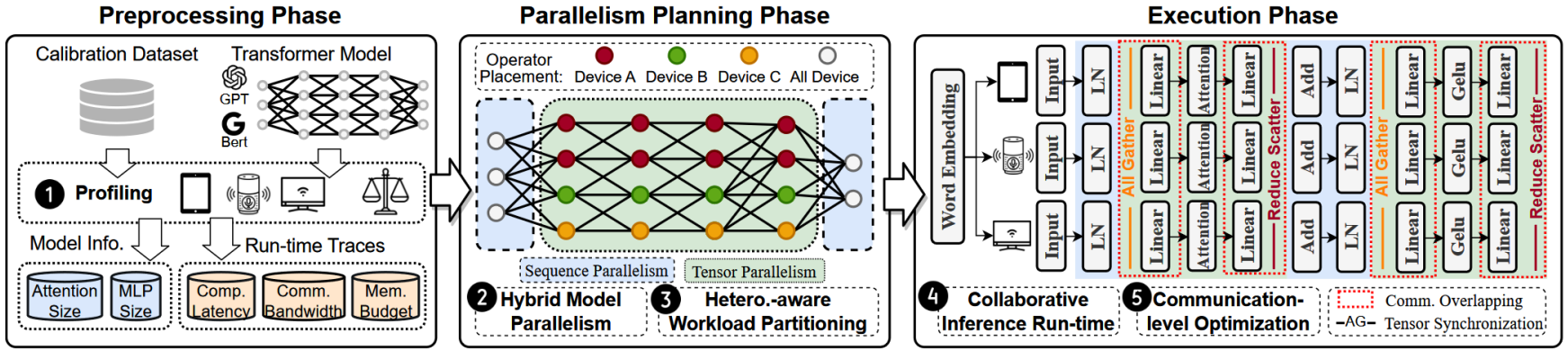
- 一个可能的边缘端协同的大语言模型推理系统的应用场景：



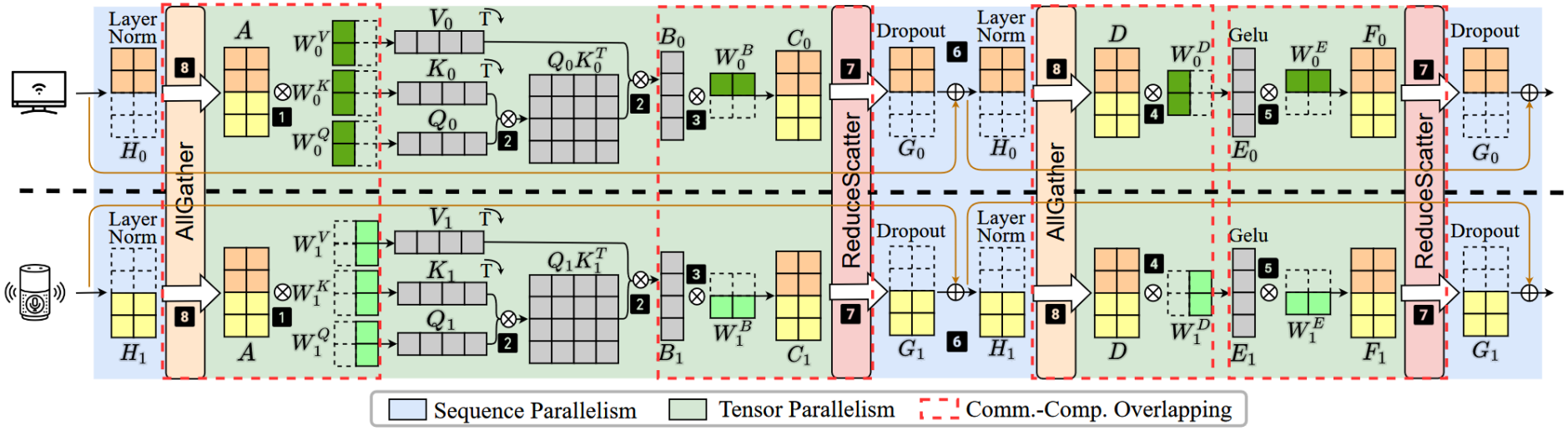
边缘协同计算赋能智能家庭示意图

边缘端协同大语言模型推理计算系统：Galaxy

- Galaxy针对**通用的Transformer架构**，包括基于编码器和解码器的模型

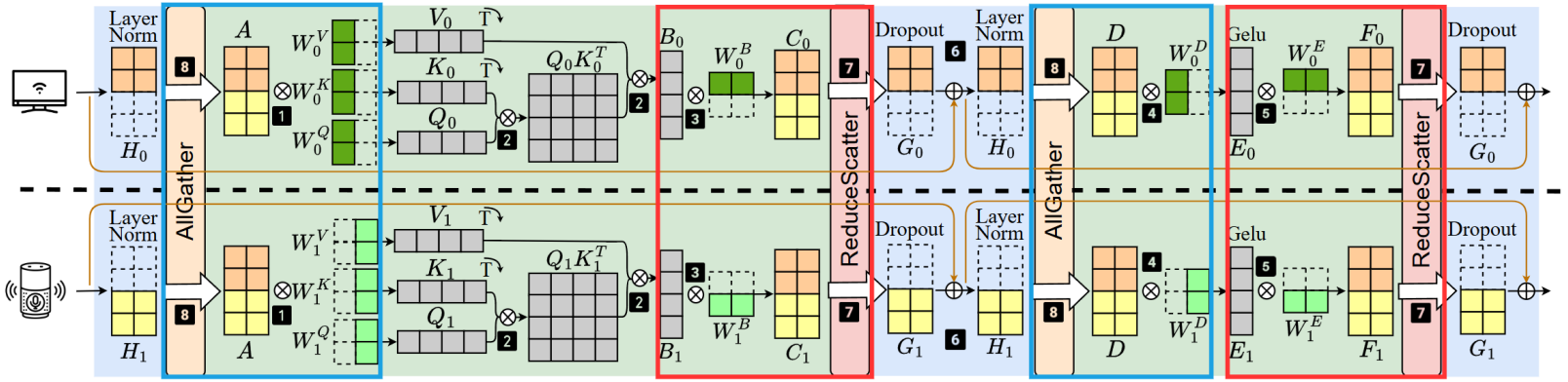


- 边缘场景通常考虑单条输入序列的推理加速，使用**张量并行和序列并行的混合并行架构**

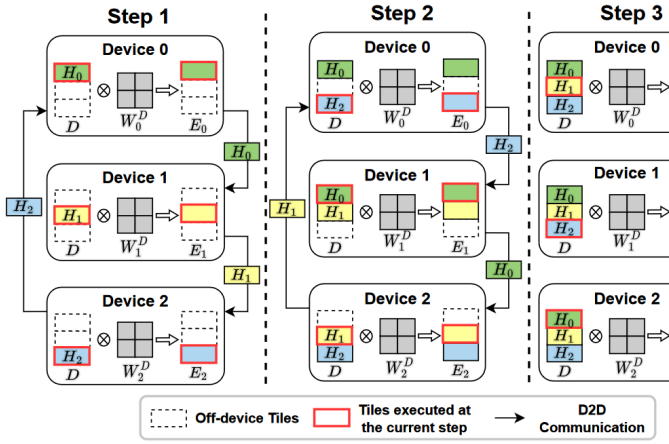


边缘端协同大语言模型推理计算系统：Galaxy

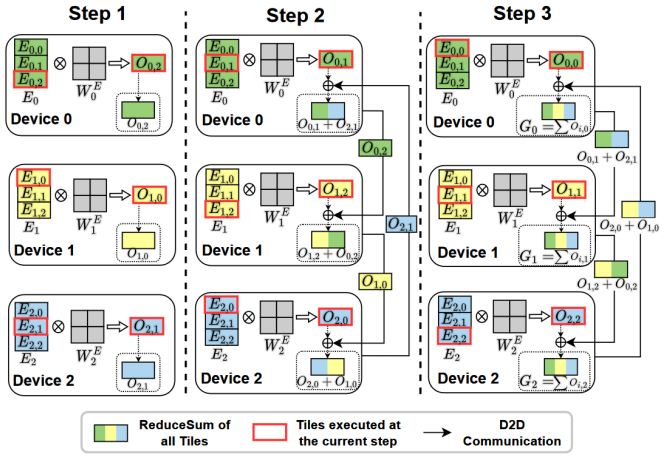
- 细粒度的通信计算重叠调度，优化集合通信的延迟开销。



AllGather 通信计算重叠　ReduceSactter通信计算重叠　AllGather 通信计算重叠　ReduceSactter通信计算重叠



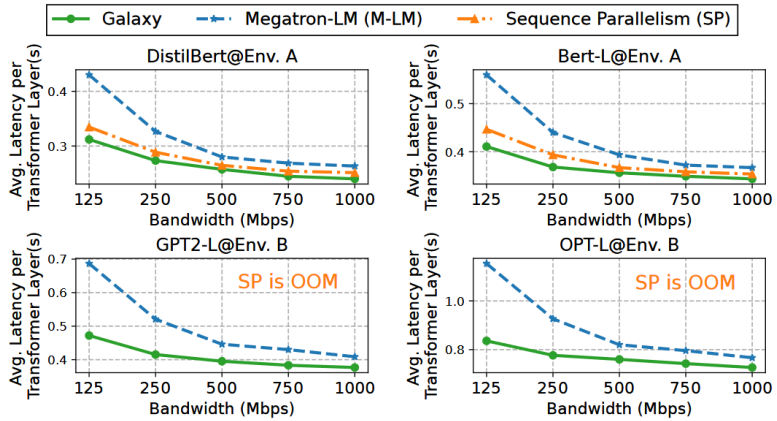
AllGather 通信计算重叠优化



ReduceScatter 通信计算重叠优化

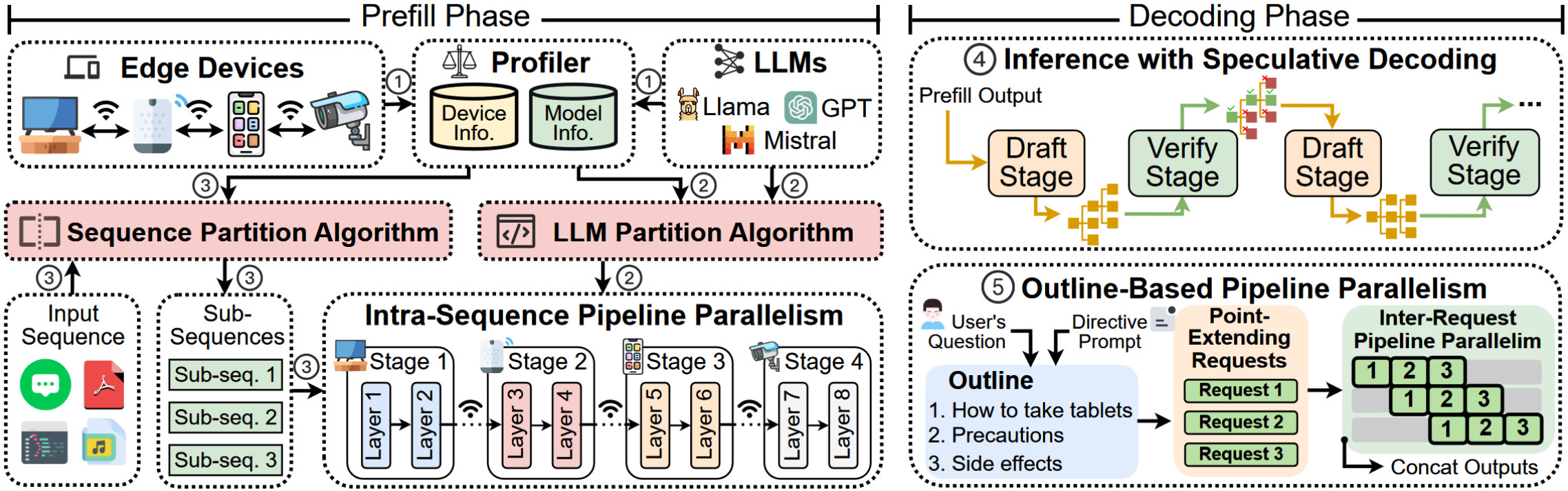
Galaxy在不同的网络条件下保持了高性能，相比于基线方法最高降低了46%的推理延迟

Model	Layers	Heads	Hidden Layer	Edge Env.	Speedup Over M-LM	Speedup Over SP
DistilBert [33]	6	12	768	A	1.37×	1.08×
Bert-L [1]	24	16	1024	A	1.36×	1.09×
				B	1.38×	1.11×
GPT2-L [12]	36	20	1280	A	1.31×	OOM
				B	1.46×	OOM
OPT-L [34]	24	16	2048	A	1.26×	OOM
				B	1.40×	OOM
				C	1.43×	OOM
OPT-XL [34]	32	32	2560	A	OOM	OOM
				B	OOM	OOM
				C	1.28×	OOM



边缘端协同大语言模型推理计算系统：Jupiter

- Jupiter针对的是**基于解码器**的生成式大语言模型边缘端协同推理。Jupiter不仅优化了**预填充阶段**（Prefilling Phase），同时还优化了**解码阶段**（Decoding Phase）

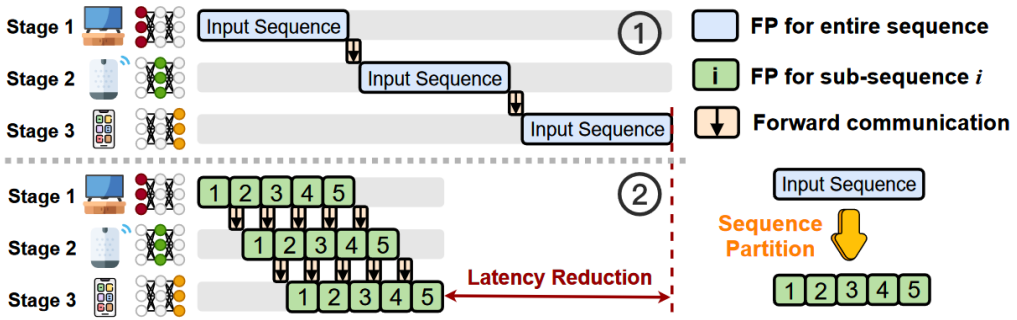


边缘端协同大语言模型推理计算系统：Jupiter

- **Prefilling:** Jupiter针对解码器模型的特点进行设计，采用了通信效率更高的流水线架构

COMM.-TO-COMP. RATIO OF VARIOUS PARALLELISM METHODS.

Model Name	Network Bandwidth	Communication-to-Computation Ratio				
		SP	TP	DT [7]	Galaxy [6]	Jupiter
Llama2-7B	100Mbps	8.16	6.96	3.48	5.19	**0.08**
	1Gbps	0.92	0.88	0.45	0.69	**0.01**
Llama2-13B	100Mbps	5.71	6.06	3.03	4.63	**0.05**
	1Gbps	0.73	0.81	0.38	0.56	**0.01**

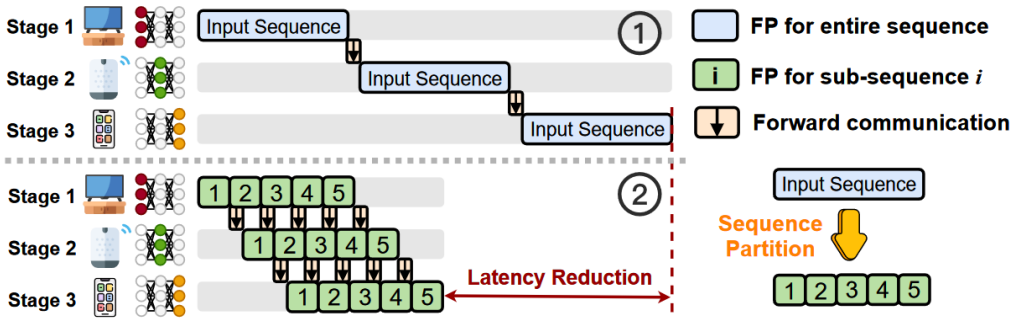


边缘端协同大语言模型推理计算系统：Jupiter

- **Prefilling:** Jupiter针对解码器模型的特点进行设计，采用了通信效率更高的流水线架构

COMM.-TO-COMP. RATIO OF VARIOUS PARALLELISM METHODS.

Model Name	Network Bandwidth	Communication-to-Computation Ratio				
		SP	TP	DT [7]	Galaxy [6]	Jupiter
Llama2-7B	100Mbps	8.16	6.96	3.48	5.19	**0.08**
	1Gbps	0.92	0.88	0.45	0.69	**0.01**
Llama2-13B	100Mbps	5.71	6.06	3.03	4.63	**0.05**
	1Gbps	0.73	0.81	0.38	0.56	**0.01**



- **Decoding:** Jupiter设计了一套适用于流水线架构的投机采样工作流，以加速自回归解码

自采样架构相比于大小模型协同的投机采样算法更加适合分布式流水线架构！

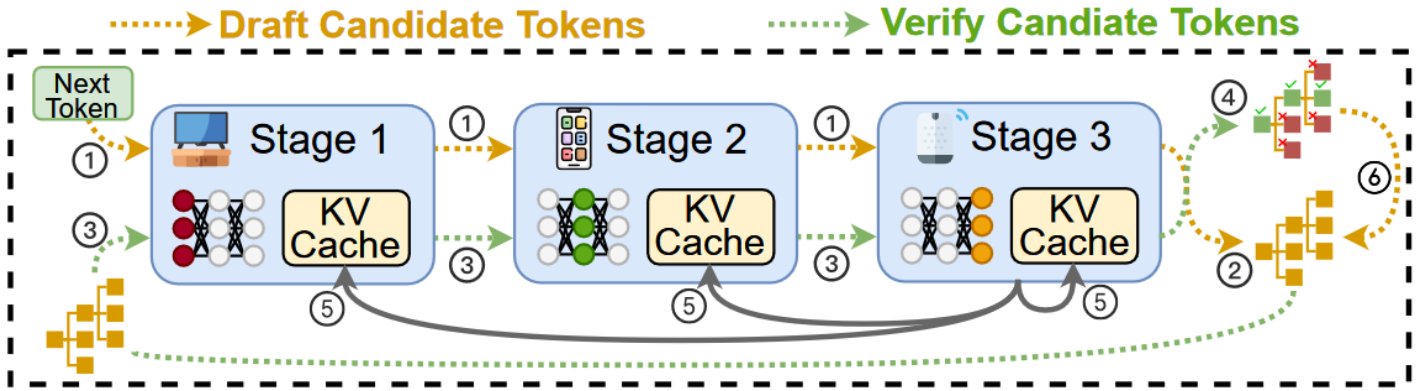


Fig. 8. A workflow of our collaborative inference with speculative decoding.

边缘端协同大语言模型推理计算系统：Jupiter

- **Decoding:** 基于大纲的流水线并行解码技术：

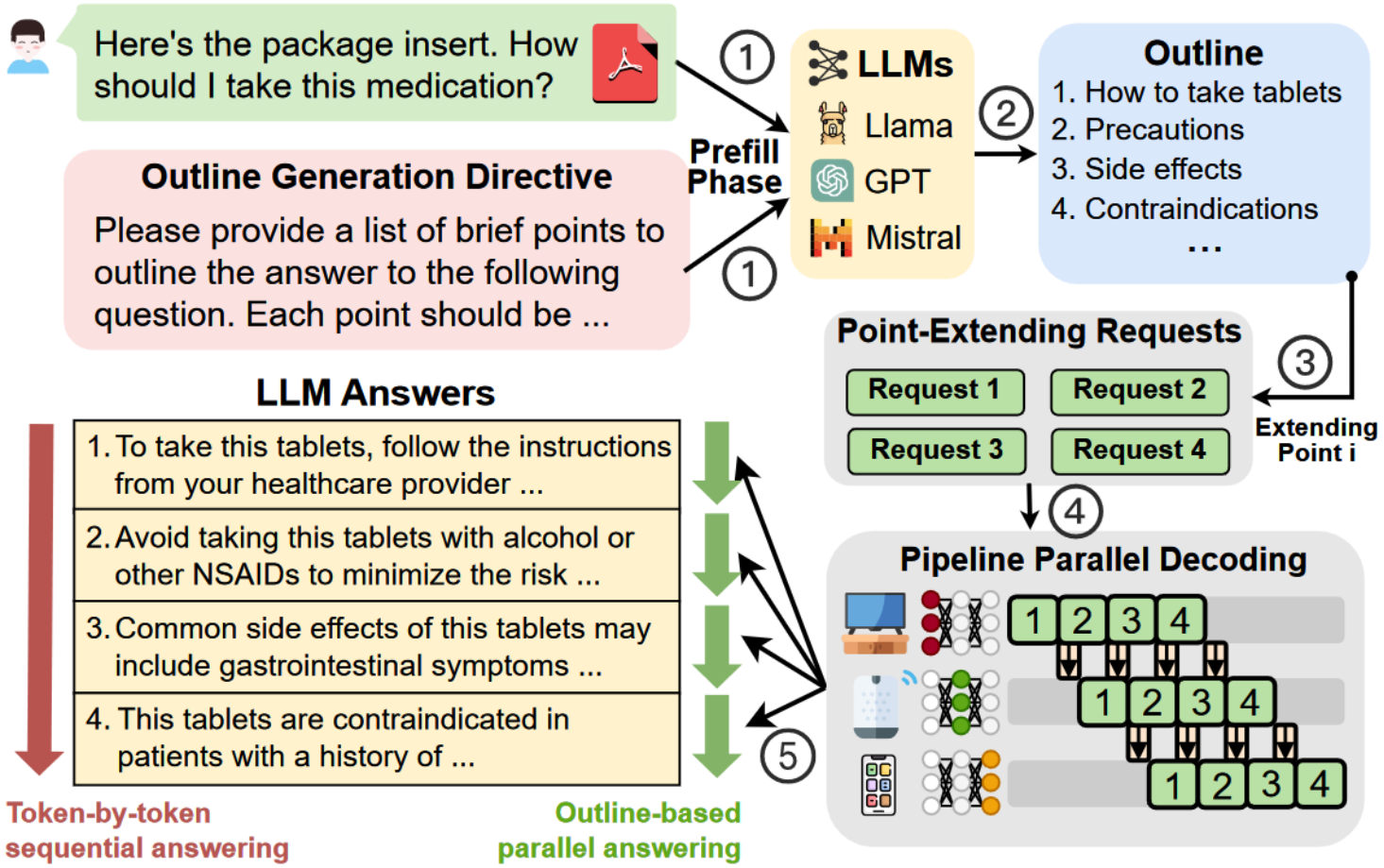


Fig. 9. An illustration of our outline-based pipeline parallel decoding.

边缘端协同大语言模型推理计算系统：Jupiter

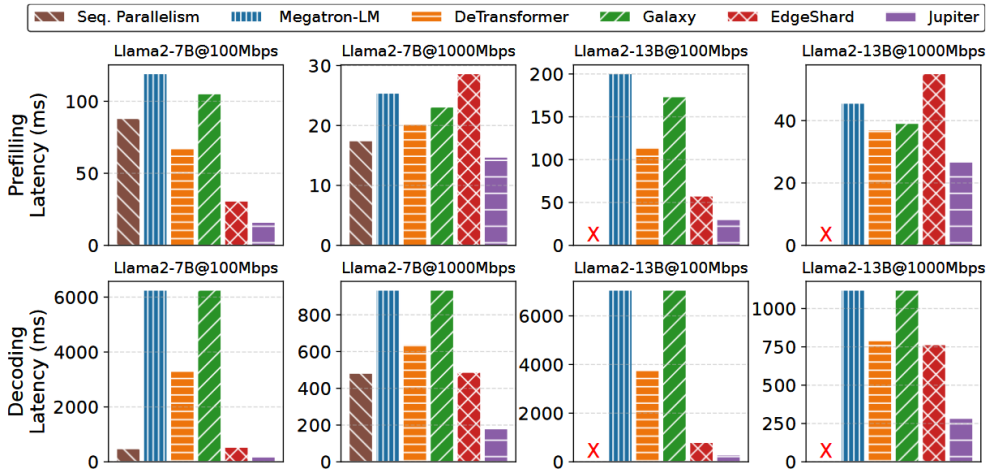
- Jupiter相比于基线方法优化了高达**26.1倍**的端到端生成延迟

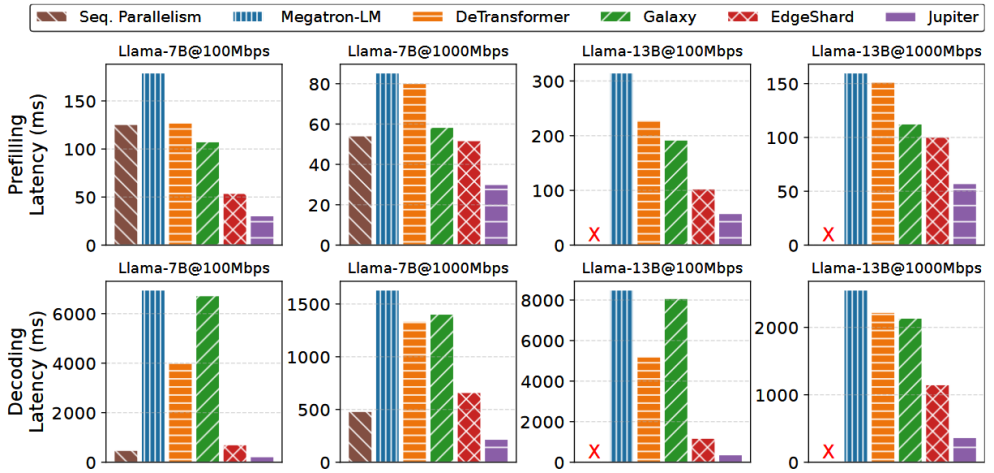
实验环境：
- ✓ 4 台Jetson Xavier NX设备
- ✓ Llama2-7B、Llama2-13B模型

生成任务：
- ✓ 平均输入序列长度: 260 tokens
- ✓ 最大生成长度: 64 tokens

基准方法：
1. Sequence Parallelism
2. Megatron-LM
3. DeTransformer
4. Galaxy
5. EdgeShard

Edge Environment	Network Bandwidth	Llama2-7B						Llama2-13B					
		SP	M-LM	DT	Galaxy	EdgeShard	Jupiter	SP	M-LM	DT	Galaxy	EdgeShard	Jupiter
Homo. Env. A	100Mbps	53.5	431.2	228.5	427.6	42.2	**16.5**	OOM	503.4	270.1	496.5	66.2	**26.3**
	500Mbps	37.4	106.9	66.4	103.9	39.0	**15.2**	OOM	130.1	83.4	125.0	63.4	**25.2**
	1Gbps	35.4	66.4	46.1	65.0	38.6	**14.9**	OOM	83.4	60.1	81.3	63.1	**24.9**
Hetero. Env. B	100Mbps	63.1	491.2	288.6	458.3	59.3	**22.4**	OOM	624.5	391.2	566.4	102.4	**38.8**
	500Mbps	47.0	167.0	126.4	142.9	56.1	**21.4**	OOM	251.2	204.5	208.0	99.7	**37.3**
	1Gbps	44.8	126.4	106.2	104.9	55.7	**20.9**	OOM	204.5	181.2	165.7	98.3	**36.8**





边缘端协同智能计算平台案例演示

使用四台英伟达Jetson Nano设备进行BERT模型的参数高效微调

使用四台英伟达Jetson Xavier NX设备协同推理Llama2-7B模型

感谢！