Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge

ASPLOS ’17, April 08-12, 2017

学生：屠晓涵

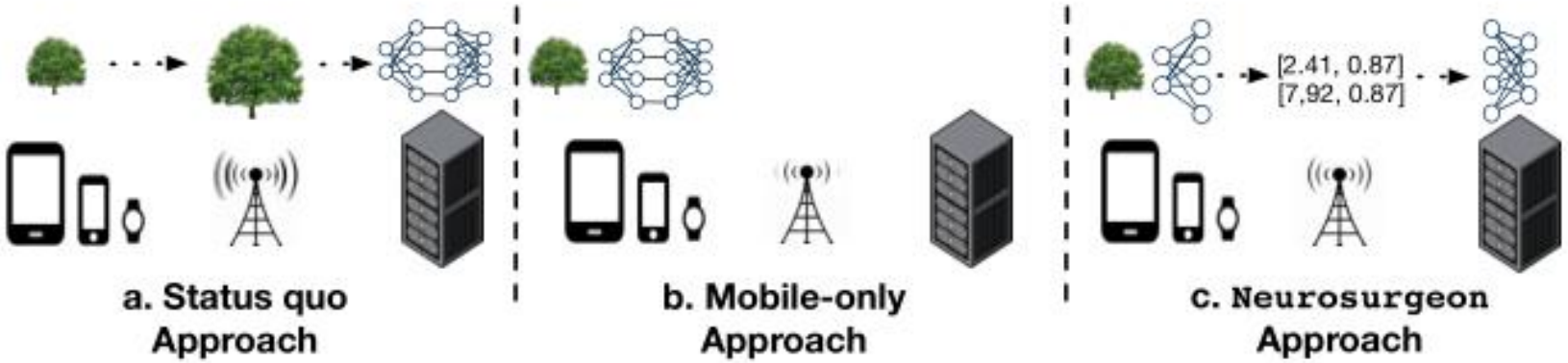
目录

- 背景和问题
- 主要工作
- 实验评估
- 总结
- 思考和后续工作

背景和问题

- 现状：对于所有使用深度学习技术来处理图像、视频而言，工业界普遍做法是利用云服务器上强大的GPU集群资源来完成应用程序的计算操作
- 导致：智能应用程序的计算能力完全依赖于Web服务商所提供的高端云服务器，移动设备（机器人）没有图像识别能力,构建的地图没有语义信息
- 问题：依赖云端，数据传输代价大，能耗大，延迟时间长

主要工作（三种方案）



- Status quo Approach: 在云服务器中远程执行所有计算（数据传输带来大代价）
- Mobile-only approach：在移动设备上执行所有计算（计算性能不如云服务器）
- Neurosurgeon approach：在云和移动设备间分割计算（轻量级的动态调度器）

三种方案的实验设备

Table 1: Mobile Platform Specifications

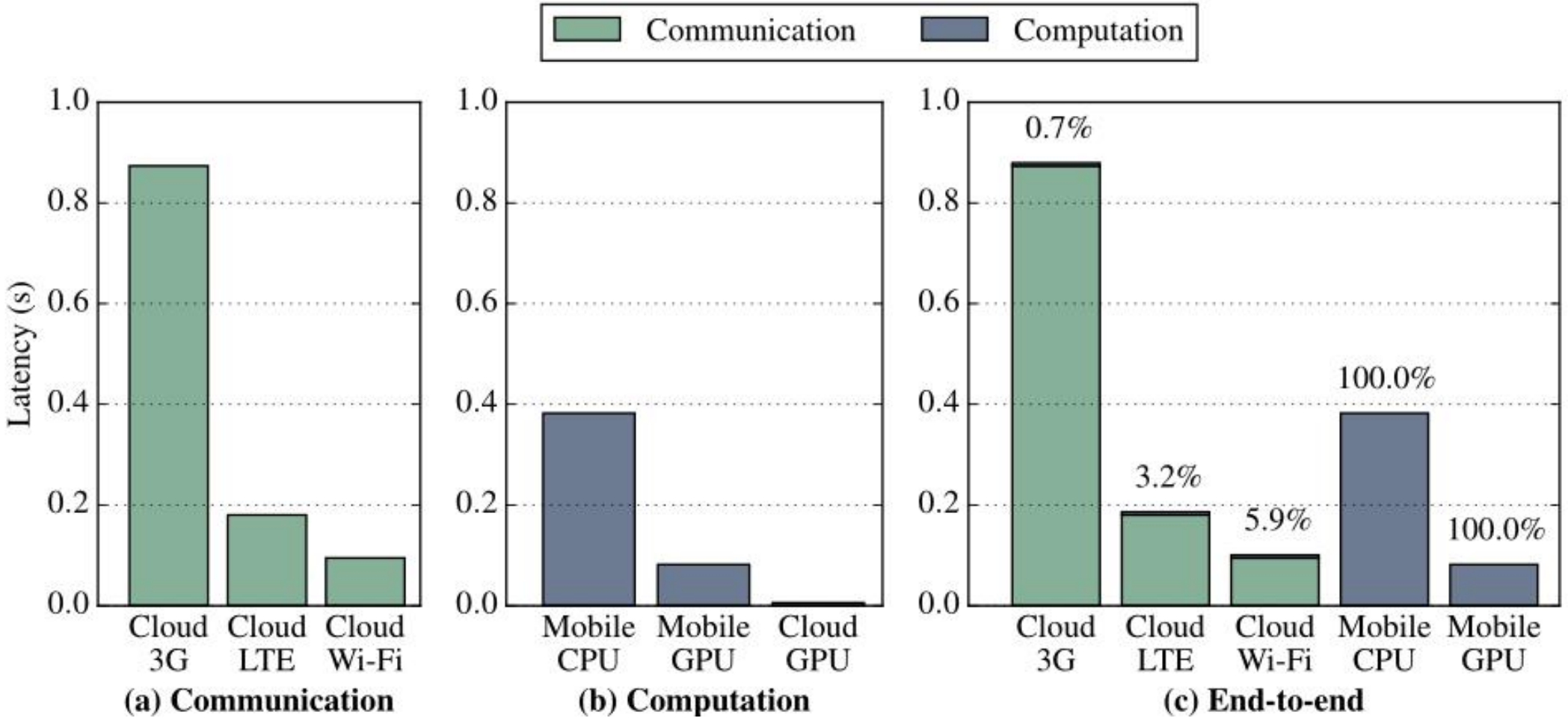
Hardware	Specifications
System	Tegra K1 SoC
CPU	4-Plus-1 quad-core ARM Cortex A15 CPU
Memory	2 GB DDR3L 933MHz
GPU	NVIDIA Kepler with 192 CUDA Cores

Table 2: Server Platform Specifications

Hardware	Specifications
System	4U Intel Dual CPU Chassis, 8 × PCIe 3.0 × 16 slots
CPU	2 × Intel Xeon E5-2620 V2, 6C, 2.10 GHz
HDD	1TB 2.5" HDD
Memory	16 × 16GB DDR3 1866MHz ECC/Server Memory
GPU	NVIDIA Tesla K40 M-Class 12 GB PCIe

- 分别运行AlexNet模型（一个深层次的CNN模型）
- 对比在云服务器上实施全部的计算操作和在移动设备上实施全部计算操作的能耗和延迟

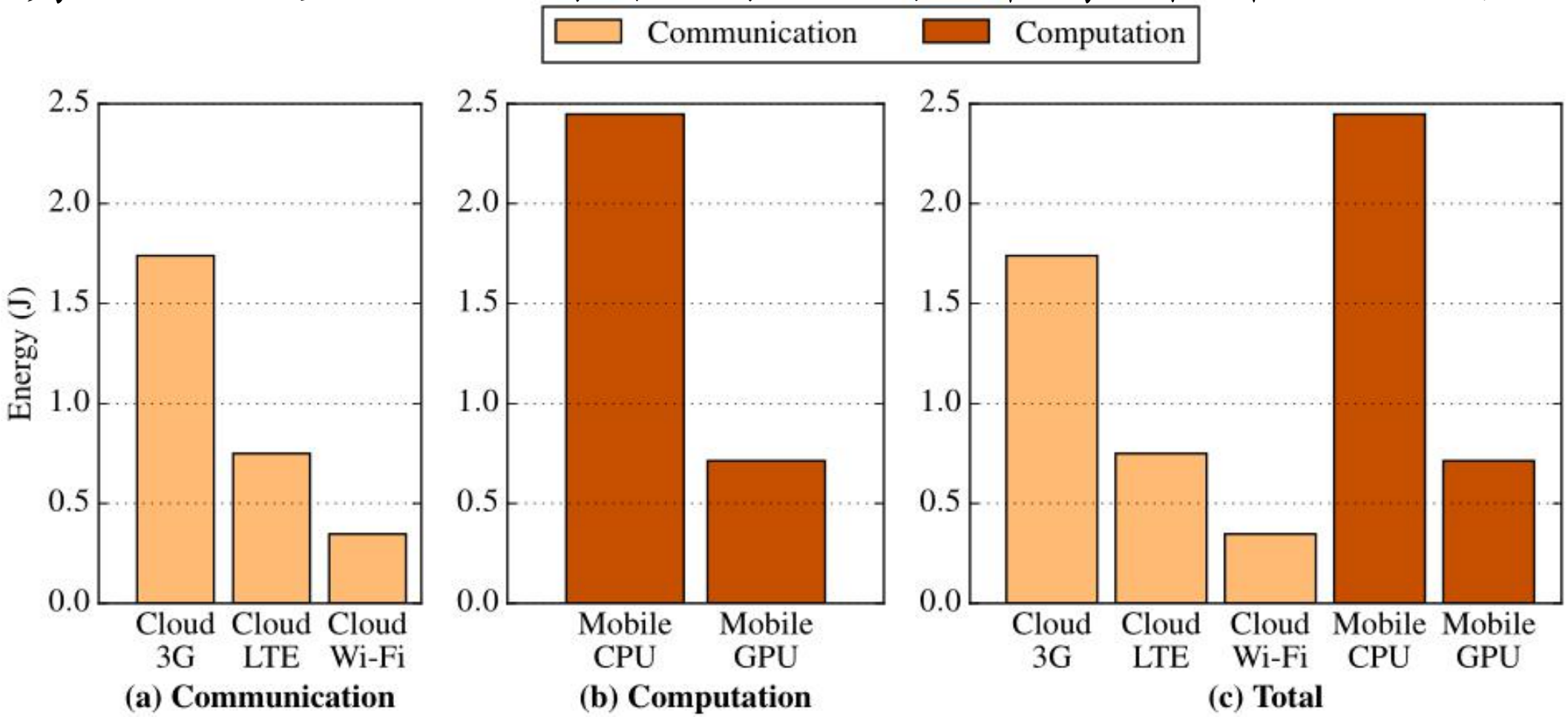
方案1和方案2的延迟时间计算：
云服务器、移动设备各自执行所有计算



- 通信延迟
- 计算延迟
- 端到端延迟

移动设备CPU处理图像的时间仍然比通过3G上传输入要快2.3倍

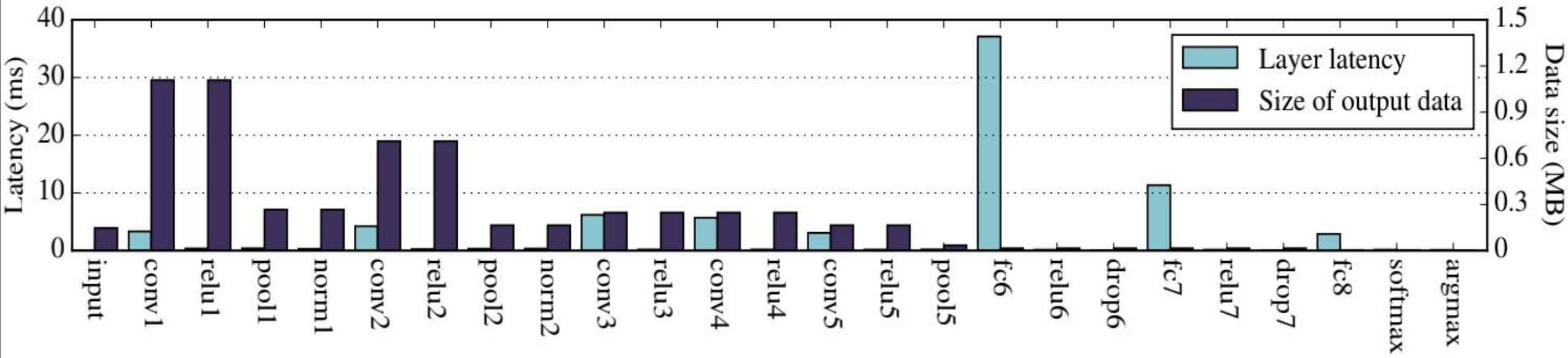
方案1和方案2的能耗计算：
云服务器、移动设备各自执行所有计算的能耗



尽管云服务器的计算能力要强于移动设备，但由于其需要进行数据的传输（在有的网络环境下的所带来的系统延迟时间和电量消耗量并不小），所以从系统的角度，完全使用云服务器进行计算的方法并不一定是最优的。

模型分析

以AlexNet为例,计算每层执行时间和数据大小



数据输出量随着层数增加而迅速递减。计算量在模型的中后部分有所增加，在全连接层时计算量达到了最高的幅度。

方案3

最佳分割点取决于模型中的拓扑层和结构层

Table 3: Benchmark Specifications

App	Abbr.	Network	Input	Layers
Image classification	IMC	AlexNet [21]	Image	24
	VGG	VGG [26]	Image	46
Facial recognition	FACE	DeepFace [27]	Image	10
Digit recognition	DIG	MNIST [28]	Image	9
Speech recognition	ASR	Kaldi [29]	Speech features	13
Part-of-speech tagging	POS	SENNA [30]	Word vectors	3
Named entity recognition	NER	SENNA [30]	Word vectors	3
Word chunking	CHK	SENNA [30]	Word vectors	3

最佳分割点随着模型的不同而变化着，因此需要一种系统能够对模型进行自动的分割并利用云服务器和移动设备进行相应计算。

方案3-自适应进行分割和优化影响因素

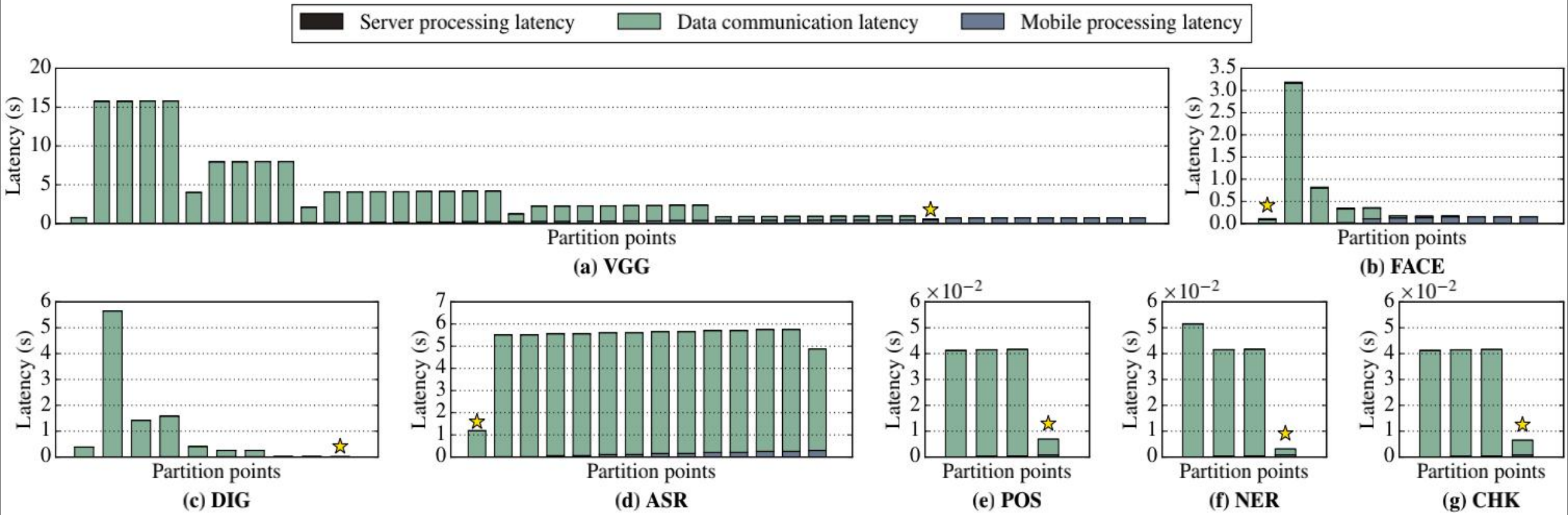
最佳分割点的影响因素：

- 静态因素，例如模型的结构；
- 动态因素，例如网络层的可连接数量、云服务器上数据中心的负载以及设备剩余可用的电量等。

- 需要自动地选择出DNN中的最佳分割点，以保证最终系统延迟时间和移动设备的电池消耗量达到最优的状态。为此设计智能切分模型的系统：Neurosurgeon

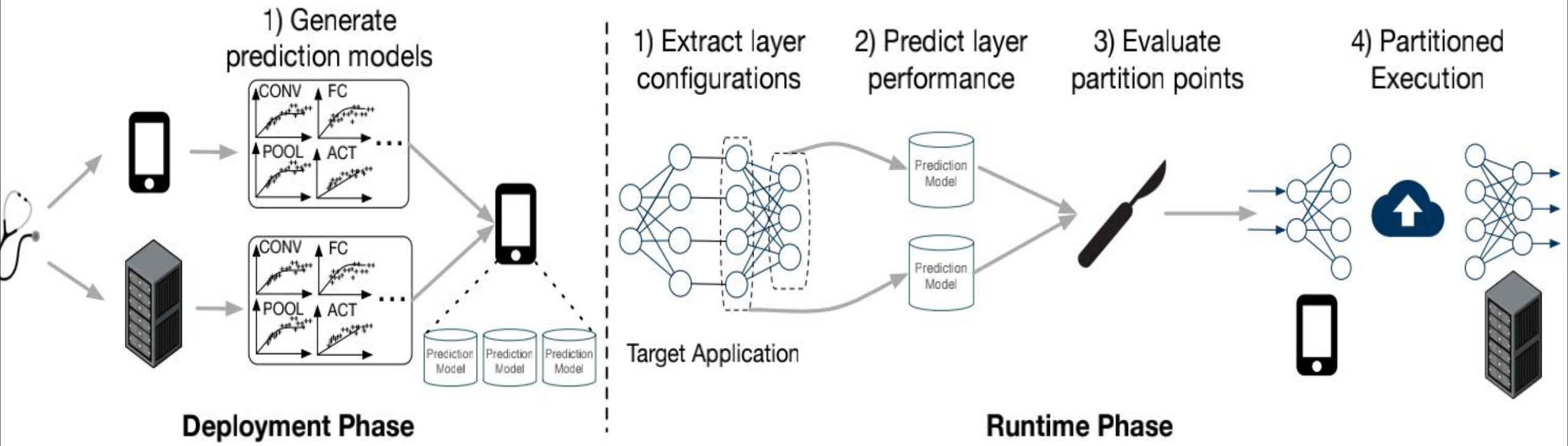
最少延迟的分割点选择

选择不同分割点时的端到端延迟



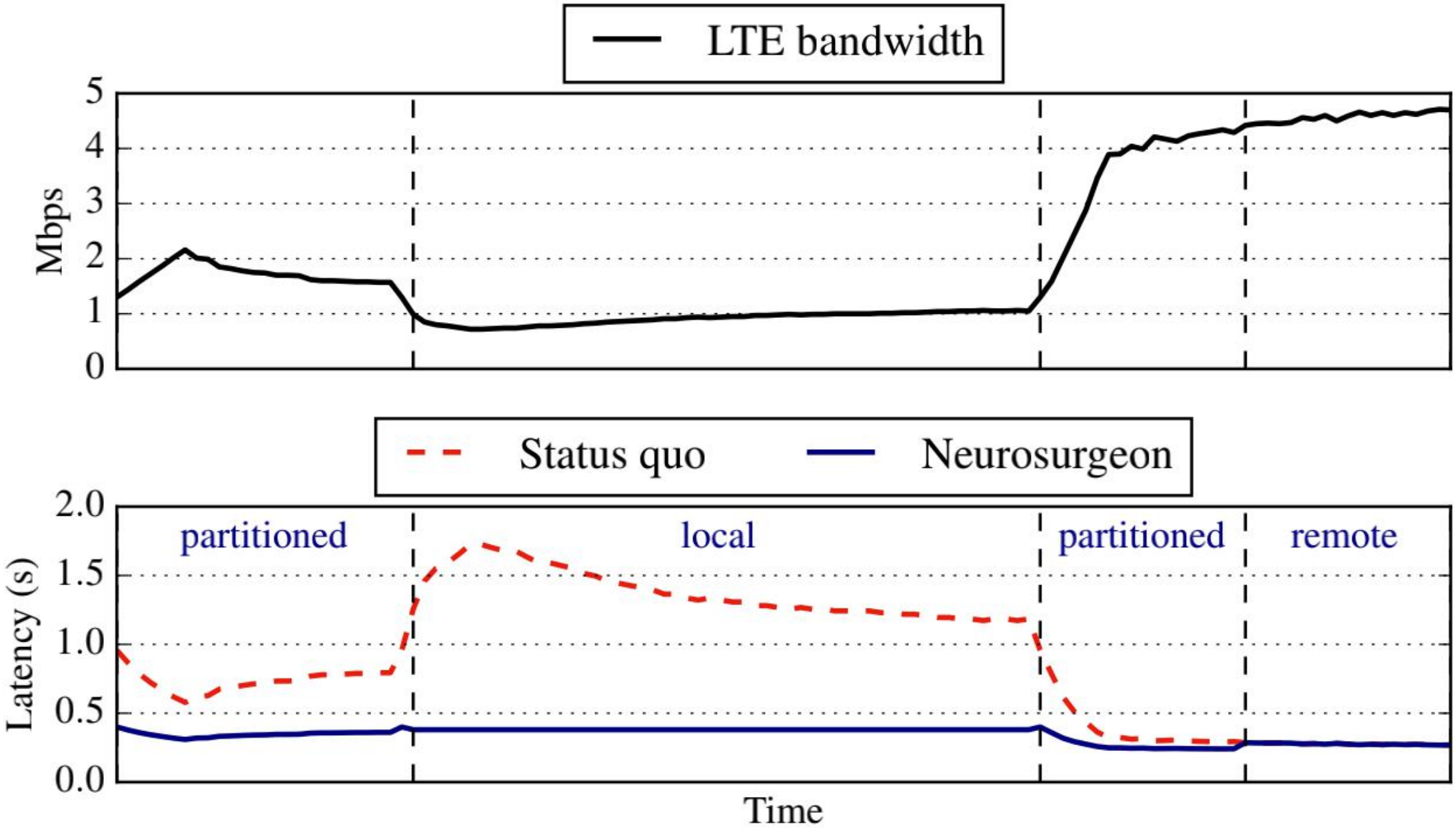
最少延迟的分割点均由星号标记

Neurosurgeon部署运行阶段

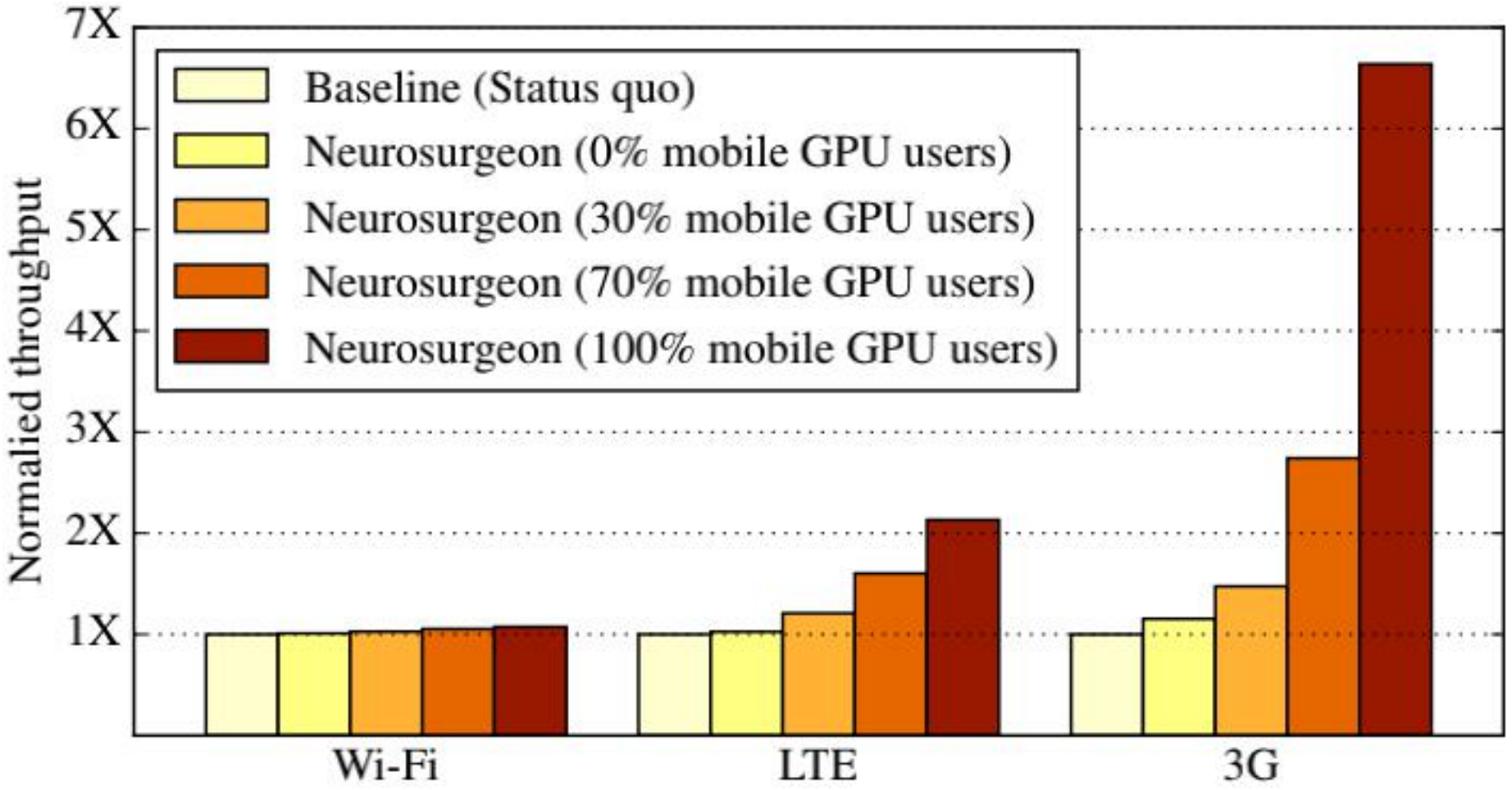


- 部署阶段-根据DNN模型中网络层的数量和类型来预测延迟时间和电量消耗，与具体DNN模型结构无关，在移动设备上创建和部署预测性能模型，不需要去执行具体的DNN模型。
- 运行阶段-Neurosuron会动态找到DNN的最佳分区点，步骤如下：1） Neurosuron分析和提取DNN架构各网络层的类型和配置；2）利用预测模型预测各网络层在移动设备和云服务器上的延迟时间和电量消耗情况；3）根据这些预测，结合当前的无线连接带宽和数据中心负载水平，Neurosurgeon选择最佳分割点 4） Neurosuron执行DNN，在移动和云之间进行分区工，实现端到端延迟时间或者电量消耗的最优化。

对无线网络变化自适应能力评估



对提高服务器吞吐量的评估



结论

- 将输入数据传输到服务器并远程执行算法并不总是最优的
- Neurosurgeon能对各种算法结构，硬件平台，无线连接有自适应能力，可以选择最佳的分割点，使系统延迟时间和移动端能耗最低
- Neurosurgeon能够根据服务器的负载情况作出适当的调整，可以选择最佳的分割点，使系统延迟时间和移动端能耗最低，数据中心吞吐量提高

思考和后续工作

- 算法没有优化，若在移动设备执行优化过的算法，数据中心平均请求查询服务时间将减少，吞吐量可提高
- Neurosurgeon利用网络层性能预测模型选择出最佳DNN分割点，回归模型简单，缺乏理论依据，可根据网络层自身配置及数据中心的负载情况进行理论分析，加上自己优化过的算法，效果可能更好
- 对于机器人：考虑关键级，关键的在移动端执行，不关键的大任务在云端执行