ASPLOS 2025 清華大學 Tsinghua University

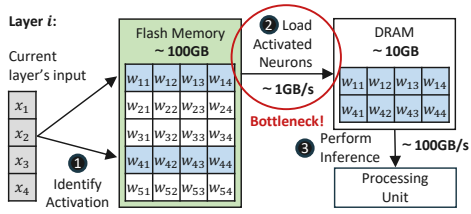
Neuralink: Fast LLM Inference on Smartphones with Neuron Co-Activation Linking

Tuowei Wang*, Ruwen Fan*, Minxing Huang, Zixu Hao, Kun Li, Ting Cao, Youyou Lu, Ju Ren+
Tsinghua University, Microsoft Research

关键词: Mobile Computing, Large Language Model, Model Sparsity, Parameter Storage

※ 研究背景

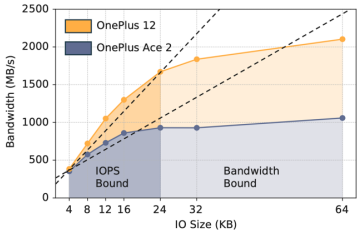
端侧大模型因其实时性、经济性、和安全性等方面的优势，正受到越来越多的关注。由于端侧设备（如：手机）的 DRAM 内存空间通常相对受限，端侧大模型部署通常采用“激活稀疏+异构存储”的方式，即将模型参数存放在 Flash 存储上，推理时动态加载部分参数到 DRAM 内存，利用模型稀疏性完成相应计算。这一方法缓解了内存压力，但是 I/O 通信成为了新的瓶颈。



Model	Compute	I/O	Total	I/O Ratio
OPT-350M	82 ms	776 ms	858 ms	90.4%
OPT-1.3B	202 ms	988 ms	1,190 ms	83.0%
OPT-6.7B	804 ms	2,224 ms	3,028 ms	73.4%
Llama-2-7B	609 ms	10,388 ms	10,997 ms	94.5%
Mistral-7B	540 ms	12,220 ms	12,760 ms	95.8%
MobiLlama-1B	230 ms	1,909 ms	2,139 ms	89.2%
Phi-2-2.7B	461 ms	1,976 ms	2,437 ms	81.1%

※ 核心观察

观察 1：端侧设备通常使用 UFS 作为存储介质，其 I/O 瓶颈主要来源于 I/O Operations Per Second (IOPS) 受限而非带宽受限。
观察 2：激活稀疏天然会引入大量不连续的访存模式，这使得 IOPS 受限的问题在端侧场景下尤为突出。
观察 3：激活稀疏下的模型神经元呈现出显著的共同激活现象，即某些神经元往往会与一组相对固定的其他神经元同时被激活。



不同读取长度下的 UFS 有效带宽

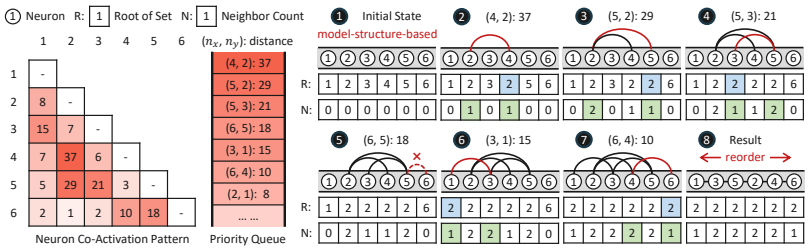
Ratio	dense	10%	20%	30%	40%
Bandwidth	1637.61	1355.35	1089.24	904.69	746.03
Latency	234.49	254.96	281.96	297.10	308.76
Speedup	-	0.92	0.83	0.79	0.76
Ratio	**50%**	**60%**	**70%**	**80%**	**90%**
Bandwidth	598.82	524.50	441.33	396.43	368.05
Latency	320.63	292.78	260.86	193.68	104.18
Speedup	0.73	0.80	0.90	1.21	2.25

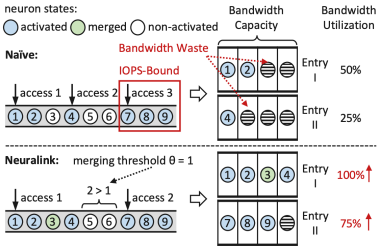
不同稀疏比下模型推理的时延和带宽，受有效带宽下降影响，稀疏比需要达到 80% 才有加速比

模型神经元共同激活矩阵

※ 关键方法

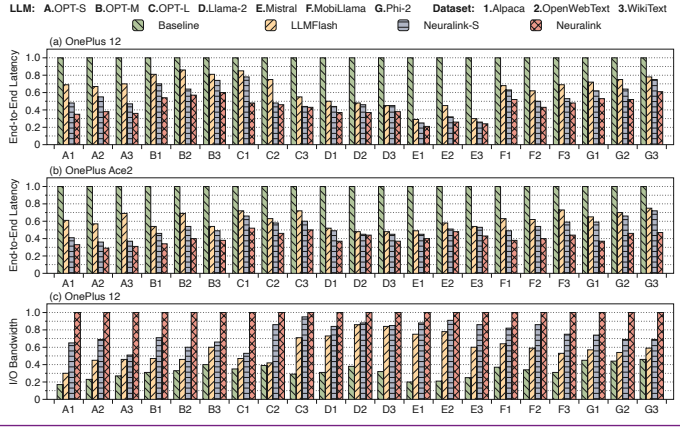
核心思想：修改存储上参数排布，将频繁被共同激活的参数排布在一起，从而提升访存连续性，减少 I/O 时延。
方法 - 离线阶段：将神经元抽象为图节点，统计神经元共同激活频率作为节点间距离，利用最短哈密顿路径算法寻找最优排布。
方法 - 在线阶段：灵活合并相近神经元链，进一步增加访存连续性；针对性设计缓存管理策略，优先缓存非连续神经元。





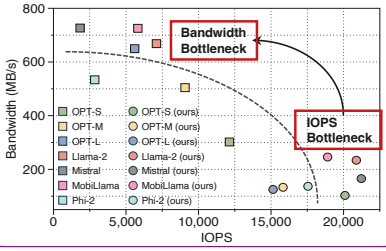
离线：神经元存储排布算法，将共同激活频次视为距离，启发式地寻找最短哈密顿路径。 在线：选择性连续读取相近神经元链。

※ 实验结果



实验结果：
(1) 3 个硬件设备，7 个大模型，8 个数据集
(2) 端到端延迟：较 llama.cpp 2.37×提升
(3) 带宽利用率：较 llama.cpp 3.28×提升

核心贡献：缓解端侧 IOPS 受限，逼近带宽上限。



作者邮箱：wtw23@mails.tsinghua.edu.cn 作者主页：https://pairshoe.github.io