SSD 卸载对 LLM 混合专家权重有害于能源效率

Kwanhee Kyung, Sungmin Yun, and Jung Ho Ahn, *Senior Member, IEEE*

摘要—大型语言模型（LLMs）应用专家混合（MoE）技术扩展到万亿参数，但需要巨大的内存容量，这激发了一线研究将专家权重从快速但较小的 DRAM（HBM）卸载到更密集的闪存 SSD。虽然 SSD 提供了具有成本效益的存储能力，但其每比特读取能耗远高于 DRAM。本文定量分析了在 LLM 推理的关键解码阶段将 MoE 专家权重卸载至 SSD 的能量影响。我们的分析比较了 DeepSeek-R1 等模型使用 SSD、CPU 内存（DDR）和 HBM 存储方案的情况，结果显示，将 MoE 权重卸载到当前的 SSD 会导致每生成标记的能量消耗显著增加（例如比 HBM 基准高出 ∼12× 倍），占据总推理能量预算的主导地位。尽管预取等技术可以有效隐藏访问延迟，但无法减轻这种基本能耗惩罚。我们进一步探索未来的技术扩展，发现 MoE 模型固有的稀疏性可能使 SSD 在能量上变得可行如果闪存读取能耗显著改进，大约提高了一个数量级。

***Index Terms*—大型语言模型，专家混合系统，推理系统，能源效率，闪存，DRAM。**

I. 介绍

大型语言模型（LLMs）通过增加更多的预训练权重参数来显著提升推理准确性 [1] 。然而，这一趋势也加大了对权重所需的内存容量以及每个代币生成时从内存加载的数据量的需求。因此，在实现高能效的同时保持高内存带宽已成为可持续、响应迅速的 LLM 推理服务的关键。典型的数据中心系统（如 NVIDIA DGX H100）采用包括设备内存（如 HBM）、CPU 内存（如 DDR）和基于 NAND Flash 的存储（SSD）在内的多层次内存架构。尽管 Flash SSD 在每比特容量成本上低于 DRAM，但它们本质上具有更高的读取能耗和更低的数据读取带宽 [2] 。

This work was partly supported by the Mobile eXperience (MX) Business, Samsung Electronics Co., Ltd and the MSIT (Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00456287) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Kwanhee Kyung, Sungmin Yun and Jung Ho Ahn are with Seoul National University, Seoul 08826, South Korea. E-mail: {kwanhee5, sungmin.yun, gajh}@snu.ac.kr. Jung Ho Ahn is the corresponding author.

专家混合（MoE）架构在 DeepSeek-R1、Mixtral 和 Llama 4 等知名模型中已广泛用于扩展模型参数同时控制训练期间的计算成本。在 MoE 模型中，专家权重构成了绝大部分参数（例如，在 DeepSeek-R1 中占 96.1%）。然而，在推理过程中，每个标记只有少数一部分专家被激活（在 DeepSeek-R1 中为 3.5%），导致稀疏访问模式。诸如 MoE-prefetching [3] 等技术通过将大量的专家权重卸载到 SSD 或 CPU 内存来利用这一点，旨在通过将其与活跃专家的计算重叠来隐藏数据传输延迟。特别是，虽然可以隐藏延迟，但在这些数据传输过程中消耗的能量并未减少，尤其是在访问能耗高的 SSDs 时更是如此。由于稀疏性在自回归解码阶段最为明显（其中一次生成一个标记），本研究重点关注分析此阶段的能量影响。

我们的分析表明，将 MoE 权重卸载到 SSD 显著增加了能源使用。在 DeepSeek-R1 上进行推理时，如果专家权重存储在 SSD 中，则每标记的能耗比保持在 HBM 中多大约 4.9×，比卸载到 CPU 内存多 3.1×（图 1）。在这种情况下，访问 SSD 上的权重所消耗的能量占总每标记能量的惊人比例达 80%，凸显了能源惩罚的严重性。

尽管目前存在这一缺点，但如果未来的 Flash 技术能够实现更好的能效（即使仍高于 DRAM），MoE 稀疏性和 SSD 容量的结合可能会提供一种比小型密集模型更高效地运行大型高精度模型的方法。因此，这项工作提供了对 MoE 模型中逐层能耗和执行时间的定量分析。我们根据权重存储位置比较了能源消耗和性能，证明了 SSD 访问的高能耗成本无法被隐藏延迟的技术所掩盖。此外，我们预测潜在的 Flash 能效改进对 SSD 卸载可行性的影响。

II. 能量比较：DRAM 与 Flash

NAND Flash 存储器的读取操作每比特消耗的能量与 DRAM 相比更多。这种差异源于底层技术：Flash 需

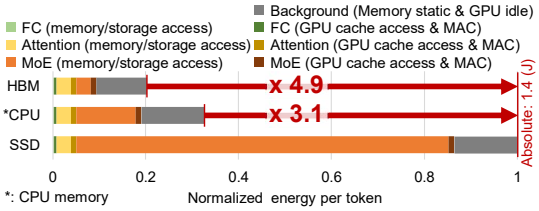

图 1. DeepSeek-R1 在预填充阶段后的第一次解码阶段中，每个标记的标准化能耗情况，对比了将 MoE 权重存储在 HBM（基线）、CPU 内存（DDR）和 SSD 上的情况。批处理大小为 1024，且每次请求包含 1024 个输入提示标记。

要施加相对较高的电压并使用复杂的机制来感测单元的阈值电压，而 DRAM 依赖于电容电荷的低电压差分检测 [4], [5]。Flash 这种较高的读取能量特性显著增加了将 LLM 权重卸载到 SSD 时观察到的能量消耗。

图 2比较了在一个代表性服务器系统（详细规格见第 IV节）中访问存储在设备内存（HBM3）、CPU 内存（DDR5-7200）和 NVMe SSD 中的数据时的能量成分（以 pJ/b 为单位）。CPU、GPU 和 SSD 通过 NVLink 互连，类似于 NVIDIA 的 Grace Hopper Superchip [6]。这些数值包括了存储/内存芯片内部消耗的能量以及外部 I/O 路径中消耗的能量。具体而言，内部 Flash 读取能量是基于 [5] 将每次读操作的能量除以页大小得出的。内部 DRAM 能量使用行业数据表 [4], [7] 和之前的研究工作 [8] 计算得出，汇总了数据传输和激活/预充电能量。外部 I/O 能量包括 NVLink 能量和内存接口能量（从数据表中计算得出或基于先前的工作进行缩放 [9]）。

在 LLM 权重完全驻留在 GPU 的 HBM 基线中，访问这些权重仅消耗 HBM 读取能量，估计为 4.2pJ/b。相比之下，即使使用如 GPUDirect Storage [10] 等技术将权重卸载到 SSD 并绕过中间复制到 CPU 内存的操作序列，也会涉及一系列操作。这包括 Flash 读取操作本身、数据写入 HBM 以及最终从 HBM 读取以供计算。总能量变为 102.4pJ/b（Flash 读取）+ 4.2pJ/b（HBM 写入）+ 4.2pJ/b（HBM 读取）= 110.8pJ/b，与仅使用 HBM 的基线相比，权重访问能量大幅增加了 ∼26 倍。

III. 大语言模型、MoE 和 MoE-预取

大语言模型： 现代的 LLM 主要由堆叠的解码器层组成。每个解码器通常包括：1) 使用查询、键和值投影来计算标记之间关系的注意力层；2) 由门控投影、上投影、激活和下投影线性变换组成的前馈网络（FFN）层；3) 层规范化和支持操作，如残差连接（参见图 3(a)）。LLM 推理主要分为两个阶段：一个预填充阶段并行处理输入

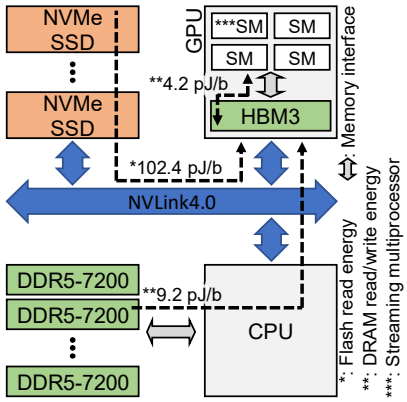

图 2. 设备内存、CPU 内存和 SSD 中存储的数据在读/写操作期间的能耗。

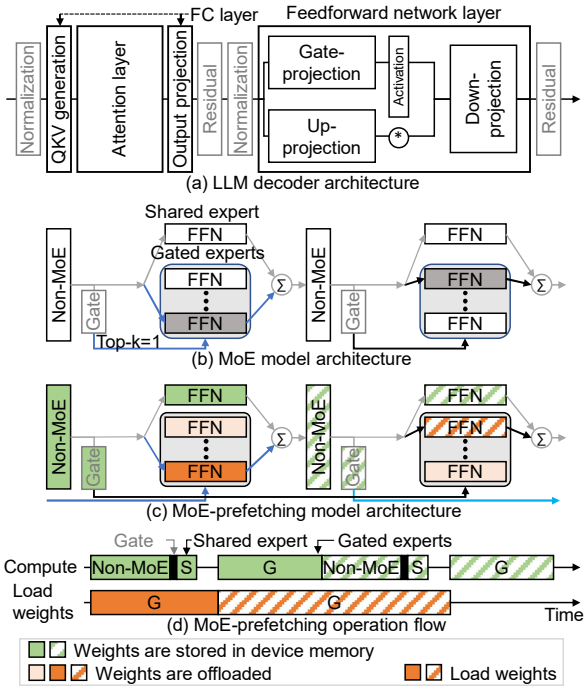

图 3. (a) 大型语言模型解码器层的高级结构。(b) 带有门控和多个专家的 MoE 层。(c) 将门控专家权重卸载后的 MoE 模型。(d) MoE 预取操作，将计算与后台权重加载重叠。

提示标记（L_{in}），以及一个解码阶段，该阶段使用先前生成的标记作为输入自回归地逐个生成输出标记（L_{out}）。

混合专家模型： MoE 架构（例如，Mixtral、DeepSeek-R1 和 Llama 4）用 MoE 层替换了大多数 FFN 层。一个 MoE 层包含多个并行的 FFNs，称为“专家”（N_{ex}）。对于每个输入标记，一个轻量级的“门控”网络动态选择这些专家中的一个小子集（top-k）作为“选定专家”。该标记仅由所选专家（以及所有标记处理的一些“共享专家”）进行处理，并且它们的输出通过根据门控值确定的加权和进行组合（参见图 3(b)）。该令牌依赖的专家激活导致稀疏性：每个令牌仅访问总 MoE 权重的一部分（例如，Mixtral 8x7B、DeepSeek-R1 和 Llama

表 I
推理系统和模型配置

Model	Param.	# Node	# GPU per node	# FFN decoder	# MoE decoder	N_{ex}	top-k
Mixtral	47B	1	4	–	32	8 (*G)	2
DeepSeek-R1	671B	4	8	3	57	256 (G) + 1 (**S)	8
Llama 4	400B	1	8	24	24	128 (G) + 1 (S)	1
Llama 3.3	70B	1	8	80	–	–	–

*G: 带门控的专家　**S: 共享专家

4 Maverick 分别为 ∼25.0%、∼3.5%和 ∼1.6%，见表 I）。然而，在批量处理请求时，会激活更多的专家，因为不同的令牌可能会选择不同的专家。

MoE-预取： 为了管理 MoE 权重的大内存占用，预取技术 [3] 将一组（通常是较大的）门控专家权重卸载到内存层次结构的较低层级，如 CPU 内存或 SSD（参见图 3(c)）。这些涉及微调第 i 个解码器的门层（dec_i），以预测在后续解码器中需要哪些专家，dec_{i+1}。当 dec_i 和 dec_{i+1} 的初始层正在执行时，系统会并行地从卸载存储（CPU 内存或 SSD）中获取 dec_{i+1} 的预测专家权重到 GPU 的设备内存（HBM）中。这种重叠旨在隐藏数据传输延迟（参见图 3(d)）。

IV. 实验设置

推理系统建模： 我们使用了一个建模工具 [1]—该工具是从 [11] 中介绍的模拟器修改而来的—它包含了当 MoE 预取操作重叠时，CPU 内存/SSD 读带宽及其对 HBM 读带宽减少的影响。计算节点镜像了 DGX H100，配备最多八个 NVIDIA H100 SXM5 80GB GPU，并通过 NVLink4.0（每方向 450GB/s）相互连接。在多节点设置中，节点通过 InfiniBand 使用八个端口进行互连，每个端口提供每方向 50GB/s 的带宽。每个 GPU 关联有 256GB 的 CPU 内存（DDR5-7200）和 NVMe SSD（参见图 2）。从 CPU 内存和 SSD 到 GPU 的聚合读取带宽匹配了 NVLink 互连的单向带宽，分别利用 DMA 技术和 GPUDirect Storage 等技术。

模型和配置： 所有实验均采用 BF16 数据类型进行模型权重和激活。第 V-A节专注于 Mixtral 8x7B（简称 Mixtral）和 DeepSeek-R1 的 MoE 模型。基线配置了足够的设备内存以容纳整个模型。第 V-B节比较了 Llama 4 Maverick MoE 模型（简称 Llama 4）与 Llama 3.3 非 MoE 模型。我们模拟了一个场景，其中只有较小的 Llama 3.3 模型能够完全适应设备内存，而 Llama 4 则需要卸载。在实验中使用的模型里，共享专家和非 MoE FFN 层都与其他全连接（FC）层归为一类。表 I总结了推理系统和模型配置。

[1]它位于 https://github.com/scale-snu/SSD-offloading

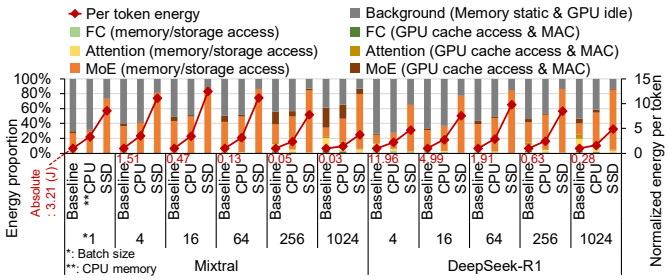

图 4. 能量分解和归一化每令牌能量。CPU 内存和 SSD 通过 NVLink5.0 与 GPU 互连。

并行性和通信： 我们采用了来自 [11] 的并行策略。MoE 层利用专家并行，将不同的专家分配给不同的 GPU。非 MoE FFN/FC 层和共享的专家在节点内使用张量并行，在节点间使用数据并行。为了同步跨 GPU 的数据，输出投影、非 MoE FFN/MoE 层以及 MoE 层中的共享专家在其操作后执行节点内的 AllReduce 操作。此外，MoE 层进行调度和合并以计算分布于不同 GPU 上的门控专家。

模拟详情： 我们使用了合成的数据集，假设对于每个解码器，在仿真过程中令牌到专家的分配遵循均匀分布。对于所有模拟请求，我们关注的是预填充阶段后第一个解码阶段，预填充阶段有 L_{in}1024 个令牌。

V. 评估

首先，我们分析了将 MoE 权重卸载到 SSDs 与 HBM 和 CPU 内存相比对能量消耗和延迟的影响，揭示了当权重卸载到 SSDs 时系统能耗显著恶化。

其次，我们探讨了在未来的 Flash 技术改进下利用 MoE 稀疏性实现节能 SSD 使用的潜力。

A. 闪存卸载对能量和延迟的影响

我们评估了 Mixtral 和 DeepSeek-R1，比较了三种门控专家权重的场景：基线（存储在 HBM 中）、卸载到 CPU 内存以及卸载到 SSD。

能耗： 我们将能量分解分为三个主要组成部分：内存/存储访问；GPU 计算，定义为缓存访问和 MAC 操作的总和；以及系统范围内的背景能耗，它结合了内存静态功耗（每 256GB CPU 内存约为 57W [4], [7]）和 GPU 空闲功率（每个 GPU 约 70W）的消耗。SSD 空闲功率可以忽略不计 [12]，因此未被包括在内。如图 4，将卸载的门控专家权重转移到 *SSD* 导致解码阶段每生成一个

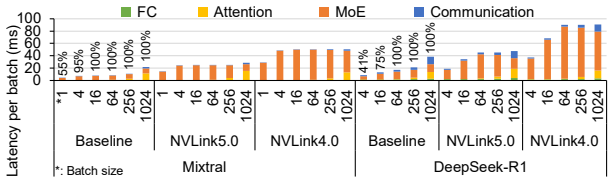

图 5. 生成单个标记的延迟分解在一个批处理中。图形上方的百分位数是每个解码器上单个 GPU 的最大活跃门控专家数量之和与所有专家被激活时等效总和的比率。

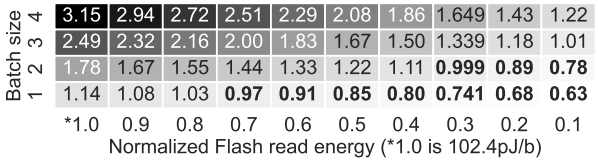

图 6. Llama 4 到 Llama 3.3 的每标记能量消耗比关于批处理大小和闪存读取能量。SSD 通过 NVLink5.0 与 GPU 相互连接。

标记所消耗的能量大幅增加。所示，与基线相比，SSD 卸载导致 Mixtral 的每令牌能耗高出 3.8× 到 12.5× 倍，DeepSeek-R1 的能耗高出 4.7× 到 9.8× 倍，具体数值随批处理大小的变化而变化。即使与卸载到 CPU 内存相比，SSD 卸载也要差得多，耗能多出 2.1× 到 3.6× 倍。这种显著的能量惩罚直接归因于从闪存读取的高能量成本。

随着批次大小的增加，加载每批激活的 MoE 专家所需的总能量趋于饱和，一旦大多数或所有独特的专家（N_{ex}）在整个批次中被要求（在我们的测试中，对于 Mixtral 大约是在 16 的批次大小时，而对于 DeepSeek-R1 则是 64）。这意味着由于摊销效应，每令牌 MoE 内存/存储访问能量随着更大批次而减少。相比之下，GPU 计算能量用于 FC、注意力和 MoE 层则随着批次大小的增加而增加，因为计算工作量增加了。尽管有摊销效果，在大批次（1024）时，SSD 访问权重放大的能源消耗仍然是主导因素：在使用 SSD 卸载的情况下，它占总能量的超过 73%（Mixtral）和 80%（DeepSeek-R1），而在基线中，访问存储在 HBM 中的 MoE 权重的能量分别降至仅 11%和 15%。重要的是，通过预取隐藏延迟并不会改变这些基础能源消耗数据。

令牌生成延迟： 虽然能量消耗较高，但可以通过有效管理来控制 SSD 卸载对延迟的影响。现代高带宽互连如 NVLink 可以减少从 SSD 和 CPU 内存传输原始数据的时间。此外，预取技术通过将其与执行其他层重叠，有效地隐藏了剩余的大部分传输延迟。图 5显示，在使用高效预取技术和类似 NVLink5.0（带宽是 NVLink4.0 两倍）这样的快速互连的情况下，卸载后的每批生成令牌端到端延迟可以与基线配置相当（将延迟惩罚降低至 Mixtral 的 1.32× 和 DeepSeek-R1 的 1.25×）。这表明了一个权衡：牺牲能效以实现可管理的延迟，并能够使用比设备内存更大的模型。

B. 来自闪存能量缩放的机会及其

第 V-A节展示了当前技术下 SSD 卸载的高能耗。在这里，我们探讨未来 Flash 读取能效的改进是否可能改变这一结论，有可能使 SSDs 对大型 MoE 模型具有吸引力。我们将大规模 MoE 模型 Llama 4（400B）在需要 SSD 卸载情况下的能效与一个较小但密集的非 MoE 模型 Llama 3.3（70B）进行比较，假设后者可以完全容纳在 HBM 中。像 Llama 4 这样的大型 MoE 模型预计将比像 Llama 3.3 这样的小型非 MoE 模型提供更高的推理精度。

关键思想是利用 *MoE* 稀疏性：如果大型 MoE 模型（由于稀疏性）每个标记访问的参数数量足够小于密集模型的总参数量，那么即使使用效率较低的 SSD，MoE 模型也可能实现更低的每标记能量消耗，前提是 SSD 的能耗惩罚有所降低。对于 Llama 4，在批处理大小为一时，每个标记仅激活了 17B 参数（占总数的 4.3%），其中这些激活权重中的 17.8% 来自卸载到 SSD 的门控专家。相比之下，Llama 3.3 始终访问其全部 70B 参数。随着 Llama 4 的批处理大小增加，更多的专家被激活，当批处理足够大以激活所有 N_{ex} 专家时，最终会访问到所有的 400B 参数。

图 6绘制了 Llama 4 与 Llama 3.3 的每令牌能量比，作为 Flash 读取能耗缩放因子的函数，在仅有 GPU 和 SSD 的系统中。只有当 Flash 读取能耗显著降低——具体来说，大约减少到其当前值的十分之一（~10pJ/b）时，Llama 4 在 SSD 上的能效才会超过 Llama 3.3 在 HBM 上的表现（比例 < 1）。这种 “SSD 胜出的情况” 在小批量尺寸（本实验中的批量大小 < 3）最为明显，在此时 MoE 稀疏度最高，表明在利用 MoE 稀疏性的前提下，仍然需要实质性的 Flash 能效技术进步，才能使 SSD 卸载与基于 HBM 的推理在能耗上具有竞争力。

VI. 讨论

我们的研究结果突显了一个关键的权衡：当前的 SSD 为 MoE 权重卸载施加了高昂的能量成本（第 V-A 节），但未来的 Flash 技术进步可能会改变这种平衡。分析（第 V-B节）表明，如果 Flash 读取能量提高到 ~10×（即 ~10pJ/b），并且有效地利用了 MoE 的稀疏性，那么 SSD 上的大型 MoE 模型可能会比小型密集型模型实现更好的能效。

此场景特别适用于移动/边缘平台，这些平台通常在小批量大小下运行，最大化 MoE 稀疏性，面临严格的电池限制，并拥有类似的 LPDDR/UFS 内存层次结构。因此，如果 Flash 达到 ∼10pJ/b 的读取能量目标，移动领域将成为有望从使用节能 Flash 卸载的大而准确的 MoE 模型中受益的主要领域。实现这一点需要专注于能源减少和稀疏数据管理的系统协同设计，在 Flash 方面持续进化。

VII. 结论

将 MoE 权重卸载到现代 Flash 固态硬盘会导致 LLM 推理能量消耗激增，因为 Flash 读取能量较高。虽然延迟通常可以隐藏，但这种能耗惩罚使得 SSD 卸载在今天从能耗角度来看是不利的。然而，未来在 Flash 读取能效方面的显著改进（大约降低 10×）结合 MoE 模型在小批量情况下的固有稀疏性，可能会使 SSDs 成为部署非常大且高精度 LLM 的能量可行选项。

参考文献

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*, 2020.

[2] K. Alizadeh, I. Mirzadeh, D. Belenko, K. Khatamifard, M. Cho, C. C. D. Mundo, M. Rastegari, and M. Farajtabar, "LLM in a Flash: Efficient Large Language Model Inference with Limited Memory," *arXiv preprint arXiv:2312.11514*, 2024.

[3] R. Hwang, J. Wei, S. Cao, C. Hwang, X. Tang, T. Cao, and M. Yang, "Pre-gated MoE: An Algorithm-System Co-Design for Fast and Scalable Mixture-of-Expert Inference," in *ISCA*, 2024.

[4] Micron, "16Gb DDR5 SDRAM Addendum," 2022.

[5] V. Mohan, S. Gurumurthi, and M. R. Stan, "FlashPower: A detailed power model for NAND flash memory," in *DATE*, 2010.

[6] NVIDIA, "Grace Hopper Superchip." [Online]. Available: https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip

[7] Micron, "Calculating Memory Power for DDR4 SDRAM," 2017.

[8] S. Wang, B. Yu, W. Xiao, F. Bai, X. Long, L. Bai, X. Jia, F. Zuo, J. Tan, Y. Guo, P. Sun, J. Zhou, Q. Zhan, S. Hu, Y. Zhou, Y. Kang, Q. Ren, and X. Jiang, "A 135 GBps/Gbit 0.66 pJ/bit Stacked Embedded DRAM with Multilayer Arrays by Fine Pitch Hybrid Bonding and Mini-TSV," in *VLSI Technology and Circuits*, 2023.

[9] M. O' Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally, "Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems," in *MICRO*, 2017.

[10] NVIDIA, "GPUDirect Storage." [Online]. Available: https://docs.nvidia.com/gpudirect-storage/index.html

[11] S. Yun, K. Kyung, J. Cho, J. Choi, J. Kim, B. Kim, S. Lee, K. Sohn, and J. Ahn, "Duplex: A Device for Large Language Models with Mixture of Experts, Grouped Query Attention, and Continuous Batching," in *MICRO*, 2024.

[12] Samsung, "9100 PRO M.2 NVMe SSD," 2025. [Online]. Available: https://semiconductor.samsung.com/consumer-storage/internal-ssd/9100-pro/