Paper Review: MMK: A Hybrid Scheduling Framework for Fine-Grained Multi-Instance GPU Sharing

Reviewer: [袁小松]
Date: [2025/5/10]

---

1. Summary of Contributions

本文提出了**MMK**，一个集成了**MIG（多实例 GPU）**、**MPS（多进程服务）**和**内核级调度**的GPU框架，以解决现有MPS+MIG配置复杂不灵活和硬件级别的粗粒度调度问题。主要贡献包括：

- **内核调度器**支持细粒度、竞争感知的 GPU 内核调度，优先处理延迟敏感的在线作业。
- 通过三个Observation及实现分析证明现有的 MIG 和 MPS 混合方法仍然存在资源争用和利用率低的问题。
- 提出了MMK，并证明了 MMK 的有效性.(e.g. 与最先进的 MIG 与 MPS 混合框架相比，并将平均作业完成时间、最大完工时间和系统吞吐量分别提高了 28%、32% 和 35%)

---

2. Strengths

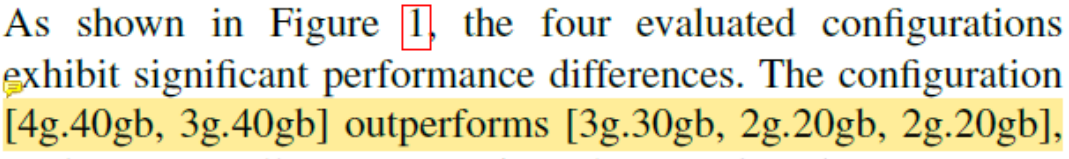
- **技术创新**：首次将 MIG、MPS 和内核调度整合到一个统一的框架中, **内核级优先级调度**尤其具有创新性。
- **低开销**：内核调度仅带来**0.08% 的运行时开销**（图 10a），使其易于部署。

3. Improvement Space & Questions

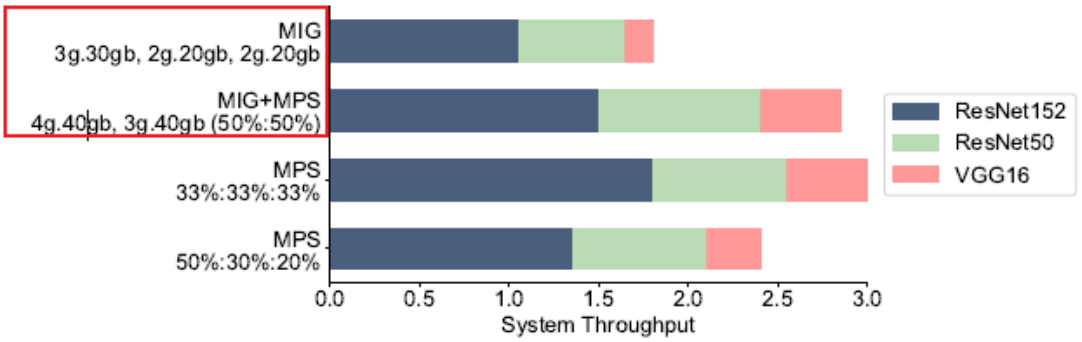
- 第一个挑战MPS+MIG的配置问题回答的没有那么清晰，**混合分区调度器** (Hybrid Partition Scheduler) 与**MIGER** 或普通的 **MPS+MIG** 服务区别在哪里?
- 对MMK在 **A100** GPU架构做了丰富的实验，有考虑在其他架构如 **H100**、**AMD** 等其他类型GPU做实验以证明MMK的**通用性**?
- 本文从内核级粒度关注了在线服务低延迟需要，提高了在线作业的优先级，取得了较好的效果，如何考虑**离线作业等待久**的问题，在实际中会有饥饿问题吗?

4. Details Problem

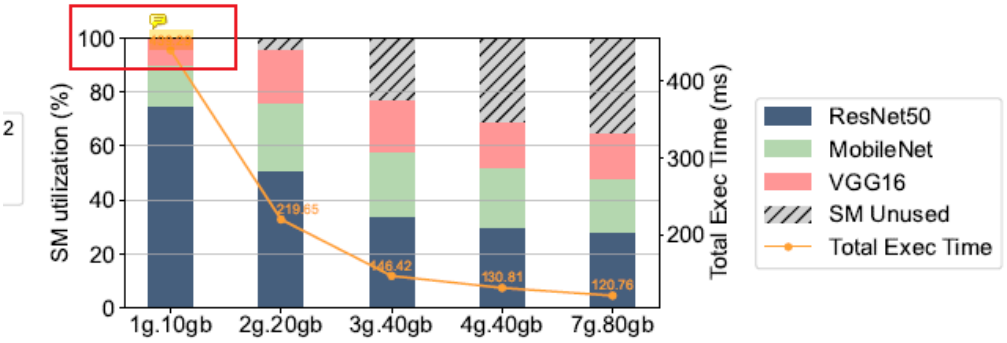
- 文中B. Observations 中

  
  As shown in Figure 1, the four evaluated configurations exhibit significant performance differences. The configuration [4g.40gb, 3g.40gb] outperforms [3g.30gb, 2g.20gb, 2g.20gb],

  [4g.40gb, 3g.40gb] 共分配了**80G**的显存且使用的是**MIG+MPS**，[3g.30gb, 2g.20gb, 2g.20gb]只分配了**70G**的显存，仅使用了**MIG**，没有遵循对照试验，后者少的10G是作为共享显存吗？需要在论文中说一下吗？：



- 图2 1g.10gb的Total Exec Time 看不清楚



ıg Fig. 2: SM utilization comparison under different MIG partitions.