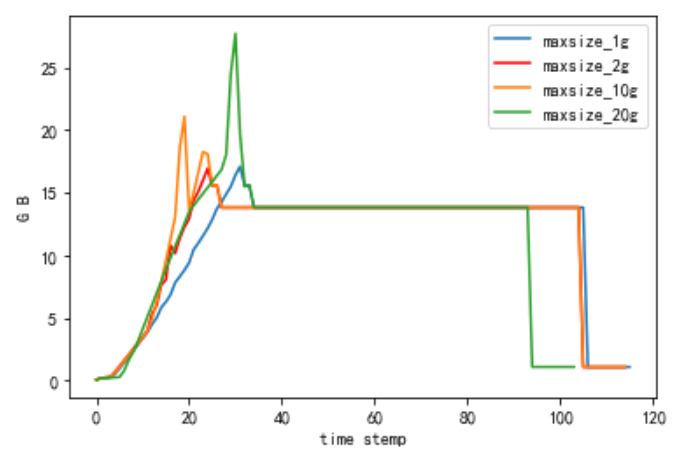
1. 资源占用问题：

大模型保存切片大小会使得初始化内存占用量出现偏差：



上图基于glm-6b 模型的初始化内存占用时序图，maxsize为模型切片大小，占用磁盘大小为约为14.4g，基于28层transformer参数逐层抽取切分。

当切片大小为2g时，加载效率和初始化内存均达到较优的水平，因此在模型部署的时候（基于transformer框架），可以切片为2g，这也是业界大模型常用的切片方案。

2.推理加速Tensorrt-LLM