探究不同负载的性能表现和资源需求（任德上）

- **实验内容：**以在jetson上使用deepstream为例，使用resnet10（物体检测）和resnet18（颜色分类）两个模型，考虑前台视频播放+后台深度模型推理叠加负载场景，探究延长视频时间、使用更复杂的模型和提升视频画质对性能和资源需求的影响。
- **实验结果：**

1) 更长的视频可以增加推理延迟占端到端延迟的比重，使调节GPU频率对端到端延迟的影响更显著。

2) 更复杂模型的引擎加载和流水线初始化阶段对CPU的需求更大；推理阶段对GPU(计算)和CPU(调度)的需求都更大。

3) 更高清的视频对硬件和CPU的需求增加，因各画质的视频都会缩放到模型网络输入大小，故对GPU需求基本不变。

初步确定负载定义方案（鲍成）

- **负载特征选取：**f_cpu、f_gpu、f_ddr、util_cpu、util_gpu、util_bw、Temp_cpu、Temp_gpu
- **负载计算过程：**

离线阶段：原型负载库构建

在多频点下采集硬件计数器，按 K 长窗标准化成序列并送入预训练 LSTM，得到高维负载嵌入向量。

在线阶段：负载向量生成与时间平滑

取当前窗口 t 的最近 K 个观测组成 X_t，标准化后送入 LSTM 编码器得到当前负载嵌入，再与原型库计算相似度。

计算过程探讨:

1）从模型定位与适用目标、表达能力与依赖条件、可解释性与工程复杂度三个角度对比了SVM和LSTM。

2）说明了单帧snapshot局限，使用高维表示的原因，以及选择LSTM的原因。

调研DDR频点→设备级带宽→应用级带宽的原理（程诗淇）

- **调节DDR频率改变的是设备级带宽**
- **调研带宽管控流程**

启动前先用 ICC 做需求聚合与预热，运行期运行期带宽型/时延型 governor 采样，以“预估保底 + 实测纠偏 + 余量”合成需求带宽，最后 NoC 用Shaper整形，用”高优先栈严格优先级，同一优先栈加权轮询”把带宽分配到各应用。

制作人：程诗淇　　审阅人：徐飞 吴旭东