ISCAS 中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周
学术论文

打破交互墙：以DLPU为中心的深度学习计算系统
Breaking the Interaction Wall: A DLPU-centric Deep Learning Computing System

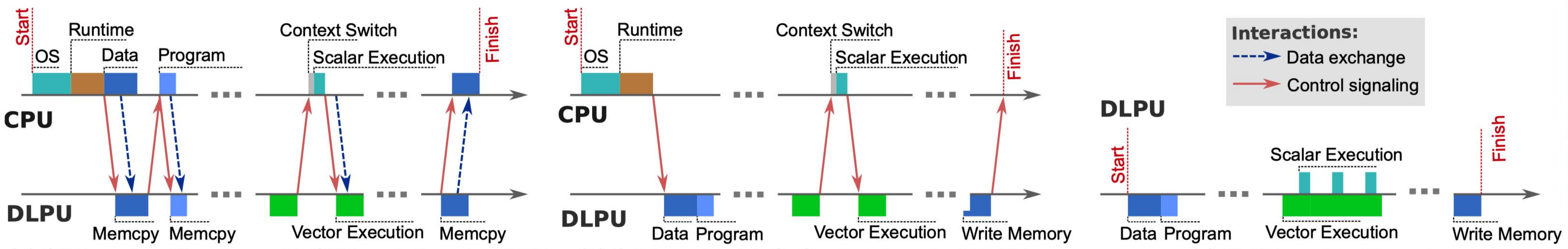
杜子东[1]，郭崎[1]，赵永威[1]，曾惜[1]，李玲[2]，程丽敏[2] 等
[1]中国科学院计算技术研究所　[2]智能软件研究中心，中国科学院软件研究所

期刊：IEEE Transactions on Computers, 2020

联系人：李玲　联系方式：18500300237 liling@iscas.ac.cn

动机

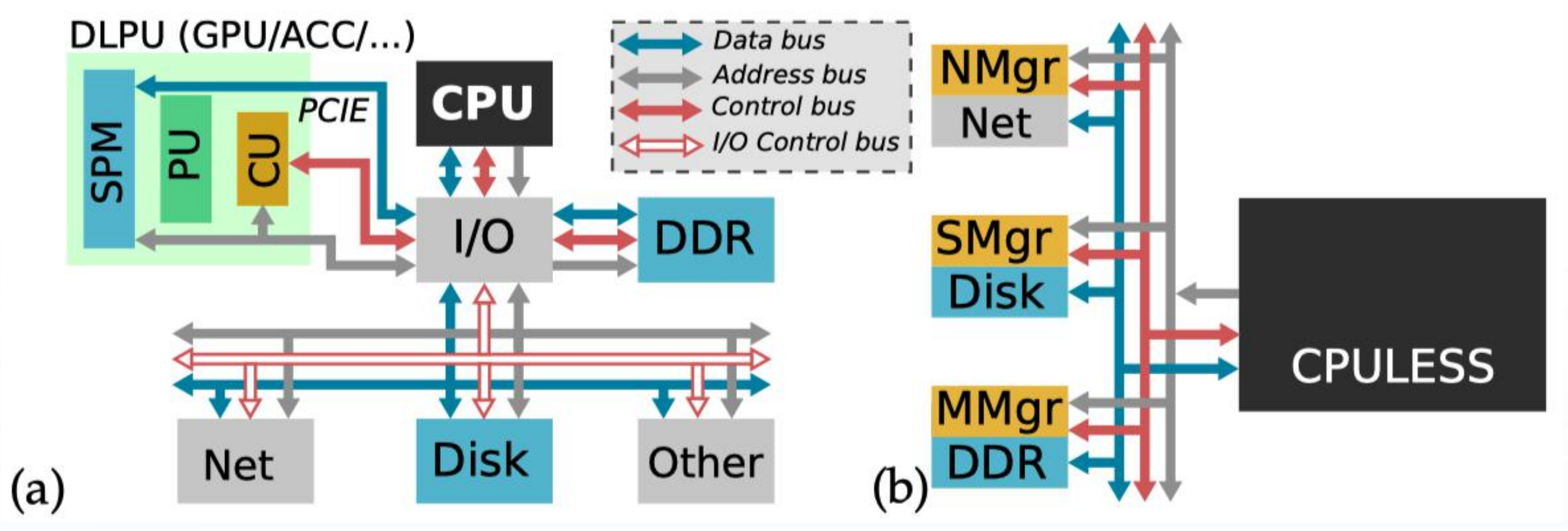
- 以CPU为中心的深度学习计算系统存在交互墙问题，导致系统效率低下
- 提出以DLPU（深度学习处理单元）为中心的深度学习计算系统，包括面向异常的编程模型（EOP）和CPULESS DLPU架构支持，可以显著提升速度、降低能耗



(a) CPU-centric DL sytem with CPU and discrete DLPU (b) CPU-centric DL SoC with integrated DLPU (c) DLPU-centric DL system

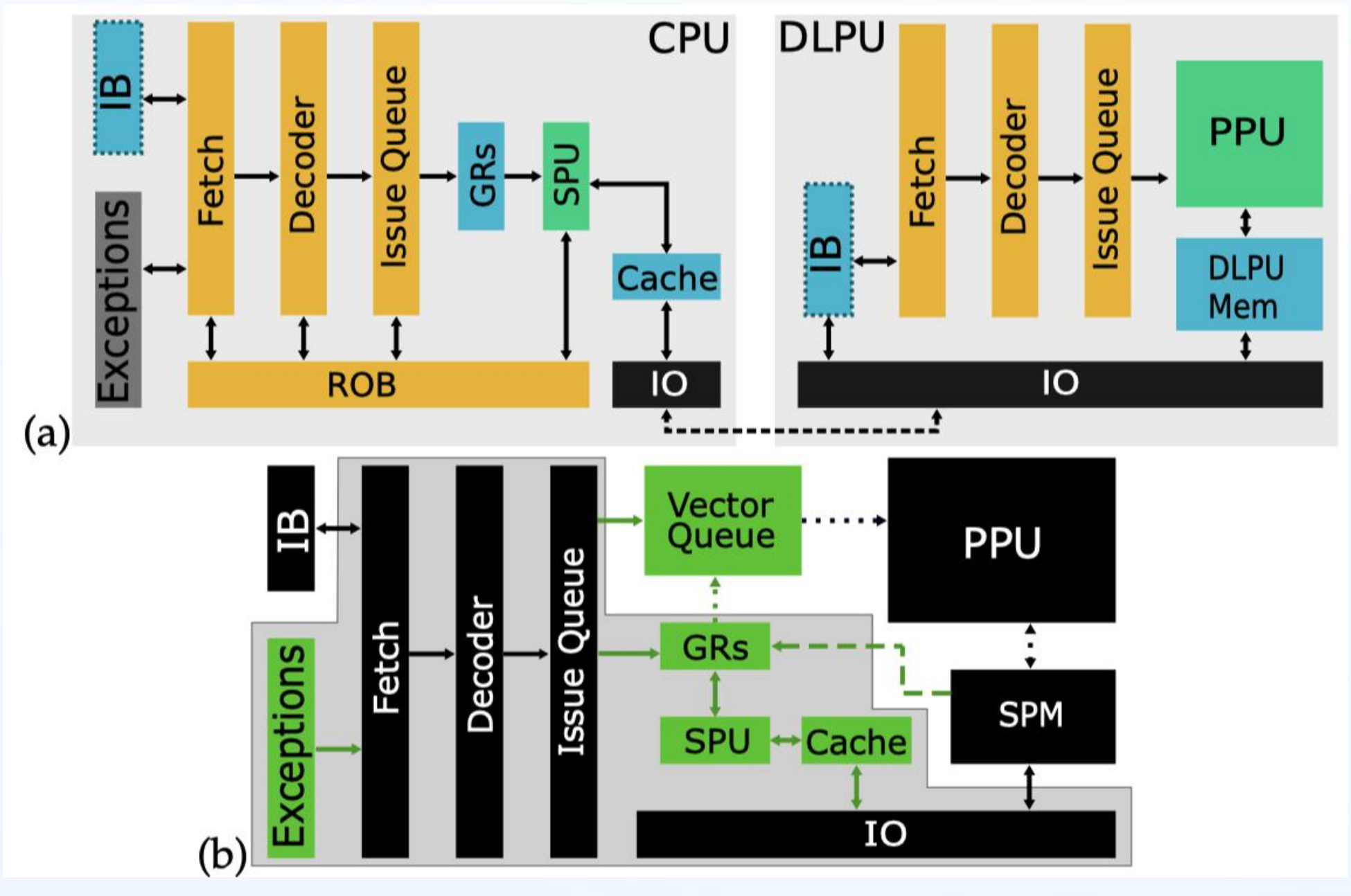
DLPU为中心的深度学习计算系统

- 深度学习计算系统架构



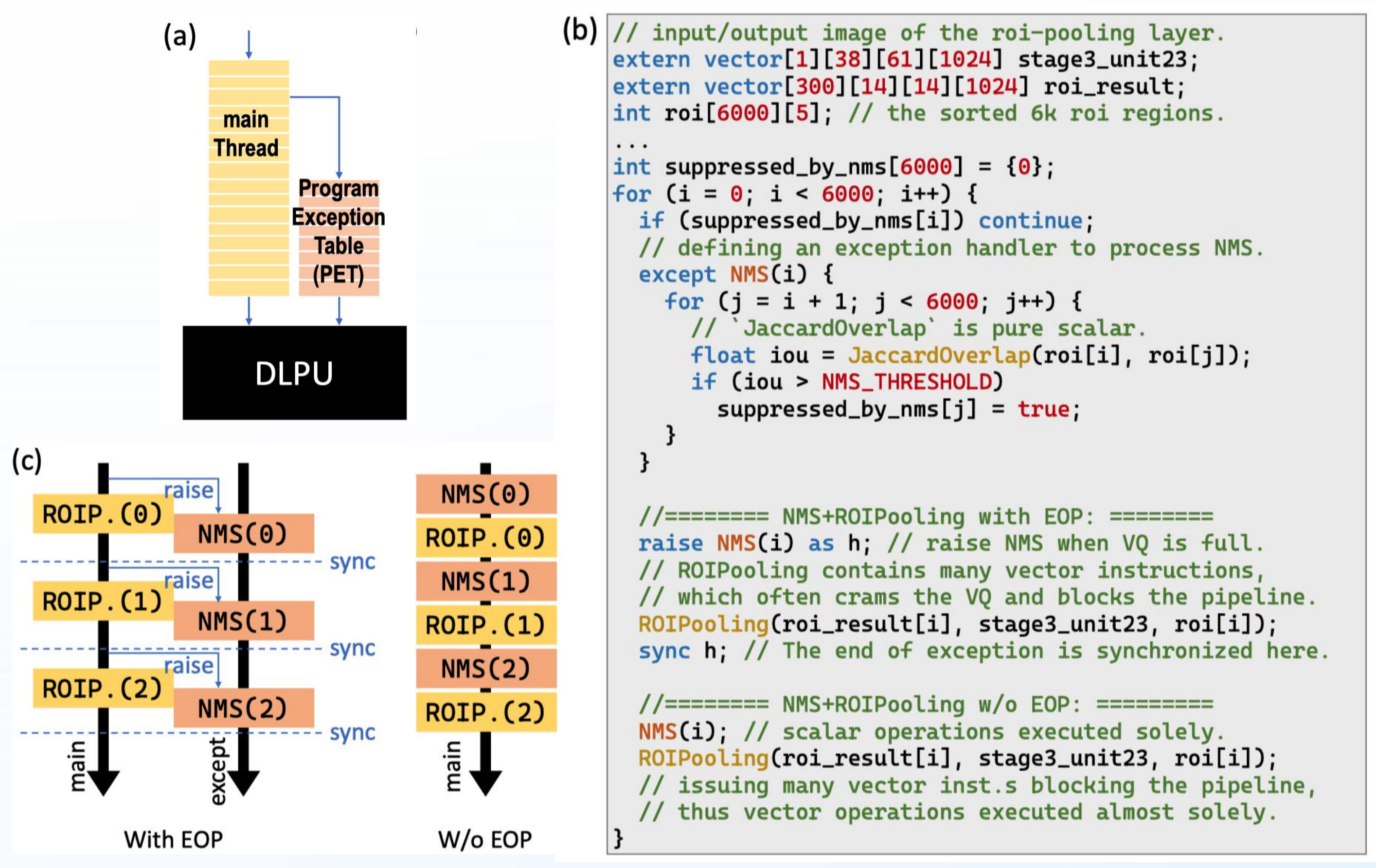
(a)主流的CPU为中心的计算系统架构
(b)以DLPU为中心的计算系统架构

- CPULESS微架构



(a) CPU为中心的微架构　(b) CPULESS 微架构

- EOP编程模型



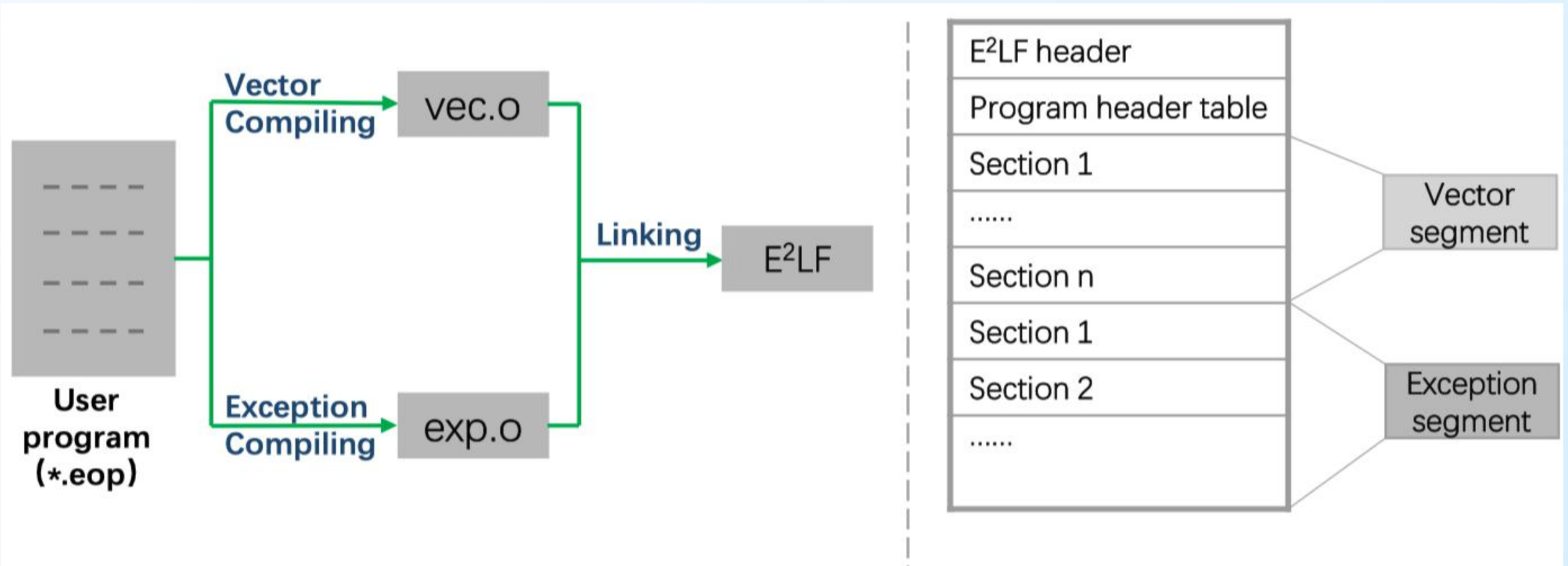
```
// input/output image of the roi-pooling layer.
extern vector[1][38][61][1024] stage3_unit23;
extern vector[300][14][14][1024] roi_result;
int roi[6000][5]; // the sorted 6k roi regions.
...
int suppressed_by_nms[6000] = {0};
for (i = 0; i < 6000; i++) {
  if (suppressed_by_nms[i]) continue;
  // defining an exception handler to process NMS.
  except NMS(i) {
    for (j = i + 1; j < 6000; j++) {
      // `JaccardOverlap` is pure scalar.
      float iou = JaccardOverlap(roi[i], roi[j]);
      if (iou > NMS_THRESHOLD)
        suppressed_by_nms[j] = true;
    }
  }

  //======== NMS+ROIPooling with EOP: ========
  raise NMS(i) as h; // raise NMS when VQ is full.
  // ROIPooling contains many vector instructions,
  // which often crams the VQ and blocks the pipeline.
  ROIPooling(roi_result[i], stage3_unit23, roi[i]);
  sync h; // The end of exception is synchronized here.

  //======== NMS+ROIPooling w/o EOP: =========
  NMS(i); // scalar operations executed solely.
  ROIPooling(roi_result[i], stage3_unit23, roi[i]);
  // issuing many vector inst.s blocking the pipeline,
  // thus vector operations executed almost solely.
}
```

(a)抽象机器模型 (b)Faster R-CNN算法的示例代码 (c)程序员视角

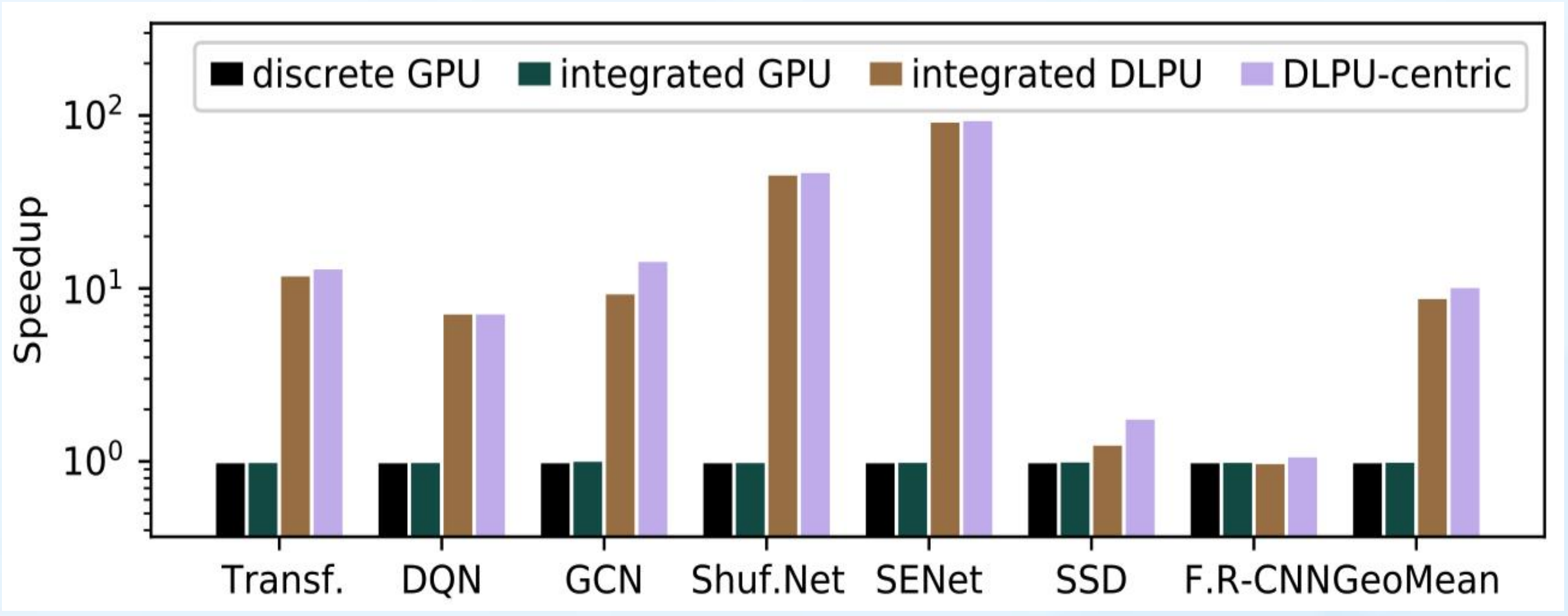
- 编译链接

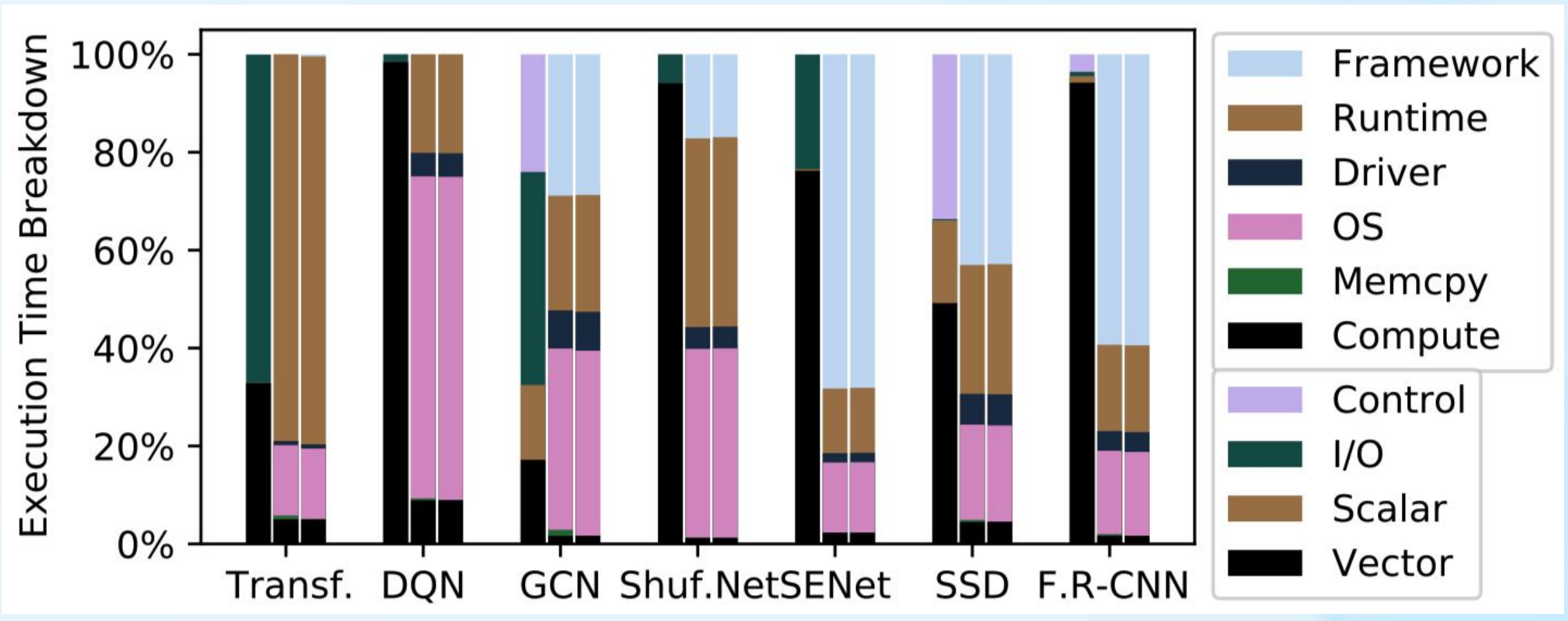


(a) Compilation process　(b) Detailed format of E²LF

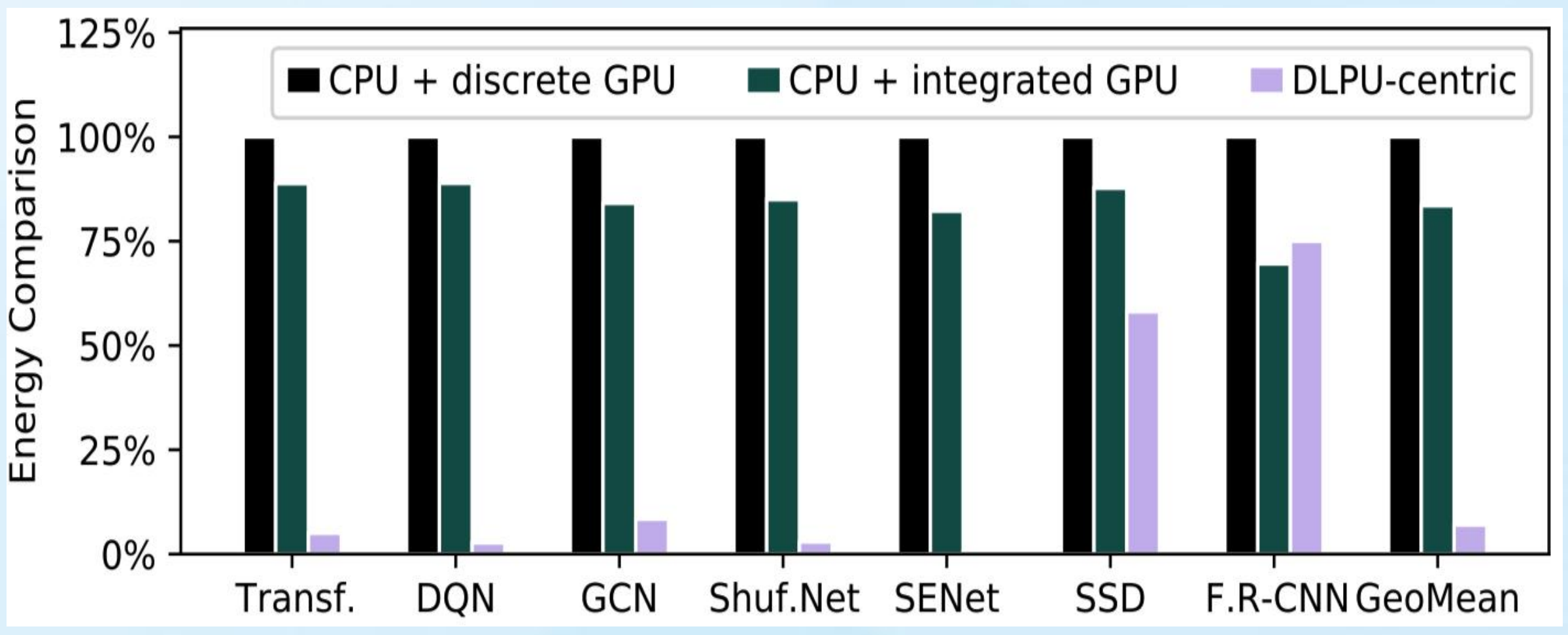
实验结果

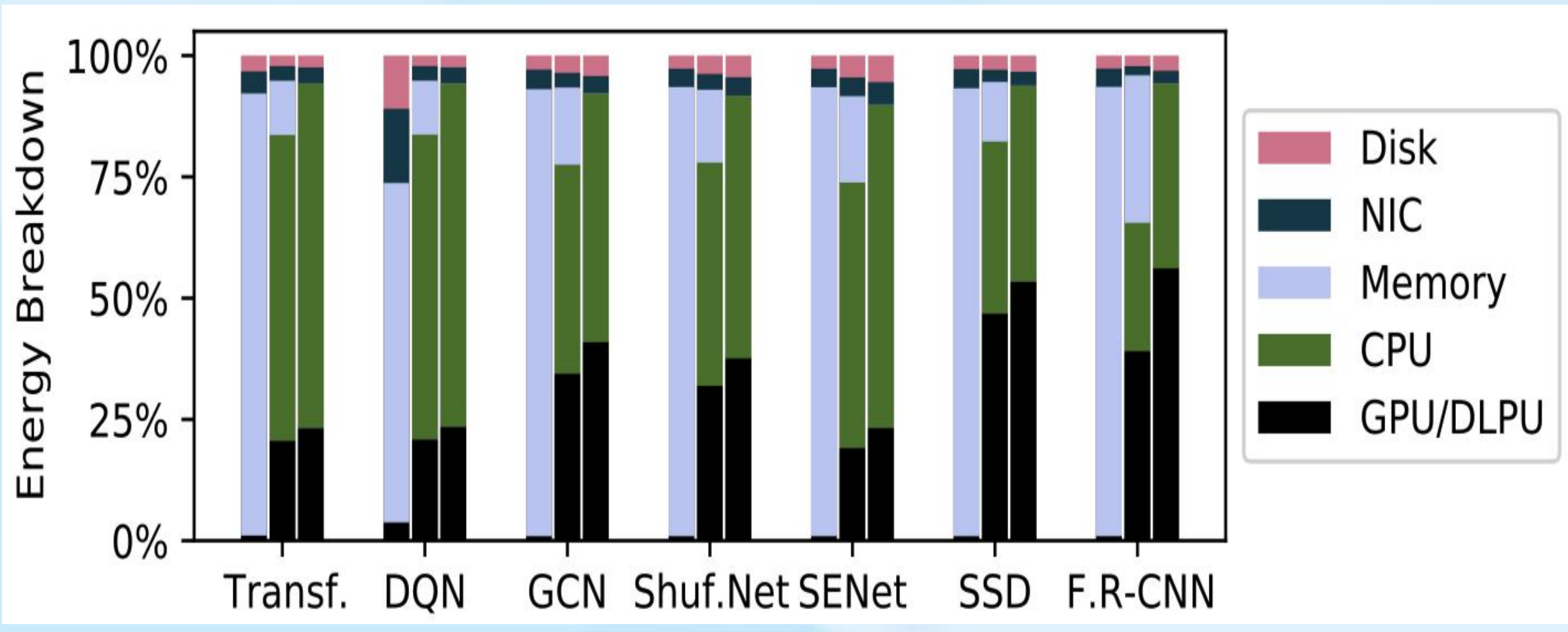
- 对比硬件：Intel Xeon 6130 CPU，Nvidia Tesla V100 GPU
- 加速比





- 能耗





相对于主流的CPU+V100系统，本文系统平均加速10.30倍，节省92.99%的能耗