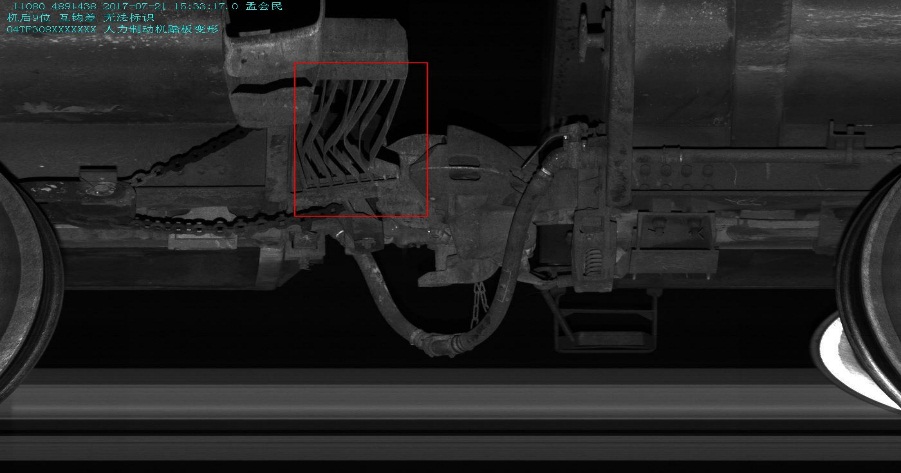
# 高维张量任一模态稀疏张量矩阵/向量乘问题

## 问题简介：

### 数据的高阶与多维特性广泛存在：

1. 列车的故障诊断数据：

铁路摄像机和传感器等采集的列车视频、音频和超声波等多种模态的数据。

1000 多个货运车辆型号、80 多万辆铁路货运列车，货运列车共有零部件 290 多个，故障种类多达 300余种。以时间轴为线索串联图像数据，形成3阶甚至更高阶的张量

1. 天文张量数据：

太阳观测数据挖掘、太阳活动对深空探测、载人航天、卫星运行、电力网络的安全有重要影响。SDO（太阳动力学观测站）每秒以10个波段拍摄太阳4096\*4096分辨率的图像，每天数据量为1.5T。以时间轴为线索串联图像数据，形成高阶的张量

### 张量介绍

0阶张量：标量

1阶张量：向量

2阶张量：矩阵

3阶张量：张量，相当于多个矩阵叠加组成的空间结构

4阶张量：张量，相当于将3阶张量看为一个元素，组成的3阶张量的向量

5阶张量：张量，相当于将3阶张量看为一个元素，组成的3阶张量的矩阵

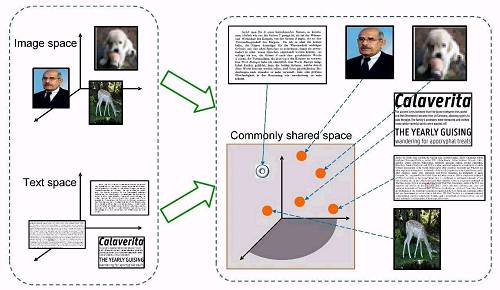
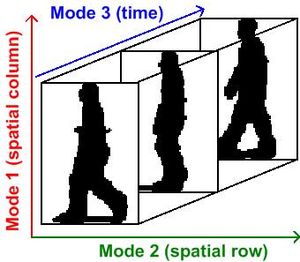
6阶张量：张量，相当于将3阶张量看为一个元素，组成的3阶张量的的空间结构

以此类推。



### 张量表示：用于多模态数据

复杂数据对象存在多个方面，通过张量建模分析和刻画数据之间的多个方面之间的相互关系



## 问题的求解算法：张量的n模态矩阵/向量积

### 张量切片张量抽取向量及切片：

### 模-n张量矩阵乘积:

一个张量 和一个矩阵 的n-mode乘积定义为:

这个定义可以写成沿mode-n展开的形式:



其性质：





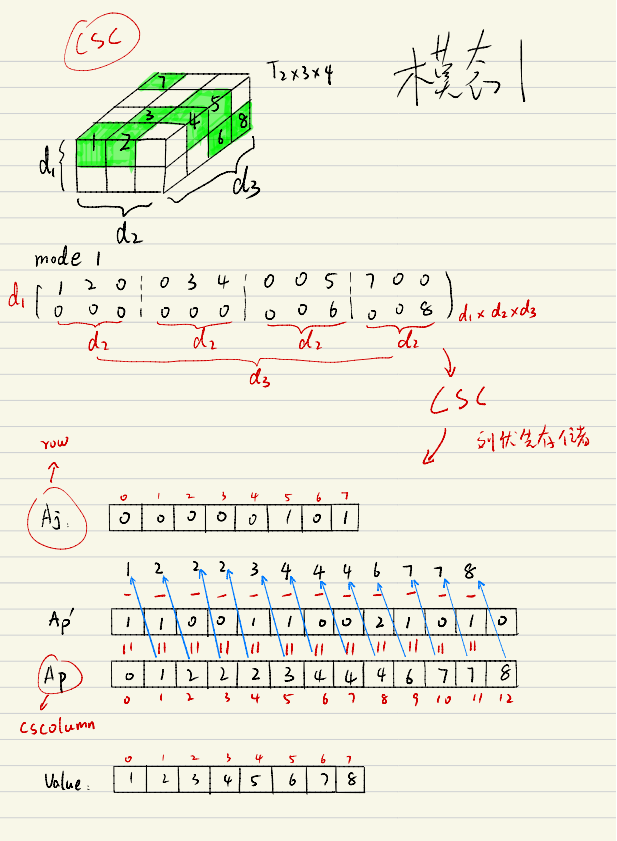
相关论文：

Optimizing Sparse Tensor Times Matrix on Multi-core and Many-core Architecture

SPLATT: Efficient and Parallel Sparse Tensor-Matrix Multiplication（S.Smith等人发表在IPDPS）

1. 高维稀疏张量任一模式的压缩存储结构：

通过将张量展开（矩阵化）后，以列优先顺序存储，再通过CSC（Compressed Column）压缩列来最终存储张量。下面以一个简单的张量的1模态压缩存储为例：



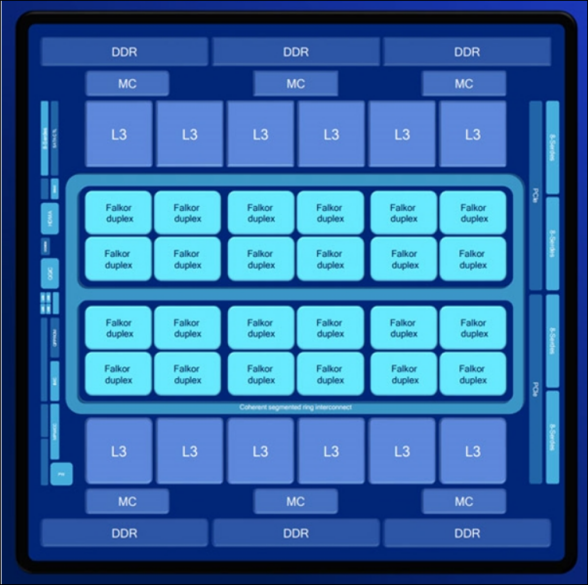
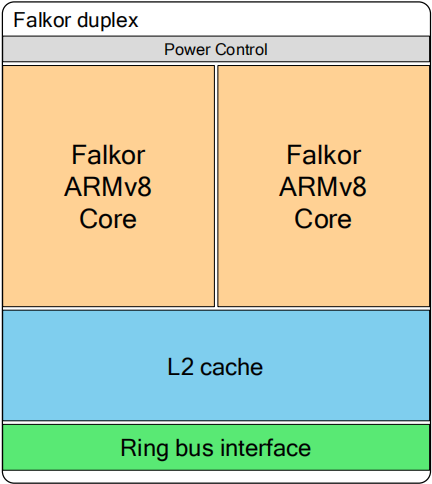
## 问题的算数密度分析：

假设稀疏张量维度为：d1×d2×d3，非零元素为nnz个，稀疏度通常<1%，一般来说每列仅有1个非零元素，那么SPTV可以近似看成：每次运算仅有两个元素相乘即：y=v\*x的类型。则由于ARM架构的采用RISC指令集，1次计算3次访存，浮点数占4字节。故计算密度位1/12。

## 问题的硬件结构概述：

目前我们基于高通Centriq系列芯片，即Qualcomm-Centriq-2434芯片，实现SPTV的CSC压缩结构的运算。

Qualcomm-Centriq-2434（高通Centriq 2400处理器系列）采用三星10纳米FinFET制程，设计为单芯片（SoC）平台级解决方案，在398mm2的面积上集成了180亿个晶体管。包含40颗高性能64位单线程Falkor自研内核（基于ARM v8，高通第五代自主架构），基础主频2.2GHz，最高可达2.5 GHz。其中，每两个内核共享512 KB二级缓存（L2 Cache），最高60MB三缓，热设计功耗110W。每个内核由总带宽为250Gbps的双向分段闭合环形总线连接。同时，支持DDR4-2667内存，最高容量768GB、6通道，32条PCIe Gen3通道和6个PCIe控制器。



## 问题的理想计算结构：

在此架构下，我们如果源源不断的进行单一乘法运算操作，有如下推论：

由2.5GHz的峰值频率可知：1个时钟周期为0.4ns。

40核单一线程模式，可同时开40个线程进行浮点计算，则需要40×(2+1)×4 = 480字节/时钟周期，即1200Gbps的访存带宽需求。

而L3级缓存与内核的连接速度也只有250Gbps贷款，更别说是主存更低的带宽。

要满足当前ARM架构下的计算强度，则需要增加缓存与主存带宽。