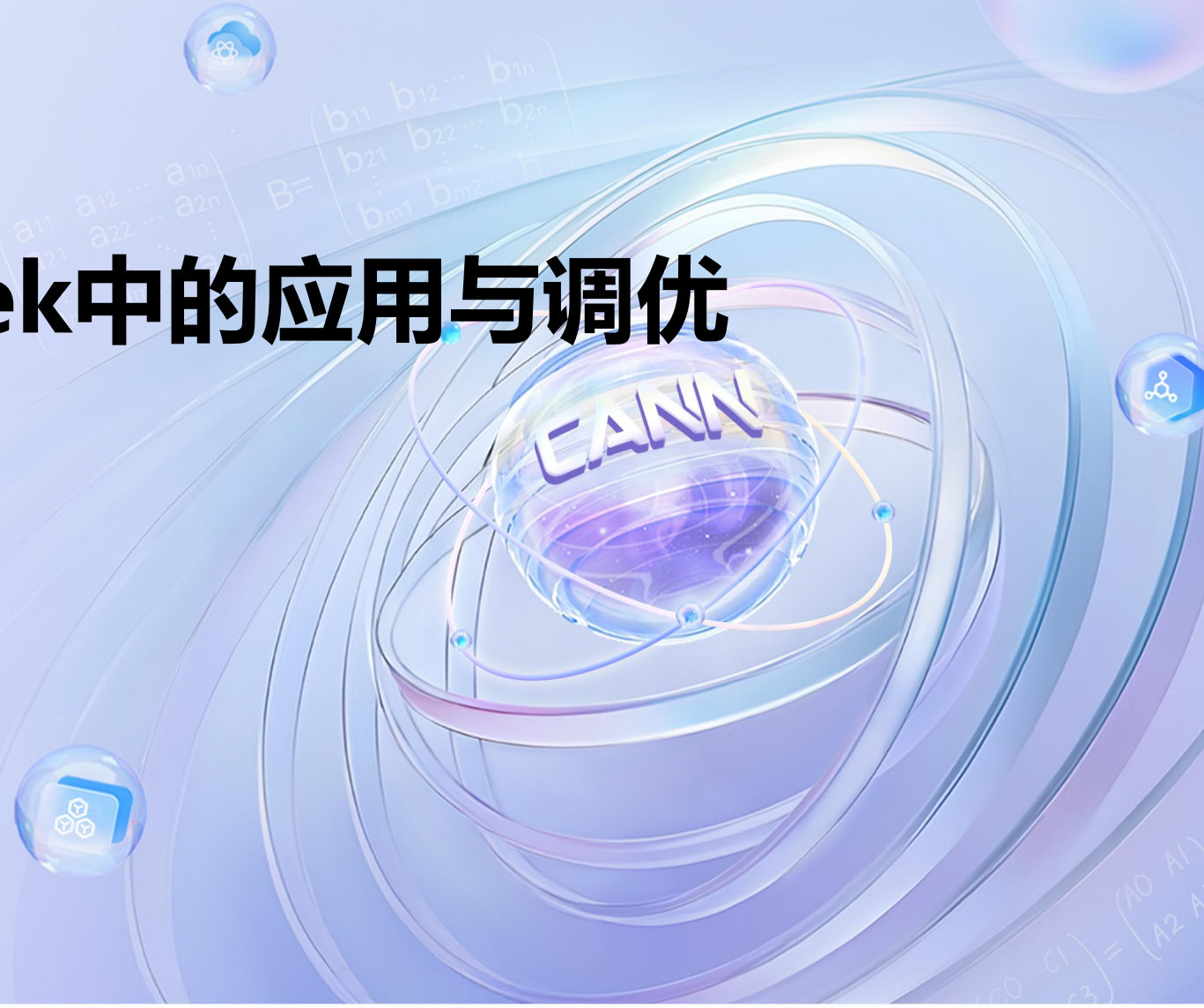
FIA算子在DeepSeek中的应用与调优



CANN

FIA算子简介

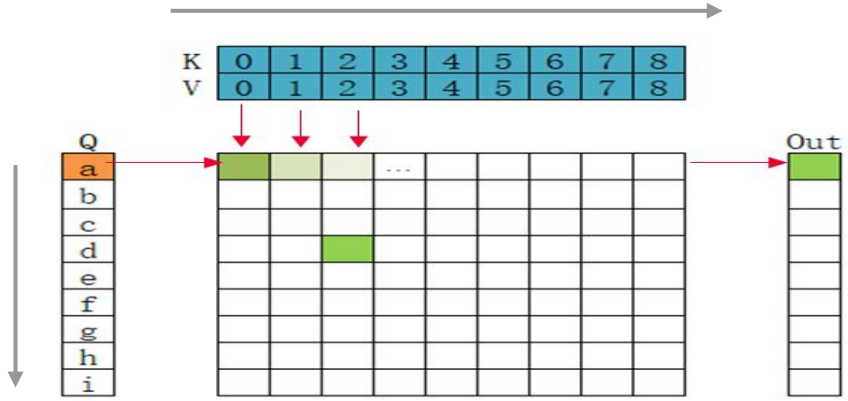
FIA（FusedInferAttentionScore）算子，是基于经典的Flash-Attention算法实现的融合算子

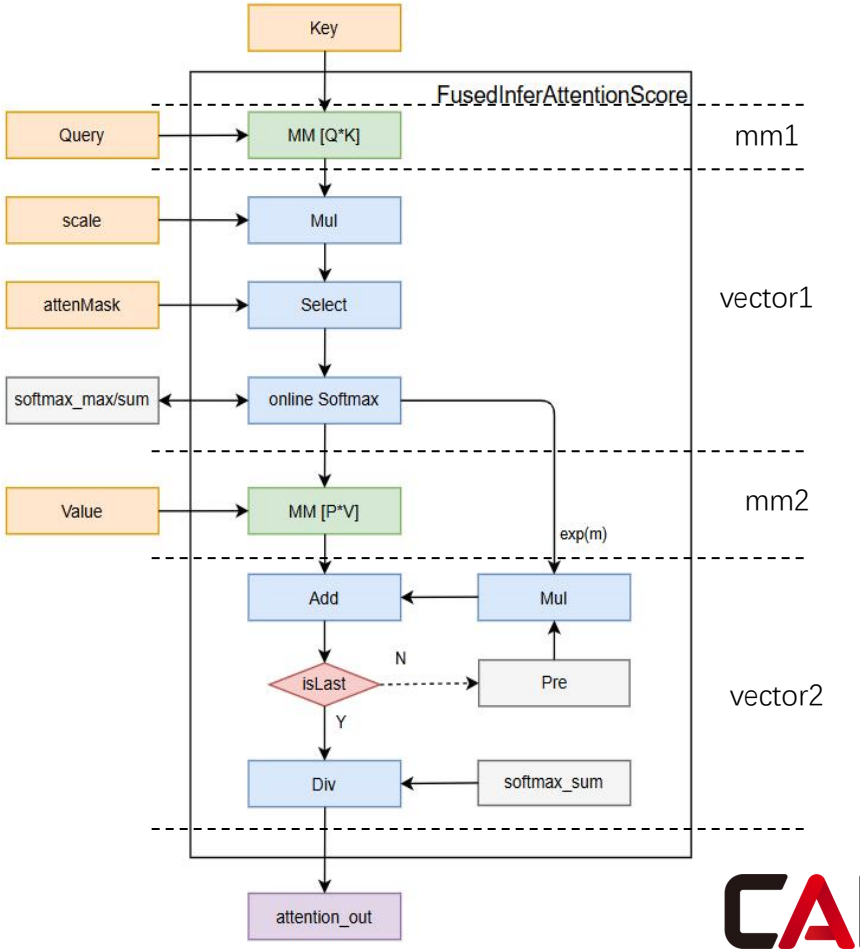
Q、K、V进行切块处理，对切块后的Q、K、V逐步计算；

切块后的Q、K做矩阵乘，进行局部softmax计算得到P，并在这个过程中，维护局部的归一化因子（最大值、累积和）；

P与切块后的V做矩阵乘，得到的结果，累加更新到历史的PV结果中；

逐步计算，最后得到完整的Attention结果；



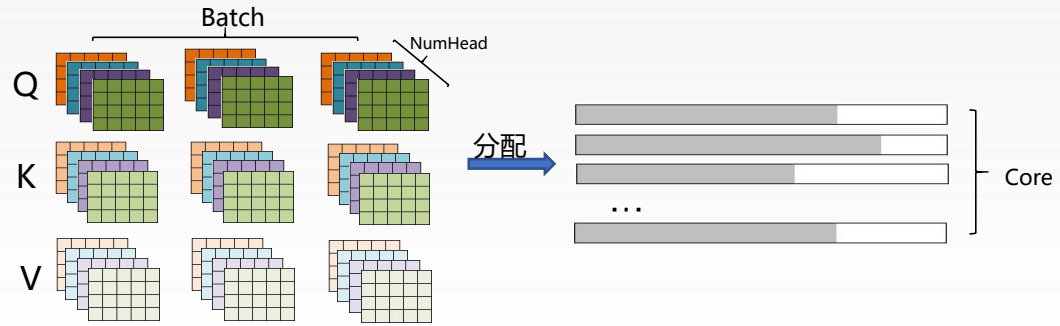


CANN

FIA算子常用优化手段

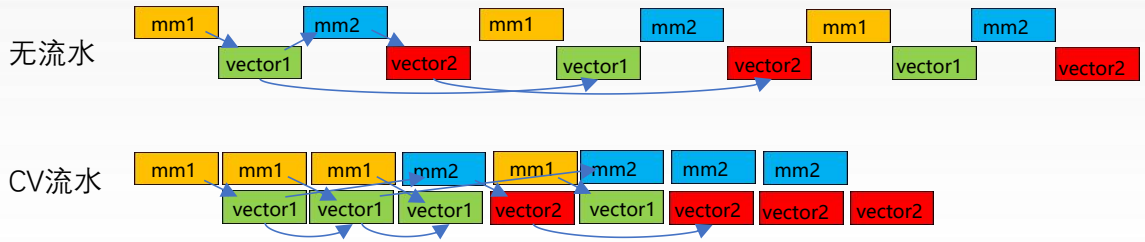
均匀分核，充分发挥算力

对Q、K、V按Batch、NumHead、Seq切块，将切块后的Q、K、V，按计算量大小，均摊到所有的核上进行计算



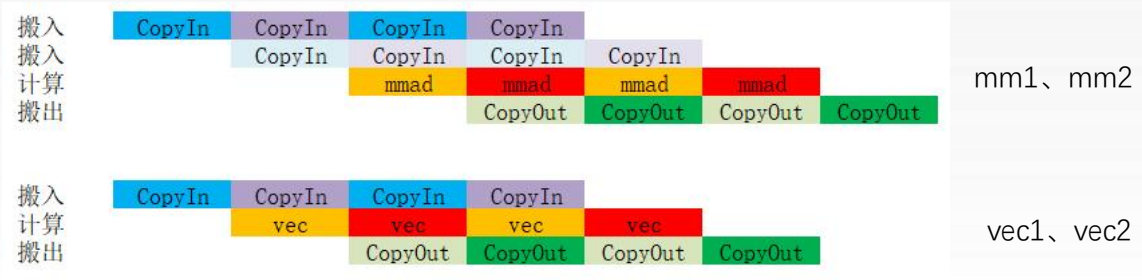
CV核间流水，Cube、Vector并行工作

FIA涉及Cube和Vector的计算，通过调整切块后mm1、vector1、mm2、vector2的执行顺序，将没有数据依赖的mm计算提前执行，从而使得没有数据依赖关系的Cube计算和Vector计算能够并行执行



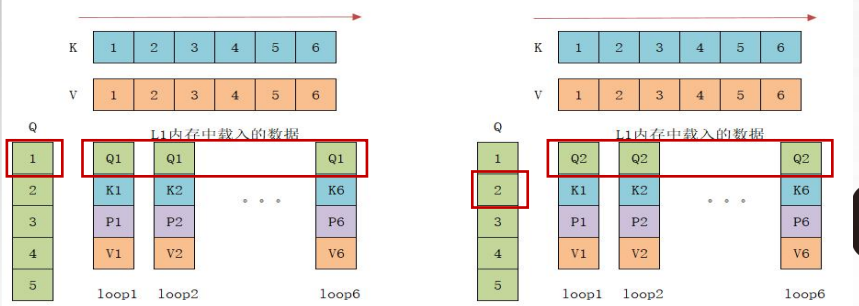
核内流水，访存与计算并行

Cube和Vector上的计算过程，都涉及数据搬入、计算、数据搬出流程，使用Double Buffer技术，使得数据搬入、计算、数据搬出能够并行执行



数据常驻，减少访存开销

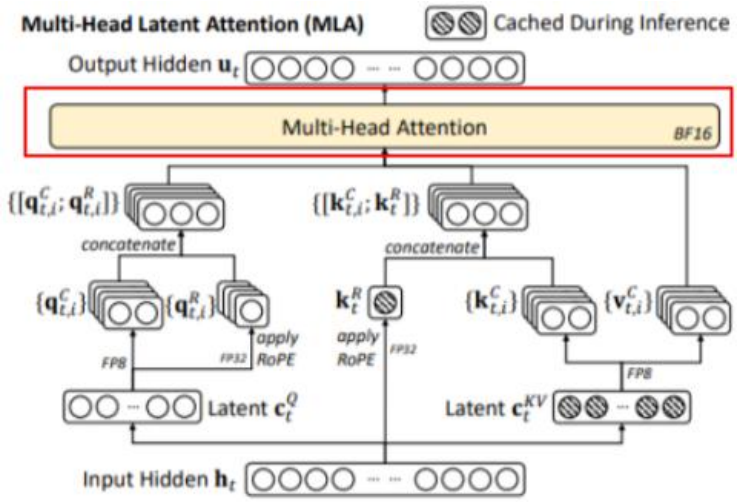
FIA对Q、K、V切块后，在遍历KV的时候，可以将Q子块常驻在内存中，K子块按遍历顺序依次载入内存，与常驻的Q子块分别进行mm1计算，避免Q子块的重复访存

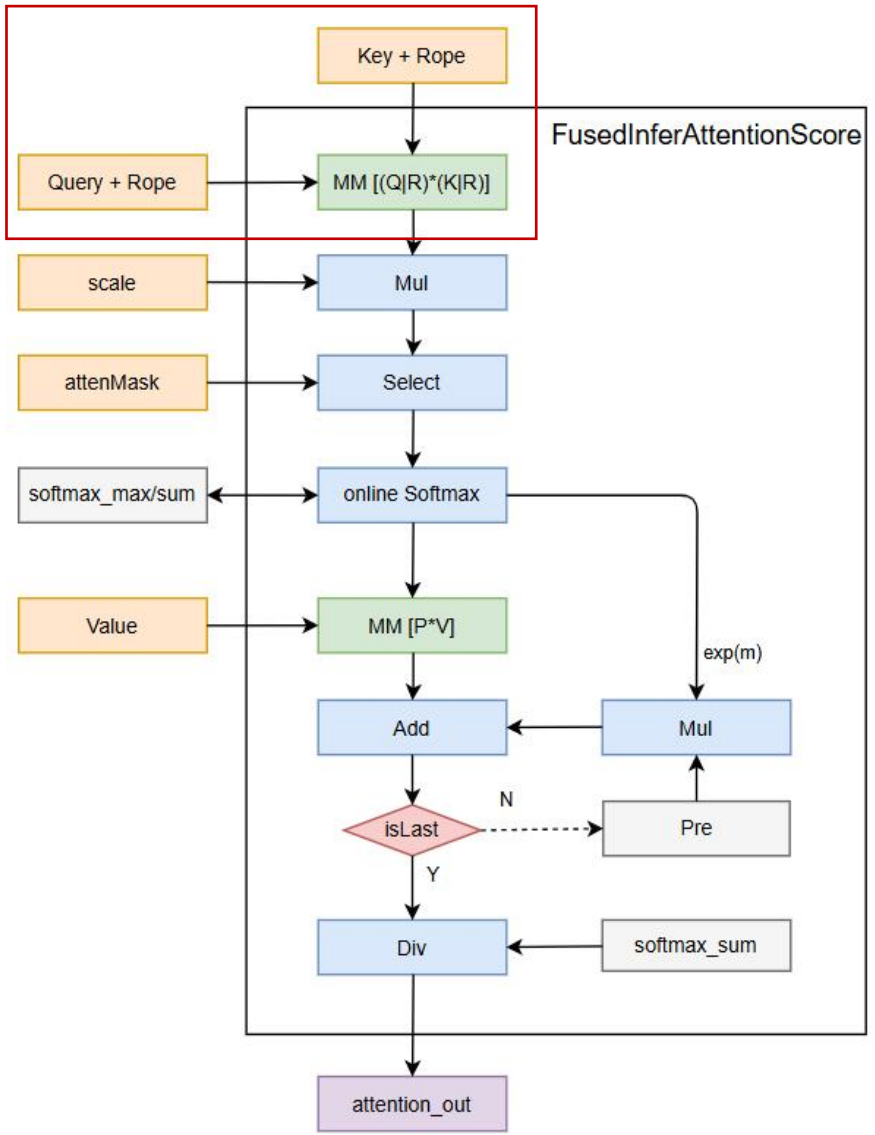


CANN

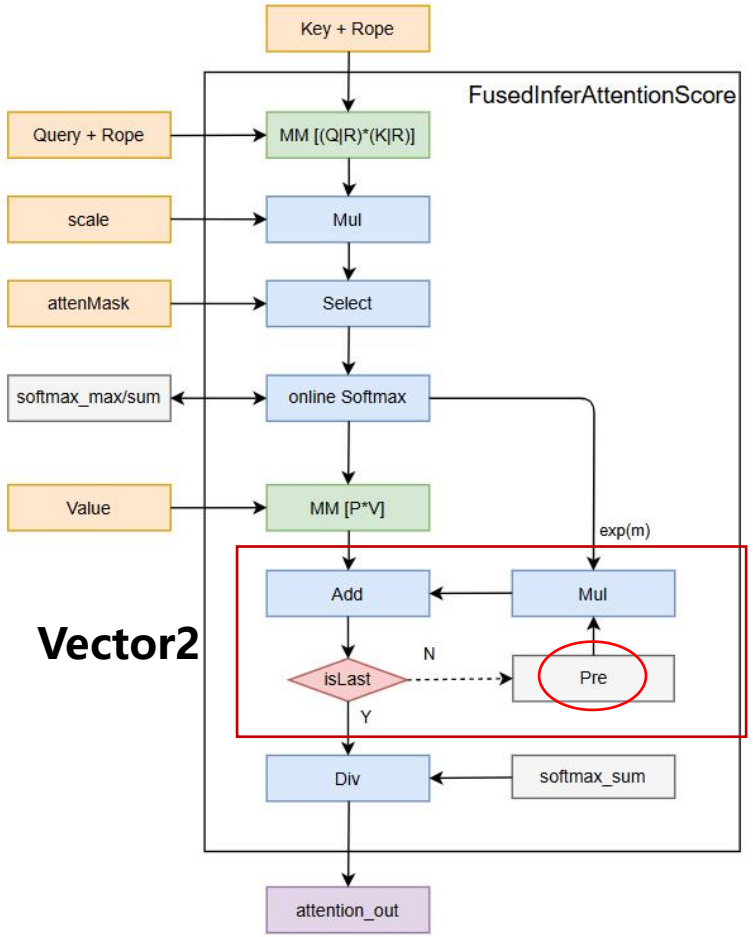
FIA算子适配MLA

DeepSeek的Decode阶段，使用了MLA技术。使用低秩联合压缩技术降低KV cache存储量；在MLA结构中增加了Rope旋转位置编码。



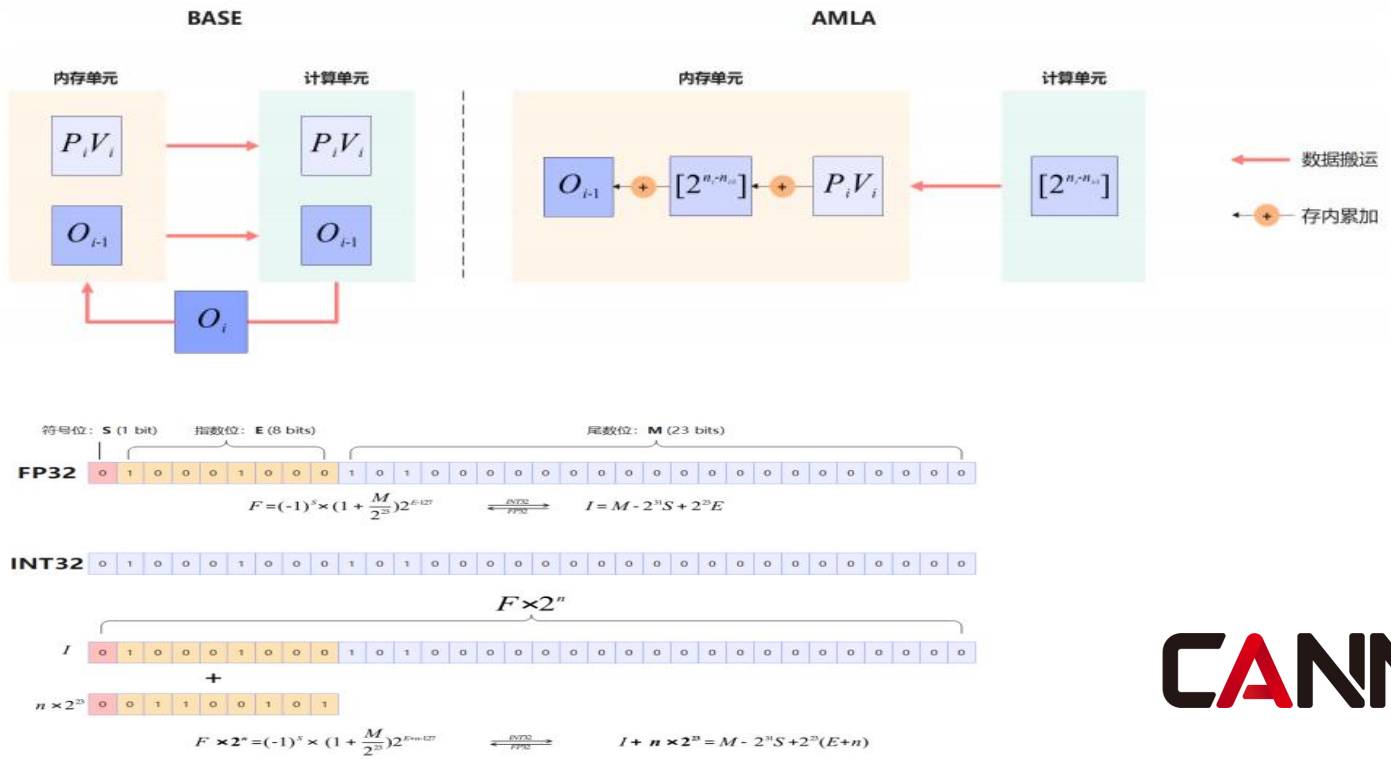


AMLA算法优化

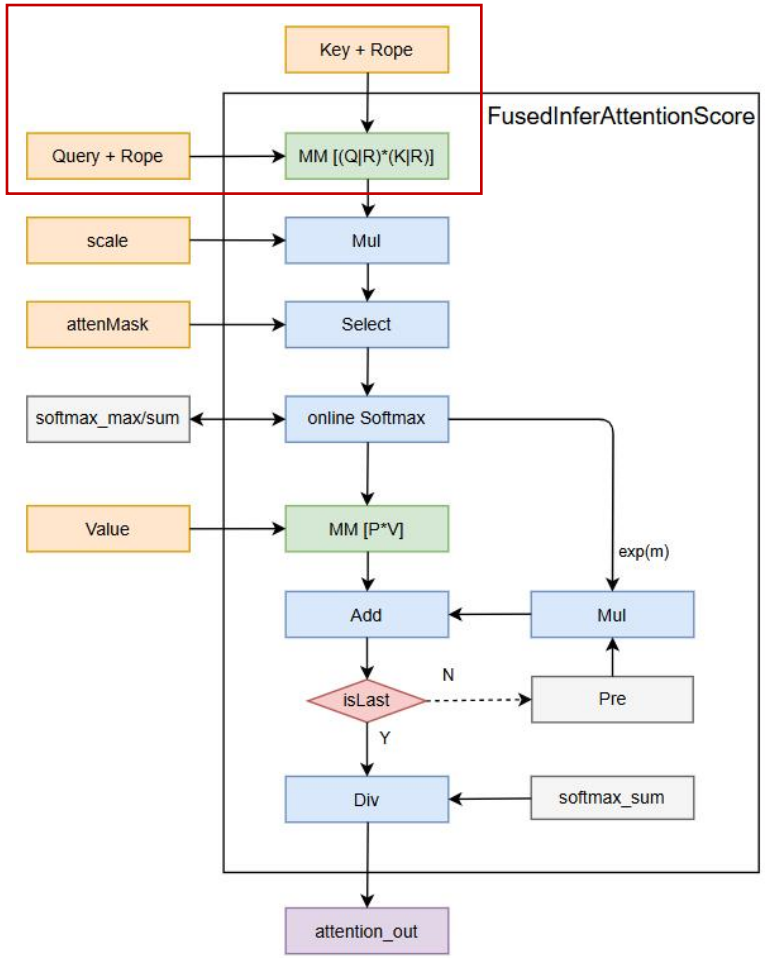


由于Q、K、V的HeadDim较大（512），Pre（Oi）的大小也会随变大。FIA在计算时，需要将 Pre（Oi）存储在GM上，在Vector2的流程中，需要先将Pre（Oi）从GM上搬入Vector的内存中，再进行乘法和加法计算，计算的结果再搬出到GM上。从而给FIA计算带来额外的访存开销。

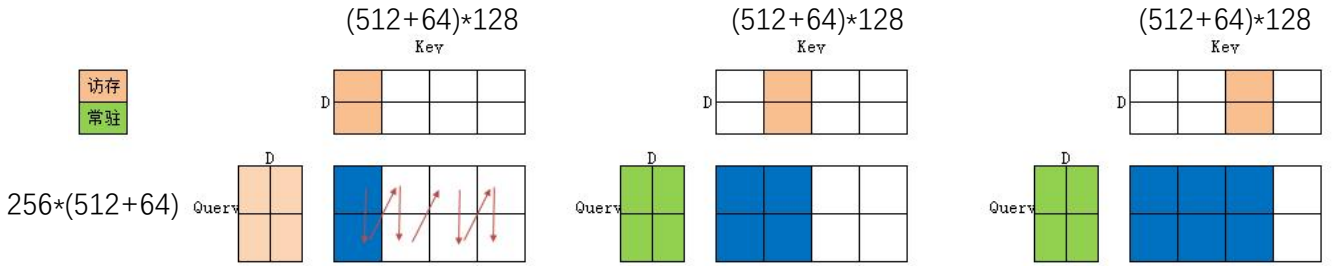
AMLA（Ascend MLA）算法利用浮点数的性质以及浮点数乘法与整数加法的等价变化，将需要在Vector上执行的乘法和加法操作，等价转换成了两次加法操作。使得Pre（Oi）可以直接使用芯片的原子累加能力，完成更新。



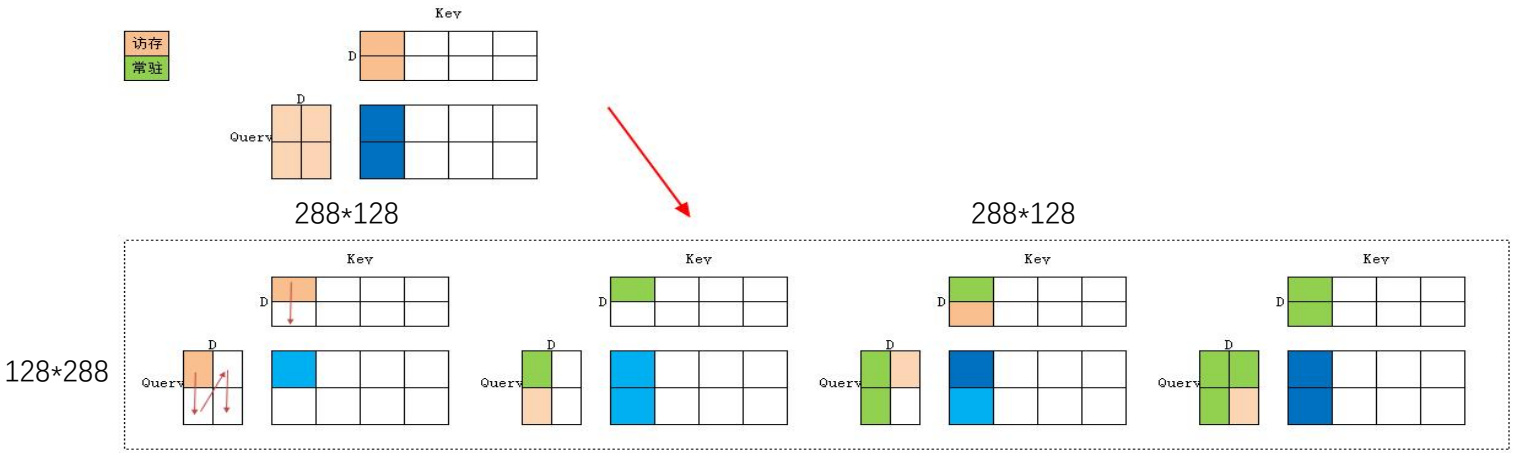
数据常驻优化



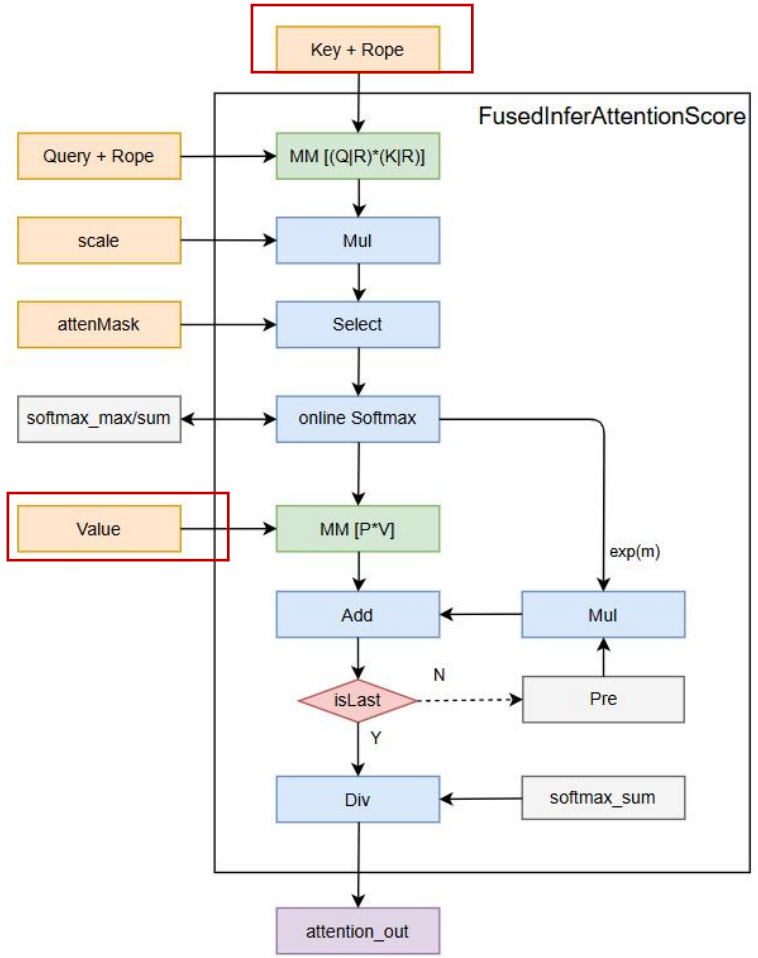
在单个Tiling块的计算过程中，Q+Qr的Tiling块大小为 256*576，K+Kr的Tiling块大小为 576*512。计算时将Q+Qr的数据(256*576)常驻内存，并将K+Kr按N轴切成4个基本块(576*128)，分别与常驻在内存中的Query做矩阵乘，减少Query数据的重复搬入



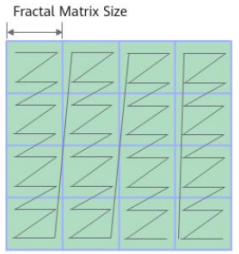
在Q+Qr与K+Kr的基本计算时，将Q+Qr按M轴和K轴均对半切的方式切成 128*288 的大小，将K+Kr按K轴对半切的方式切成 288*128 的大小，并且调整矩阵乘法的遍历顺序，先遍历M轴，再遍历K轴，使得被切块后的Key能做一个小常驻，减少对Key的重复搬入



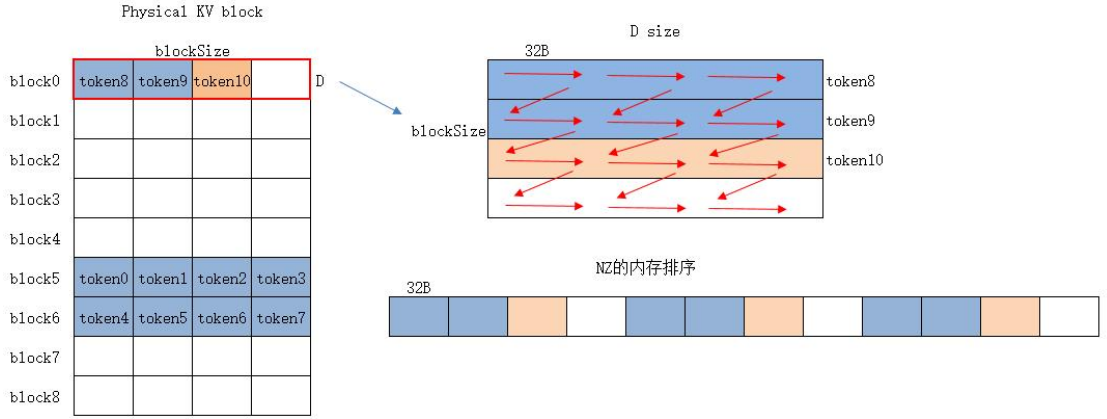
PA-NZ优化方案



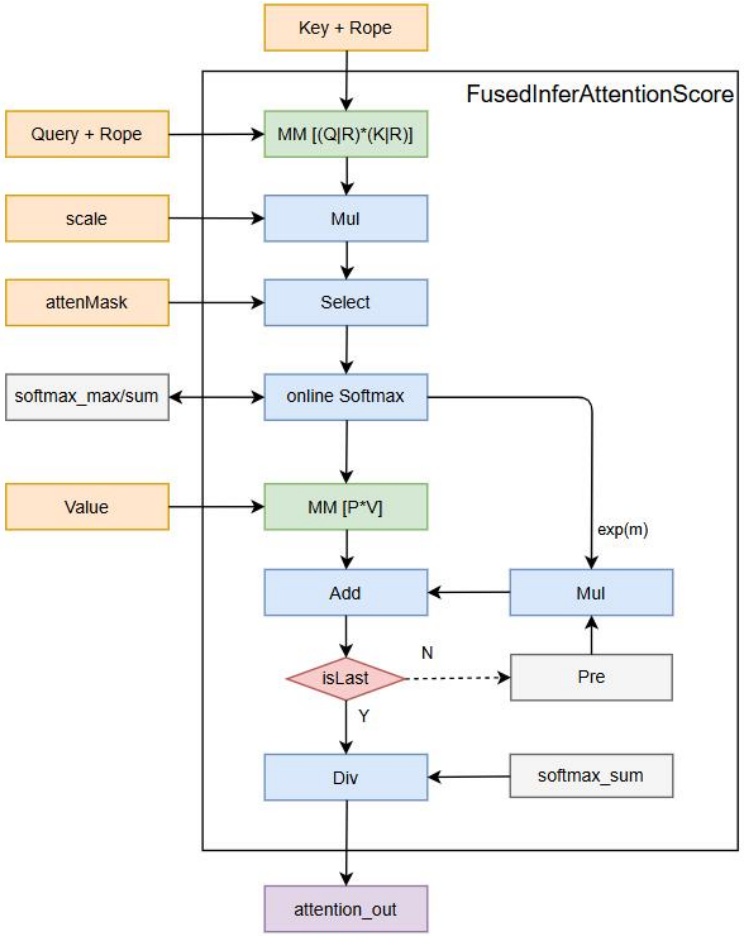
NZ 数据格式是一种专为提升计算性能而设计的特殊数据排布格式，在昇腾芯片中被广泛使用，能最大化挖掘 Cube 计算单元的性能潜力，在矩阵乘法操作中带来更高的计算效率

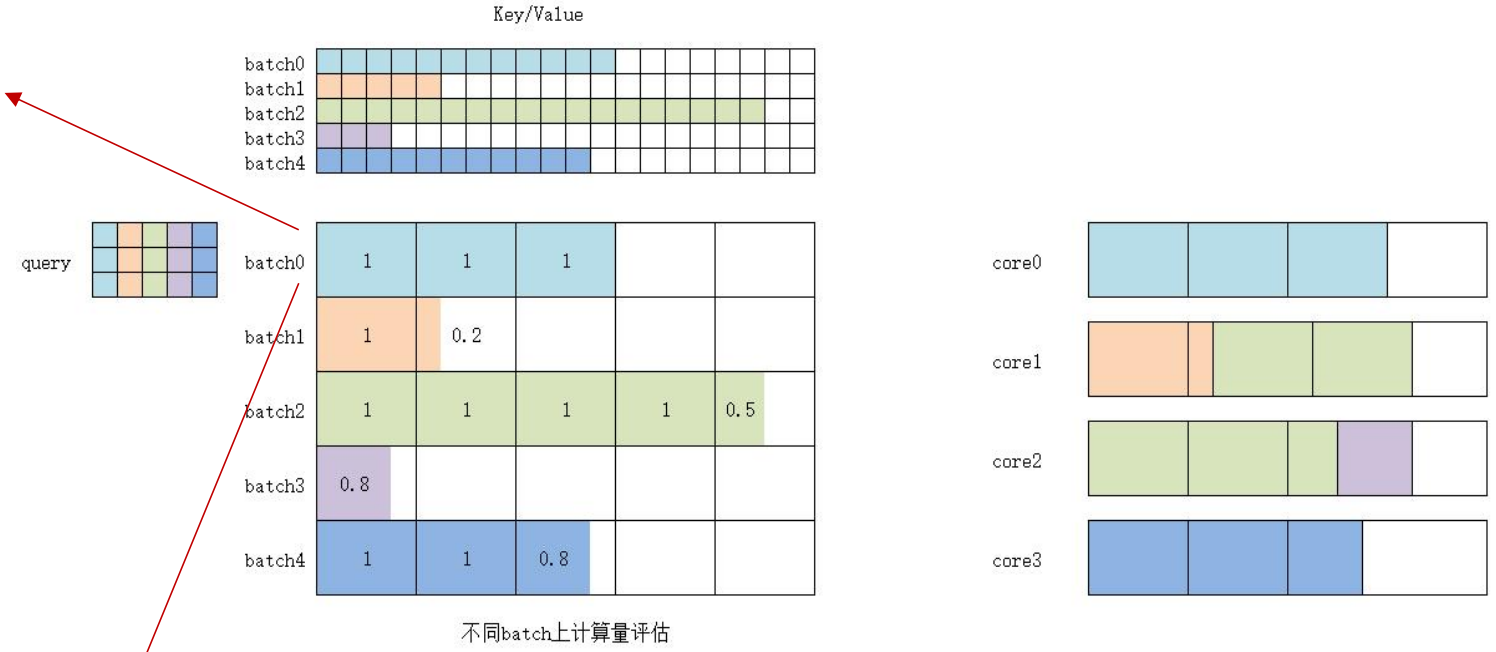


FIA算子直接按NZ格式将Key、Value数据从片上内存读入到Cube计算单元的内存中，避免访存过程中的ND、NZ数据格式转换，从而提升FIA算子的访存效率



负载均衡优化方案





不同Batch的计算耗时不相同。FIA根据实际Key/Value数据量大小（actual-sequence），对每个Batch的计算量（耗时）进行评估。

按Tiling块粒度，将每个Tiling块的计算均摊到多个核上进行计算。

并且支持利用Flash-Decoding技术，将Key/Value切分到不同核上进行计算。

CANN

Thank you.

社区愿景：打造开放易用、技术领先的AI算力新生态

社区使命：使能开发者基于CANN社区自主研究创新，构筑根深叶茂、跨产业协同共享共赢的CANN生态

Vision: Building an Open, Easy-to-Use, and Technology-leading AI Computing Ecosystem

Mission: Enable developers to independently research and innovate based on the CANN community and build a win-win CANN ecosystem with deep roots and cross-industry collaboration and sharing.

上CANN社区获取干货

关注CANN公众号获取资讯

https://gitcode.com/cann