| Operator | Input/Data Precision | Weight Precision | Bias/Accumulator Precision | Output Precision | constrains of params | Source File | Offline analysis File |
|---|---|---|---|---|---|---|---|
| ArgMax | Int8 | - | - | Int32(value,index) | 仅最后轴；rank>1时shape[0]=1；输出两维2；workspace 8B。 | executor/core/ops/argmax.c；executor/core/ops/venus/ | tools/tpacker/graph_analysis/ops/ArgMax.py |
| AvgPool2dInt | Int8 | - | Accumulator Int32 | Int8 | NCHW，batch=1，非ceil；stride 1/2/4，kernel>=stride；Int32累加，按平台分块使用SM。 | executor/core/ops/avgpool2dint.c；executor/core/ops/venus/avgpool2dint.h | tools/tpacker/graph_analysis/ops/Pool.py |
| BatchNorm2dInt | Int8 | Weight Int8 | Bias/Accumulator Int32 | Int8 | NCHW；weight/bias长度=C，bias scale=qx+qw；0<=qx+qw-qo<=63；输入输出SM。 | executor/core/ops/batchnorm2dint.c；executor/core/op | tools/tpacker/graph_analysis/ops/BatchNorm2dInt.py |
| BmmInt | Int8 | RHS Int8 | Accumulator Int32 | Int8 | 两输入同dtype、rank 2/3且K匹配；0<=qx+qy-qo<=63；PSRAM输出需SM workspace。 | executor/core/ops/bmmint.c；executor/core/ops/venus/ | tools/tpacker/graph_analysis/ops/BmmInt.py |
| Concat | Int8 | - | - | Int8 | 输入同rank/dtype，除axis外shape相同；输出同dtype；按平台限制scale移位，PSRAM需SM。 | executor/core/ops/concat.c；executor/core/ops/venus/concat.h | tools/tpacker/graph_analysis/ops/Concat.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Conv1dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | N,C,W；dilation=1，stride 1/2/4，kernel>=stride；Int32 bias/累加；权重离线压缩并按SM容量分块。 | executor/core/ops/conv1dint.c；executor/core/ops/venus/conv1dint.h | tools/tpacker/graph_analysis/ops/venus/Conv1dInt.py |
| | Int8 | Int4 packed | Bias/Accumulator Int32 | Int16 | | | |
| | Int8 | Int4 packed | Bias/Accumulator Int32 | Int32 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int16 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int32 | | | |
| Conv2dInt | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | NCHW，batch=1；stride 1/2/4，pad<kernel；0<=qx+qw-qo<=63；Int32 bias/累加，权重重排及PSRAM分块计入SM。 | executor/core/ops/conv2dint.c；executor/core/ops/venus/conv2dint.h | tools/tpacker/graph_analysis/ops/venus/Conv2dInt.py |
| ConvTranspose2dInt | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | NCHW，batch=1，group=1，dilation=1；stride 1/2/4，output_padding<stride；Int32累加，权重翻置重排。 | executor/core/ops/convtranspose2dint.c；executor/core/ops/venus/deconv2dint.h | tools/tpacker/graph_analysis/ops/venus/ConvTranspose2dInt.py |
| Flatten | Any byte dtype | - | - | Same as input | axis合法；仅重解释shape，必要时连续拷贝；dtype/scale不变。 | executor/core/ops/flatten.c；executor/core/ops/venus/f | tools/tpacker/graph_analysis/ops/Flatten.py |
| Gather | Any byte dtype | Indices Int64 | - | Same as data | indices在[Int32/Int64，axis合法；输出scale与data相同；按元素字节宽度搬运。 | executor/core/ops/gather.c；executor/core/ops/venus/gather.h | tools/tpacker/graph_analysis/ops/Gather.py |
| | Any byte dtype | Indices Int32 | - | Same as data | | | |
| GluInt | Int8 | Gate Int8 | Accumulator Int32 | Int8 | split轴长度为正偶数；输入Int8；sigmoid查表由shift=<63；SM/PSRAM按平台分块。 | executor/core/ops/gluint.c；executor/core/ops/venus/gl | tools/tpacker/graph_analysis/ops/GLUInt.py |
| GRUInt | Int8 | Weights Int8 | Bias/Accumulator Int32 | Int8 | 单层、batch=1；输入/权重Int8，bias/门累加Int32；输出scale=hidden scale；状态及workspace在SM。 | executor/core/ops/gruint.c；executor/core/ops/venus/gr | tools/tpacker/graph_analysis/ops/GRUInt.py |
| iqAdd | Int8 | RHS Int8 | Accumulator Int8 | Int8 | 两shape且zero point=0；按qx-qo/qy-qo移位；PSRAM或重标定使用SM workspace。 | executor/core/ops/iqadd.c；executor/core/ops/venus/iqadd.h | tools/tpacker/graph_analysis/ops/iqAdd.py |
| iqDiv | Int32 | RHS Int32 vector | Accumulator Int32 | Int32 | 输入输出SM；向量仅Int32；标量lhs/rhs/y同dtype且除数为正2次幂；0<=qo-(qx-qy)<=63。 | executor/core/ops/iqdiv.c；executor/core/ops/venus/iqdiv.h | tools/tpacker/graph_analysis/ops/iqDiv.py |
| | Int8 | Scalar Int8 | - | Int8 | | | |
| | Int16 | Scalar Int16 | - | Int16 | | | |
| | Int32 | Scalar Int32 | - | Int32 | | | |
| iqMul | Int8 | RHS Int8 same/scalar/NC11 | Accumulator Int32 | Int8 | 输入同dtype、zero point=0；仅同shape/标量及对拼NC11广播；0<=qx+qy-qo<=63；PSRAM分块到SM。 | executor/core/ops/iqmul.c；executor/core/ops/venus/iqmul.h | tools/tpacker/graph_analysis/ops/iqMul.py |
| | Int16 | RHS Int16 same/scalar | Accumulator Int32 | Int16 | | | |
| | Int32 | RHS Int32 same/scalar | Accumulator Int32 | Int32 | | | |
| iqPad | Int8 | Pads Int64; Fill Int8 | - | Int8 | Int8，rank 1..4；pads为成对Int64，constant/reflect/replicate按平台限制；PSRAM按需暂存。 | executor/core/ops/iqpad.c；executor/core/ops/venus/iq | tools/tpacker/graph_analysis/ops/iqPad.py |
| iqSigmoid | Int16 | - | Accumulator Int16 | Int8 | zero point=0，输出固定Int8 Q7；输入转查表Q域，shift按平台限制；PSRAM分块使用SM。 | executor/core/ops/iqsigmoid.c；executor/core/ops/venus/iqsigmoid.h | tools/tpacker/graph_analysis/ops/iqSigmoid.py |
| iqSub | Int8 | RHS Int8 | Accumulator Int8 | Int8 | 两输入同shape Int8；qx-qo、qy-qo在平台范围；PSRAM或重标定需SM。 | executor/core/ops/iqsub.c；executor/core/ops/venus/iq | tools/tpacker/graph_analysis/ops/iqSub.py |
| iqSum | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴且长度>0；输出保留维1并同dtype；0<=qx-qo<=63；输入输出SM。 | executor/core/ops/iqsum.c；executor/core/ops/venus/iq | tools/tpacker/graph_analysis/ops/iqSum.py |
| iqTanh | Int16 | - | Accumulator Int16 | Int8 | 输入转查表Q域，输出Int8 Q7；zero point=0；PSRAM按平台分块到SM。 | executor/core/ops/iqtanh.c；executor/core/ops/venus/ | tools/tpacker/graph_analysis/ops/iqTanh.py |
| iqVar | Int8 | - | Accumulator Int32 | Int8 | Int8，归约最后/倒数第二轴；shift=2*qx-qo按平台限制；统计/转置使用SM workspace。 | executor/core/ops/iqvar.c；executor/core/ops/venus/iq | tools/tpacker/graph_analysis/ops/iqVar.py |
| LayerNormInt | Int8 | Gamma Int8/Int16 | Bias/Accumulator Int32 | Int8 | 尾部1/2维归一化；gamma/bias大小匹配，bias Int32；0<=qx<=15且0<=15+qw-qo<=63；统计缓冲在SM。 | executor/core/ops/layernormint.c；executor/core/ops/v | tools/tpacker/graph_analysis/ops/LayerNormInt.py |
| LinearInt | Int8 | Int8 packed | Bias Int8或Int16或Int32; Accumulator Int32 | Int8 | transB=1，输入rank 1..平台上限；weight为rank2，bias匹配输出宽度；shift按输出位宽限制；pack/DMA/转置计入SM。 | executor/core/ops/linearint.c；executor/core/ops/venus/linearint.h | tools/tpacker/graph_analysis/ops/LinearInt.py |
| | Int8 | Int8 packed | Bias Int8或Int16或Int32; Accumulator Int32 | Int16 | | | |
| | Int8 | Int8 packed | Bias Int8或Int16或Int32; Accumulator Int32 | Int32 | | | |
| LogSoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25并按输出scale重标定；workspace按轴长分块。 | executor/core/ops/logsoftmaxint.c；executor/core/ops/v | tools/tpacker/graph_analysis/ops/LogSoftmaxInt.py |
| LSTMInt | Int8 | Weights Int8 | Bias/Accumulator Int32; Cell Int32 Q15 | Int8 | 单层单向、rank3、batch=1；输入/权重Int8，bias Int32，cell Int32 Q15；状态与workspace在SM。 | executor/core/ops/lstmint.c；executor/core/ops/venus/ | tools/tpacker/graph_analysis/ops/LstmInt.py |
| MaxPool | Int8 | | | Int8 | NCHW Int8、非ceil；stride 1/2/4，kernel>=stride；大tensor/PSRAM按平台分块到SM。 | executor/core/ops/maxpool.c；executor/core/ops/venus/ | tools/tpacker/graph_analysis/ops/Pool.py |
| PReLU | Int8 | Slope Int32(attr) | - | Int8 | 输入输出shape/dtype相同且在SM；slope/post_shift在[0..63]且和<=63。 | executor/core/ops/prelu.c；executor/core/ops/venus/prelu.h | tools/tpacker/graph_analysis/ops/activations.py |
| | Int16 | Slope Int32(attr) | - | Int16 | | | |
| | Int32 | Slope Int32(attr) | - | Int32 | | | |
| ReLU | Int8 | - | - | Int8 | 0<=qo-qx<=63；SM路径按表支持；任一端PSRAM时仅Int8->Int8，workspace最多64KiB。 | executor/core/ops/relu.c；executor/core/ops/venus/relu.h | |
| | Int8 | - | - | Int16 | | | |
| | Int8 | - | - | Int32 | | | |
| | Int16 | - | - | Int8 | | | |
| | Int16 | - | - | Int16 | | | |
| | Int16 | - | - | Int32 | | | |
| | Int32 | - | - | Int8 | | | |
| | Int32 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Requant | Int8 | - | - | Int8 | 输入Int8、zero point=0；scale为2次幂，位宽/shift按平台限制；同位宽输入输出浪SM。 | executor/core/ops/requant.c；executor/core/ops/venus/requant.h | tools/tpacker/graph_analysis/ops/Requant.py |
| | Int8 | - | - | Int16 | | | |
| | Int8 | - | - | Int32 | | | |
| Reshape | Any byte dtype | - | - | Same as input | 常量shape可离线解析且元素数不变；dtype/scale不变，通常别名。 | executor/core/ops/reshape.c；executor/core/ops/venus | tools/tpacker/graph_analysis/ops/Reshape.py |
| ShuffleChannel | - | - | - | - | 不支持：虽有源码和未导入分析器文件，但缺少执行注册及分析器导入，普通模型不可达。 | executor/core/ops/shufflechannel.c；executor/core/ops | unsupported：ShuffleChannel.py 未导入，且 operator_list.h 未注册 |
| Slice | Any byte dtype | starts/ends/axes/steps Int32 | - | Same as input | 参数为单元素常量Int32/Int64；axis合法，step=1；dtype/scale不变，非连续PSRAM使用SM。 | executor/core/ops/slice.c；executor/core/ops/venus/slic | tools/tpacker/graph_analysis/ops/Slice.py |
| SoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25、输出由Q15重标定；workspace含exp2指数/求和缓冲。 | executor/core/ops/softmaxint.c；executor/core/ops/ven | tools/tpacker/graph_analysis/ops/SoftmaxInt.py |
| Split | Int8 | - | - | Int8 | axis合法，各split长度之和等于轴长；输出dtype/scale同输入；PSRAM按段使用SM/DMA。 | executor/core/ops/split.c；executor/core/ops/venus/split.h | tools/tpacker/graph_analysis/ops/Split.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Tile | Int8 | Repeats Int64 | - | Int8 | repeats为常量Int64且长度=rank、各值>0；dtype/scale不变；输出/重复块按需使用SM。 | executor/core/ops/tile.c；executor/core/ops/venus/tile.h | tools/tpacker/graph_analysis/ops/Tile.py |
| topN | Int8 | Index offset Int64 | - | Int16(value,index) | 仅最后轴、max_num=1、shape[0]=1；offset为Int64；输出值/索引对，workspace按平台固定，workspace 8B。 | executor/core/ops/topn.c；executor/core/ops/venus/topn | tools/tpacker/graph_analysis/ops/topN.py |
| topN2 | Int16(value,index) | - | - | Int16(value,index) | 输入rank3且首维2，仅最后轴、max_num=1；输入输出scale相同并位于SM；workspace按平台固定，workspace 8B。 | executor/core/ops/topn2.c；executor/core/ops/venus/to | tools/tpacker/graph_analysis/ops/topN2.py |
| Transpose | Int8 | - | - | Same as input | rank/perm须满足平台白名单；dtype/scale不变；矩阵容量及PSRAM搬进由SM workspace托管。 | executor/core/ops/transpose.c；executor/core/ops/venus/transpose.h | tools/tpacker/graph_analysis/ops/Transpose.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Quant | Float32 | - | - | Int8 | 模型边界转换；输入输出shape相同，zero point=0，scale为正2次幂；当前有效输出仅Int8，不能原地复用。 | executor/core/ops/quant.c；executor/core/comm/utils.c | tools/tpacker/graph_analysis/ops/Quant.py |
| Dequant | Int8\|UInt8\|Int32 | - | - | Float32 | 模型边界转换；输入输出shape相同，zero point=0，scale指数为整数且在[0,30]；Float32输出不提供进入算通计算算子。 | executor/core/ops/dequant.c；executor/core/comm/util | tools/tpacker/graph_analysis/ops/Dequant.py |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ArgMax | Int8 | - | - | Int32(value,index) | 仅最后轴；rank>1时shape[0]=1；输出首维2；workspace 8B。 | executor/core/ops/argmax.c；executor/core/ops/arcs/argmax.h | tools/tpacker/graph_analysis/ops/ArgMax.py |
| | Int32 | - | - | Int32(value,index) | | | |
| AvgPool2dInt | Int8 | - | Accumulator Int32 | Int8 | NCHW、batch=1、平ceil；stride 1/2/4，kernel>=stride；Int32累加，按平台分块使用SM。 | executor/core/ops/avgpool2dint.c；executor/core/ops/arcs/avgpool2dint.h | tools/tpacker/graph_analysis/ops/Pool.py |
| BatchNorm2dInt | Int8 | Weight Int8 | Bias/Accumulator Int32 | Int8 | NCHW；weight/bias长度=C，bias scale=qx+qw；0<=qx+qw-qo<=63；输入输出SM。 | executor/core/ops/batchnorm2dint.c；executor/core/ops/arcs/batchnorm2dint.h | tools/tpacker/graph_analysis/ops/BatchNorm2dInt.py |
| BmmInt | Int8 | RHS same dtype | Accumulator Int32 | Int8 | 两输入同dtype、rank 2/3且K匹配；0<=qx+qy-qo<=63；PSRAM输出需SM workspace。 | executor/core/ops/bmmint.c；executor/core/ops/arcs/bmmint.h | tools/tpacker/graph_analysis/ops/BmmInt.py |
| | Int8 | RHS same dtype | Accumulator Int32 | Int32 | | | |
| | Int32 | RHS same dtype | Accumulator Int32 | Int8 | | | |
| | Int32 | RHS same dtype | Accumulator Int32 | Int32 | | | |
| Concat | Int8 | - | - | Int8 | 输入同rank/dtype，除axis外shape相同；输出同dtype；按平台限制scale移位，PSRAM需SM。 | executor/core/ops/concat.c；executor/core/ops/arcs/concat.h | tools/tpacker/graph_analysis/ops/Concat.py |
| | Int32 | - | - | Int32 | | | |
| Conv1dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | N,C,W；dilation=1，stride 1/2/4，kernel>=stride；Int32 bias/累加；权重离线压缩并按SM容量分块。 | executor/core/ops/conv1dint.c；executor/core/ops/arcs/conv1dint.h | tools/tpacker/graph_analysis/ops/arcs/Conv1dInt.py |
| | Int8 | Int4 packed | Bias/Accumulator Int32 | Int32 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int32 | | | |
| Conv2dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8/Int32 | NCHW、batch=1；stride 1/2/4，pad<kernel；0<=qx+qw-qo<=63；Int32 bias/累加，权重重排及PSRAM分块计入SM。ARCS固定Int8、4/8-bit权重、输出Int8/Int32。 | executor/core/ops/conv2dint.c；executor/core/ops/arcs/conv2dint.h | tools/tpacker/graph_analysis/ops/arcs/Conv2dInt.py |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8/Int32 | | | |
| ConvTranspose2dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | NCHW、batch=1、group=1、dilation=1；stride 1/2/4，output_padding<stride；Int32累加，权重翻转重排。 | executor/core/ops/convtranspose2dint.c；executor/core/ops/arcs/deconv2dint.h | tools/tpacker/graph_analysis/ops/arcs/ConvTranspose2dInt.py |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| FFNInt | Int8 | Weights Int8 | Bias/Accumulator Int32 | Int8 | runtime-only：有ARCS wrapper/backend，但无普通离线analysis，普通ONNX打包不可达；精度及workspace须由预构造资源满足。 | executor/core/ops/ffnint.c；executor/core/ops/arcs/ffnint.h | runtime-only：无普通离线 analysis |
| Flatten | Any byte dtype | - | - | Same as input | axis合法；仅重解释shape，必要时连续拷贝；dtype/scale不变。 | executor/core/ops/flatten.c；executor/core/ops/arcs/flatten.h | tools/tpacker/graph_analysis/ops/Flatten.py |
| Gather | Any byte dtype | Indices Int64 | - | Same as data | indices仅Int32/Int64，axis合法；输出scale与data相同；按元素字节宽度搬运。 | executor/core/ops/gather.c；executor/core/ops/arcs/gather.h | tools/tpacker/graph_analysis/ops/Gather.py |
| | Any byte dtype | Indices Int32 | - | Same as data | | | |
| GluInt | Int8 | Gate Int8 | Accumulator Int32 | Int8 | split轴长度为正偶数；输入Int8；sigmoid及乘法shift<=63；SM/PSRAM按平台分块。 | executor/core/ops/gluint.c；executor/core/ops/arcs/gluint.h | tools/tpacker/graph_analysis/ops/GLUInt.py |
| GRUInt | Int8 | Weights Int8 | Bias/Accumulator Int32 | Int8 | 单层、batch=1；输入/权重Int8，bias/门累加Int32；输出scale=hidden scale；状态及workspace在SM。 | executor/core/ops/gruint.c；executor/core/ops/arcs/gruint.h | tools/tpacker/graph_analysis/ops/GRUInt.py |
| IqAdd | Int8 | RHS Int8 | Accumulator Int8 | Int8 | 同shape且zero point=0；按qx-qo/qy-qo移位；PSRAM或重标定使用SM workspace。 | executor/core/ops/iqadd.c；executor/core/ops/arcs/iqadd.h | tools/tpacker/graph_analysis/ops/IqAdd.py |
| IqDiv | Int32 | RHS Int32 vector | Accumulator Int32 | Int32 | 输入输出SM；向量仅Int32；标量lhs/rhs/Y同dtype且除数为正2次幂；0<=qo-(qx-qy)<=63。 | executor/core/ops/iqdiv.c；executor/core/ops/arcs/iqdiv.h | tools/tpacker/graph_analysis/ops/IqDiv.py |
| | Int8 | Scalar Int8 | - | Int8 | | | |
| | Int32 | Scalar Int32 | - | Int32 | | | |
| IqMul | Int8 | RHS Int8 same/scalar/NC11 | Accumulator Int32 | Int8 | 输入同dtype、zero point=0；Int8支持同shape/标量/NC11广播，Int32仅同shape/标量；0<=qx+qy-qo<=63；PSRAM分块到SM。 | executor/core/ops/iqmul.c；executor/core/ops/arcs/iqmul.h | tools/tpacker/graph_analysis/ops/IqMul.py |
| | Int32 | RHS Int32 same/scalar | Accumulator Int32 | Int32 | | | |
| IqPad | Int8 | Pads Int64; Fill Int8 | - | Int8 | Int8，rank 1..4；pads为成对Int64，constant/reflect/replicate按平台限制；PSRAM按需暂存。 | executor/core/ops/iqpad.c；executor/core/ops/arcs/iqpad.h | tools/tpacker/graph_analysis/ops/IqPad.py |
| IqSigmoid | Int8/Int32 | - | Accumulator Int32 | Int8 | zero point=0，输出固定Int8 Q7；输入转查表Q域，shift按平台限制；PSRAM分块使用SM。 | executor/core/ops/iqsigmoid.c；executor/core/ops/arcs/iqsigmoid.h | tools/tpacker/graph_analysis/ops/IqSigmoid.py |
| IqSub | Int8 | RHS Int8 | Accumulator Int8 | Int8 | 两输入同shape Int8；qx-qo、qy-qo在平台范围；PSRAM或重标定需SM。 | executor/core/ops/iqsub.c；executor/core/ops/arcs/iqsub.h | tools/tpacker/graph_analysis/ops/IqSub.py |
| IqSum | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴且长度>0；输出保留维1并同dtype；0<=qx-qo<=63；输入输出SM。 | executor/core/ops/iqsum.c；executor/core/ops/arcs/iqsum.h | tools/tpacker/graph_analysis/ops/IqSum.py |
| IqTanh | Int32 Q27 | - | Accumulator Int32 | Int8 Q7 | 输入转查表Q域，输出Int8 Q7；zero point=0；PSRAM按平台分块到SM。 | executor/core/ops/iqtanh.c；executor/core/ops/arcs/iqtanh.h | tools/tpacker/graph_analysis/ops/IqTanh.py |
| IqVar | Int8 | - | Accumulator Int32 | Int8 | Int8，归约最后/倒数第二轴；shift=2*qx-qo按平台限制；统计/转置使用SM workspace。 | executor/core/ops/iqvar.c；executor/core/ops/arcs/iqvar.h | tools/tpacker/graph_analysis/ops/IqVar.py |
| LayerNormInt | Int8 | Gamma Int8 | Bias/Accumulator Int32 | Int8 | 末尾1..2维归一化；gamma/bias大小匹配，bias Int32；0<=qx<=15且0<=15+qw-qo<=63；统计缓冲在SM。 | executor/core/ops/layernormint.c；executor/core/ops/arcs/layernormint.h | tools/tpacker/graph_analysis/ops/LayerNormInt.py |
| LinearInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | transB=1，输入rank 1..平台上限；weight为rank2，bias匹配输出宽度；shift按输出位宽限制；pack/DMA/转置计入SM。 | executor/core/ops/linearint.c；executor/core/ops/arcs/linearint.h | tools/tpacker/graph_analysis/ops/LinearInt.py |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int32 | | | |
| LogSoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25并按输出scale重标定；workspace按轴长分块。 | executor/core/ops/logsoftmaxint.c；executor/core/ops/arcs/logsoftmaxint.h | tools/tpacker/graph_analysis/ops/LogSoftmaxInt.py |
| LSTMInt | Int8 | Weights Int8 | Bias/Accumulator Int32; Cell Int32 Q15 | Int8 | 单层单向、rank3、batch=1；输入/权重Int8，bias Int32，cell Int32 Q15；状态与workspace在SM。 | executor/core/ops/lstmint.c；executor/core/ops/arcs/lstmint.h | tools/tpacker/graph_analysis/ops/LstmInt.py |
| MaxPool | Int8 | - | - | Int8 | NCHW Int8、平ceil；stride 1/2/4，kernel>=stride；大tensor/PSRAM按平台分块到SM。 | executor/core/ops/maxpool.c；executor/core/ops/arcs/maxpool.h | tools/tpacker/graph_analysis/ops/Pool.py |
| MultiHeadAttentionInt | Int8 | Weights Int8 | Bias/Accumulator Int32 | Int8 | runtime-only：无普通离线analysis/设备入口；必须提供workspace，精度及布局由预构造ARCS资源保证。 | executor/core/ops/multiheadattentionint.c；executor/core/ops/arcs/multiheadattentionint.h | runtime-only：无普通离线 analysis |
| PReLU | Int8 | Slope Int32(attr) | - | Int8 | 输入输出shape/dtype相同且在SM；slope/post_shift对0..63且和<=63。 | executor/core/ops/prelu.c；executor/core/ops/arcs/prelu.h | tools/tpacker/graph_analysis/ops/activations.py |
| | Int32 | Slope Int32(attr) | - | Int32 | | | |
| ReLU | Int8 | - | - | Int8 | 0<=qo-qx<=63；SM路径按表支持；任一端PSRAM时仅Int8->Int8，workspace最多64KiB。 | executor/core/ops/relu.c；executor/core/ops/arcs/relu.h | |
| | Int8 | - | - | Int32 | | | |
| | Int32 | - | - | Int8 | | | |
| | Int32 | - | - | Int32 | | | |
| ReLUx | Int8 | - | - | Int8 | 输入输出SM；输出Int8；threshold装入Int8，shift 0..63。 | executor/core/ops/relux.c；executor/core/ops/arcs/relux.h | |
| | Int32 | - | - | Int8 | | | |
| Requant | Int8 | - | - | Int8/Int32 | 输入Int8、zero point=0；scale为2次幂，位宽/shift按平台限制；同位宽输入输出浩SM。 | executor/core/ops/requant.c；executor/core/ops/arcs/requant.h | tools/tpacker/graph_analysis/ops/Requant.py |
| Reshape | Any byte dtype | - | - | Same as input | 常量shape可离线解析且元素数不变；dtype/scale不变，通常别名。 | executor/core/ops/reshape.c；executor/core/ops/arcs/reshape.h | tools/tpacker/graph_analysis/ops/Reshape.py |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ShuffleChannel | - | - | - | - | 不支持：虽有源码和未导入分析器文件，但缺少执行注册及分析器导入。普通模型不可达。 | executor/core/ops/shufflechannel.c；executor/core/ops/venus/shufflechannel.h | unsupported：ShuffleChannel.py 未导入，且 operator_list.h 未注册 |
| Slice | Any byte dtype | starts/ends/axes/steps Int32\|Int64 | - | Same as input | 参数为单元素常量Int32/Int64；axis合法，step=1；dtype/scale不变，非连续/PSRAM使用SM。 | executor/core/ops/slice.c；executor/core/ops/arcs/slice.h | tools/tpacker/graph_analysis/ops/Slice.py |
| SoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25、输出由Q15重标定；workspace含Int32指数/求和缓冲。 | executor/core/ops/softmaxint.c；executor/core/ops/arcs/softmaxint.h | tools/tpacker/graph_analysis/ops/SoftmaxInt.py |
| | Int8 | - | Accumulator Int32 | Int32 | | | |
| SparifyFFNInt | Int8 | Weights Int4\|Int8 + Mask | Bias/Accumulator Int32 | Int8 | runtime-only：无普通离线analysis/设备入口；稀疏布局、精度及workspace须由预构造ARCS资源保证。 | executor/core/ops/sparifyffnint.c；executor/core/ops/arcs/sparifyffnint.h | runtime-only：无普通离线 analysis |
| Split | Int8 | - | - | Int8 | axis合法，各split长度之和等于轴长；输出dtype/scale同输入；PSRAM按段使用SM/DMA。 | executor/core/ops/split.c；executor/core/ops/arcs/split.h | tools/tpacker/graph_analysis/ops/Split.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Tile | Int8 | Repeats Int64 | - | Int8 | repeats为常量Int64且长度=rank、各值>0；dtype/scale不变；输出/重复块按需使用SM。 | executor/core/ops/tile.c；executor/core/ops/arcs/tile.h | tools/tpacker/graph_analysis/ops/Tile.py |
| topN | Int8 | Index offset Int64 | - | Int32(value,index) | 仅最后轴、max_num=1、shape[0]=1；offset为Int64；输出值/索引对，workspace按平台固定。workspace 16B。 | executor/core/ops/topn.c；executor/core/ops/arcs/topn.h | tools/tpacker/graph_analysis/ops/topN.py |
| topN2 | Int32(value,index) | - | - | Int32(value,index) | 输入rank3且首维2，仅最后轴、max_num=1；输入输出scale相同并位于SM；workspace按平台固定。workspace 16B。 | executor/core/ops/topn2.c；executor/core/ops/arcs/topn2.h | tools/tpacker/graph_analysis/ops/topN2.py |
| Transpose | Int8 | - | - | Same as input | rank/perm须满足平台白名单；dtype/scale不变；矩阵容量及PSRAM搬运由SM workspace约束。 | executor/core/ops/transpose.c；executor/core/ops/arcs/transpose.h | tools/tpacker/graph_analysis/ops/Transpose.py |
| | Int32 | - | - | Int32 | | | |
| Quant | Float32 | - | - | Int8 | 模型边界转换；输入输出shape相同，zero point=0，scale为正2次幂；当前有效输出仅Int8，不能原地复用。 | executor/core/ops/quant.c；executor/core/comm/utils.c | tools/tpacker/graph_analysis/ops/Quant.py |
| Dequant | Int8\|UInt8\|Int32 | - | - | Float32 | 模型边界转换；输入输出shape相同，zero point=0，scale指数为整数且在[0,30]；Float32输出不得继续进入普通计算算子。 | executor/core/ops/dequant.c；executor/core/comm/utils.c | tools/tpacker/graph_analysis/ops/Dequant.py |

| Operator | Input/Data Precision | Weight Precision | Bias/Accumulator Precision | Output Precision | constrains of params | Source File | Offline analysis File |
|---|---|---|---|---|---|---|---|
| ArgMax | Int8 | - | - | Int32(value,index) | 仅最后轴；rank>1时shape[0]=1；输出首维2；workspace BB。 | executor/core/ops/argmax.c；executor/core/ops/venusA/argmax.h | tools/tpacker/graph_analysis/ops/ArgMax.py |
| | Int16 | - | - | Int32(value,index) | | | |
| | Int32 | - | - | Int32(value,index) | | | |
| AvgPool2dInt | Int8 | - | Accumulator Int32 | Int8 | NCHW、batch=1、非ceil；stride 1/2/4，kernel>=stride；Int32累加，按平台分块使用SM。 | executor/core/ops/avgpool2dint.c；executor/core/ops/venusA/avgpool2dint.h | tools/tpacker/graph_analysis/ops/Pool.py |
| BatchNorm2dInt | Int8 | Weight Int8 | Bias/Accumulator Int32 | Int8 | NCHW；weight/bias长度=C，bias scale=qx+qw；0<=qx+qw-qo<=63；输入输出SM。 | executor/core/ops/batchnorm2dint.c；executor/core/ops/venusA/batchnorm2dint.h | tools/tpacker/graph_analysis/ops/BatchNorm2dInt.py |
| BmmInt | Int8 | RHS same dtype | Accumulator Int32 | Int8 | 两输入同dtype、rank 2/3且K匹配；0<=qx+qy-qo<=63；PSRAM输出需SM workspace。 | executor/core/ops/bmmint.c；executor/core/ops/venusA/bmmint.h | tools/tpacker/graph_analysis/ops/BmmInt.py |
| | Int8 | RHS same dtype | Accumulator Int32 | Int16 | | | |
| | Int8 | RHS same dtype | Accumulator Int32 | Int32 | | | |
| | Int16 | RHS same dtype | Accumulator Int32 | Int8 | | | |
| | Int16 | RHS same dtype | Accumulator Int32 | Int16 | | | |
| | Int16 | RHS same dtype | Accumulator Int32 | Int32 | | | |
| | Int32 | RHS same dtype | Accumulator Int32 | Int8 | | | |
| | Int32 | RHS same dtype | Accumulator Int32 | Int16 | | | |
| | Int32 | RHS same dtype | Accumulator Int32 | Int32 | | | |
| Concat | Int8 | - | - | Int8 | 输入同rank/dtype，除axis外shape相同；输出同dtype；按平台限制scale移位，PSRAM需SM。 | executor/core/ops/concat.c；executor/core/ops/venusA/concat.h | tools/tpacker/graph_analysis/ops/Concat.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Conv1dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | N,C,W；dilation=1，stride 1/2/4，kernel>=stride；Int32 bias/累加；权重离线压缩并按SM容量分块。 | executor/core/ops/conv1dint.c；executor/core/ops/venusA/conv1dint.h | tools/tpacker/graph_analysis/ops/venusA/Conv1dInt.py |
| | Int8 | Int4 packed | Bias/Accumulator Int32 | Int16 | | | |
| | Int8 | Int4 packed | Bias/Accumulator Int32 | Int32 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int16 | | | |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int32 | | | |
| Conv2dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | NCHW、batch=1；stride 1/2/4，pad<kernel；0<=qx+qw-qo<=63；Int32 bias/累加，权重重排及PSRAM分块计入SM。 | executor/core/ops/conv2dint.c；executor/core/ops/venusA/conv2dint.h | tools/tpacker/graph_analysis/ops/venusA/Conv2dInt.py |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| ConvTranspose2dInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | NCHW、batch=1、group=1、dilation=1；stride 1/2/4，output_padding<stride；Int32累加，权重翻转重排。 | executor/core/ops/convtranspose2dint.c；executor/core/ops/venusA/deconv2dint.h | tools/tpacker/graph_analysis/ops/venusA/ConvTranspose2dInt.py |
| | Int8 | Int8 packed | Bias/Accumulator Int32 | Int8 | | | |
| Flatten | Any byte dtype | - | - | Same as input | axis合法；仅重解释shape，必要时连续拷贝；dtype/scale不变。 | executor/core/ops/flatten.c；executor/core/ops/venusA/flatten.h | tools/tpacker/graph_analysis/ops/Flatten.py |
| Gather | Any byte dtype | Indices Int64 | - | Same as data | indices仅Int32/Int64，axis合法；输出scale与data相同；按元素字节宽度搬运。 | executor/core/ops/gather.c；executor/core/ops/venusA/gather.h | tools/tpacker/graph_analysis/ops/Gather.py |
| | Any byte dtype | Indices Int32 | - | Same as data | | | |
| QGelu | Int8 | - | Accumulator Int32 Q27 | Int8 | 仅VenusA；输入SM、输出Int8；scale指数qx[-3.90]/qo[-36,57]；Int8时68/元素、Int16/32约48/元素workspace。 | executor/core/ops/gelu.c；executor/core/ops/venusA/gelu.h | tools/tpacker/graph_analysis/ops/Gelu.py |
| | Int16 | - | Accumulator Int32 Q27 | Int8 | | | |
| | Int32 | - | Accumulator Int32 Q27 | Int8 | | | |
| GluInt | Int8 | Gate Int8 | Accumulator Int32 | Int8 | split轴长度为正偶数；输入Int8；sigmoid及乘法shift-<=63；SM/PSRAM按平台分块。 | executor/core/ops/gluint.c；executor/core/ops/venusA/gluint.h | tools/tpacker/graph_analysis/ops/GLUInt.py |
| | Int8 | Gate Int8 | Accumulator Int32 | Int16 | | | |
| | Int8 | Gate Int8 | Accumulator Int32 | Int32 | | | |
| GRUInt | Int8 | Weights Int8 | Bias/Accumulator Int32 | Int8 | 单层、batch=1；输入/权重Int8，bias/门累加Int32；输出scale=hidden scale；状态及workspace在SM。 | executor/core/ops/gruint.c；executor/core/ops/venusA/gruint.h | tools/tpacker/graph_analysis/ops/GRUInt.py |
| IqAdd | Int8 | RHS same dtype/same shape | Accumulator same dtype | Int8 | 同shape且zero point=0；按qx-qo/qy-qo移位；PSRAM或重标定使用SM workspace。VenusA离线拒绝标量，仅同shapeInt8/16/32匹配输入输出。 | executor/core/ops/iqadd.c；executor/core/ops/venusA/iqadd.h | tools/tpacker/graph_analysis/ops/IqAdd.py |
| | Int16 | RHS same dtype/same shape | Accumulator same dtype | Int16 | | | |
| | Int32 | RHS same dtype/same shape | Accumulator same dtype | Int32 | | | |
| IqDiv | Int32 | RHS Int32 vector | Accumulator Int32 | Int32 | 输入输出SM；向量仅Int32；标量lhs/rhs/Y同dtype且除数为正2次幂；0<=qo-(qx-qy)<=63。 | executor/core/ops/iqdiv.c；executor/core/ops/venusA/iqdiv.h | tools/tpacker/graph_analysis/ops/IqDiv.py |
| | Int32 | Scalar Int32 | - | Int32 | | | |
| IqMul | Int8 | RHS Int8 same/scalar/NC11 | Accumulator Int32 | Int8 | 输入同dtype、zero point=0；仅同shape/标量及同时NC11广播；0<=qx+qy-qo<=63；PSRAM分块到SM。VenusA仅Int8支持NCHW*NC11；不支持任意rank广播。 | executor/core/ops/iqmul.c；executor/core/ops/venusA/iqmul.h | tools/tpacker/graph_analysis/ops/IqMul.py |
| | Int16 | RHS Int16 same/scalar | Accumulator Int32 | Int8 | | | |
| | Int16 | RHS Int16 same/scalar | Accumulator Int32 | Int16 | | | |
| | Int32 | RHS Int32 same/scalar | Accumulator Int32 | Int32 | | | |
| IqPad | Int8 | Pads Int64; Fill Int8(0) | - | Int8 | Int8，rank 1..4；pads为成对Int64，constant/reflect/replicate按平台限制；PSRAM按需暂存。 | executor/core/ops/iqpad.c；executor/core/ops/venusA/iqpad.h | tools/tpacker/graph_analysis/ops/IqPad.py |
| IqSigmoid | Int8 | - | Accumulator Int32 | Int8 | zero point=0，输出固定Int8 Q7；输入转查表Q域，shift按平台限制；PSRAM分块使用SM。 | executor/core/ops/iqsigmoid.c；executor/core/ops/venusA/iqsigmoid.h | tools/tpacker/graph_analysis/ops/IqSigmoid.py |
| | Int16 | - | Accumulator Int32 | Int8 | | | |
| | Int32 | - | Accumulator Int32 | Int8 | | | |
| IqSub | Int8 | RHS Int8 | Accumulator Int8 | Int8 | 两输入同shape Int8；qx-qo、qy-qo在平台范围；PSRAM或重标定需SM。 | executor/core/ops/iqsub.c；executor/core/ops/venusA/iqsub.h | tools/tpacker/graph_analysis/ops/IqSub.py |
| IqSum | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴且长度>0；输出保留维1并同dtype；0<=qx-qo<=63；输入输出SM。 | executor/core/ops/iqsum.c；executor/core/ops/venusA/iqsum.h | tools/tpacker/graph_analysis/ops/IqSum.py |
| | Int16 | - | Accumulator Int32 | Int16 | | | |
| | Int32 | - | Accumulator Int32 | Int32 | | | |
| IqTanh | Int32 Q27 | - | Accumulator Int32 | Int8 Q7 | 输入转查表Q域，输出Int8 Q7；zero point=0；PSRAM按平台分块到SM。 | executor/core/ops/iqtanh.c；executor/core/ops/venusA/iqtanh.h | tools/tpacker/graph_analysis/ops/IqTanh.py |
| IqVar | Int8 | - | Accumulator Int32 | Int8 | Int8，归约最后/倒数第二轴；shift=2*qx-qo按平台限制；统计/转置使用SM workspace。 | executor/core/ops/iqvar.c；executor/core/ops/venusA/iqvar.h | tools/tpacker/graph_analysis/ops/IqVar.py |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| LayerNormInt | Int8 | Gamma Int8 | Bias/Accumulator Int32 | Int8 | 末部1/2维归一化；gamma/bias大小匹配，bias Int32；0<=qx<=15且0<=15+qw-qo<=63；统计缓冲在SM。 | executor/core/ops/layernormint.c；executor/core/ops/venusA/layernormint.h | tools/tpacker/graph_analysis/ops/LayerNormInt.py |
| LinearInt | Int8 | Int4 packed | Bias/Accumulator Int32 | Int8 | transB=1，输入rank 1..平台上限；weight为rank2，bias匹配输出宽度；shift按输出位宽限制；pack/DMA/转置计入SM。 显式组合：(8,8w4)->8；(8,8w8)->8/16/32；(8,32w32)->32；(16,16w16)->8/16/32；(32,32w32)->8/32；(32,8w8)->8。 | executor/core/ops/linearint.c；executor/core/ops/venusA/linearint.h | tools/tpacker/graph_analysis/ops/LinearInt.py |
| | Int8 | Int16 | Weight same dtype | Bias/Accumulator Int32 | Int8 | Int16 | Int32 |
| | Int8 | Int32 | Bias/Accumulator Int32 | Int32 | | | |
| | Int32 | Int32 | Bias/Accumulator Int32 | Int8 | Int32 | | |
| | Int32 | Int8 | Bias/Accumulator Int32 | Int8 | | | |
| LogSoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25并按输出scale重标定；workspace按输长分块。 | executor/core/ops/logsoftmaxint.c；executor/core/ops/venusA/logsoftmaxint.h | tools/tpacker/graph_analysis/ops/LogSoftmaxInt.py |
| LSTMInt | Int8 | Weights Int8 | Bias/Accumulator Int32; Cell Int32 Q15 | Int8 | 单层单向、rank3、batch=1；输入/权重Int8、bias Int32，cell Int32 Q15；状态与workspace在SM。 | executor/core/ops/lstmint.c；executor/core/ops/venusA/lstmint.h | tools/tpacker/graph_analysis/ops/LstmInt.py |
| MaxPool | Int8 | - | - | Int8 | NCHW Int8、非ceil；stride 1/2/4，kernel>=stride；大tensor/PSRAM按平台分块到SM。 | executor/core/ops/maxpool.c；executor/core/ops/venusA/maxpool.h | tools/tpacker/graph_analysis/ops/Pool.py |
| PReLU | Int8 | Slope Int32(attr) | - | Int8 | 输入输出shape/dtype相同且在SM；slope/post_shift对0..63首和<=63。 | executor/core/ops/prelu.c；executor/core/ops/venusA/prelu.h | tools/tpacker/graph_analysis/ops/activations.py |
| | Int16 | Slope Int32(attr) | - | Int16 | | | |
| | Int32 | Slope Int32(attr) | - | Int32 | | | |
| ReLU | Int8 | - | - | Int8 | 0<=qo-qx<=63；SM路径按表支持；任一端PSRAM时仅Int8->Int8，workspace最多64KiB。 | executor/core/ops/relu.c；executor/core/ops/venusA/relu.h | |
| | Int8 | - | - | Int16 | | | |
| | Int8 | - | - | Int32 | | | |
| | Int16 | - | - | Int8 | | | |
| | Int16 | - | - | Int16 | | | |
| | Int16 | - | - | Int32 | | | |
| | Int32 | - | - | Int8 | | | |
| | Int32 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| ReLUx | Int8 | - | - | Int8 | 输入输出SM；输出Int8；threshold装入Int8，shift 0..63。 | executor/core/ops/relux.c；executor/core/ops/venusA/relux.h | |
| Requant | Int8 | - | - | Int8 | 输入Int8、zero point=0；scale为2次幂，位宽/shift按平台限制；同位宽输入输出须SM。 | executor/core/ops/requant.c；executor/core/ops/venusA/requant.h | tools/tpacker/graph_analysis/ops/Requant.py |
| | Int8 | - | - | Int16 | | | |
| | Int8 | - | - | Int32 | | | |
| Reshape | Any byte dtype | - | - | Same as input | 常量shape可离线解析且元素数不变；dtype/scale不变，通常别名。 | executor/core/ops/reshape.c；executor/core/ops/venusA/reshape.h | tools/tpacker/graph_analysis/ops/Reshape.py |
| Slice | Any byte dtype | starts/ends/axes/steps Int32/Int64 | - | Same as input | 参数为单元素常量Int32/Int64；axis合法，step=1；dtype/scale不变，非连续/PSRAM使用SM。 | executor/core/ops/slice.c；executor/core/ops/venusA/slice.h | tools/tpacker/graph_analysis/ops/Slice.py |
| SoftmaxInt | Int8 | - | Accumulator Int32 | Int8 | 仅最后轴、长度1..2048；zero point=0；输入转Q25、输出由Q15重标定；workspace含Int32指数/求和缓冲。 | executor/core/ops/softmaxint.c；executor/core/ops/venusA/softmaxint.h | tools/tpacker/graph_analysis/ops/SoftmaxInt.py |
| | Int8 | - | Accumulator Int32 | Int16 | | | |
| | Int8 | - | Accumulator Int32 | Int32 | | | |
| | Int16 | - | Accumulator Int32 | Int8 | | | |
| | Int16 | - | Accumulator Int32 | Int16 | | | |
| | Int16 | - | Accumulator Int32 | Int32 | | | |
| | Int32 | - | Accumulator Int32 | Int8 | | | |
| | Int32 | - | Accumulator Int32 | Int16 | | | |
| | Int32 | - | Accumulator Int32 | Int32 | | | |
| Split | Int8 | - | - | Int8 | axis合法，各split长度之和等于轴长；输出dtype/scale同输入；PSRAM按段使用SM/DMA。 | executor/core/ops/split.c；executor/core/ops/venusA/split.h | tools/tpacker/graph_analysis/ops/Split.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| QSwish | Int8 | - | Accumulator Int32 Q27 | Int8 | 仅VenusA；输入/输出SM，输出Int8；scale指数qx[-3,90]/qo[-36,57]；Int8约6B/元素、Int16/32约4B/元素workspace。 | executor/core/ops/swish.c；executor/core/ops/venusA/swish.h | tools/tpacker/graph_analysis/ops/Swish.py |
| | Int16 | - | Accumulator Int32 Q27 | Int8 | | | |
| | Int32 | - | Accumulator Int32 Q27 | Int8 | | | |
| Tile | Int8 | Repeats Int64 | - | Int8 | repeats为常量Int64且长度=rank。各值>0；dtype/scale不变；输出/重复块按需使用SM。 | executor/core/ops/tile.c；executor/core/ops/venusA/tile.h | tools/tpacker/graph_analysis/ops/Tile.py |
| topN | Int8 | Index offset Int64 | - | Int32(value,index) | 仅最后轴，max_num=1，shape[0]=1；offset为Int64；输出值/索引对，workspace按平台固定，workspace 16B。 | executor/core/ops/topn.c；executor/core/ops/venusA/topn.h | tools/tpacker/graph_analysis/ops/topN.py |
| topN2 | Int32(value,index) | - | - | Int32(value,index) | 输入rank3且首维2，仅最后轴、max_num=1；输入输出scale相同并位于SM；workspace按平台固定，workspace 16B。 | executor/core/ops/topn2.c；executor/core/ops/venusA/topn2.h | tools/tpacker/graph_analysis/ops/topN2.py |
| Transpose | Int8 | - | - | Same as input | rank/perm须满足平台白名单；dtype/scale不变；矩阵容量及PSRAM搬运由SM workspace约束。 | executor/core/ops/transpose.c；executor/core/ops/venusA/transpose.h | tools/tpacker/graph_analysis/ops/Transpose.py |
| | Int16 | - | - | Int16 | | | |
| | Int32 | - | - | Int32 | | | |
| Quant | Float32 | - | - | Int8 | 模型边界转换；输入输出shape相同，zero point=0，scale为正2次幂；当前有效输出仅Int8，不能原地复用。 | executor/core/ops/quant.c；executor/core/comm/utils.c | tools/tpacker/graph_analysis/ops/Quant.py |
| Dequant | Int8\|UInt8\|Int32 | - | - | Float32 | 模型边界转换；输入输出shape相同，zero point=0，scale指数为整数且在[0,30]；Float32输出不得继续进入普通计算算子。 | executor/core/ops/dequant.c；executor/core/comm/utils.c | tools/tpacker/graph_analysis/ops/Dequant.py |