

Stable Diffusion XL Turbo UNet FP32 512x512

Shamith Achanta

06.03.2024

1 Assumptions

- The set of operators that have the same output memory size are likely to be fused and computed as a single operator to reduce the number of times the output needs to be read from memory. Hence, the total memory of the blocks in red are not counted in the analysis.
- The on-chip memory on the NPU is a parameter. In this analysis, the on-chip memory is set to 4 MB and data (weights + output) with memory size greater than the on-chip memory will need to be stored in the Last-level cache (if-any) or Main Memory

Figure 1: Optimization 1

| Node | Operator | Memory (in Bytes) | Output Size | Inputs Memory (in Bytes) | Weights and Bias Memory (in Bytes) | Output Memory (in Bytes) | Weights and Bias Memory (in MB) | Output Memory (in MB) | Memory (in MB) |
|----------------------|----------|----------------------|----------------|--------------------------------|---|--------------------------------|--|--------------------------|-------------------|
| /down_blorReshape | | 2621440 | 655360 | 2621472 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorTranspose | | 2621440 | 655360 | 2621440 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| Constant_ζConstant | | 8 | 1 | 0 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| /down_blorUnsqueeze | | 8 | 1 | 8 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| Constant_ζConstant | | 8 | 1 | 0 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| /down_blorUnsqueeze | | 8 | 1 | 8 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| Constant_ζConstant | | 8 | 1 | 0 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| /down_blorUnsqueeze | | 8 | 1 | 8 | 0 | 8 | 0 | 8.00E-06 | 8.00E-06 |
| /down_blorConcat | | 24 | 3 | 24 | 0 | 24 | 0 | 2.40E-05 | 2.40E-05 |
| /down_blorReshape | | 2621440 | 655360 | 2621464 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorMatMul | | 4259840 | 655360 | 2621440 | 1638400 | 2621440 | 1.6384 | 2.62144 | 4.25984 |
| /down_blorAdd | | 2624000 | 655360 | 2621440 | 2560 | 2621440 | 0.00256 | 2.62144 | 2.624 |
| /down_blorDiv | | 2621440 | 655360 | 2621440 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorAdd | | 2621440 | 655360 | 5242880 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorReduceMean | | 4096 | 1024 | 2621440 | 0 | 4096 | 0 | 0.004096 | 0.004096 |
| /down_blorSub | | 2621440 | 655360 | 2625536 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorPow | | 2621440 | 655360 | 2621440 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorReduceMean | | 4096 | 1024 | 2621440 | 0 | 4096 | 0 | 0.004096 | 0.004096 |
| /down_blorAdd | | 4096 | 1024 | 4096 | 0 | 4096 | 0 | 0.004096 | 0.004096 |
| /down_blorSqrt | | 4096 | 1024 | 4096 | 0 | 4096 | 0 | 0.004096 | 0.004096 |
| /down_blorDiv | | 2621440 | 655360 | 2625536 | 0 | 2621440 | 0 | 2.62144 | 2.62144 |
| /down_blorMul | | 2624000 | 655360 | 2621440 | 2560 | 2621440 | 0.00256 | 2.62144 | 2.624 |
| /down_blorAdd | | 2624000 | 655360 | 2621440 | 2560 | 2621440 | 0.00256 | 2.62144 | 2.624 |
| /down_blorMatMul | | 4259840 | 655360 | 2621440 | 1638400 | 2621440 | 1.6384 | 2.62144 | 4.25984 |
| /down_blorMatMul | | 5440000 | 49280 | 630784 | 5242880 | 197120 | 5.24288 | 0.19712 | 5.44 |
| /down_blorMatMul | | 5440000 | 49280 | 630784 | 5242880 | 197120 | 5.24288 | 0.19712 | 5.44 |
| /down_blorShape | | 24 | 3 | 2621440 | 0 | 24 | 0 | 2.40E-05 | 2.40E-05 |

2 Operator Memory Distribution

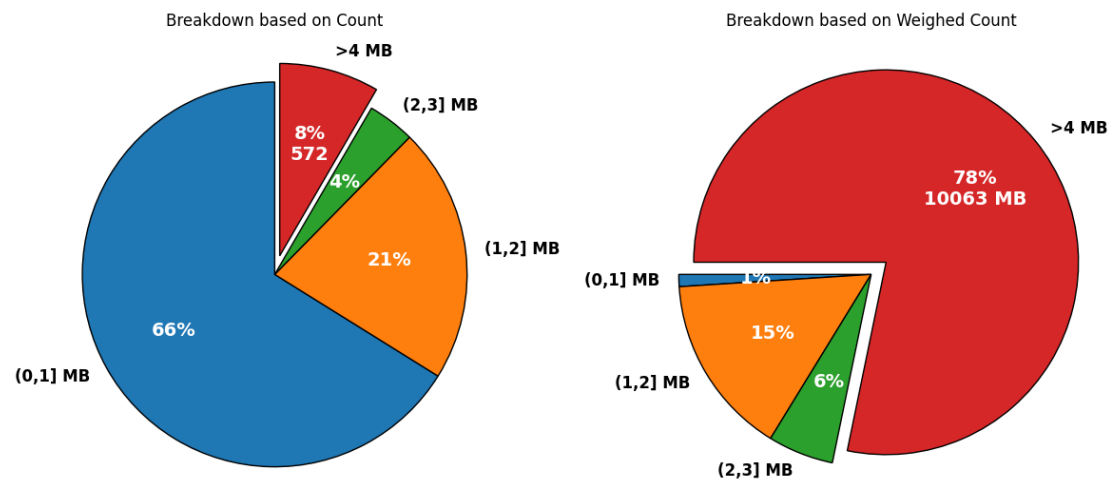
- Output + Weight matrices above on-chip memory size for an operator need to be stored in the Main Memory or last-level cache (if-any)
- Total memory of all operators that have memory size $>$ on-chip memory size is 10 GB

Figure 2: Operator Memory Distribution

SDXL Turbo UNet FP32 512x512

Should Weights + Output of an Operator
be stored in Main Memory during single inference?

If memory size of the Operator > 4 MB
(on-chip memory) with no NPU or Last-level cache



3 Memory Requirement of Individual Operators

Operators that have weights + output memory size $>$ on-chip memory size

Figure 3: Memory Requirement of Individual Operators

