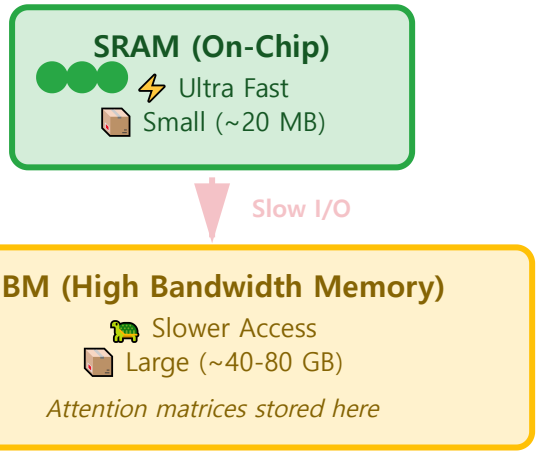
Flash Attention: IO-Aware Optimization

GPU Memory Hierarchy



Standard Attention:

- ✗ Load full attention matrix
- ✗ Multiple HBM reads/writes
- ✗ $O(n^2)$ memory bottleneck

Flash Attention:

- ✓ Block-wise computation
- ✓ Minimize HBM access
- ✓ O(n) memory usage

Flash Attention v1

- Tiling and recomputation
- Block-wise attention
- 2-4x speedup

Flash Attention v2

- Better parallelization
- Improved work partitioning
- 5-9x faster than standard

Medical Impact

- Real-time patient analysis
- Affordable long-context
- Clinical deployment ready

Key Innovation

- Fuse operations in SRAM
- Avoid materialization
- 3-10x less HBM access

🚀 Revolutionary Efficiency

Flash Attention achieves exact attention with O(n) memory and 3-10x speedup by optimizing GPU memory access patterns, making long-context medical AI practical. **The bottleneck is I/O, not computation!**