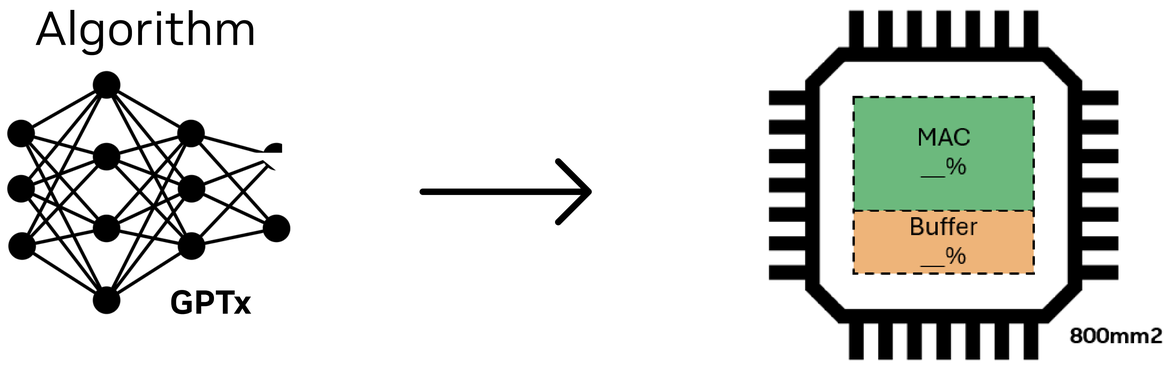
Mind the Gap: Attainable Data Movement and Operational Intensity Bounds for Tensor Algorithms

Qijing Huang, Po-An Tsai, Joel S Emer, Angshuman Parashar
{jennyhuang, poant aparashar}@nvidia.com, emer@csail.mit.edu

A Key Design Challenge



How to provision chip area between storage and compute?

DSE ?

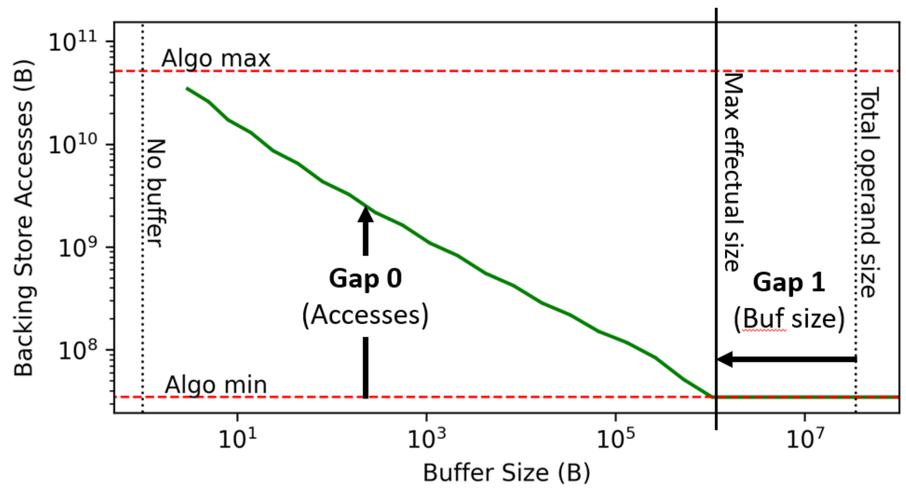
- Expensive
- Lack of insights

Roofline Analysis ?

- No buffer size suggestion

Buffer storage affects the **achievable data movement**

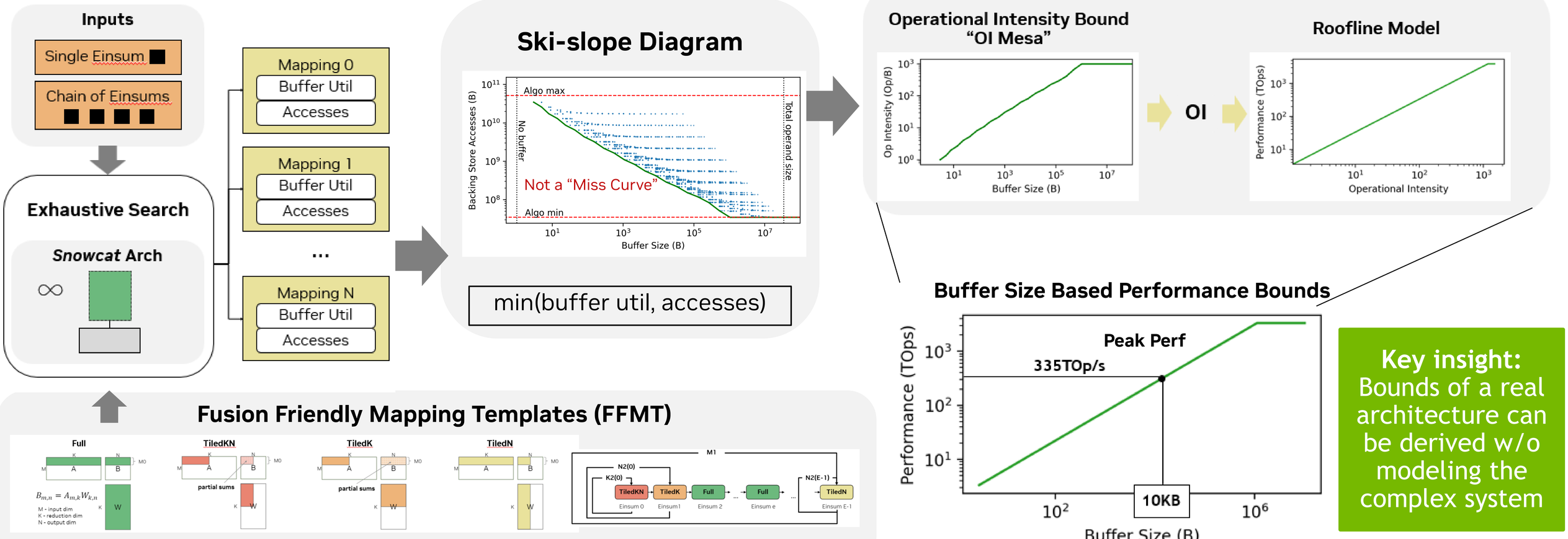
Data Movement vs. Buffer Size Gaps



[Gap 0] Given a buffer capacity, what is the minimal attainable *data access count*?
[Gap 1] How much additional capacity is required to achieve full data reuse?
[rate of change of Gap 0] How does an algorithm benefit from incremental increase in buffer capacity?

1 2

Our **Orojenesis** Flow for Early-Stage DSE



Key insight: Bounds of a real architecture can be derived w/o modeling the complex system

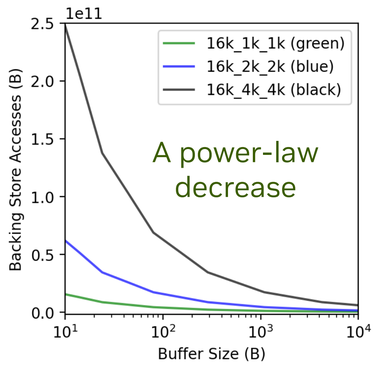
- Fusion demands consistent **order** and **granularity** of tiles

3

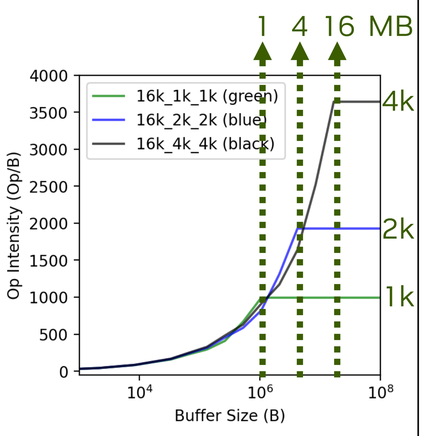
4

1. Provides valuable design insights

Data Movement Bound



OI Bound



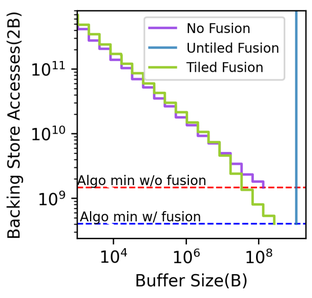
Observations :

1. The optimal OI of a GEMM is limited by its **smallest dimension**
2. The maximal effectual buffer size of a GEMM is approximately **the size of its smallest operand**

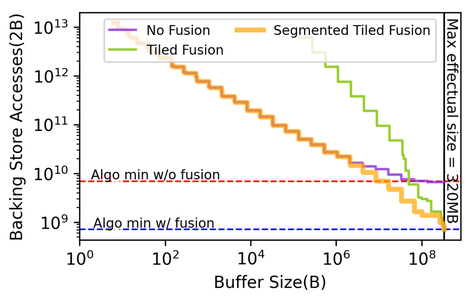
5

2. Comprehends complex fusion space

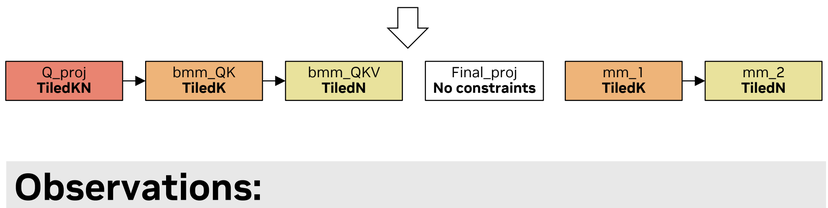
Fusing 32k x 4k x 16k, 32k x 16k x 4k



Fusing GPT-6.7b Block w/ 6 Einsums



Chain Segmentation



Observations:

1. Bounds w/ fusion do not always outperform the unfused bounds
2. Fusing all layers can be suboptimal

6

3. Offers fast area tradeoff in early DSE

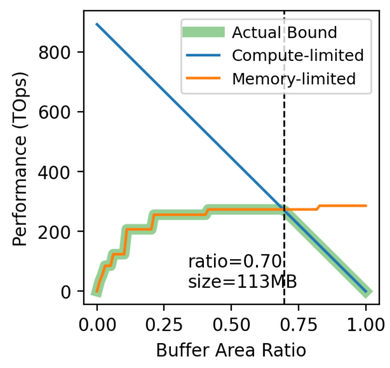
Algorithm

GTP3-6.7b Prefilling
Seq len=2k, batch size=16, feature dim=4k

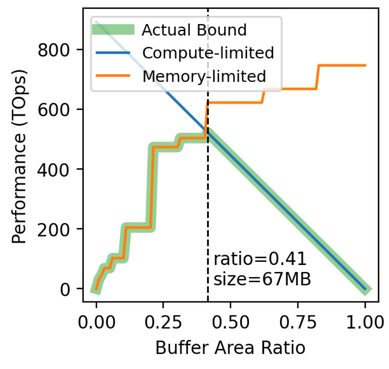
HW Specs

529 um^2 in 40nm
2.59 um^2/ B, 332.5 um^2/MAC
149GB/s DRAM BW @700MHz

Unfused



Fused



Observations:

1. Performance is a **concave** function wrt to the buffer area ratio
2. Fused workload demands **lower** buffer area but leads to better performance