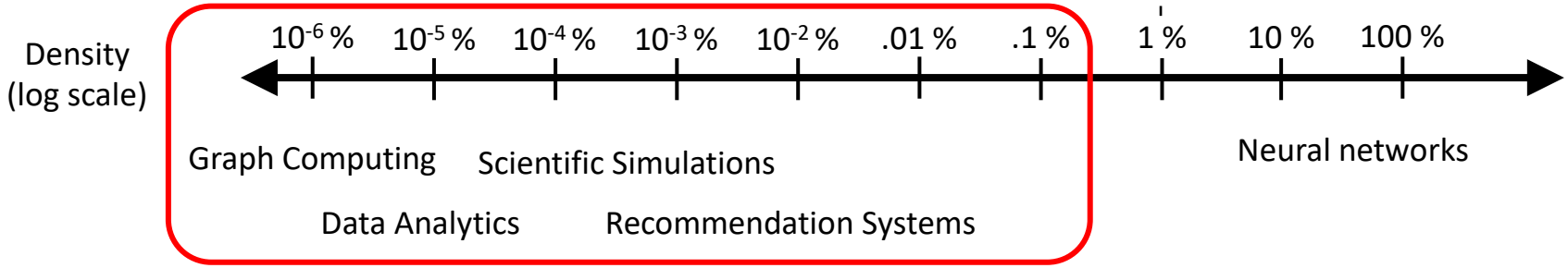
Accelerating Sparse Tensor Algebra by Overbooking Buffer Occupancy

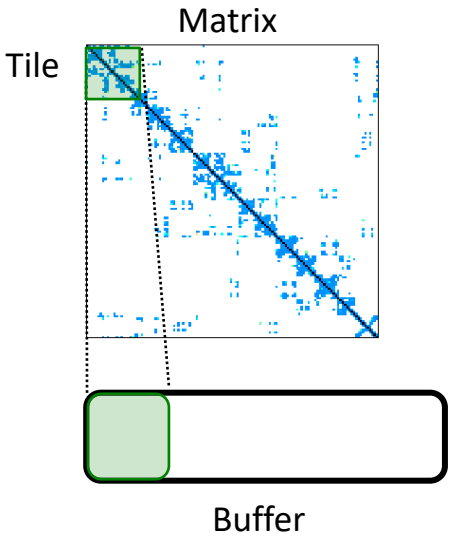
Fisher Xue, Nellie Wu, Joel Emer, Vivienne Sze

RESEARCH LABORATORY OF ELECTRONICS AT MIT
MTL microsystems technology laboratories massachusetts institute of technology

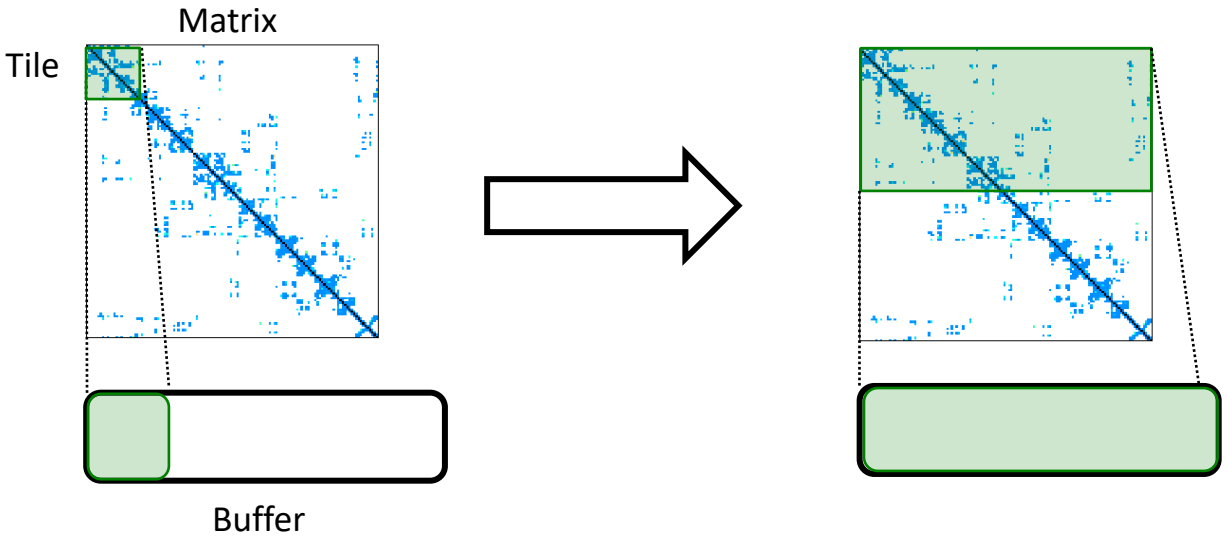
Many applications operate on highly sparse tensors



Better tiling improves on-chip data reuse

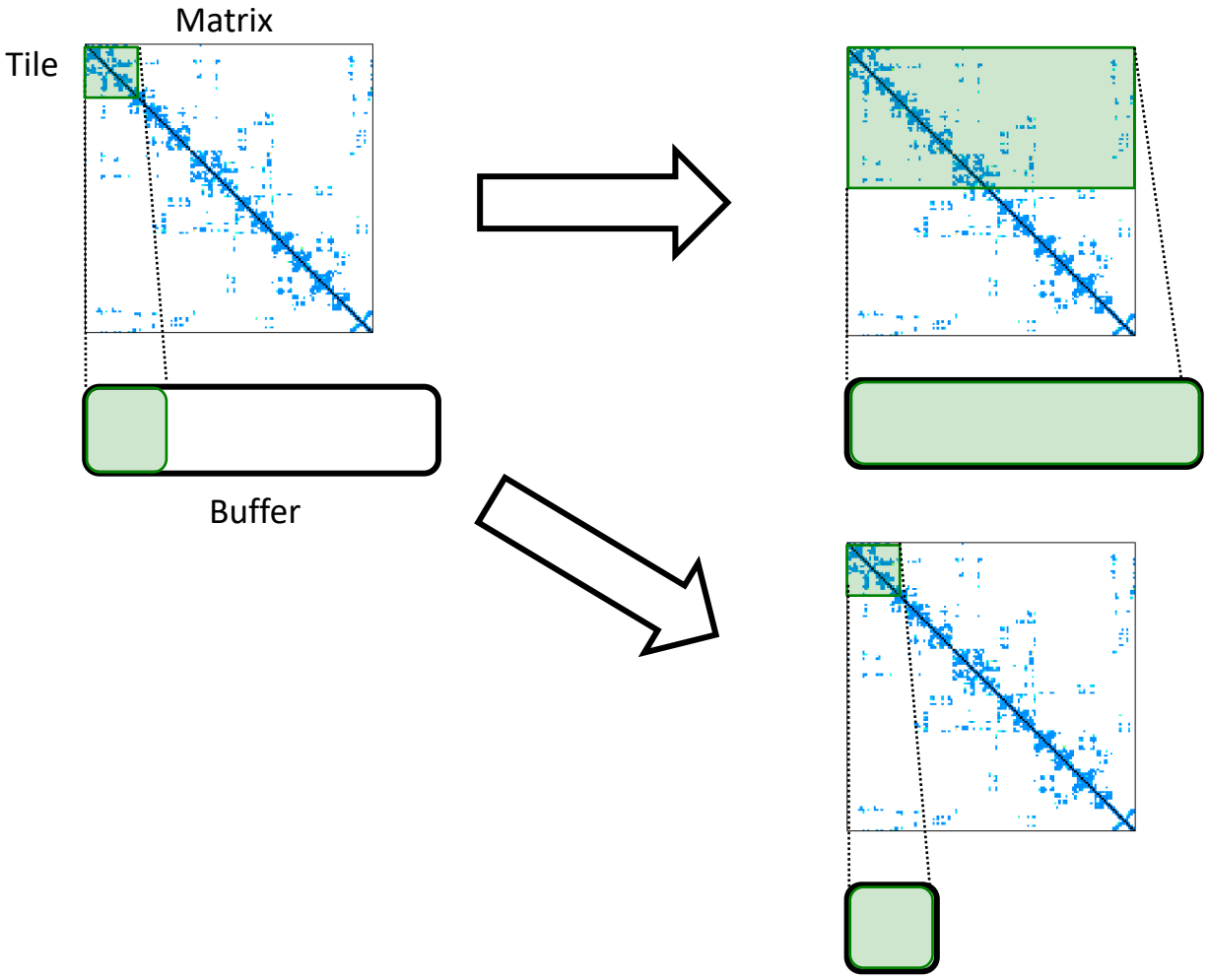


Better tiling improves on-chip data reuse



Larger tile
⇒ More data reuse

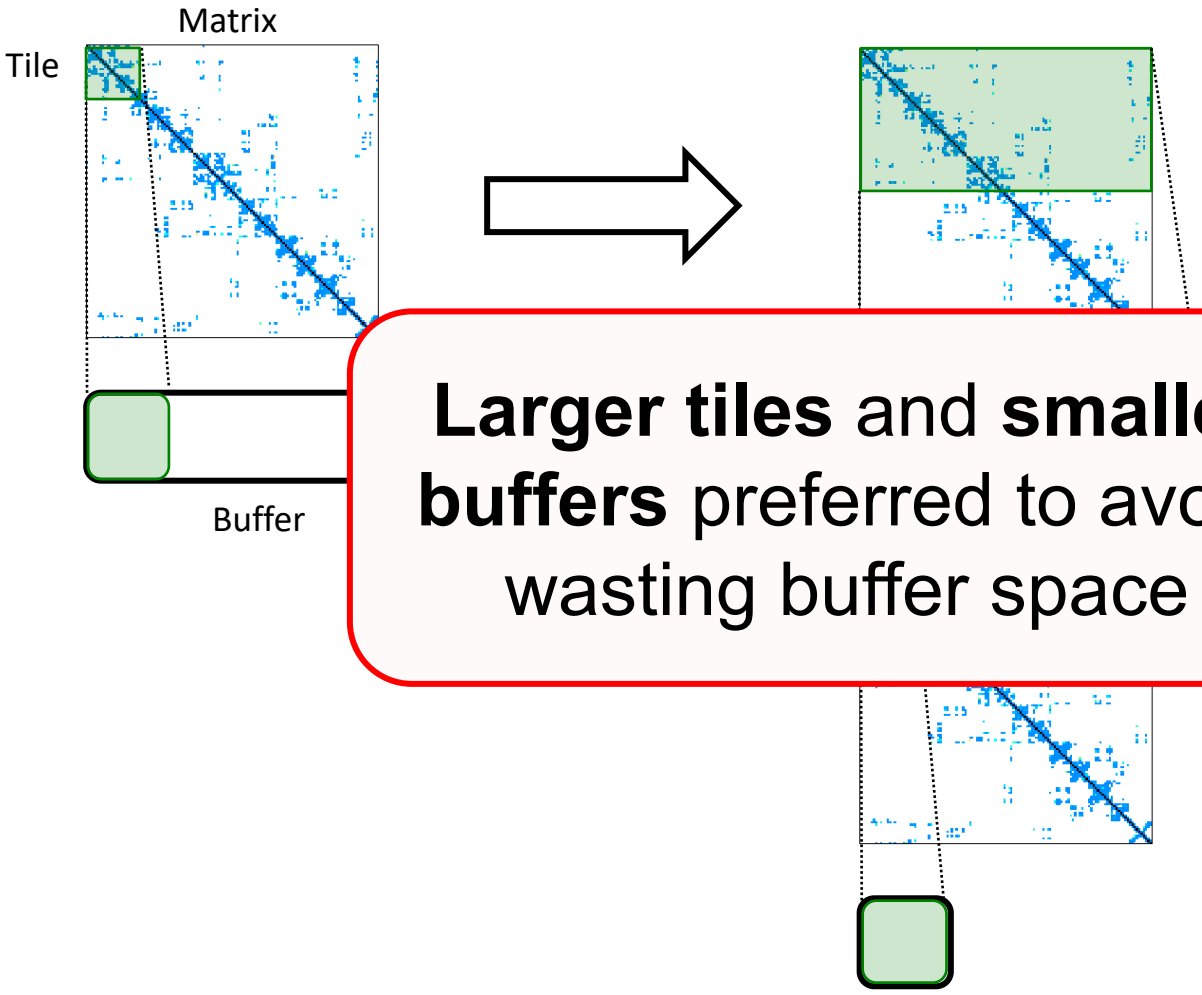
Better tiling improves on-chip data reuse



Larger tile
⇒ More data reuse

Smaller buffer
⇒ Lower cost

Better tiling improves on-chip data reuse

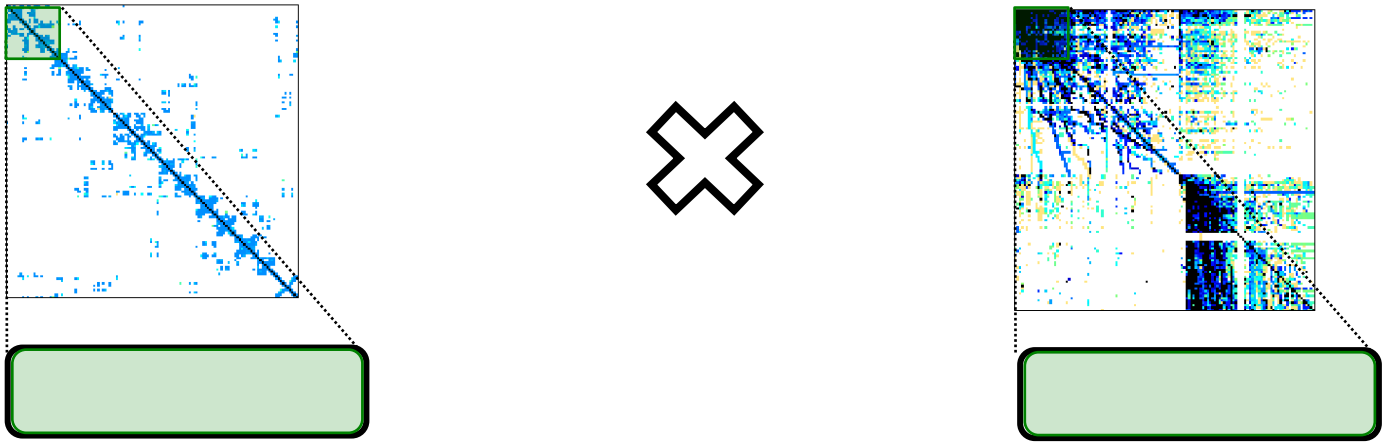


Larger tile
⇒ More data reuse

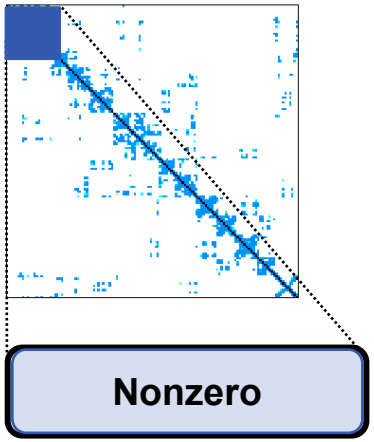
Larger tiles and **smaller buffers** preferred to avoid wasting buffer space

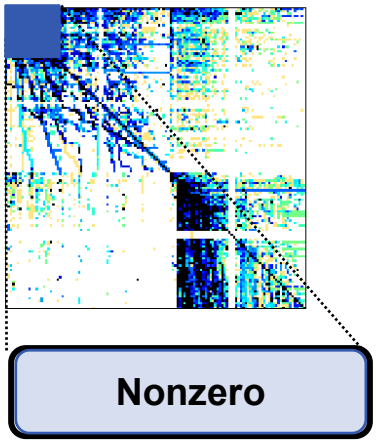
Smaller buffer
⇒ Lower cost

Coordinate-space tiling

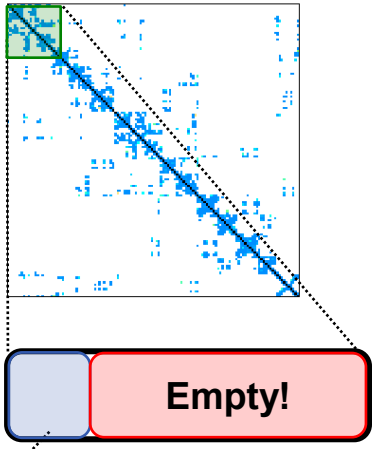


Coordinate-space tiling

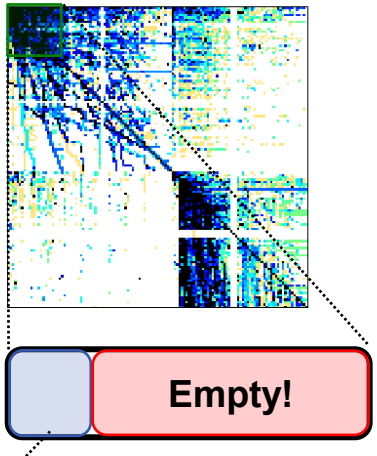




Coordinate-space tiling

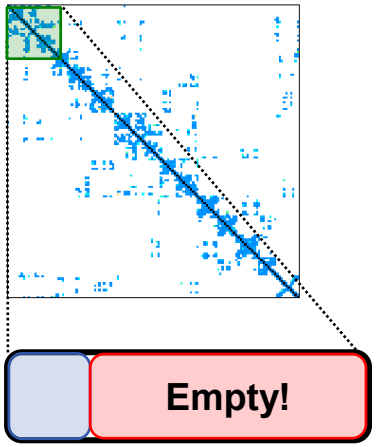


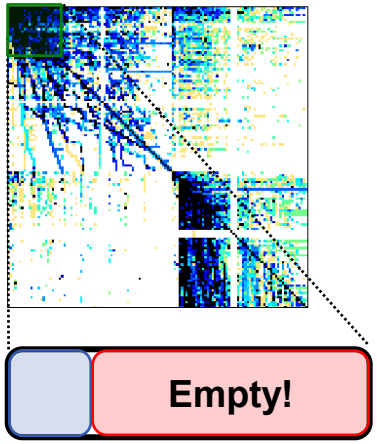
Nonzero



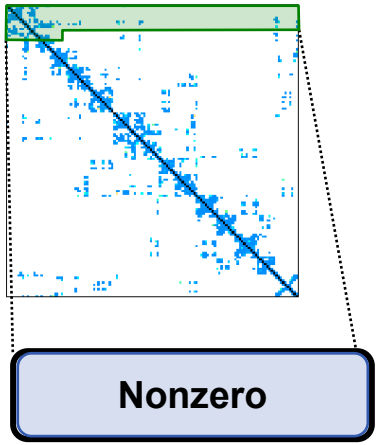
Nonzero

Coordinate-space tiling

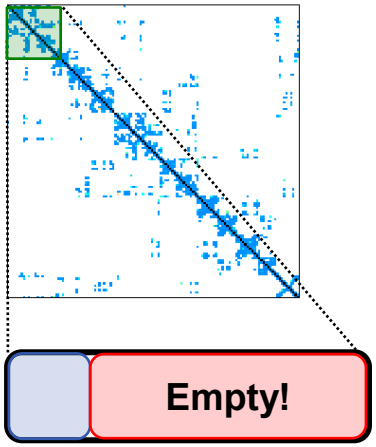


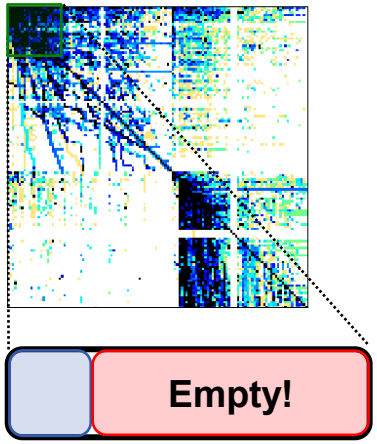


Position-space tiling

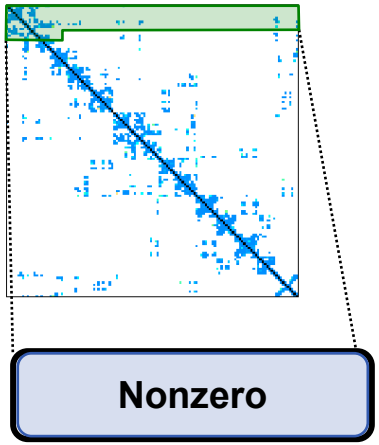


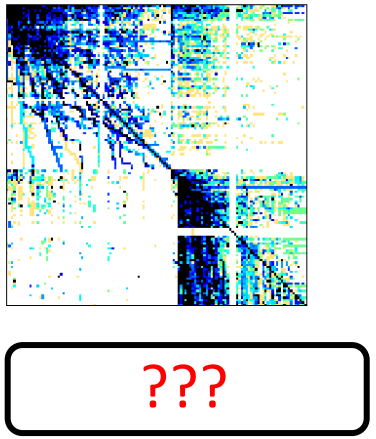
Coordinate-space tiling



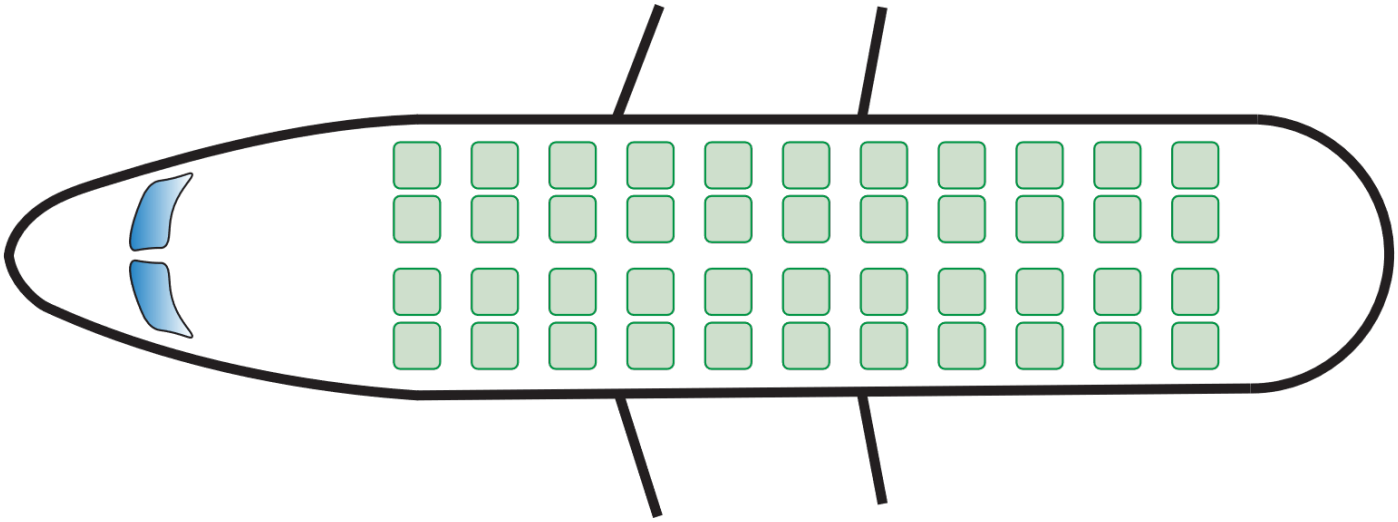


Position-space tiling

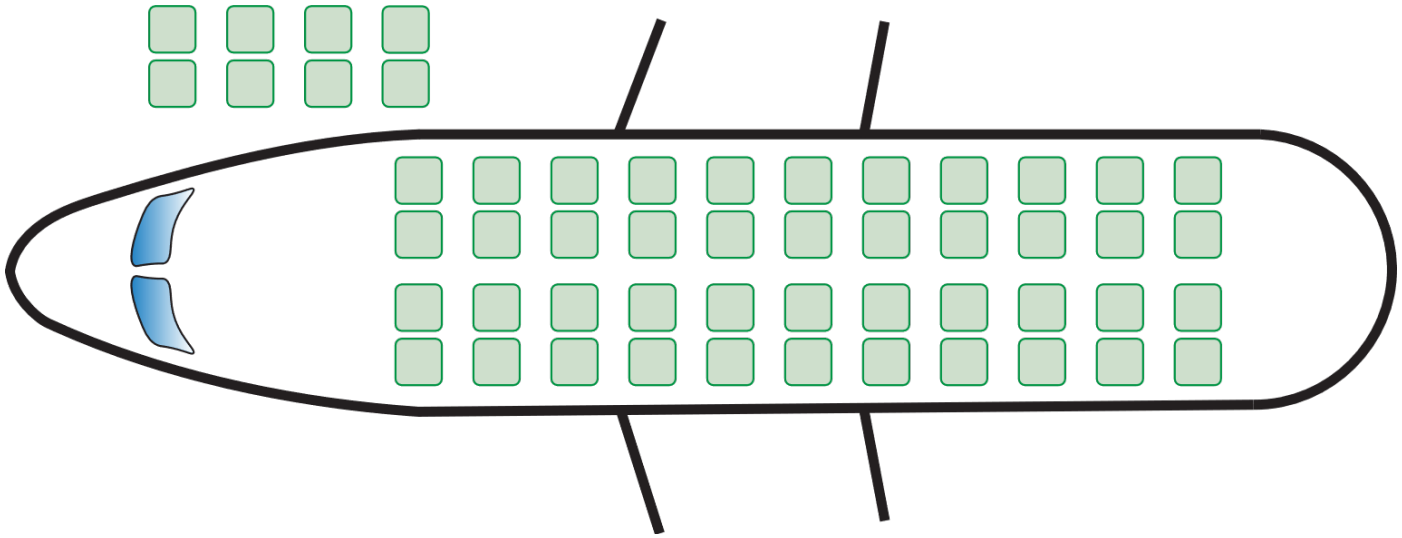




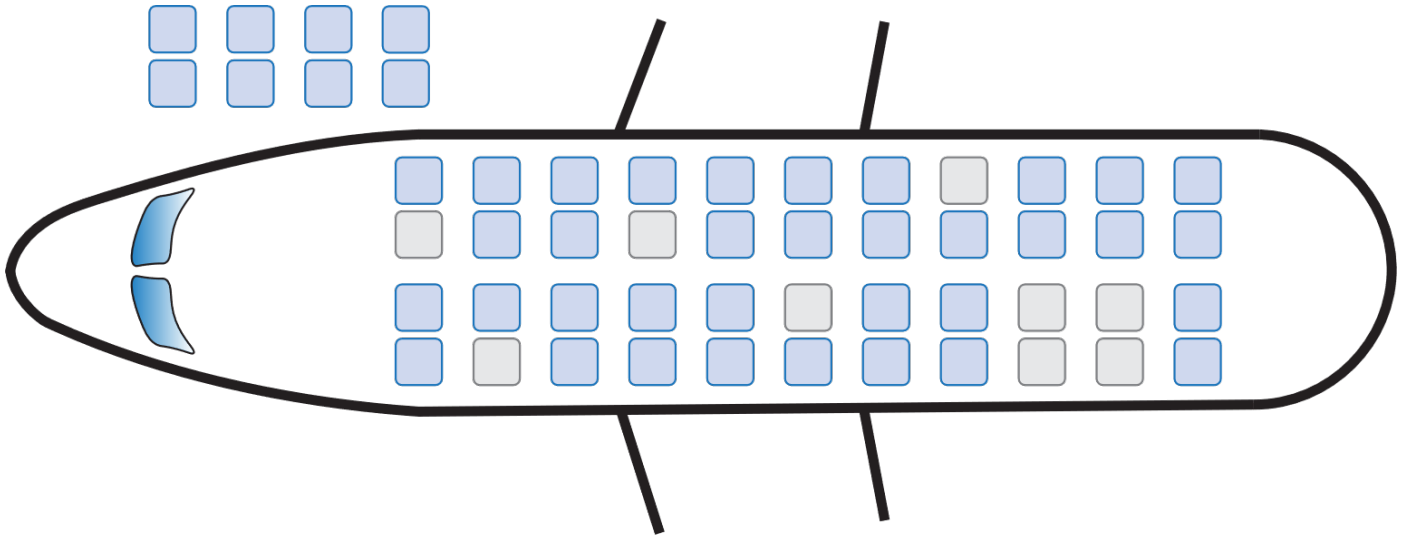
Overbooking



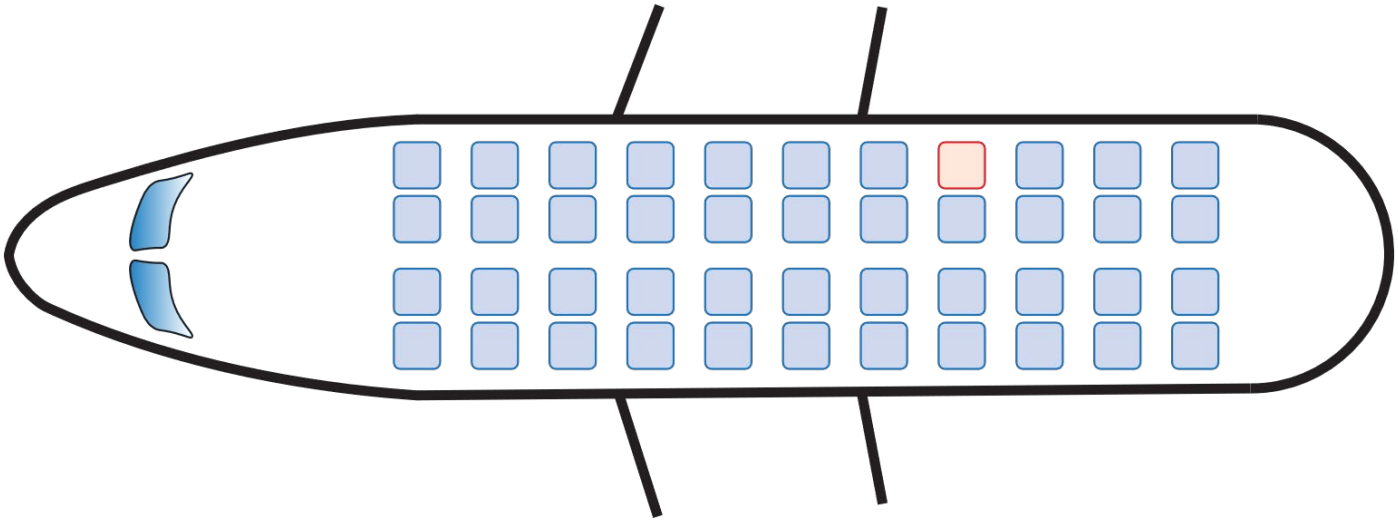
Overbooking



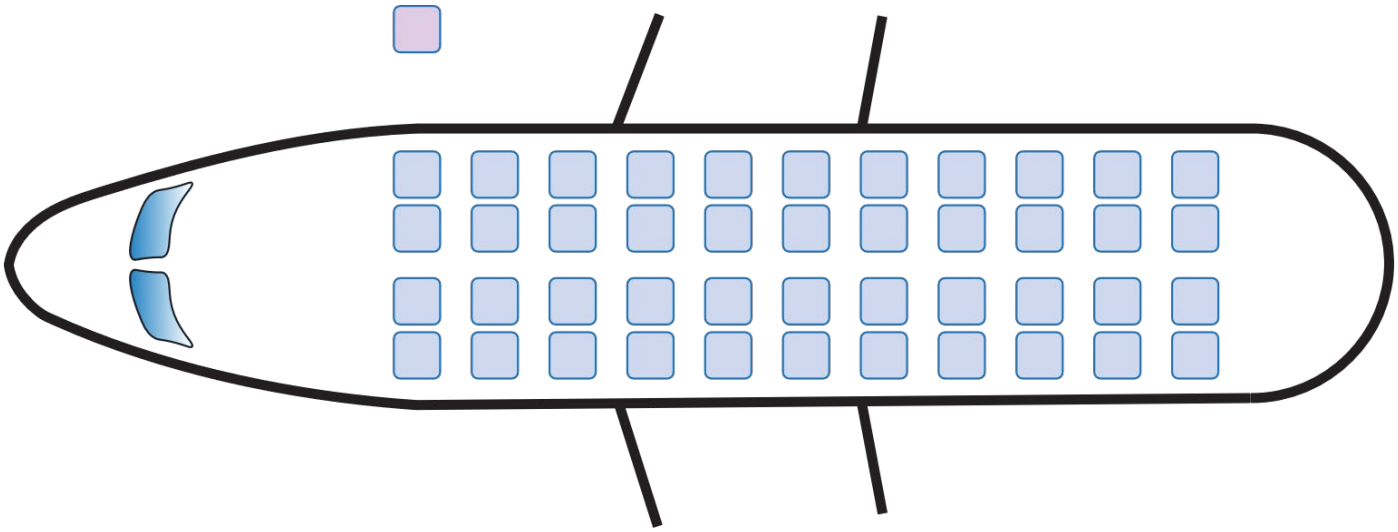
Overbooking



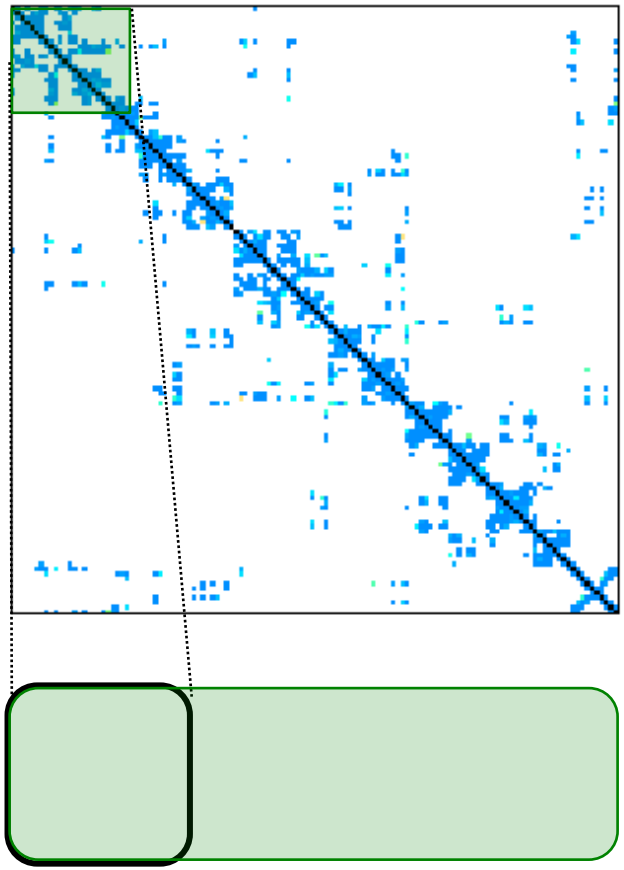
Overbooking



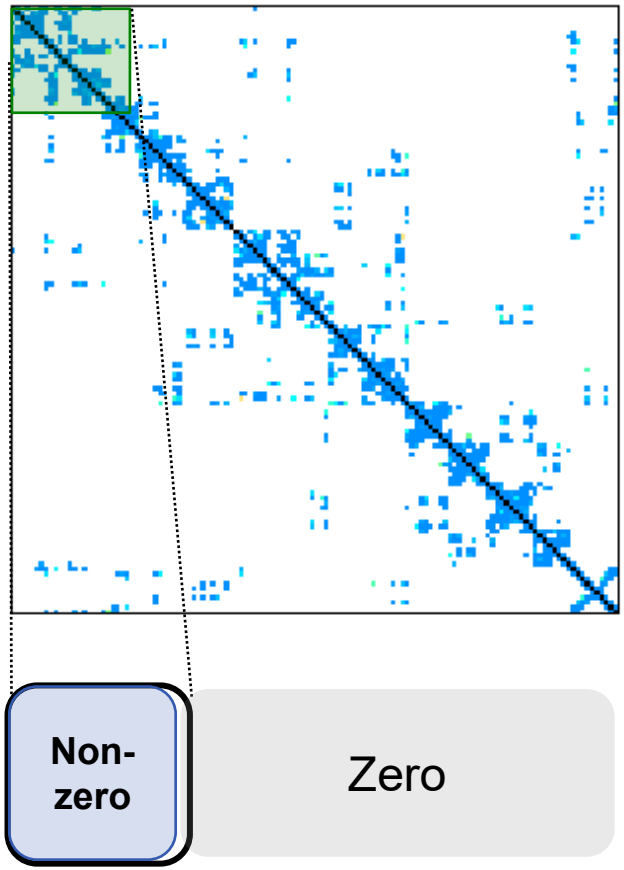
Overbooking



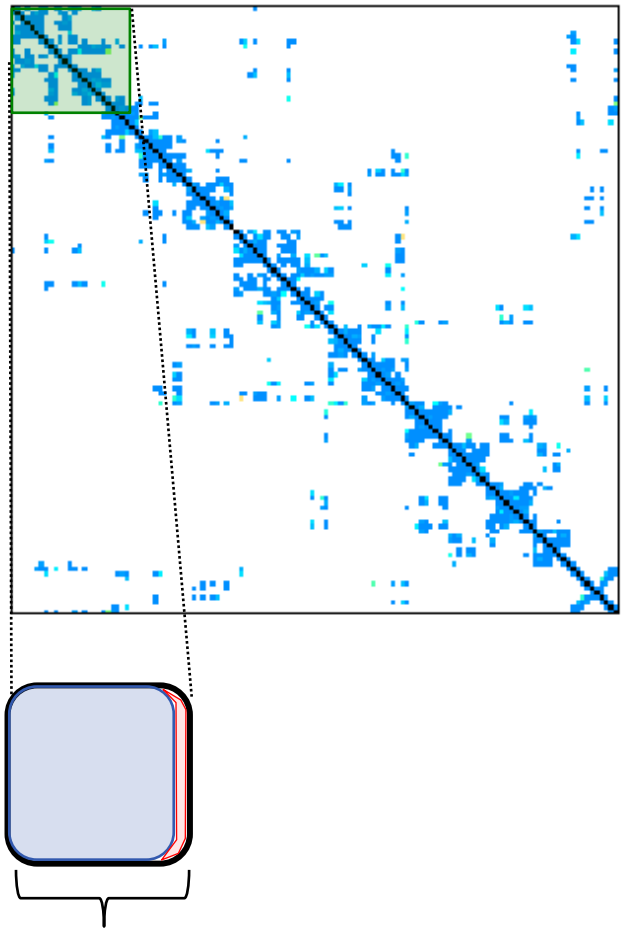
Overbooking buffers for sparse matrices



Overbooking buffers for sparse matrices



Overbooking buffers for sparse matrices



1300x smaller buffer than prior work
2.4x smaller than optimal coordinate-space tiling