mobileye™

Planning Tile & Fuse Transform in MLIR

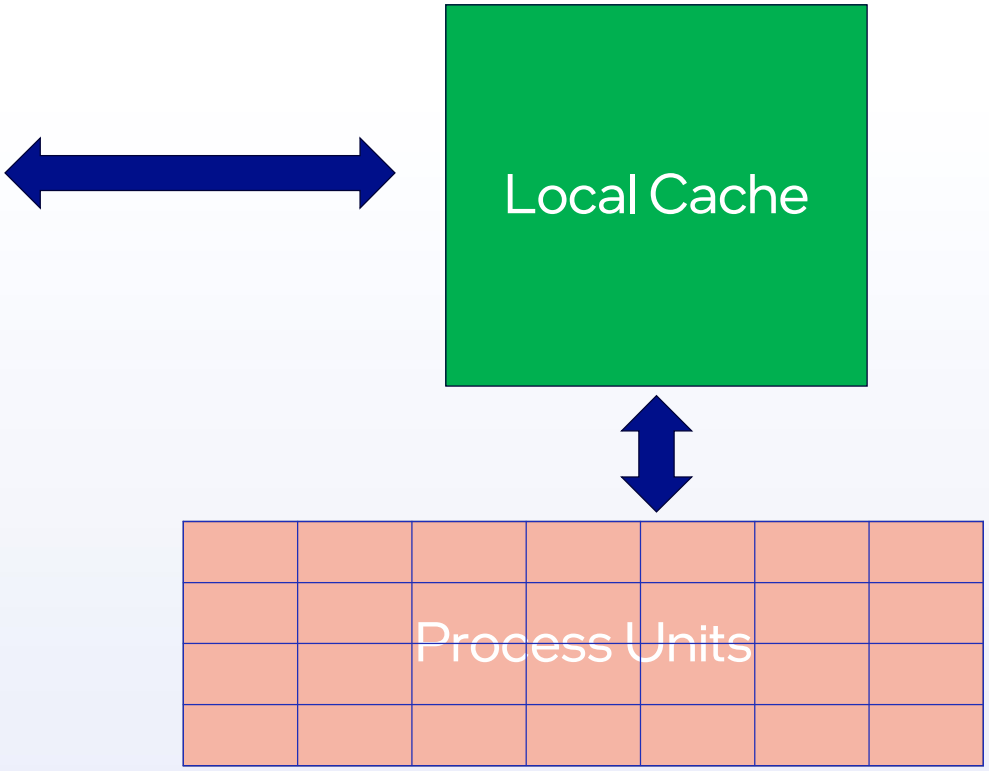
April 8, 2025

Aviad Cohen, Algorithm engineer

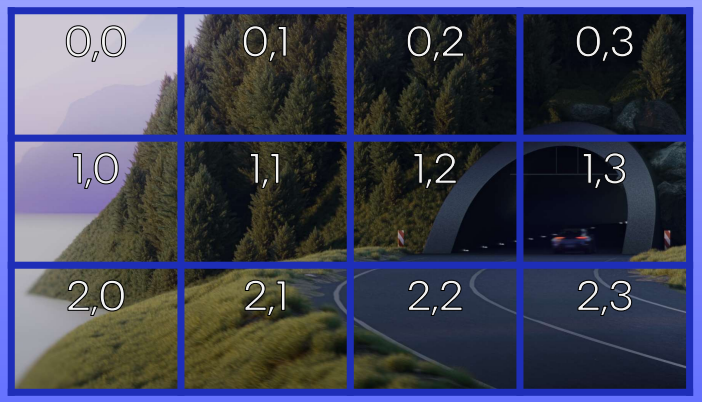
1 Picture, 1000 words

What is tiling all about?

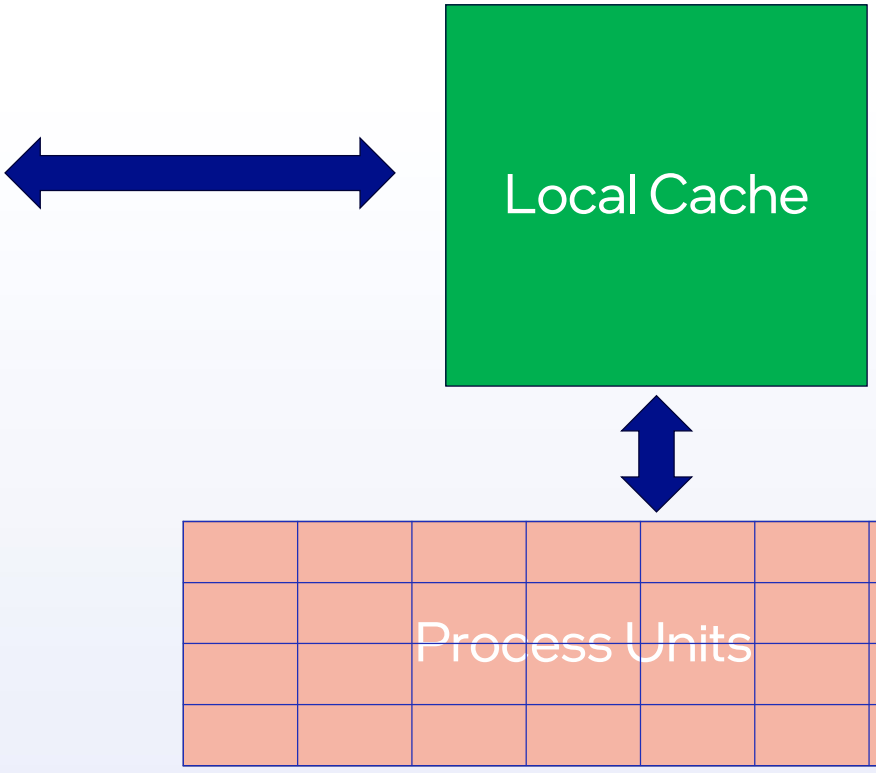
Main Memory



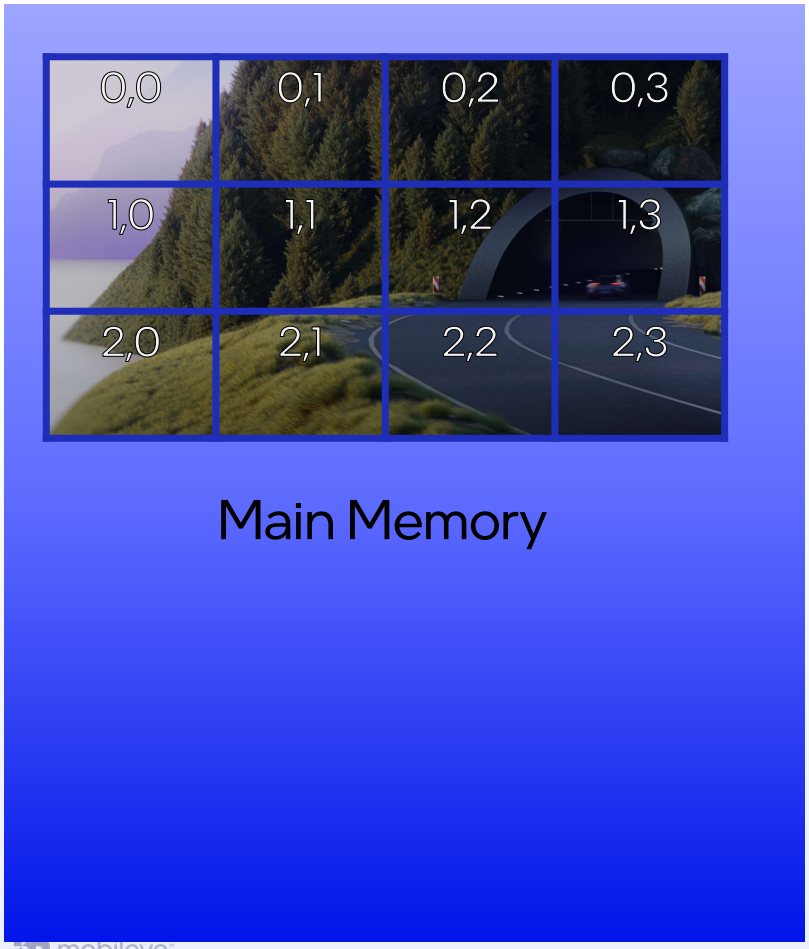
What is tiling all about?



Main Memory



What is tiling all about?



What is tiling all about?

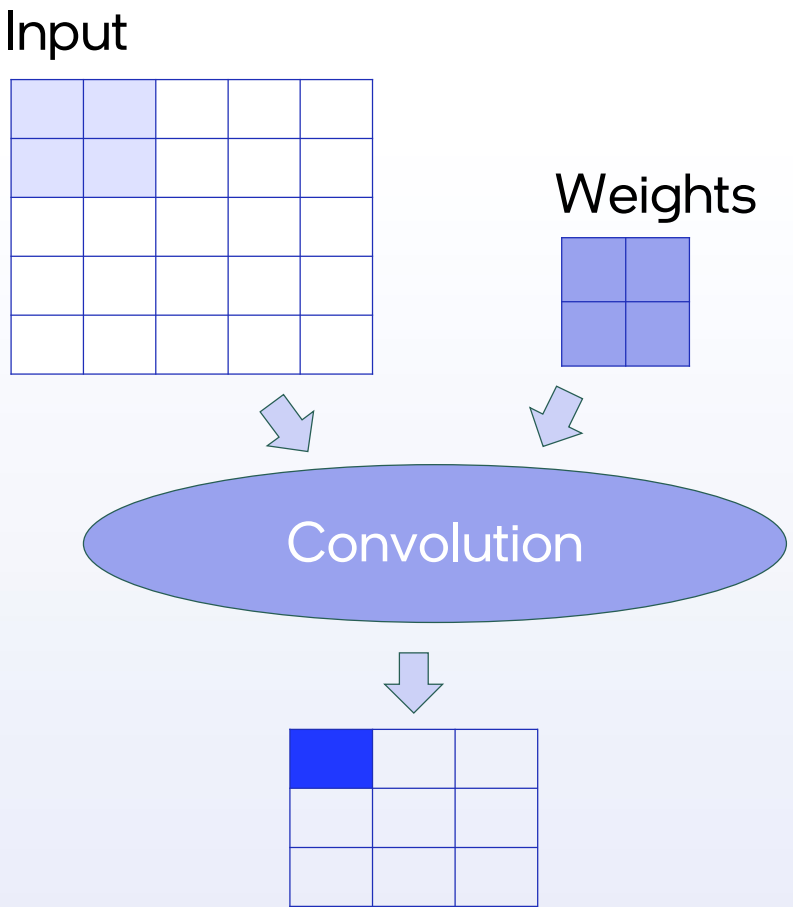
What is tiling all about?

Goal – Reach High Performance

- How?
  - Optimize cache usage by keeping data localized.

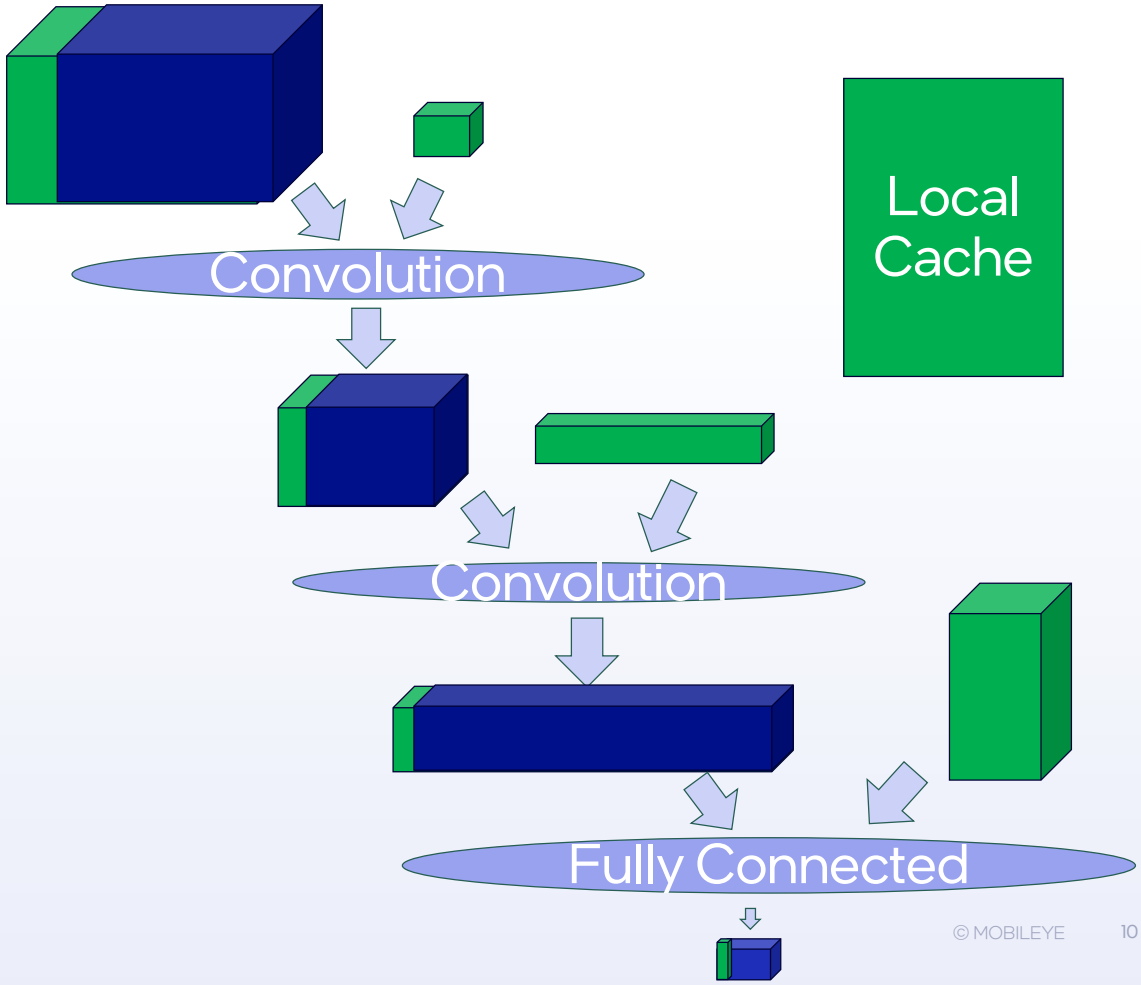
Goal – Reach High Performance

- How?
  - Optimize cache usage by keeping data localized.
  - Retain essential data within the cache for all tiles.



Goal – Reach High Performance

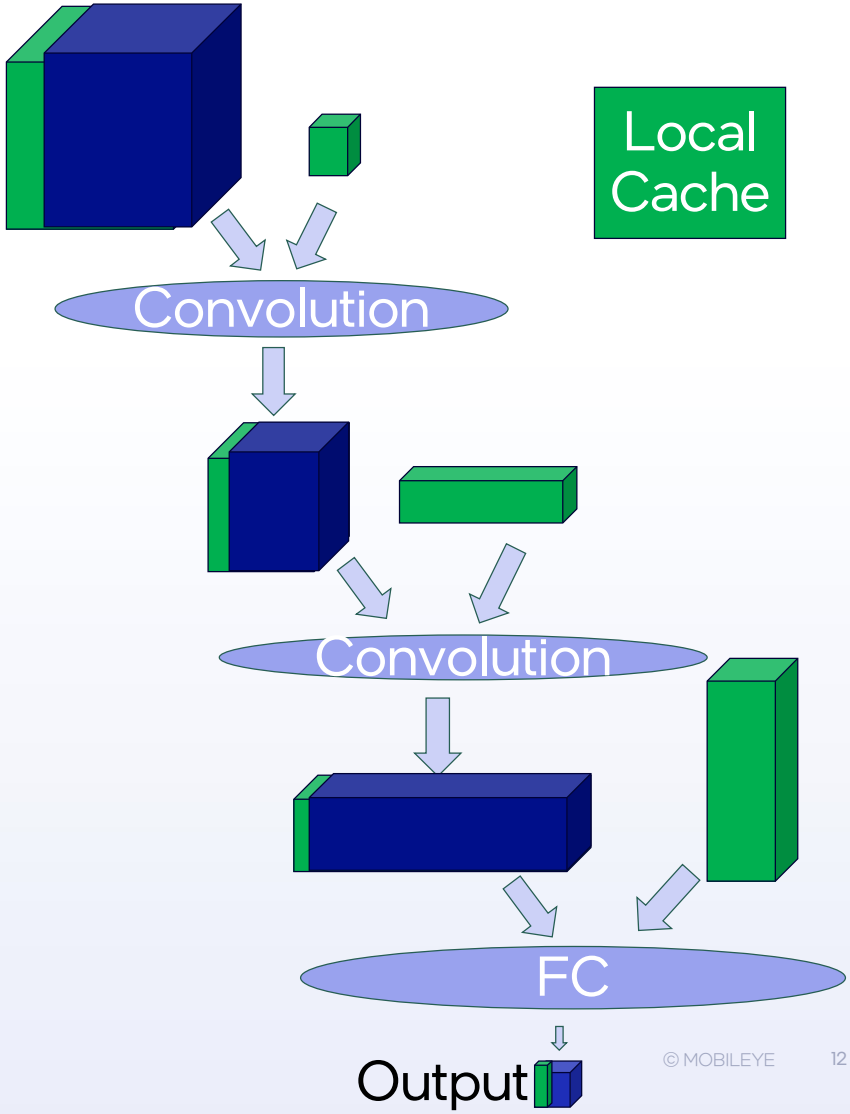
- How?
  - Optimize cache usage by keeping data localized.
  - Retain essential data within the cache for all tiles.
  - Utilize larger tiles to minimize control flow overhead.



Fortunately, transform already exists!

Tile & Fuse - The transform

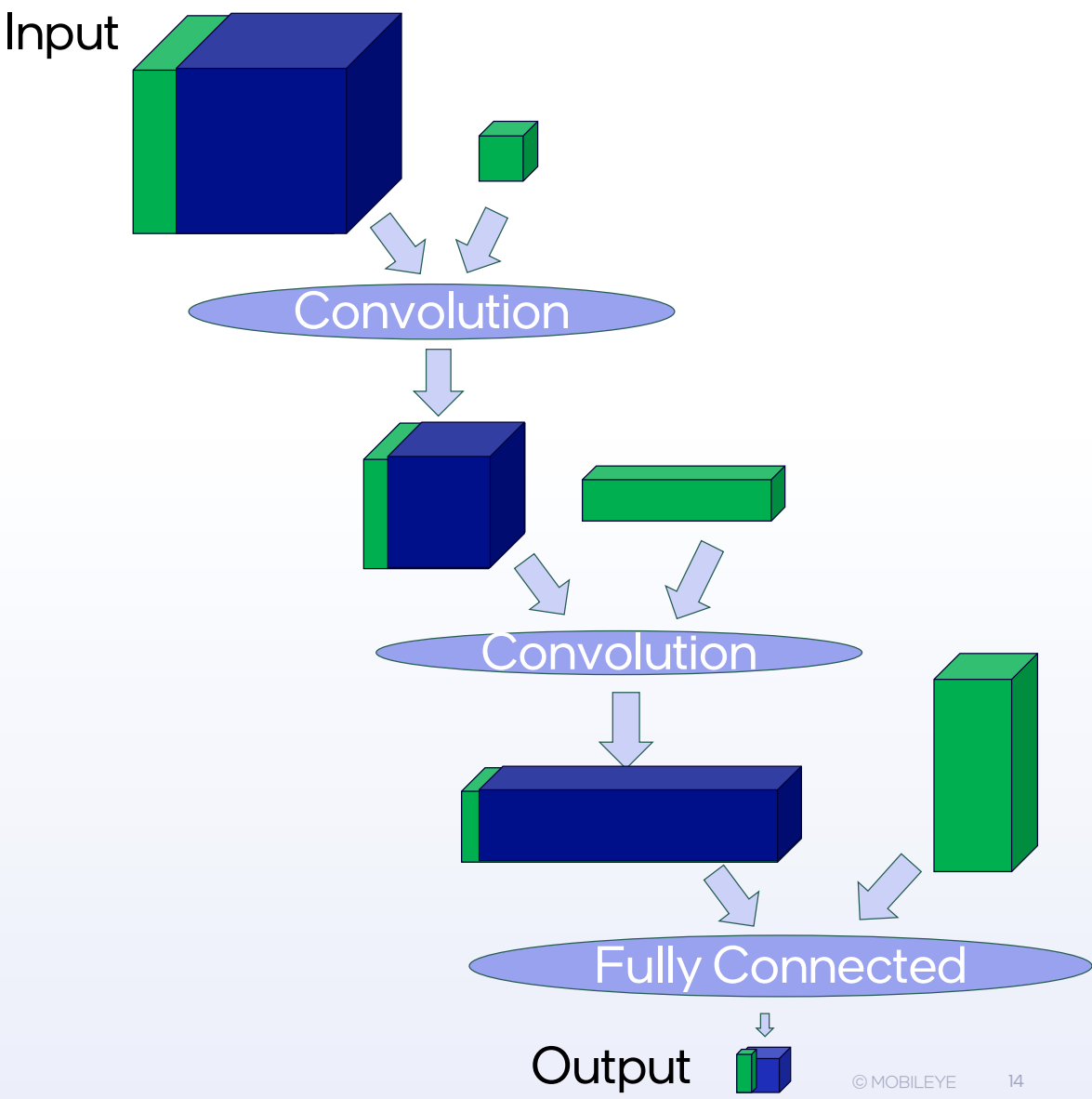
- Demands Tiling Interface.
- General flow:
  - Tile a root operation into loops.
  - Fuse consumer/producer operations into the existing loops.
- Can be controlled by a control function.



How to the select the tile size?

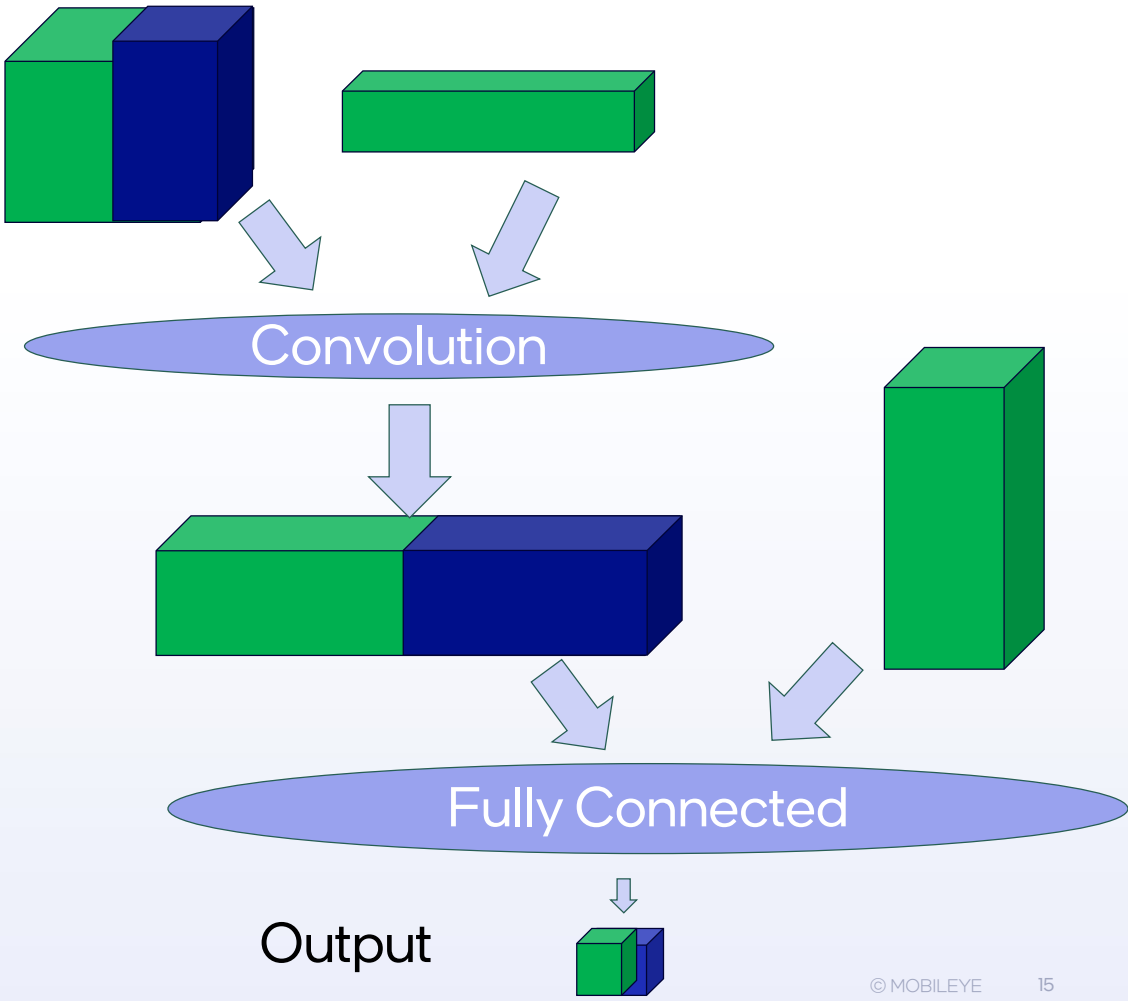
Vertical approach

Increase number of fused operations over tile size



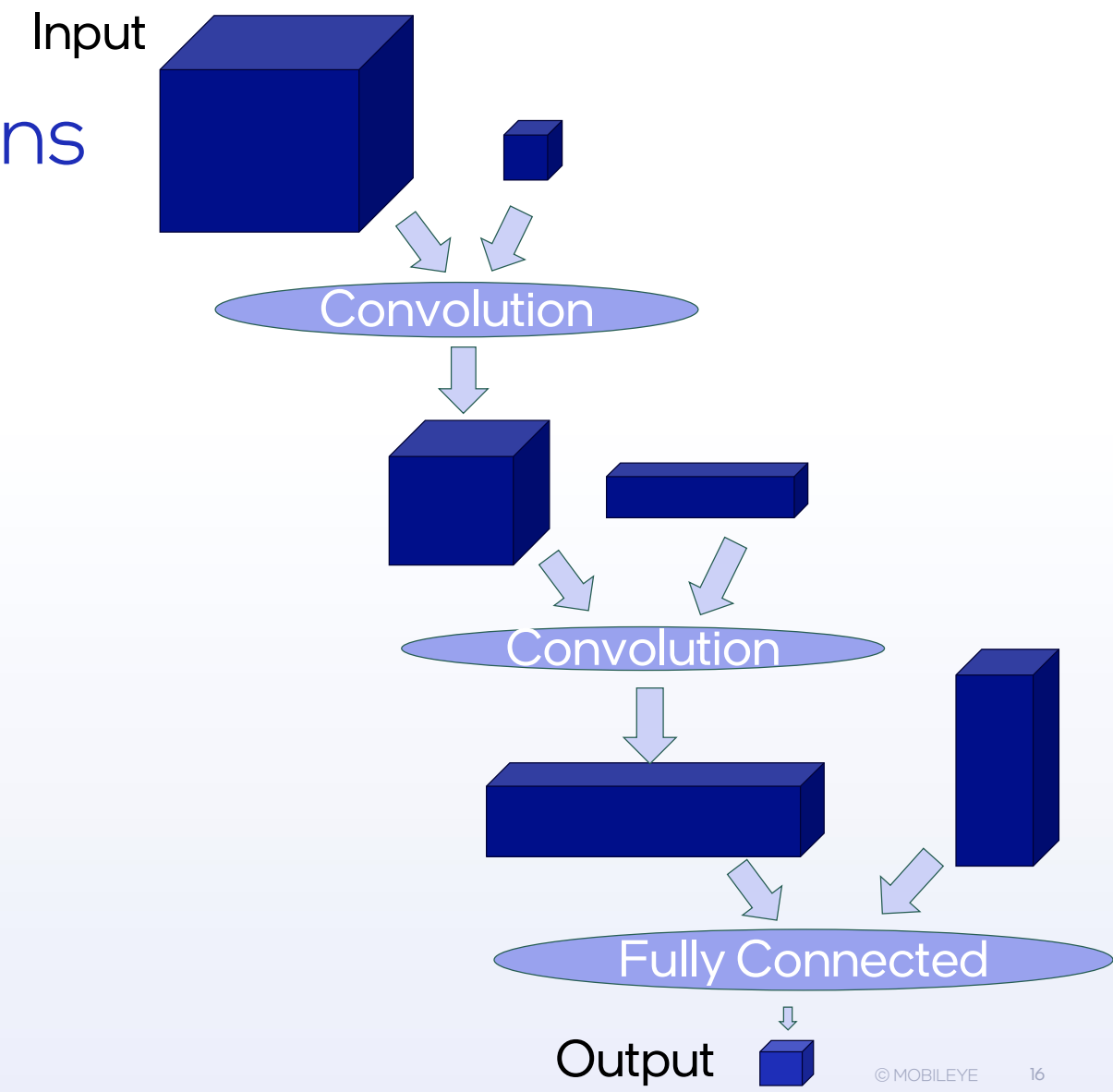
Horizontal Tiling

Increase tile size over number of fused operations



Planning Considerations

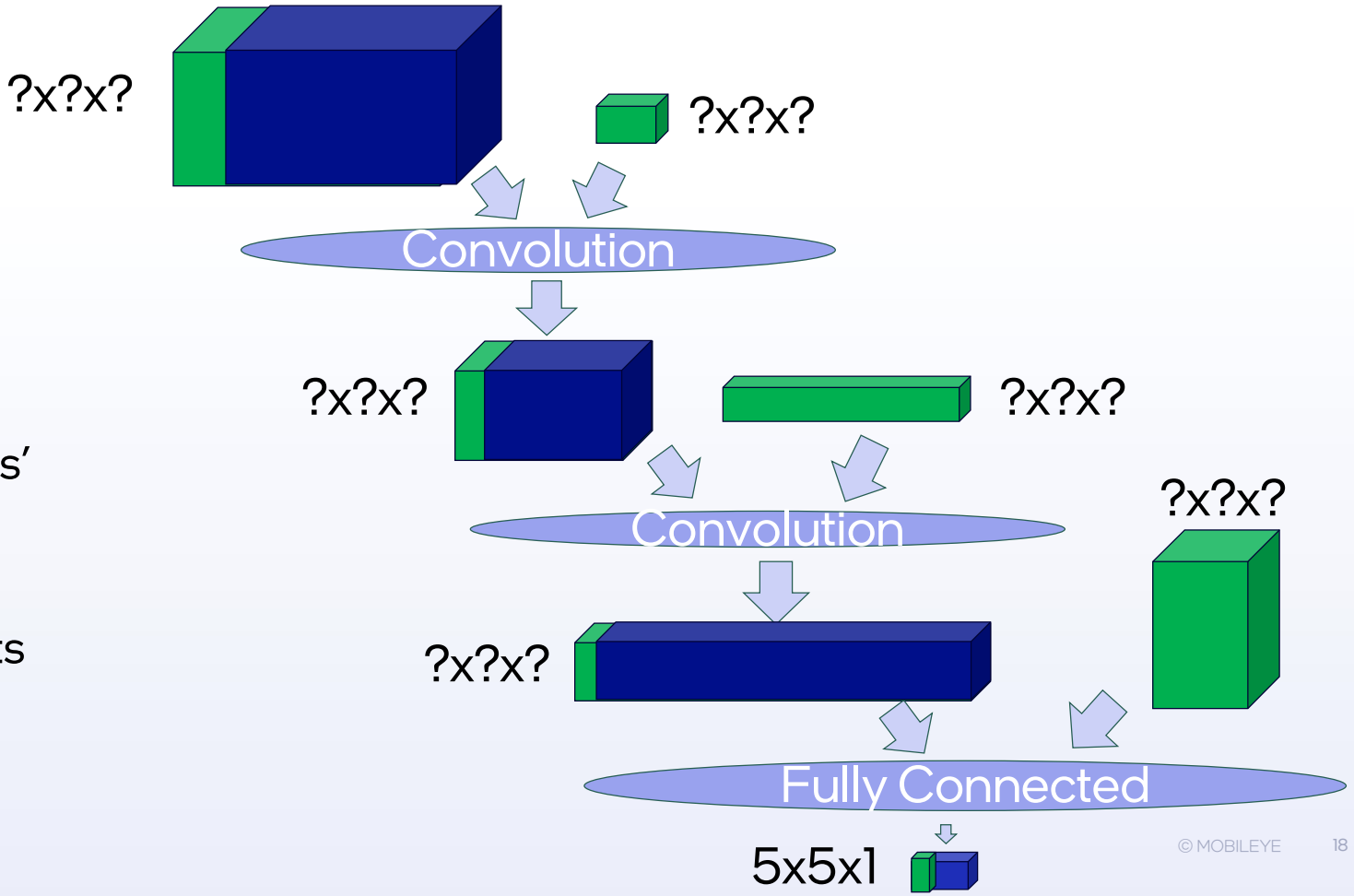
- Find the sweet spot between tile size and fusion to minimize bandwidth
- Overlapping tiles
- Scheduling approaches
- Too big control flow



So, what is missing?

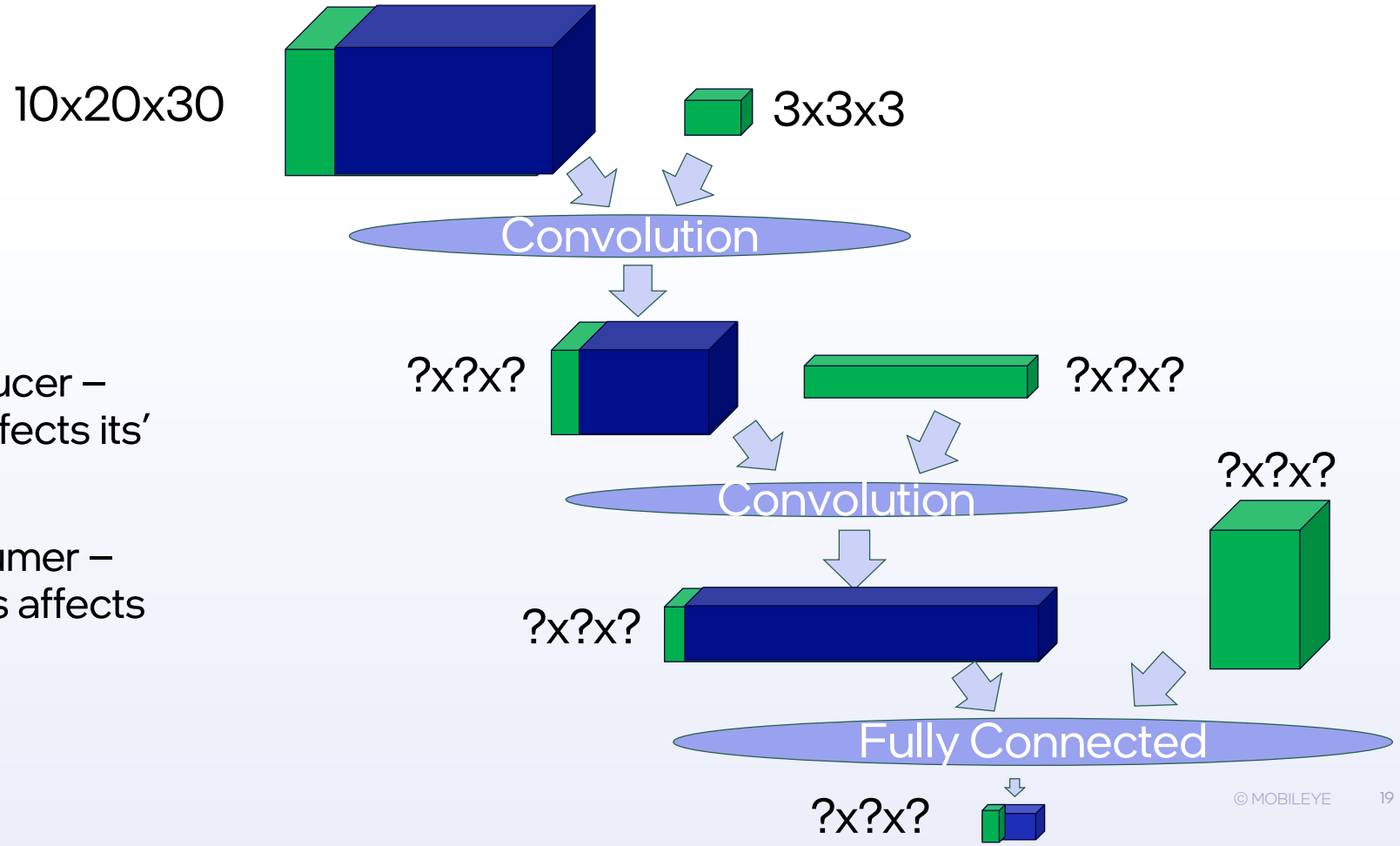
Proposal – Fusion Interface

- Given tiled producer – how its' results affects its' consumers?
- Given tiled consumer – how its' operands affects its' producers?



Proposal – Fusion Interface

- Given tiled producer – how its' results affects its' consumers?
- Given tiled consumer – how its' operands affects its' producers?



RFC - Fusion Analysis Interface for Compute Operations

- **Additional method can be added, let's enhance the interface together!**

- For more details follow the RFC:

https://discourse.llvm.org/t/fusion-analysis-interface-for-compute-operations/85743

Thank you!

Aviad Cohen
Aviad.cohen2@mobileye.com

mobileye™