MEMORY HIERARCHY FOR APPROXIMATE COMPUTING

BY – MOHD SALMAN KHAN

WHY APPROXIMATE ?

Pi = 3.14159265359 ...

Pi = 3.14

APPROXIMATE COMPUTING – WHAT DOES IT MEAN ?

To relax data precision and/or tolerate some level of loss of data quality to gain energy efficiency.

APPROXIMATE COMPUTING – WHERE CAN IT BE USED ?

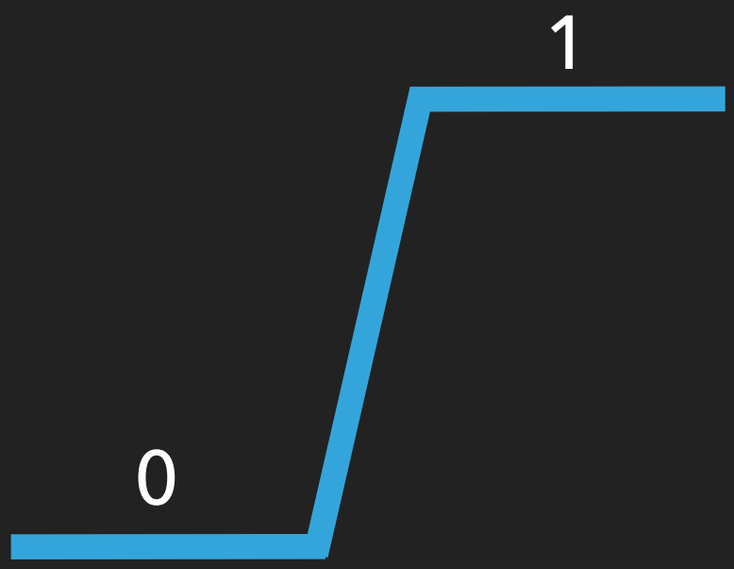
100 ERRONEOUS PIXELS

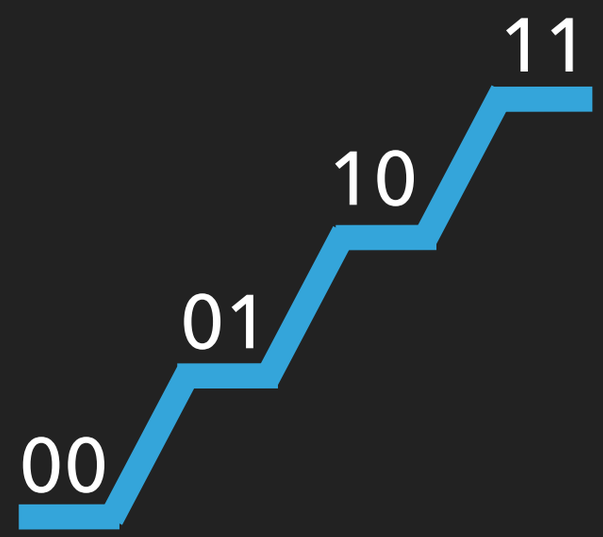
=

MULTI-LEVEL CELLS – FOR CACHES IN GENERAL

NAND and NOR Flash Cells

Phase Change Memory (PCM) devices





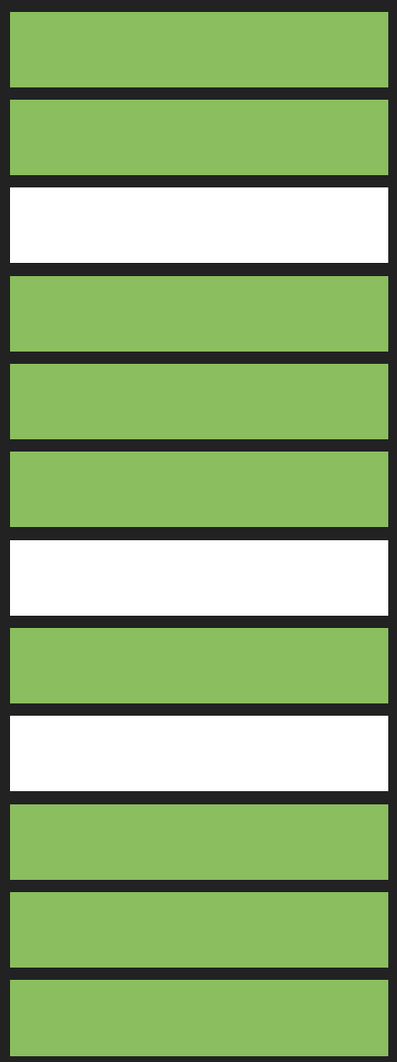
WHAT DO WE GET ?

- 2 times Information in 1 cell space
- Energy consumption reduced by half
- More Error Prone - slight voltage variation
- Can be converted to Regular Binary system
- Tunable Approximate Cache - to store both approximate and precise data.

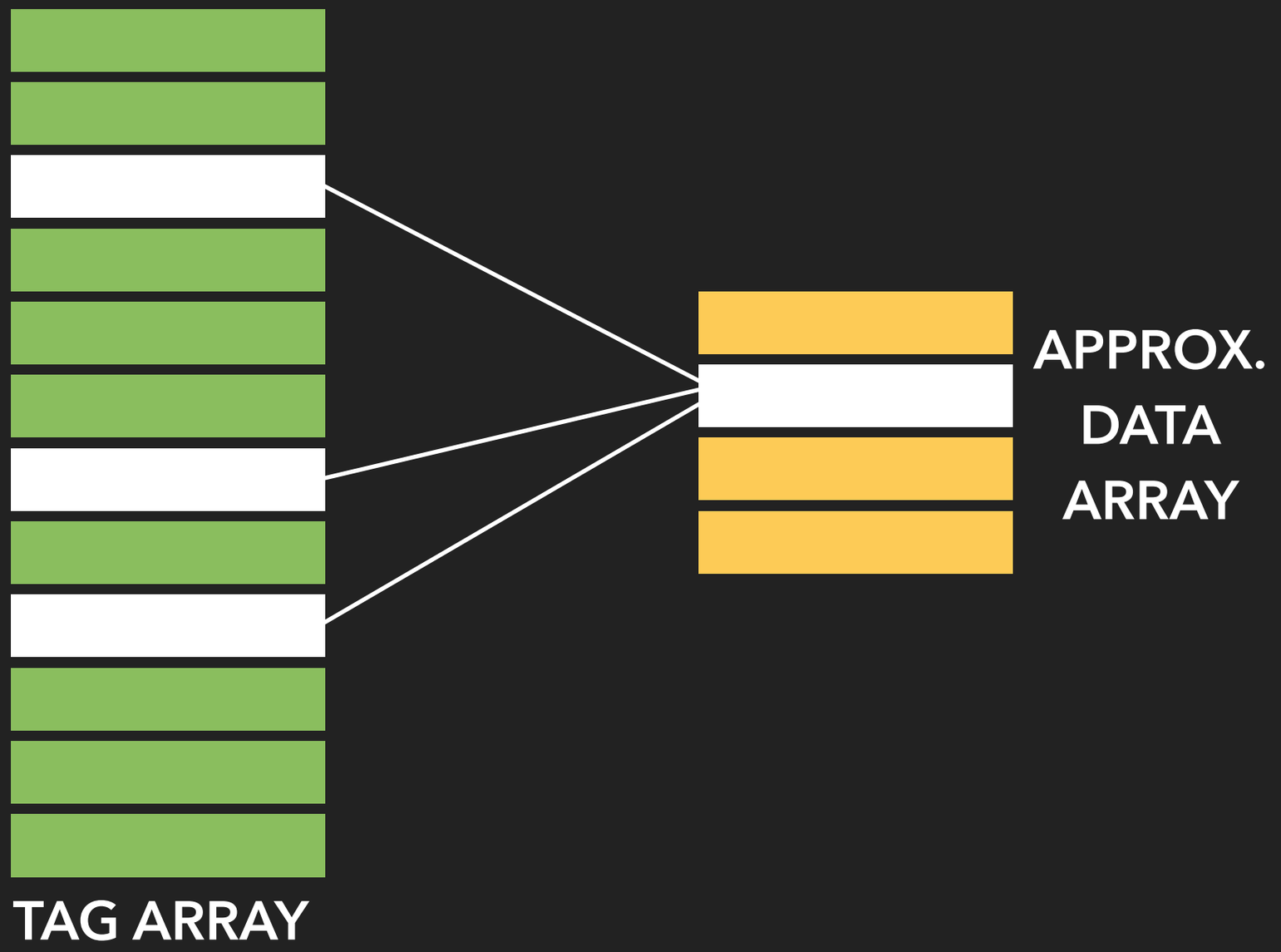
APPROXIMATE COMPUTING AT LLC

- Modern processors contain large last level caches (LLCs)
- They consume substantial energy and area
- We observe that a large fraction of cache values exhibit approximate similarity
- Values across cache blocks are not identical but are similar.
- LLC architecture: the Doppelgänger cache.

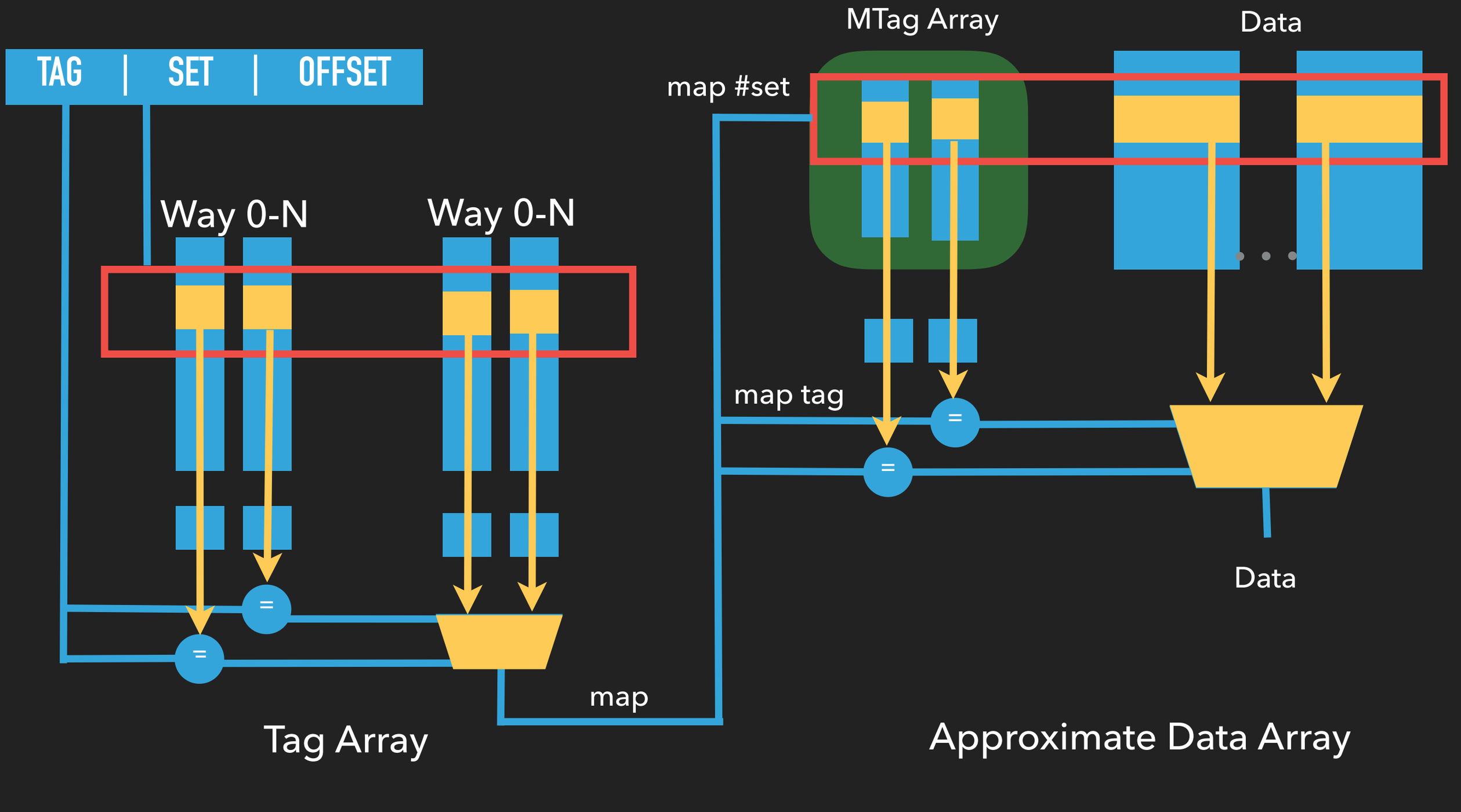
DOPPELGÄNGER CACHE OVERVIEW



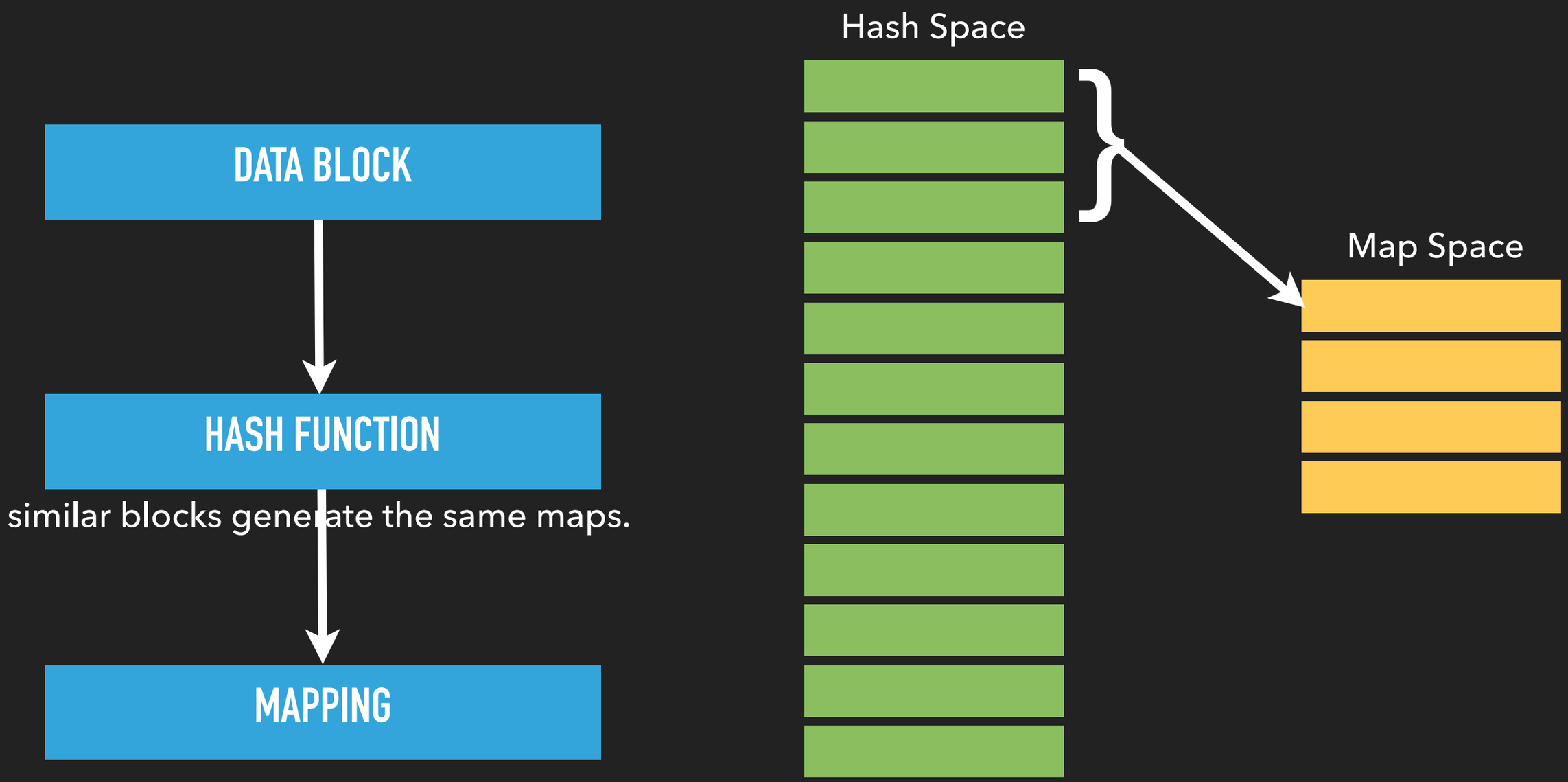
DOPPELGÄNGER CACHE OVERVIEW



DOPPELGÄNGER CACHE ARCHITECTURE – LOOKUPS



DOPPELGÄNGER CACHE ARCHITECTURE – SIMILARITY CAPTURE



The mapping converts hash to an M-bit integer value

WHAT DO WE GET ?

- Tag lookup latency increases.
- The data array latency decreases.
- Combined MTag and data access latencies are lower.
- Provide energy and data saving.

MEMORY HIERARCHY FOR APPROXIMATE COMPUTING

?