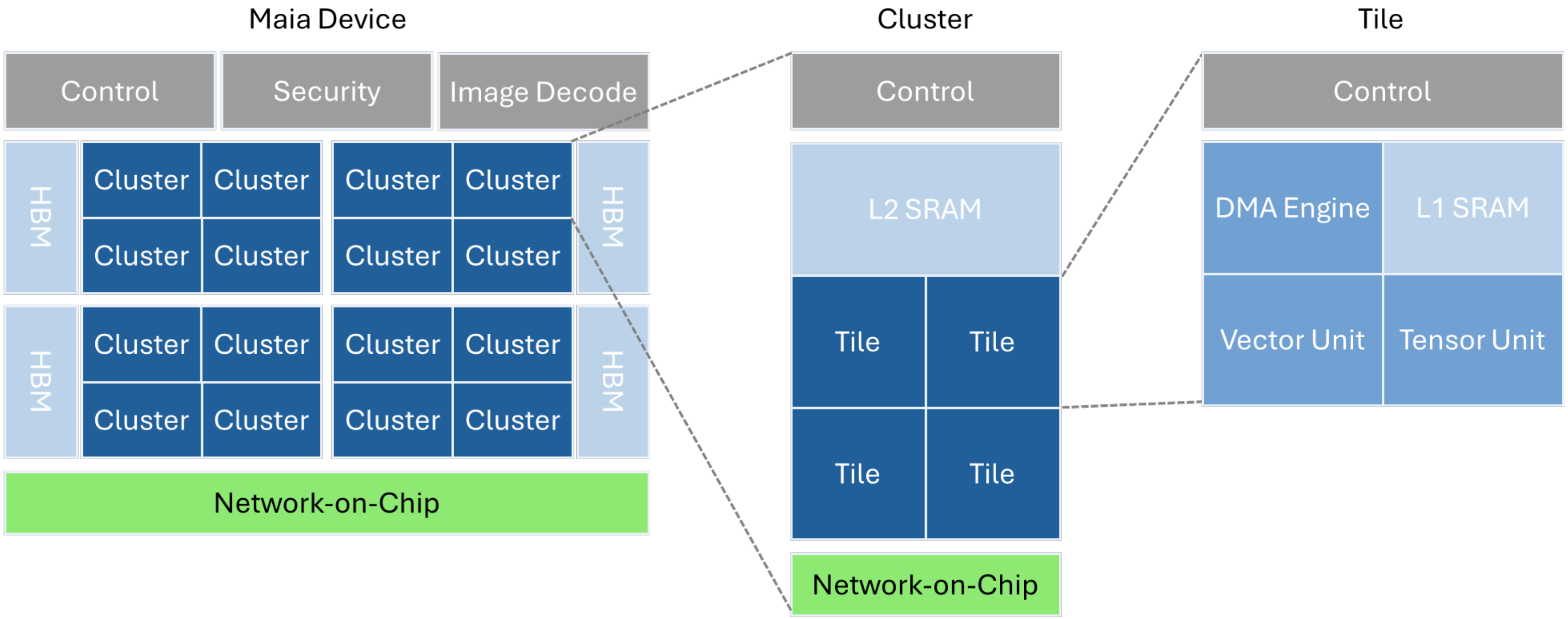
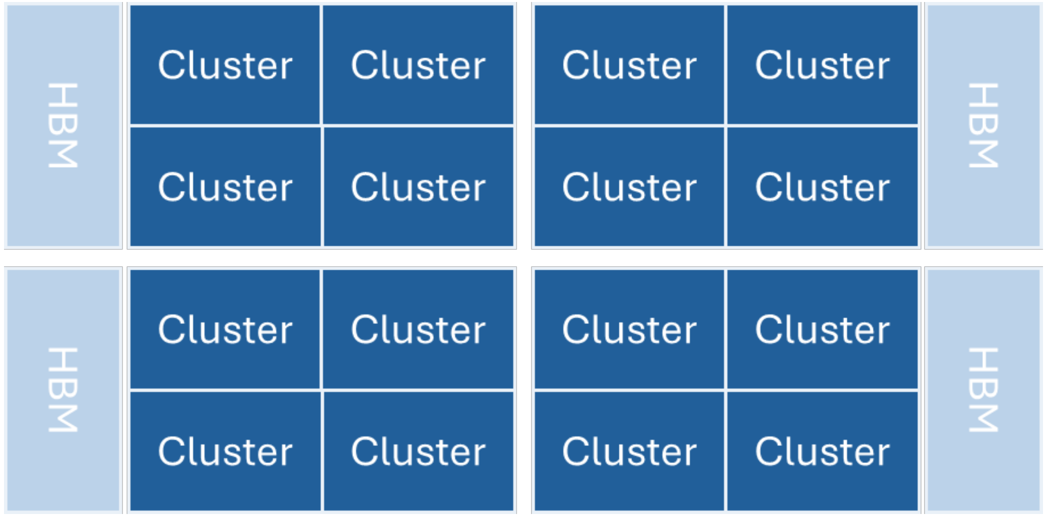
Triton for Azure Maia

Ian Baird

Maia-100 Device Architecture



Triton for Maia-100



HBM Cluster Cluster Cluster Cluster HBM

Cluster Cluster Cluster Cluster

HBM Cluster Cluster Cluster Cluster HBM

Cluster Cluster Cluster Cluster

Compiler Responsibilities

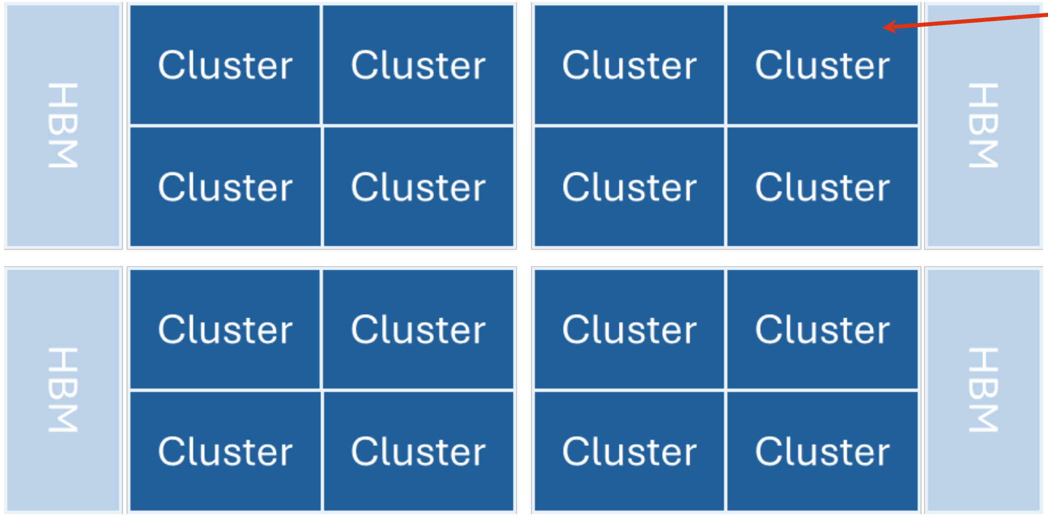
Splitting cluster work among tiles, optimize tile performance

Maia processor orchestration, semaphore management

Data movement in memory hierarchy, overlap copy and compute

Optimal codegen

Triton for Maia-100



```
import triton
import triton.language as tl

@triton.jit
def matmul_kernel( … ):
    pid =
tl.program_id(axis=0)
```

Compiler Responsibilities

Splitting cluster work among tiles, optimize tile performance

Maia processor orchestration, semaphore management

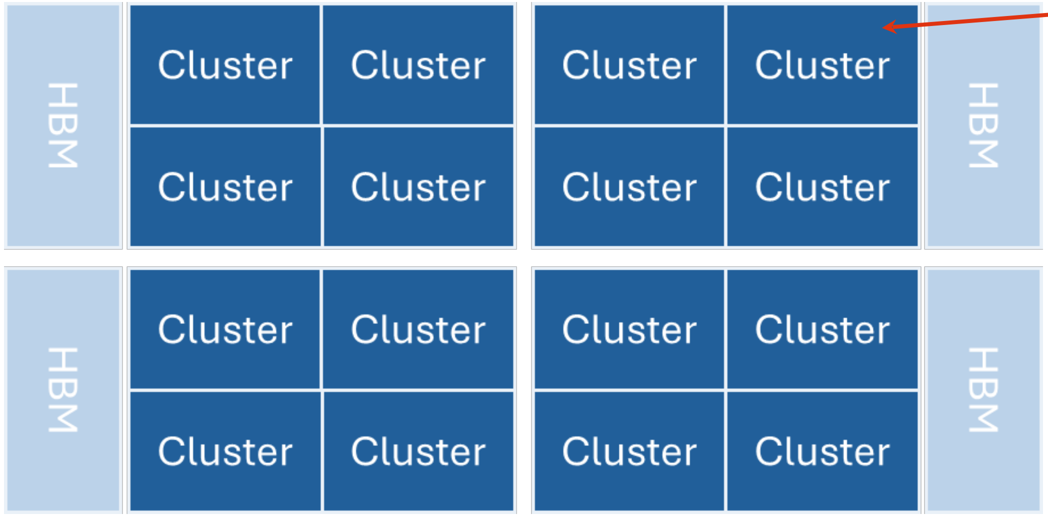
Data movement in memory hierarchy, overlap copy and compute

Optimal codegen

Triton for Maia-100

```
import triton
import triton.language as tl

@triton.jit
def matmul_kernel( … ):
    pid =
tl.program_id(axis=0)
```



Compiler Responsibilities

Splitting cluster work among tiles, optimize tile performance

Maia processor orchestration, semaphore management

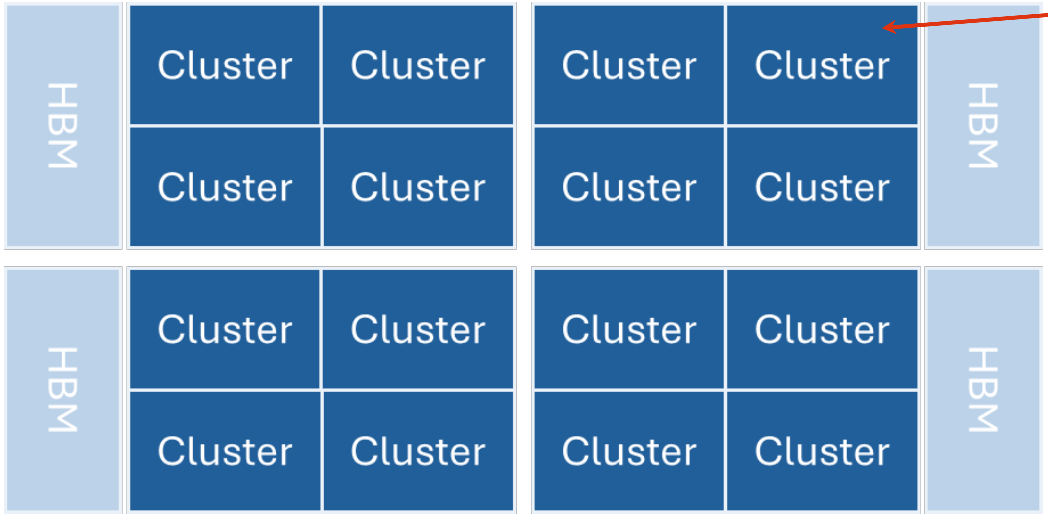
Data movement in memory hierarchy, overlap copy and compute

Optimal codegen

Triton for Maia-100

```
import triton
import triton.language as tl


@triton.jit
def matmul_kernel( … ):
    pid =
tl.program_id(axis=0)
```



Compiler Responsibilities

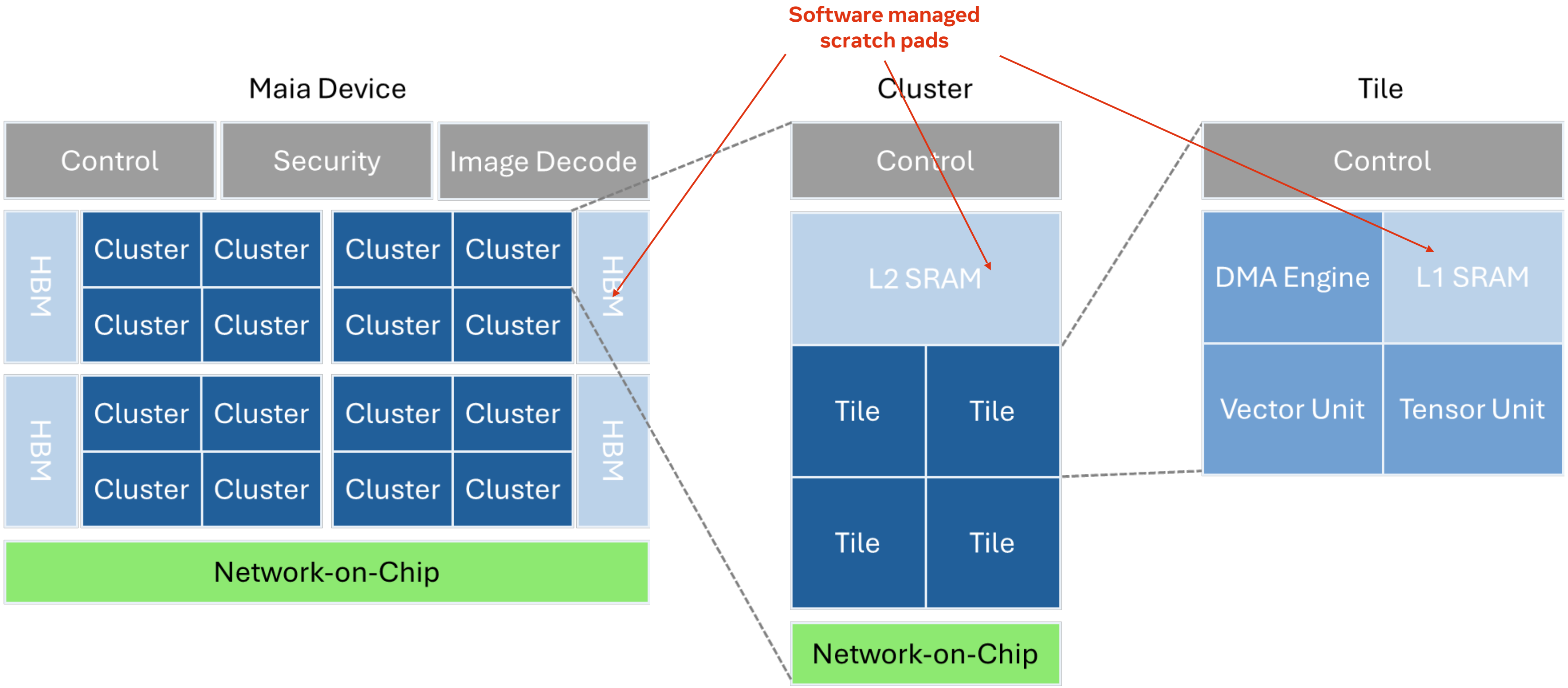
Splitting cluster work among tiles, optimize tile performance

Maia processor orchestration, semaphore management

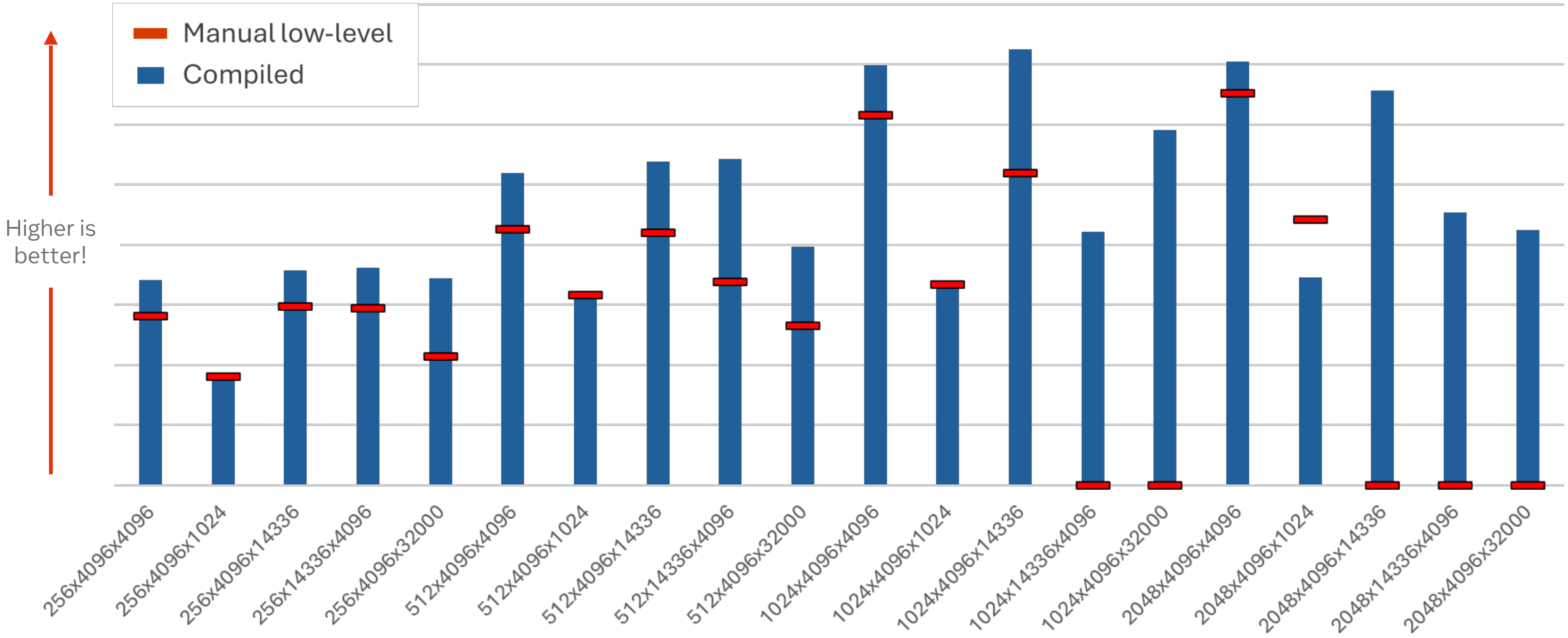
Data movement in memory hierarchy, overlap copy and compute

Optimal codegen

Maia-100 Device Architecture



Matmul Performance on Mistral 7B Shapes



Thank you!