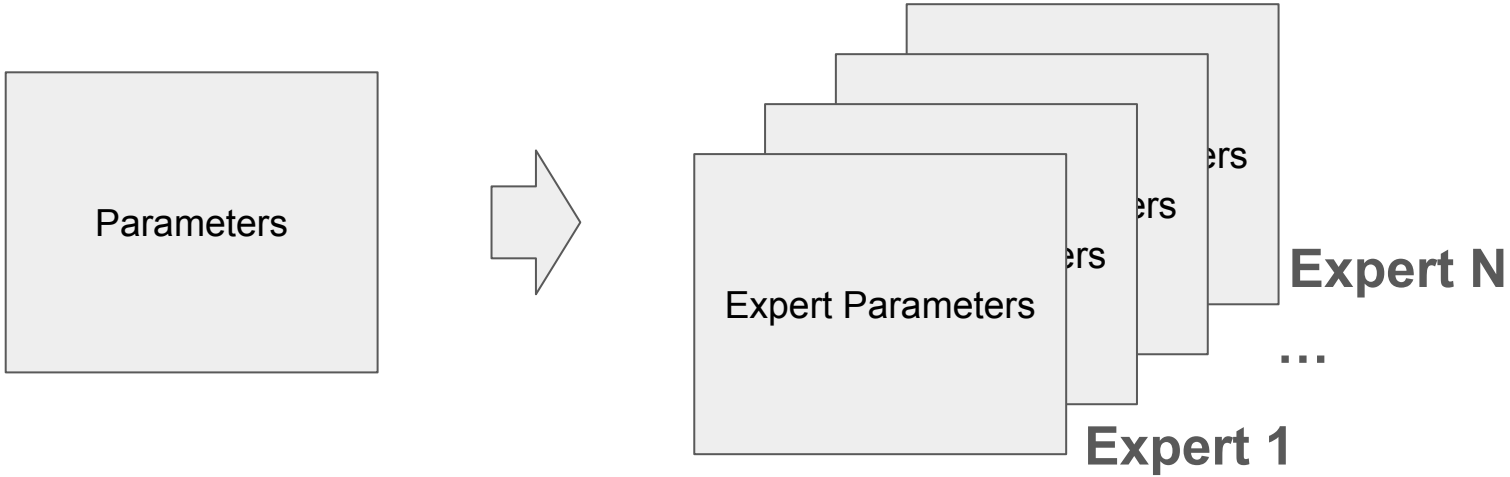
Implementing MoNDE

2024 Summer

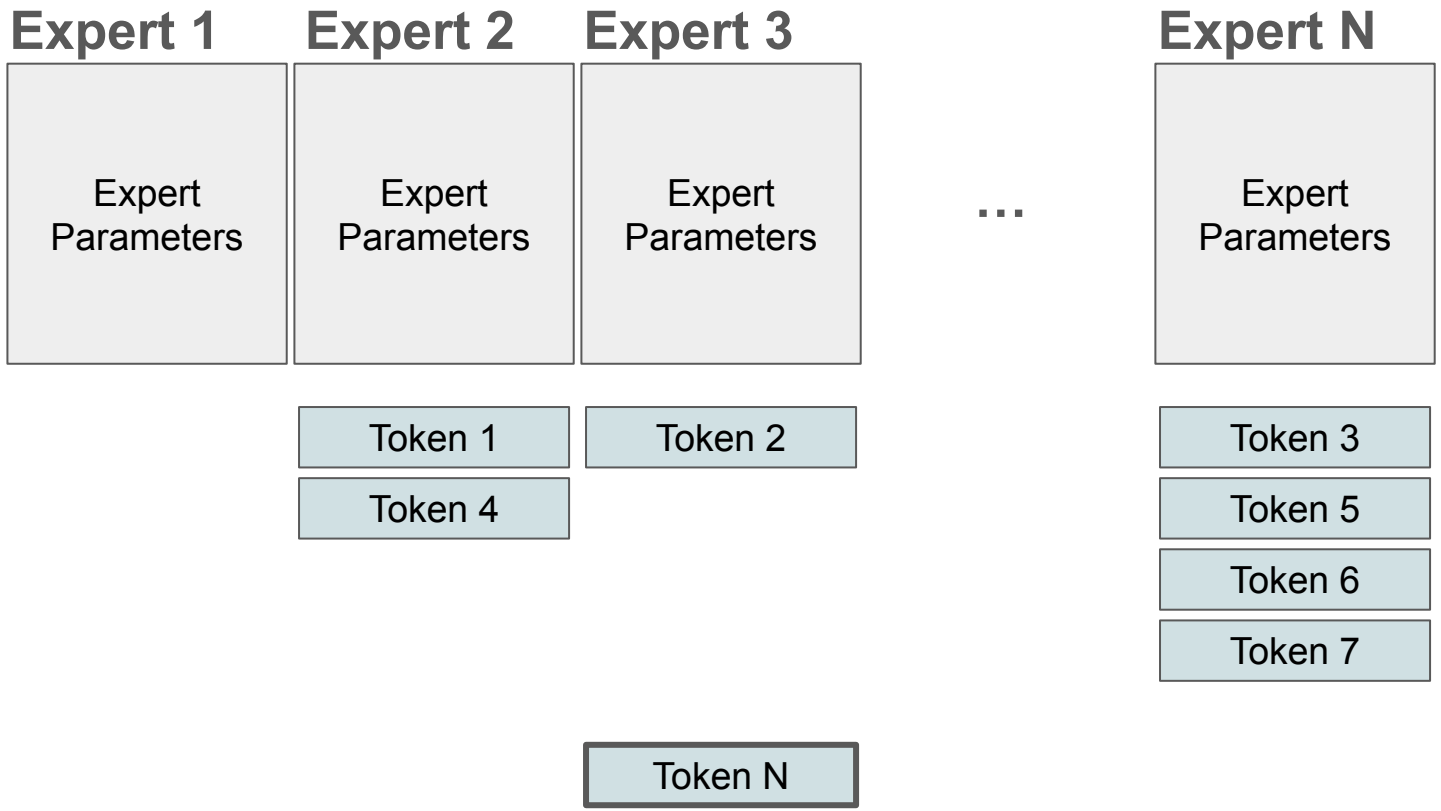
Inho Park

Mixture of Experts

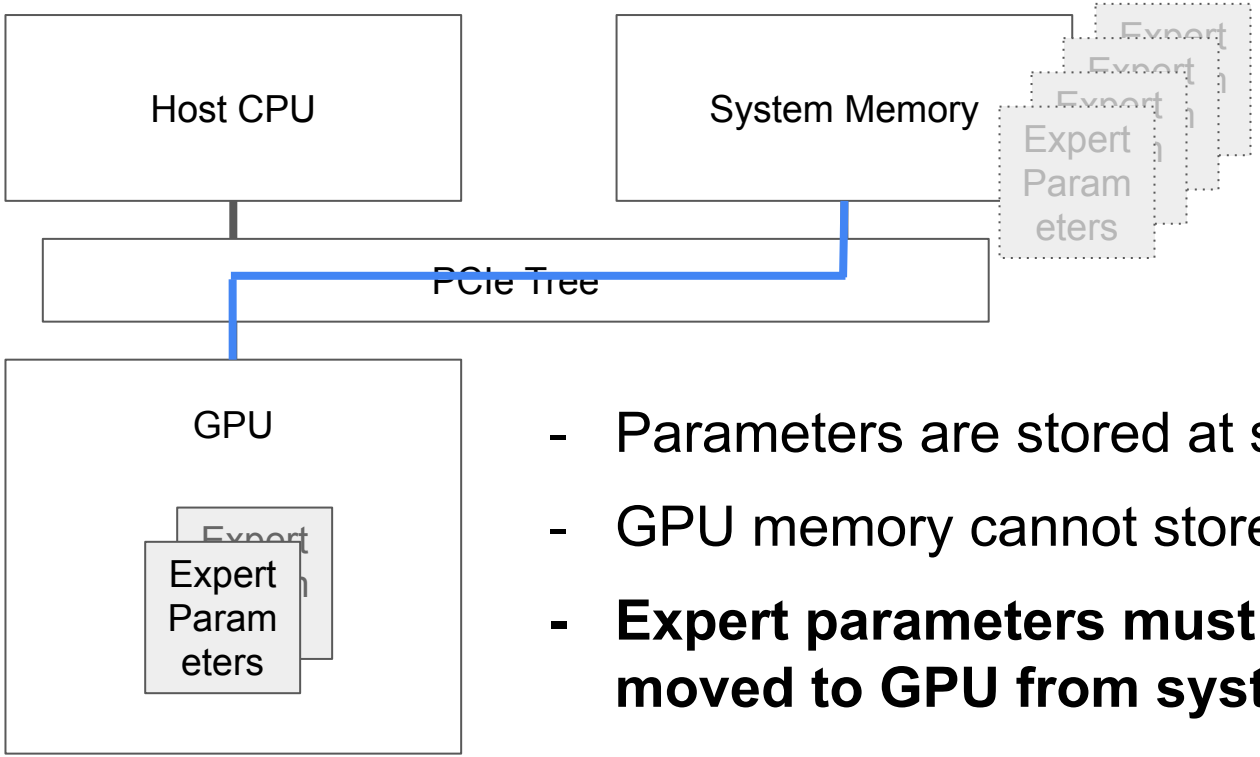


- **Increase number of parameters** while **restraining calculation burden**
- Increase training speed

Mixture of Experts

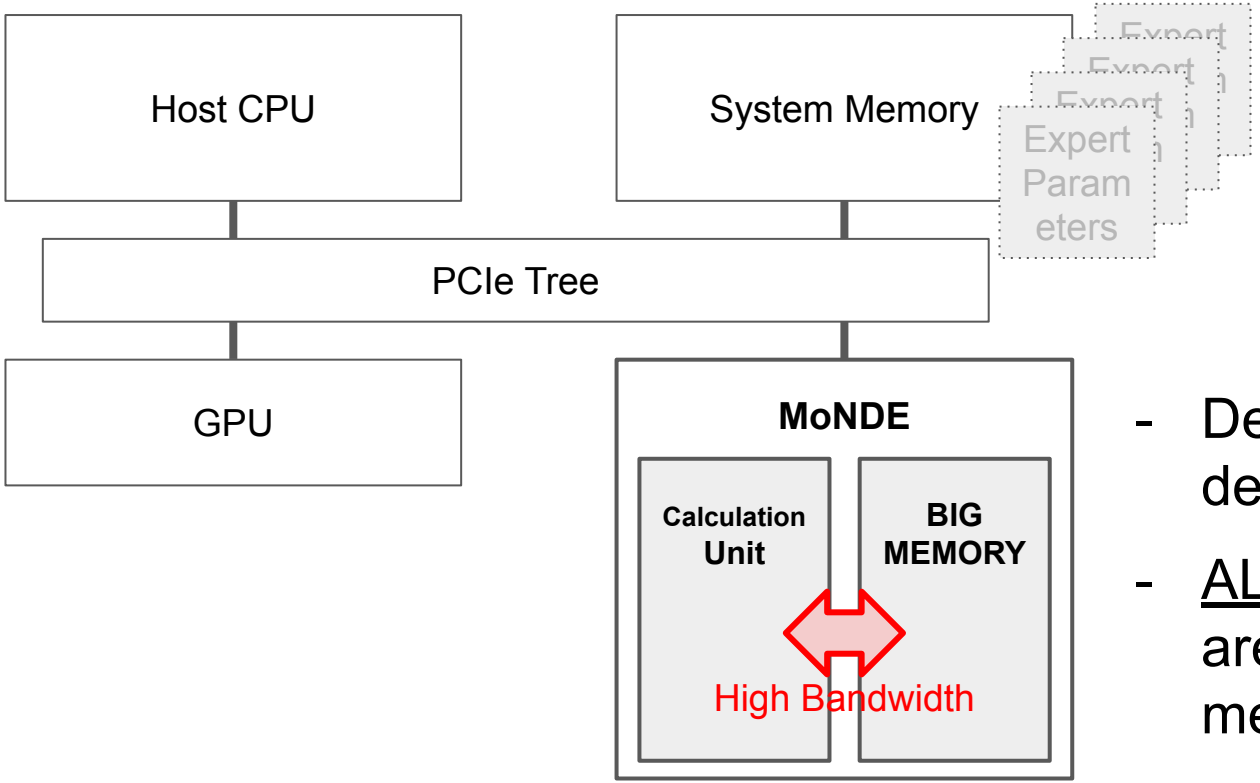


Weights must be moved to GPU from Memory



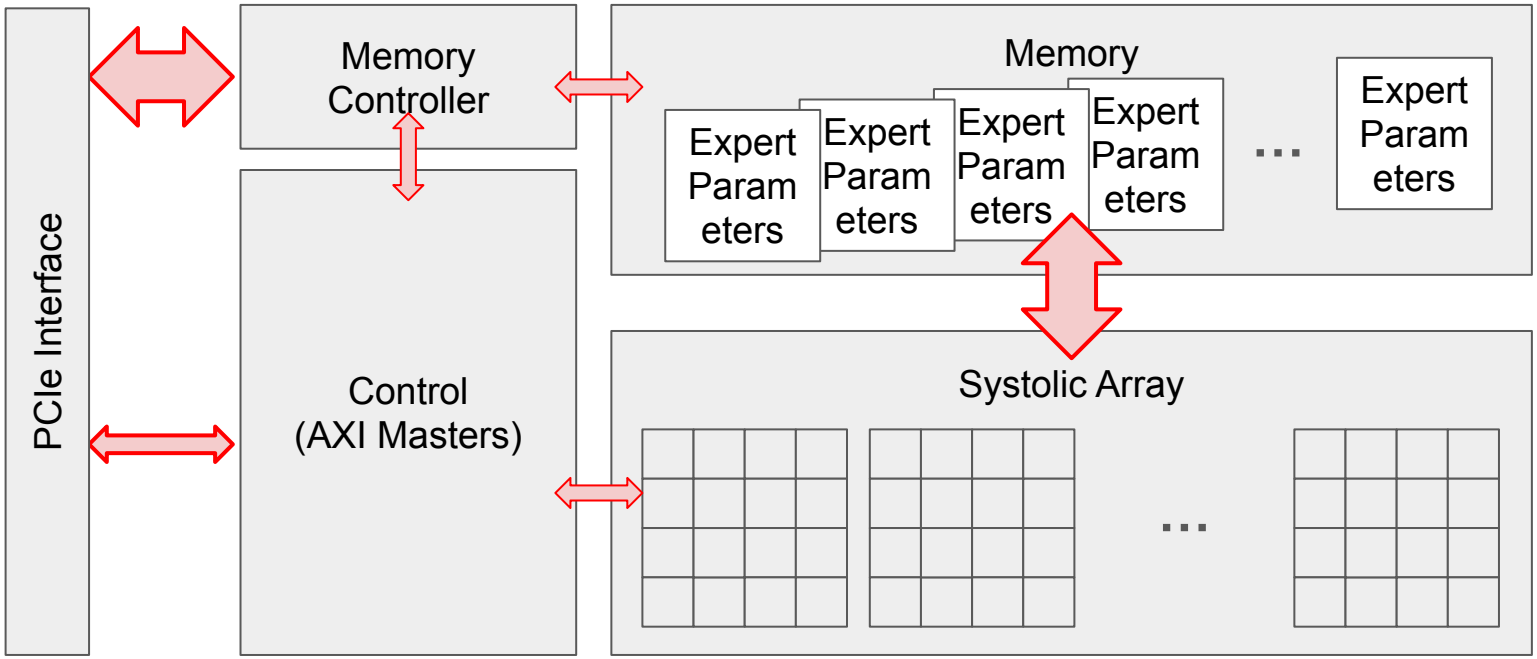
- Parameters are stored at system memory
- GPU memory cannot store all parameters
- **Expert parameters must be repeatedly moved to GPU from system memory**

Weights must be moved to GPU from Memory



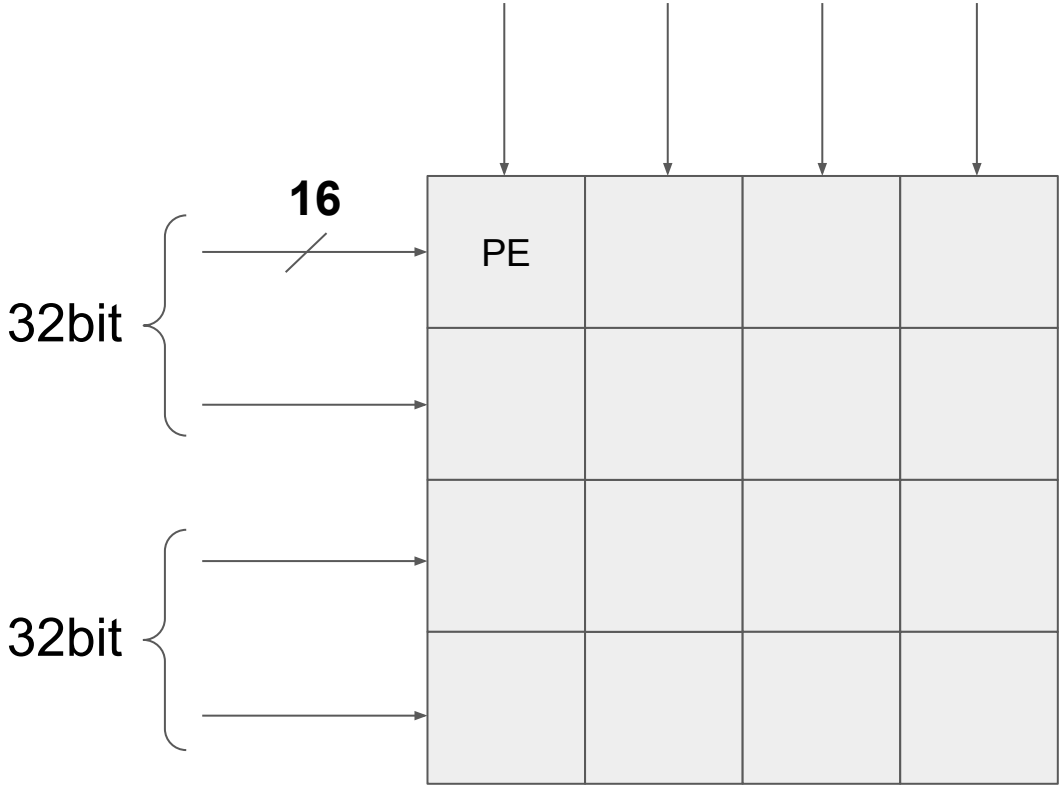
- Device entirely dedicated for MoE
- ALL Expert Parameters are stored at MoNDE's memory

Overall Structure

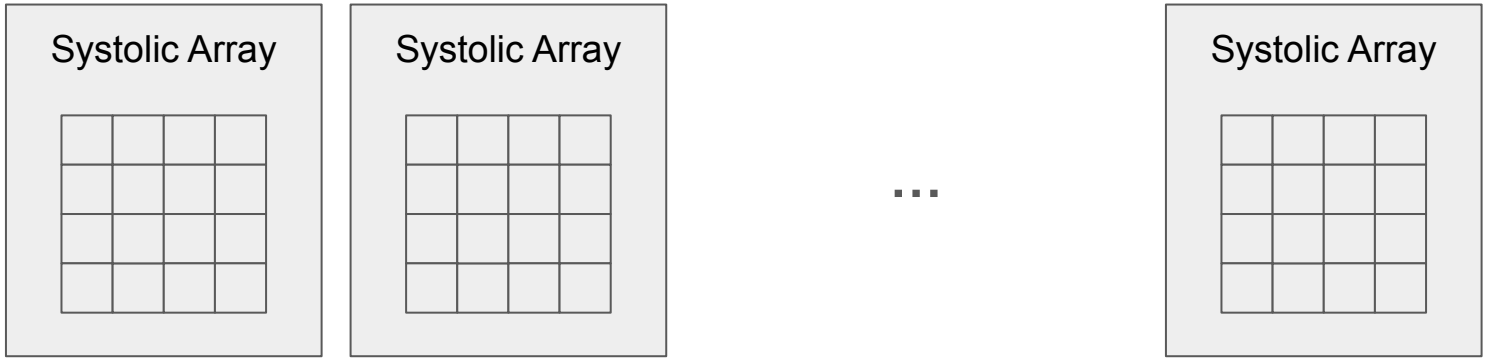


Systolic Array

“2 x 32bit Bus”

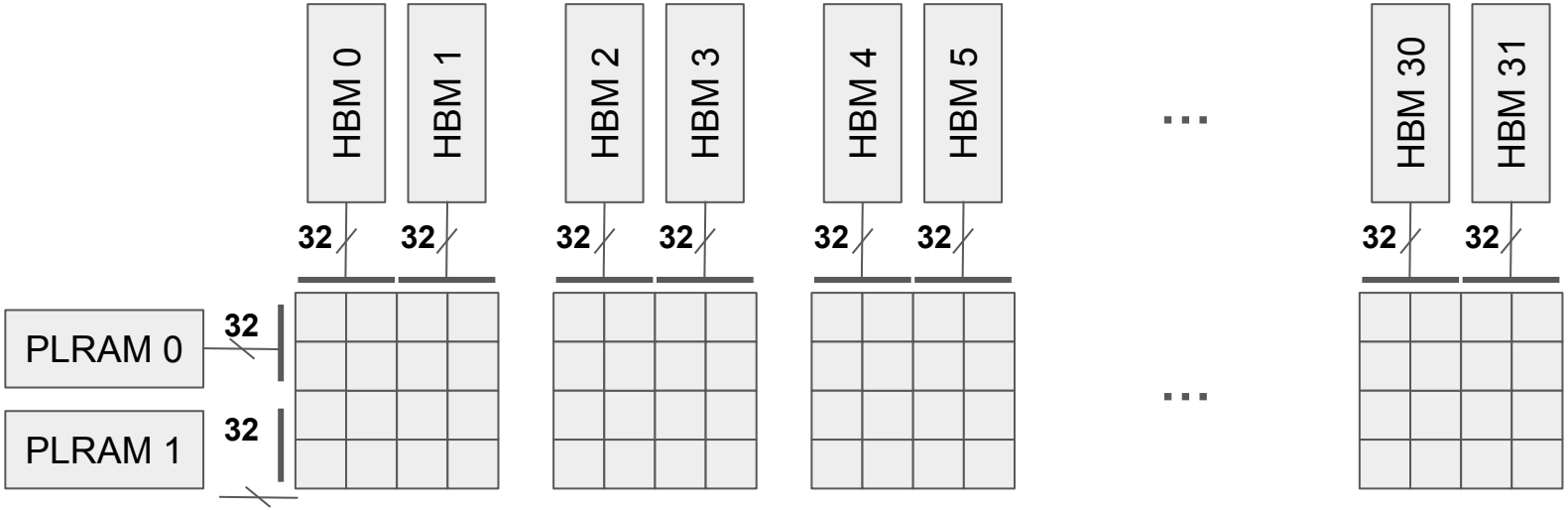


Systolic Array



Due to limitation of U280’s capabilities,
16 systolic arrays were implemented

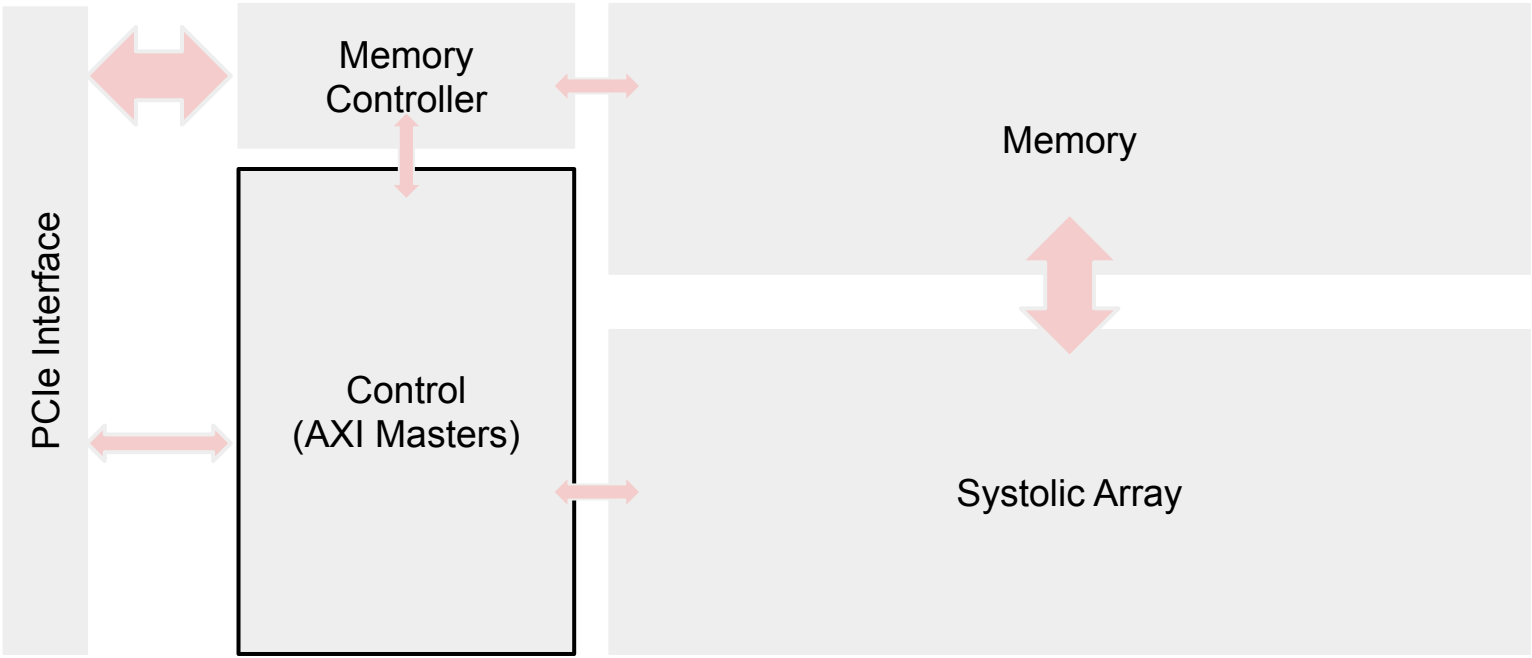
Memory



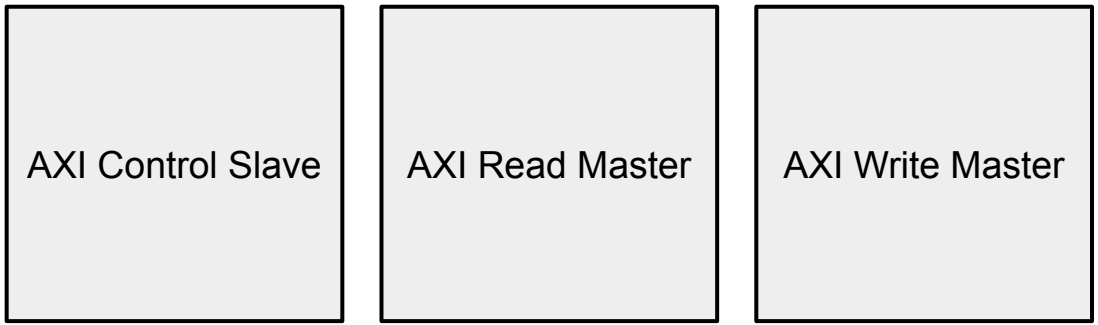
Each systolic array is connected to two HBM pseudo channels, 32 channels are used in total.

Since HBMs can be connected to maximum 32 channels, activation data are stored at PLRAMs (each 128KB).

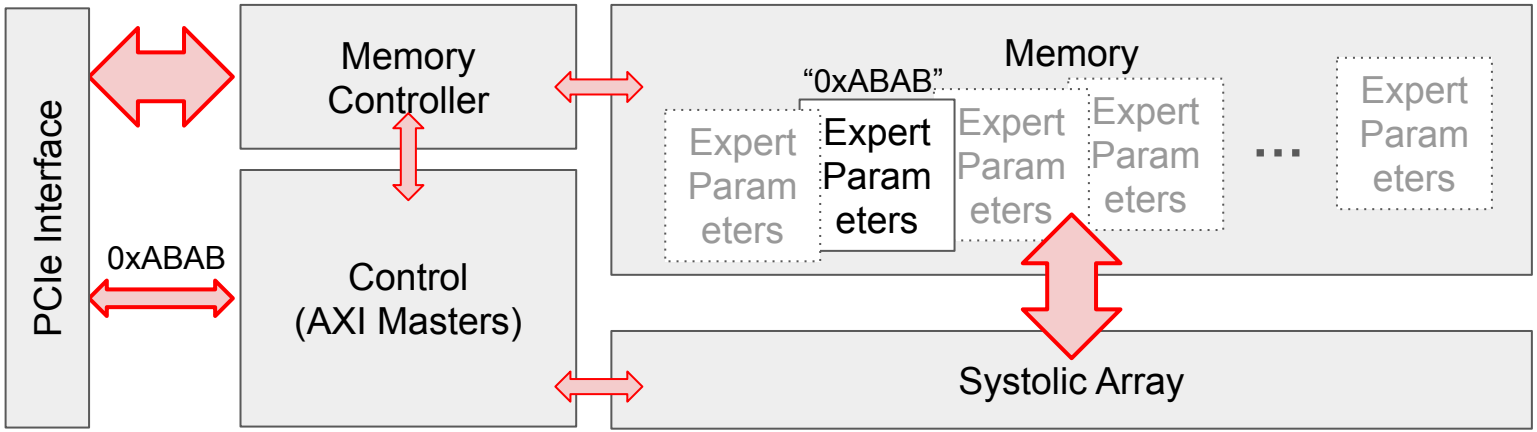
Control


PCIe Interface
Memory Controller
Memory
Control (AXI Masters)
Systolic Array

Control

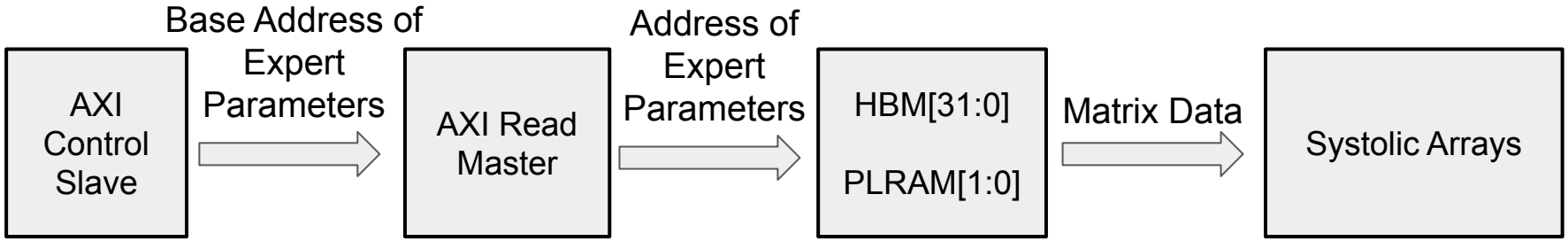


AXI Control Slave



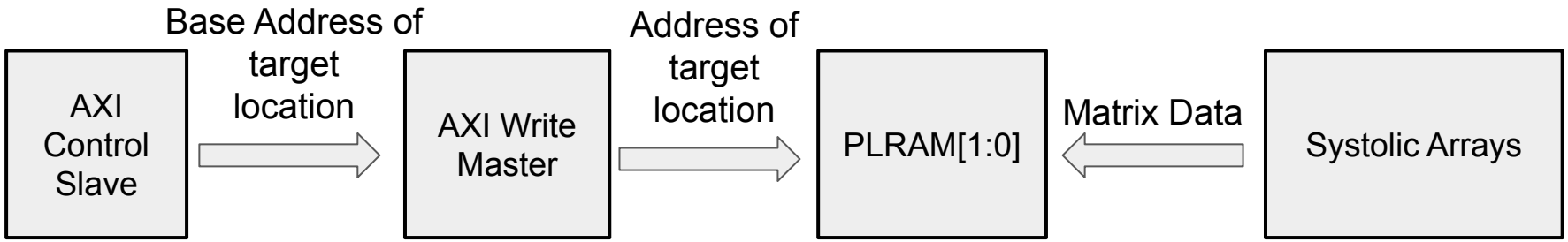
1. The address of expert parameters that should be used for this token is passed.
2. Read&Write control signals (ap_start, ap_done)

AXI Read Master



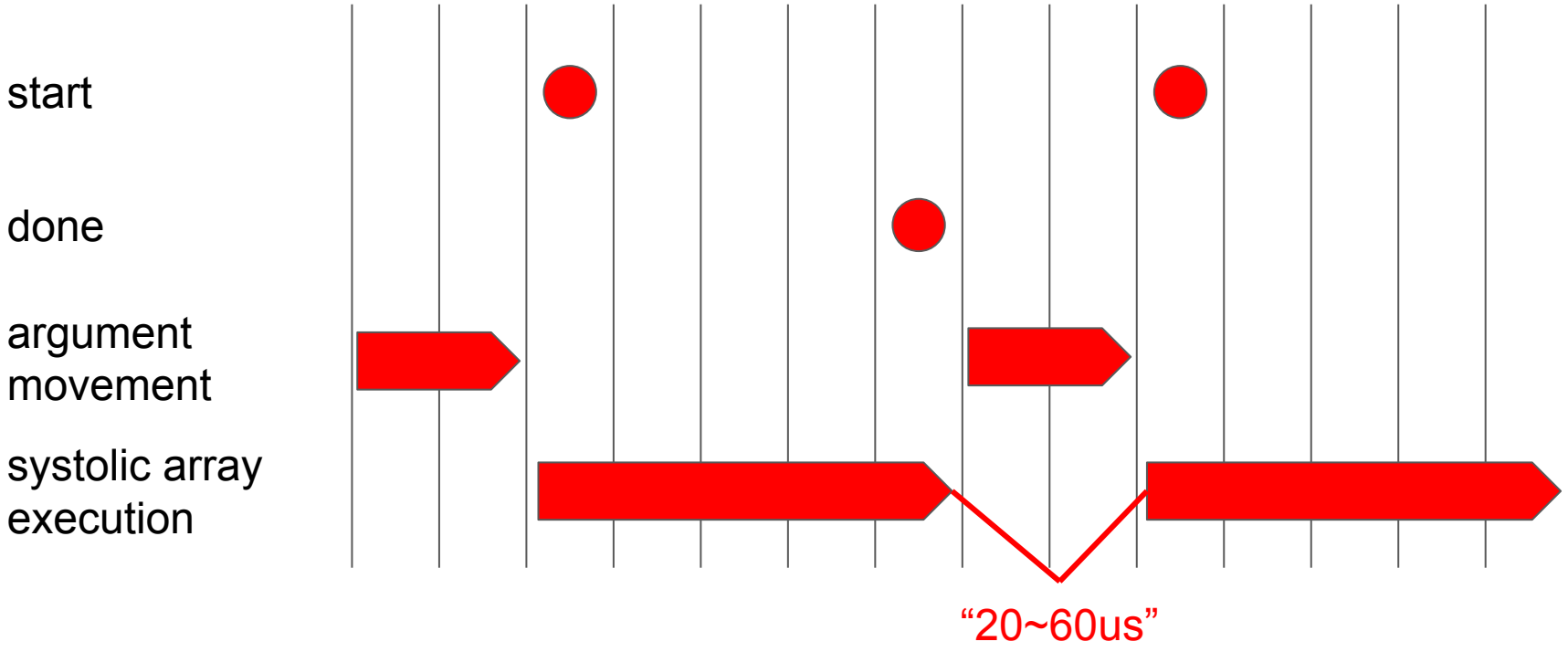
1. Base address of 34(32 HBM, 2 PLRAM) memory blocks are passed to Read Master
2. Data is read from Memory and passed to systolic arrays

AXI Write Master

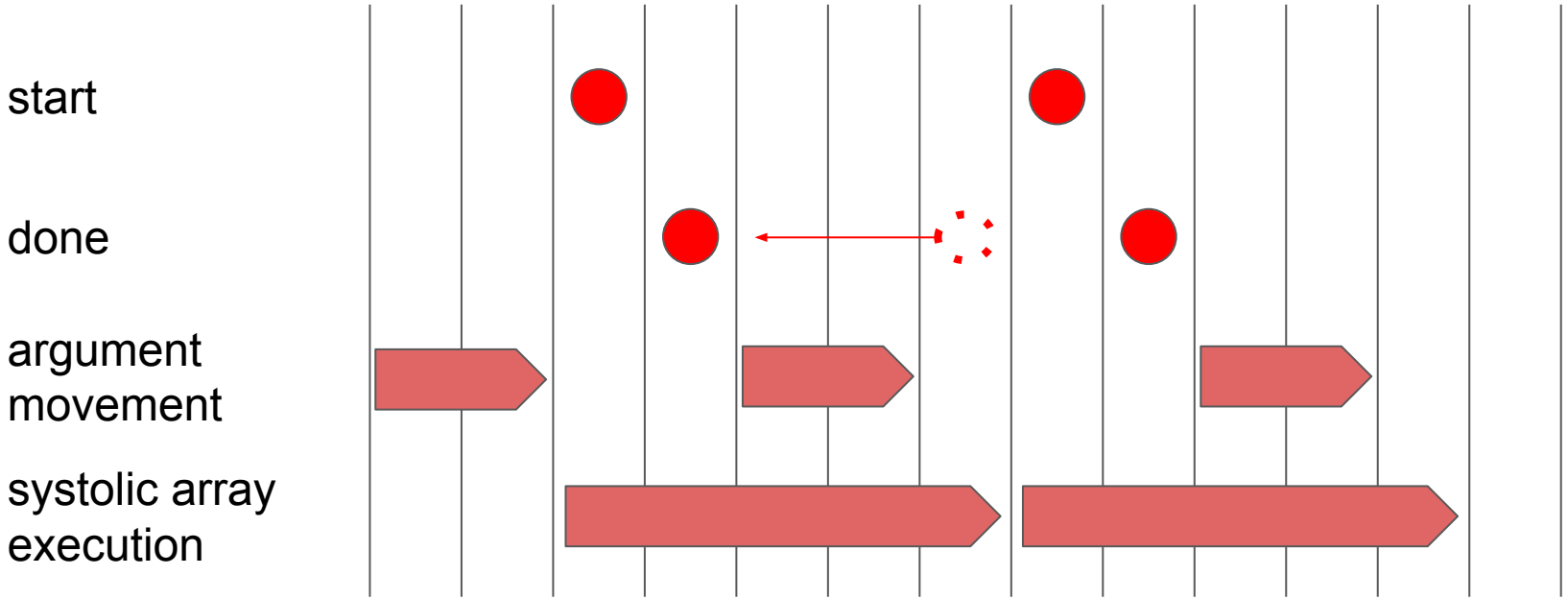


1. Base address of 2 memory blocks are passed to Write Master
2. Calculation Results are passed and written from systolic arrays to PLRAM

Timeline (Basic Control)

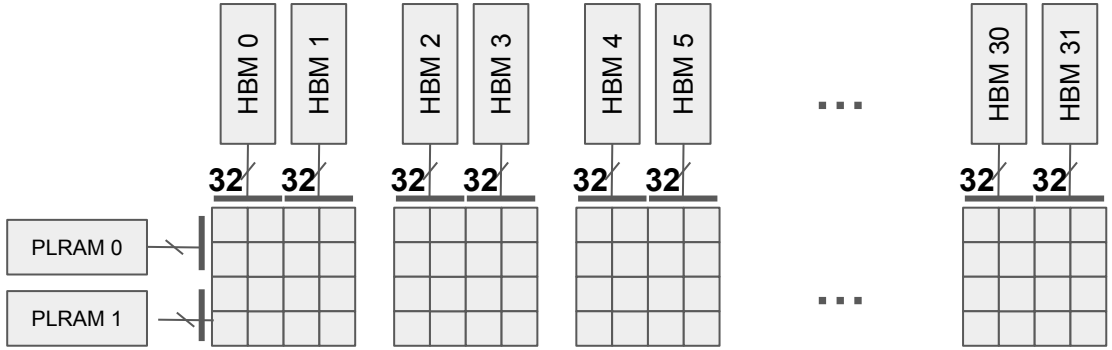


Timeline



- read arguments as soon as systolic arrays start execution

Performance



Ideal Bandwidth

- HBM:

  32ports x 4byte/ports x 133.9MHz = 17.1392 GB/s

- PLRAM:

  2ports x 4byts/ports x 133.9MHz = 1.0712 GB/s

→ 17.1392 GB/s + 1.0712 GB/s = **18.2104 GB/s**

Performance

Measured Bandwidth

matrix A : 4 x 30,000
matrix B : 30,000 x 64
matrix C : 64 x 64

✓ to check **performance**:

measured execution time of 10,000 times of A x B

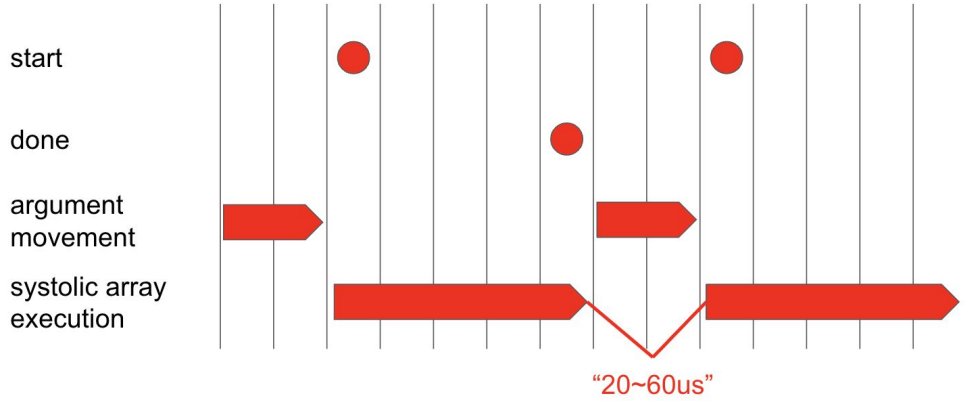
✓ to check **accuracy**:

checked calculation result of A x B x C.

Performance

Results

- **Ideal Bandwidth: <u>18.2104 GB/s</u>**
- Basic Control: 8.988 GB/s (49.3%)
- **Improved Control: 17.3943 GB/s (95.5%)**

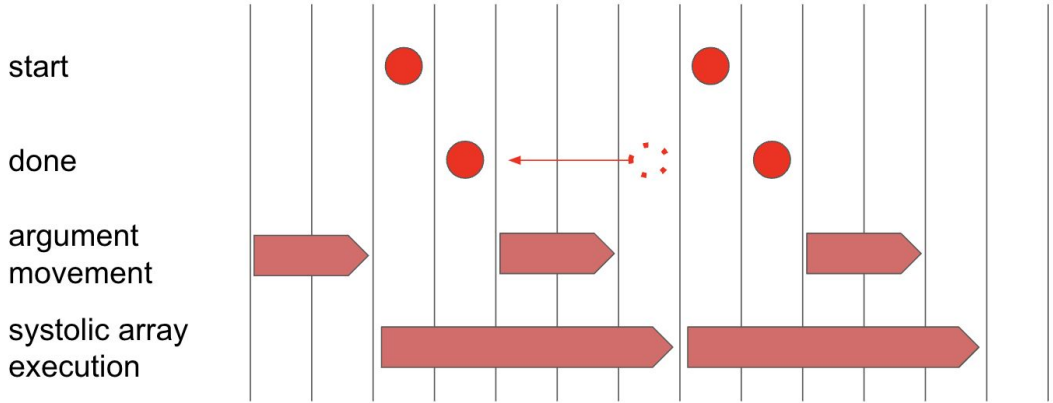


Performance

```
[Test Info]
Matrix A: 4 x 30000
Matrix B: 30000 x 64
Matrix C: 64 x 64
Number of Repeats: 10000

Bandwidth: 17.3943GB/s

TEST FINISHED
```

Results

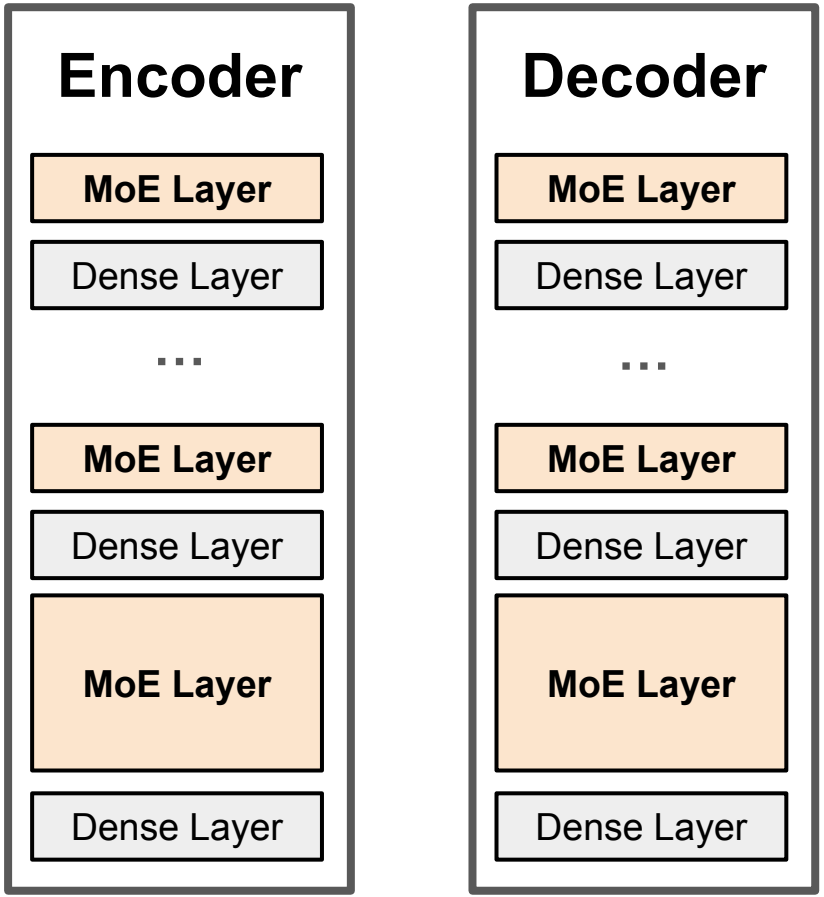
- **Ideal Bandwidth: <u>18.2104 GB/s</u>**
- Basic Control: 8.988 GB/s (49.3%)
- **Improved Control: 17.3943 GB/s (95.5%)**



Performance: Demo

MoE model for *summarizing task* using *‘google/switch-base-8’*

Performance: Demo



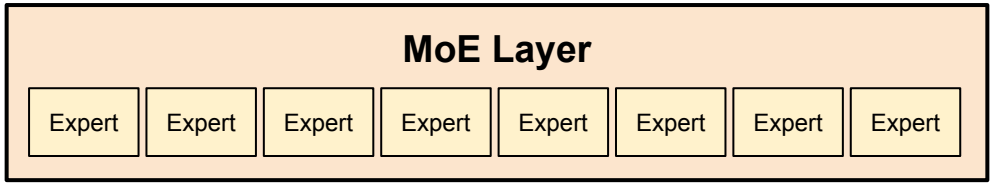
Encoder

- **6** Dense layers
- **6** MoE layers

Decoder

- **6** Dense layers
- **6** MoE layers

Performance: Demo

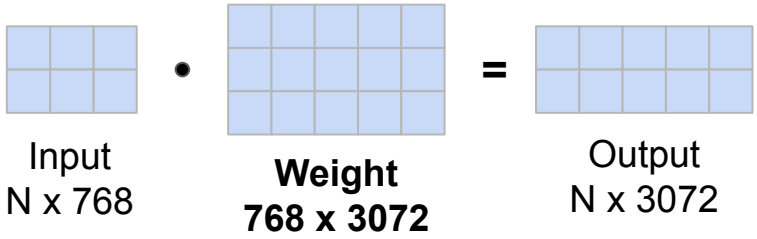


Each MoE Layer

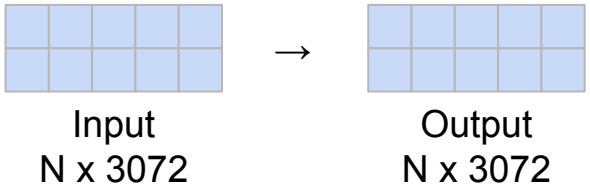
8 Experts

Performance: Demo

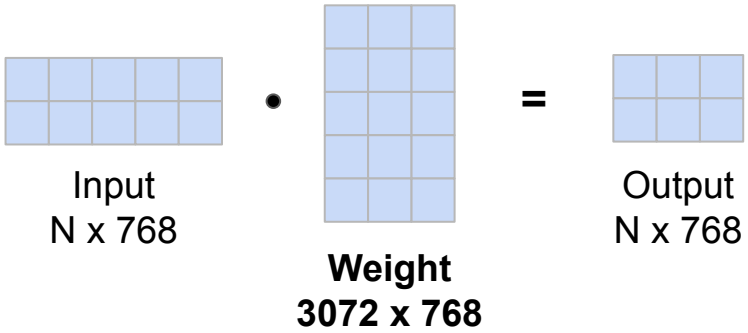
1. Linear



2. ReLU



3. Linear

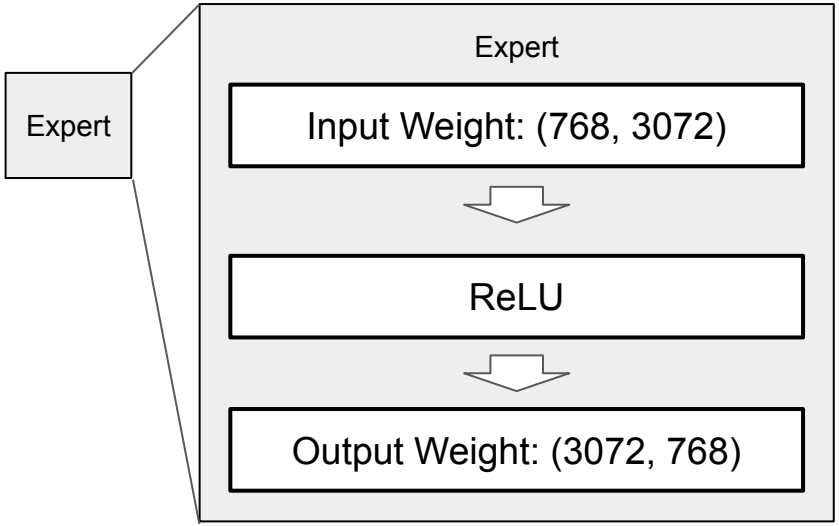


Performance: Demo

Size of MoE parameters

1. Each Expert's parameter count:
   - Matrix 1 : 768 x 3072 = 2,359,296
   - Matrix 2 : 3072 x 768 = 2,359,296
   - Total data size = 2 x 2,359,296 x 2bytes = 9MB
2. Each Layer's Data size:
   - 8 expert x 9MB/expert = 72 MB/layer
3. Total Data size:
   - 12 layers x 72MB/layer = **864MB of MoE parameters**

Performance: Demo



0. Expert Data is Uploaded in HBM

1. Activation is moved to PLRAM
2. Calculation: matrix mult. + ReLU
3. Result is stored in PLRAM

4. Calculation
5. Result is stored in PLRAM

6. Final Data is moved to Host

Performance: Demo

[Tokens per Expert]

Layer	1	2	3	4	5	6	7	8	Experts Used
1	10	7	9	10	6	6	16	11	8
2	2	12	3	14	14	16	8	6	8
3	10	9	9	5	20	8	5	9	8
4	17	6	11	0	3	30	0	8	6
5	2	1	40	4	8	9	2	9	8
6	3	4	15	13	11	15	6	8	8

Performance: Demo

CPU

```
...Programmed Successfully: device[0]

...Uploading Weights (Host -> FPGA)
...Uploading Weights Done: 0s

...Inference Start
...Inference Done: 19.265466s

Result: <pad> Peter and Elizabeth were in a party in Paris.</s>
```

CPU: 19.26s

MoNDE

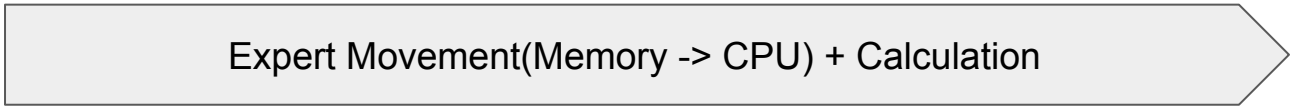
```
...Programmed Successfully: device[0]

...Uploading Weights (Host -> FPGA)
...Uploading Weights Done: 9s

...Inference Start
...Inference Done: 7.179557s

Result: <pad> Peter and Elizabeth were in a party in Paris.</s>

...Inference Done: 7.505218s

Result: <pad> Peter and Elizabeth were in a party in Paris.</s>
```

Transfer Weights: 9s

MoNDE: 7.18s

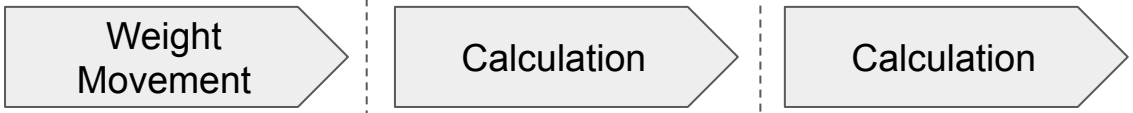
Performance: Demo

1. CPU



19.09s

2. MoNDE



10s | 7.18s | 7.50s