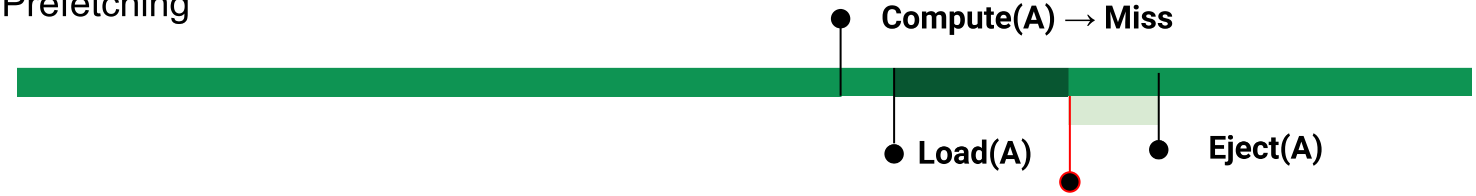
Improving Hardware Utilization in MAPLE

Victor Gao - 1000523613
Leo Han - 1004921677

Background

No Prefetching

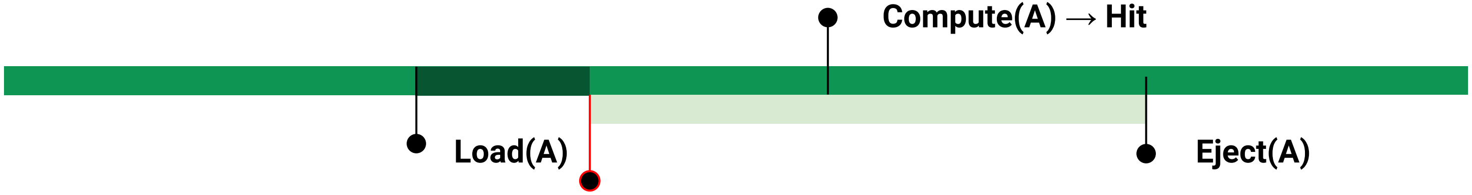


Compute(A) → Miss

Load(A)

Eject(A)

W/ Prefetching



Compute(A) → Hit

Load(A)

Eject(A)

LIMA - Loop of Indirect Memory Accesses

```
for (i = K; i < N; i++) {
        data = A[i];
        …
}
```

```
For (i = K; i < N; i++) {
        data = A[B[i]];
        …
}
```

			0	1	2	3	4	5	6
7	8	9	10	11					

	2					9			
				4					
	3								
			0						
	6	7		5		10		8	
						1	11		

DAE - Decoupled Access and Execute

Original

```
For (i = K; i < N; i++) {
        data = A[B[i]];
        res[i] += data * C[i];
        …
}
```

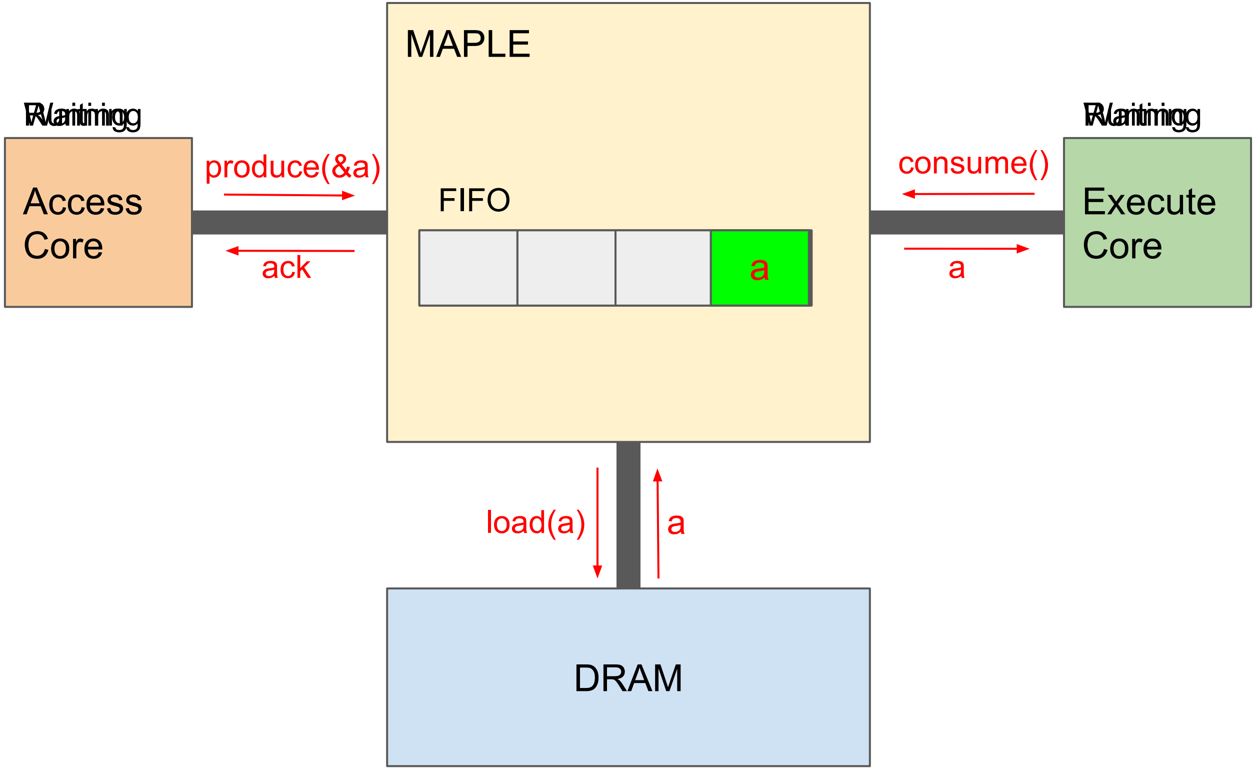
Access Thread

```
For (i = K; i < N; i++) {
        produce(&A[B[i]]);
}
```

Execute Thread

```
For (i = K; i < N; i++) {
        data = consume();
        res[i] += data * C[i];
        …
}
```

MAPLE - Memory Access Parallel-Load Engine



MAPLE Benefits

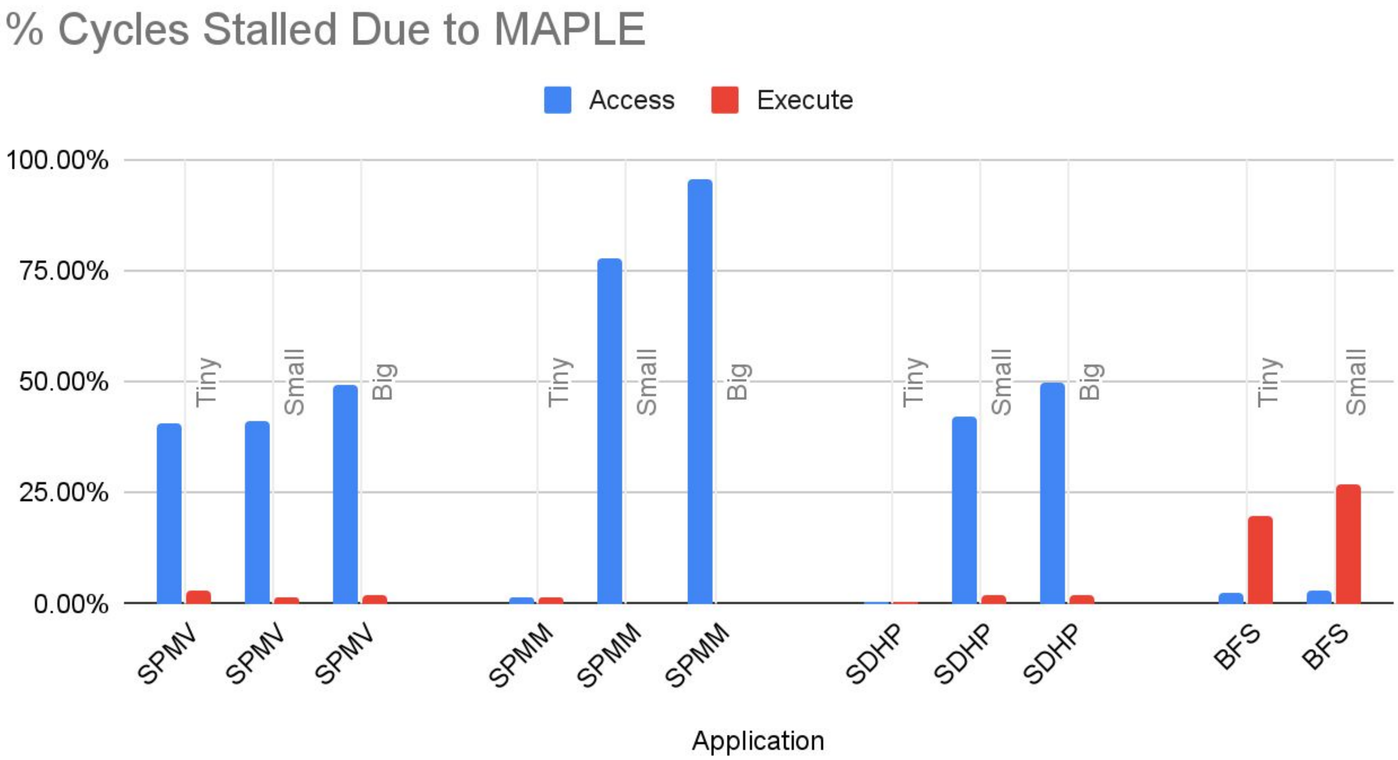
- Offloading memory access latency from access core
- Single MAPLE core can support multiple access/execute cores
- No architectural changes needed in cores and no specialized ISA

Motivation

Imbalance in workload will result in frequent stall in the less loaded thread.

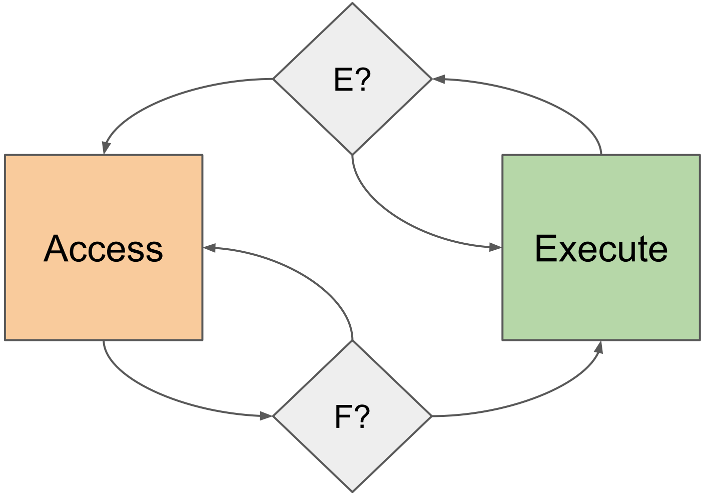
a. If access is less loaded, it will stall when FIFO is full.
b. If execute is less loaded, it will stall when FIFO is empty.

Profiling Stalls



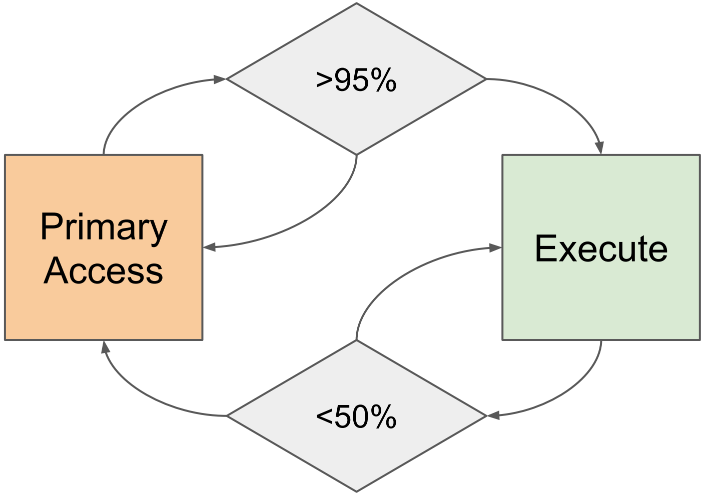
Proposed Solution

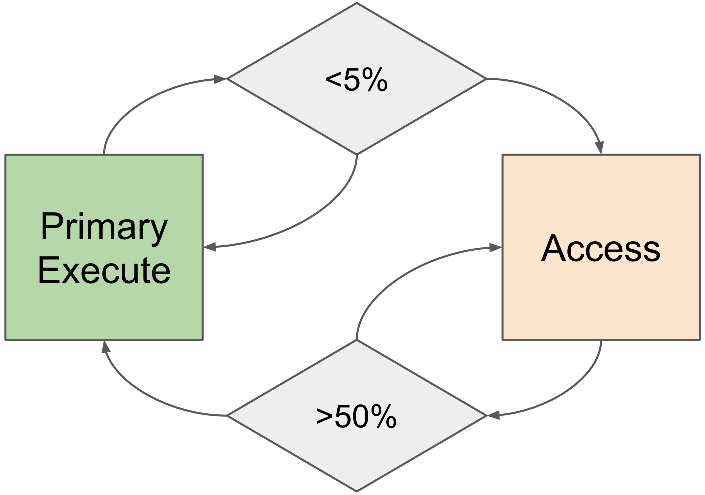
Symmetric Switching



Proposed Solution

Asymmetric Switching





Experiment

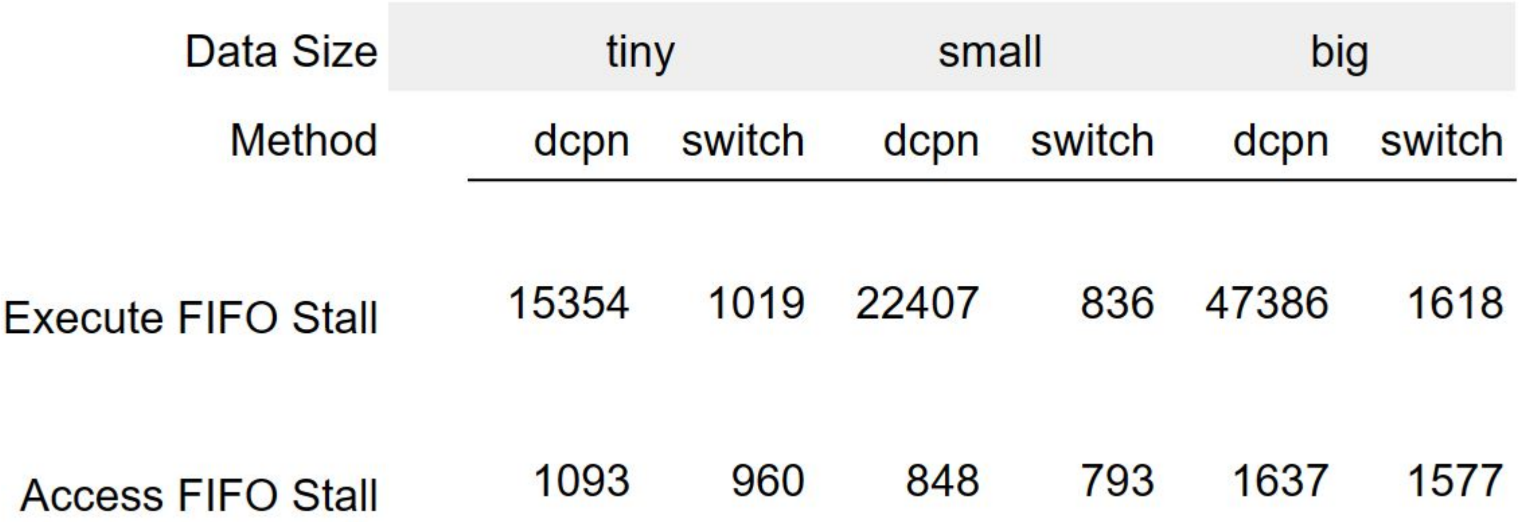
Limitation

- So far in simulation, haven’t got the access and execute threads to coordinate properly when switching is added.

Work Around

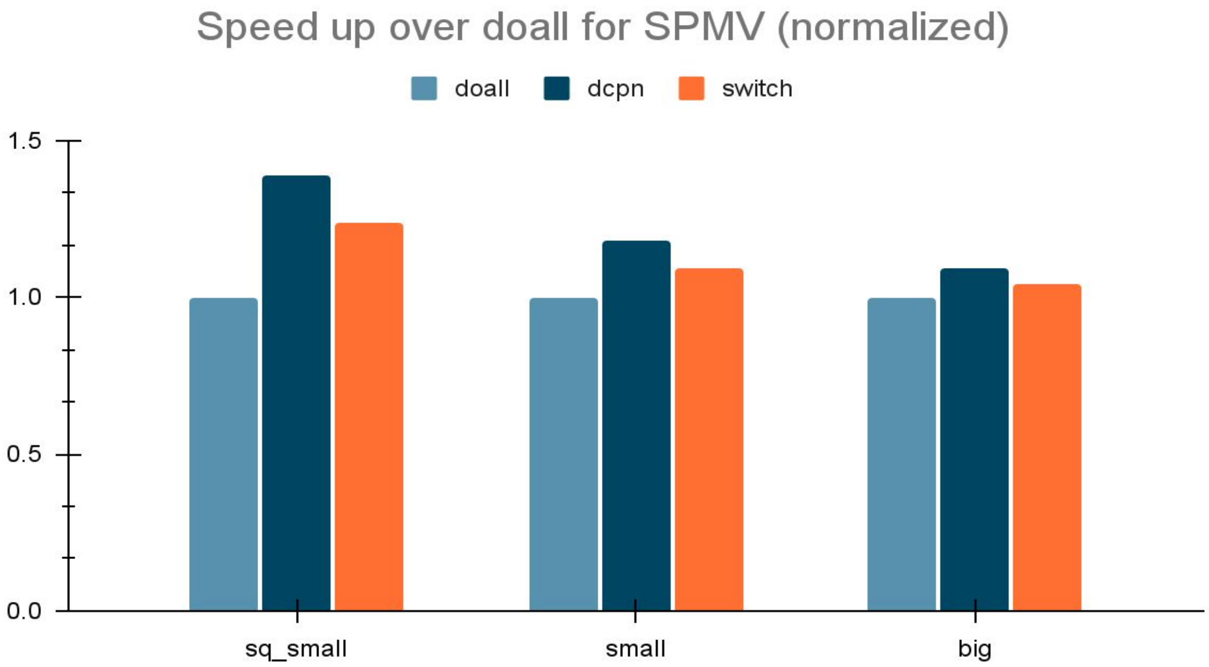
- Run the access and execute on the same thread with switching depends on the level of the FIFO (perfect prefetching).

Preliminary Results



Data Size	tiny		small		big	
Method	dcpn	switch	dcpn	switch	dcpn	switch
Execute FIFO Stall	15354	1019	22407	836	47386	1618
Access FIFO Stall	1093	960	848	793	1637	1577

Preliminary Results



What is Next?

- Synchronizing role-switching between parallel cores
- Implementing and collecting results on other workloads (SPMM, SDHP, BFS)
- Implemented and test with coarser-grained polling of FIFO status
- Exploring hardware interrupt-based role switching

Questions?