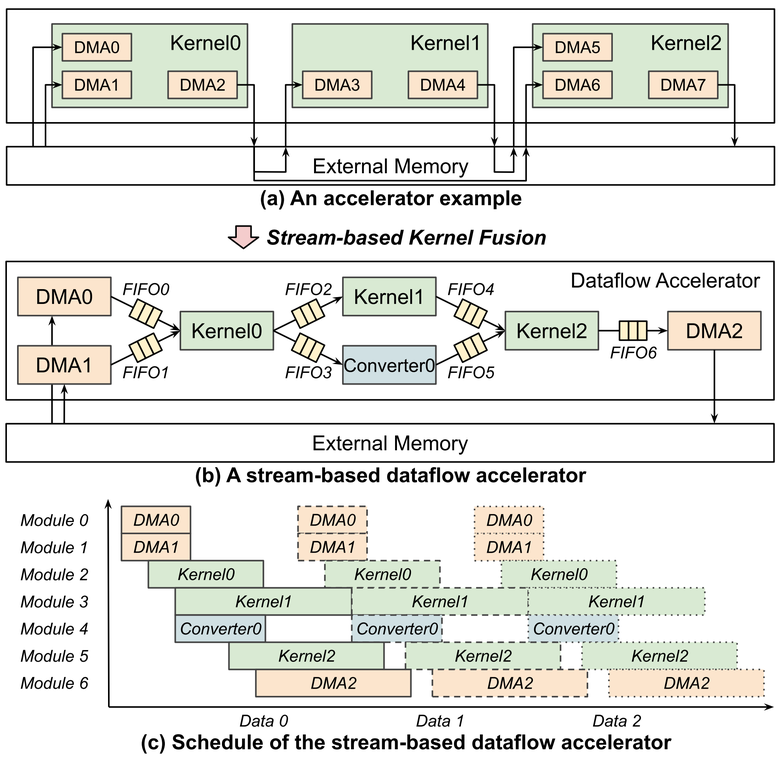
StreamTensor: Make Tensors Stream in Dataflow Accelerators for LLMs

Hanchen Ye[1], Deming Chen[1,2] *[1]University of Illinois at Urbana-Champaign; [2]Inspirit IoT, Inc.*

Dataflow Architecture: Advantages and Pitfalls

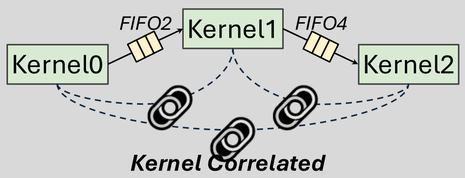


Reduce on-chip buffer utilization ✅
Reduce external memory access ✅
Reduce latency & throughput ✅

Pitfall 1 ❌ *External Memory Access*

- Best external memory layout matching stream pattern?
- Widen external memory interfaces to maximize bandwidth?

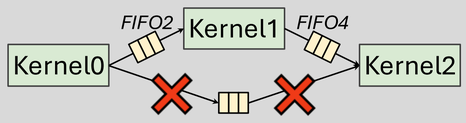
Pitfall 2 ❌ *Inter-kernel Correlation*



Kernel Correlated

- Kernel tiling & parallelization?
- Local buffer partition?
- Kernel loop permutation?

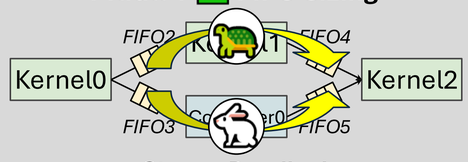
Pitfall 3 ❌ *Kernel Fusion*



Stream Layout/Order Mismatch

- Possible to stream?
- Minimal stream buffer size?
- Global fusion strategy?

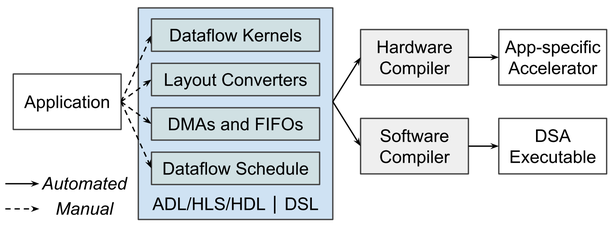
Pitfall 4 ❌ *FIFO Sizing*



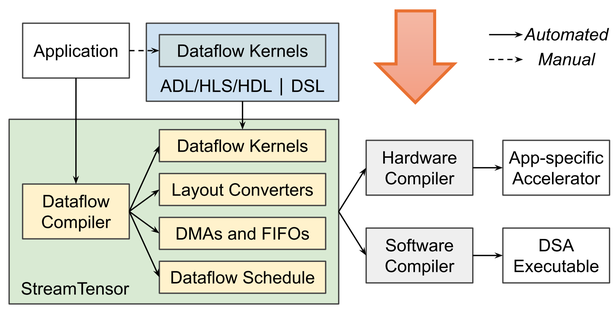
Stream Deadlock

- Latency mismatch on different paths between two kernels
- Minimal FIFO size to avoid deadlock or kernel stall?

Design Paradigm Shift

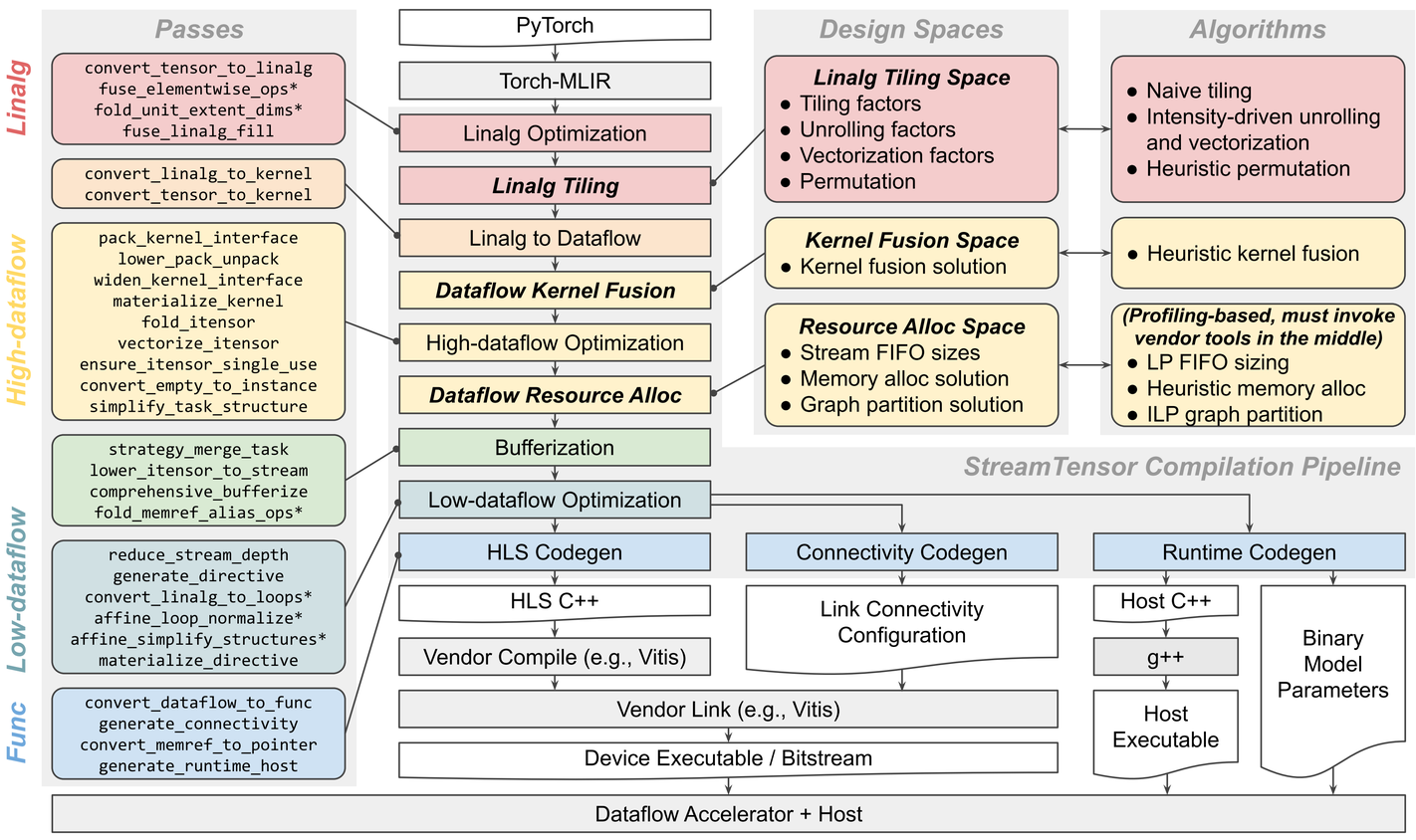


- Manual dataflow kernels, layout converters, DMAs, and FIFOs design
- Difficult to find solutions of the pitfalls

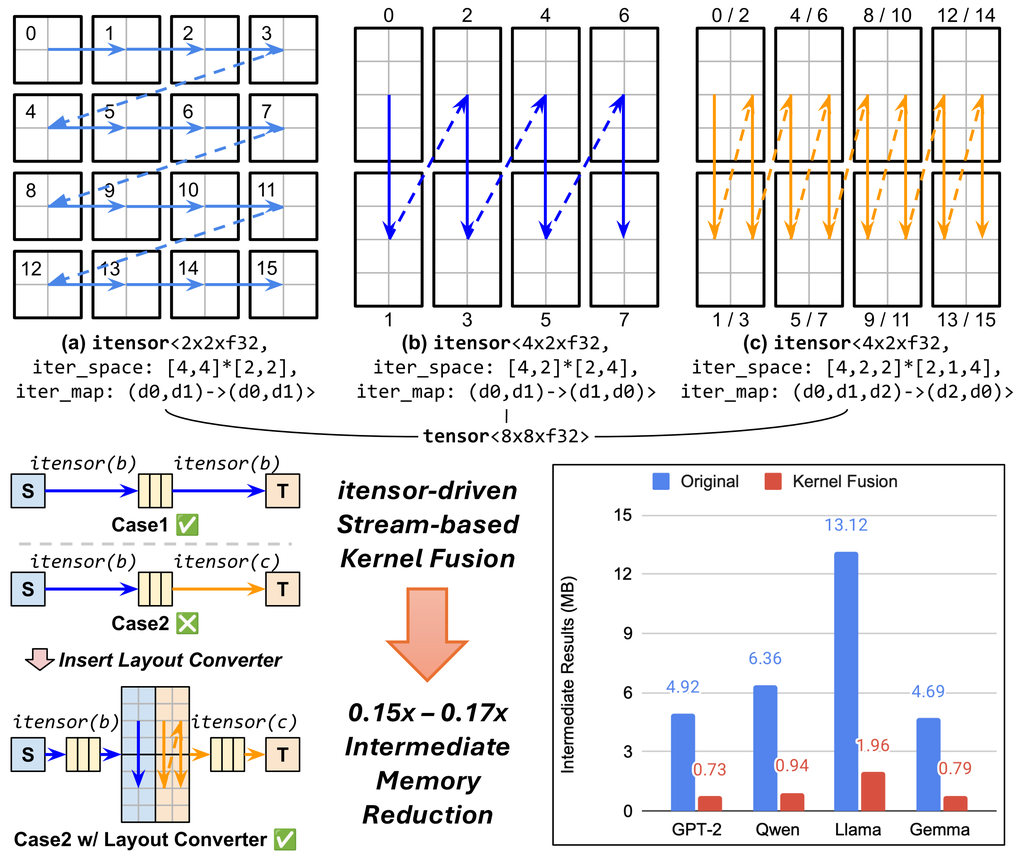


- Automate the generation of dataflow kernels, layout converters, DMAs, and FIFOs
- Resolve pitfalls through systematic DSE
- Support auto-tuning and external library

StreamTensor Framework



Iterative Tensor (itensor) Typing System



StreamTensor Evaluation Results on LLMs

[Input Len: Output Len]	Ours Latency (ms)	Ours TTFT (ms)	Ours Speed (token/s)	Allo [15] (Ratio of $\frac{Ours}{Allo}$) Latency (ms)	Allo TTFT (ms)	Allo Speed (token/s)	DFX [29] (Ratio of $\frac{Ours}{DFX}$) Latency (ms)	DFX TTFT (ms)	DFX Speed (token/s)
[32:32]	194.99	34.59	199.51	238.32 (0.82x)	81.50 (0.42x)	204.05 (0.98x)	350.00 (0.56x)	177.20 (0.20x)	185.19 (1.08x)
[64:64]	358.24	61.27	215.51	476.64 (0.75x)	162.99 (0.38x)	204.05 (1.06x)	694.70 (0.52x)	349.10 (0.18x)	185.19 (1.16x)
[128:128]	696.65	125.35	224.05	953.28 (0.73x)	325.98 (0.38x)	204.05 (1.10x)	1384.00 (0.50x)	692.80 (0.18x)	185.19 (1.21x)
[256:256]	1387.76	272.85	229.61	1906.56 (0.73x)	651.96 (0.42x)	204.05 (1.13x)	2800.00 (0.50x)	1417.60 (0.19x)	185.19 (1.24x)
Geo. Mean	-	-	-	**0.76x**	**0.40x**	**1.06x**	**0.52x**	**0.19x**	**1.17x**

[Input Len: Output Len]	Ours Latency (ms)	Ours TTFT (ms)	Ours Speed (token/s)	A100 (Ratio of $\frac{Ours}{A100}$) Latency (ms)	A100 TTFT (ms)	A100 Speed (token/s)	2080Ti (Ratio of $\frac{Ours}{2080Ti}$) Latency (ms)	2080Ti TTFT (ms)	2080Ti Speed (token/s)
[32:32]	194.99	34.59	199.51	291.16 (0.67x)	8.72 (3.97x)	113.30 (1.76x)	518.46 (0.38x)	24.98 (1.38x)	64.85 (3.08x)
[64:64]	358.24	61.27	215.51	567.41 (0.63x)	8.76 (6.99x)	114.56 (1.88x)	1010.81 (0.35x)	25.23 (2.43x)	64.94 (3.32x)
[128:128]	696.65	125.35	224.05	1118.28 (0.62x)	8.65 (14.49x)	115.35 (1.94x)	3969.76 (0.18x)	25.26 (4.96x)	32.45 (6.90x)
[256:256]	1387.76	272.85	229.61	2227.79 (0.62x)	8.53 (31.99x)	115.35 (1.99x)	7914.23 (0.18x)	25.23 (10.81x)	32.45 (7.08x)
Geo. Mean	-	-	-	**0.64x**	10.65x	**1.89x**	**0.25x**	3.67x	**4.73x**

0.64x Latency and 1.89x Speed vs. NVIDIA A100 ✅
0.76x Latency and 1.06x Speed vs. SOTA FPGA Accelerator ✅
Improved Design Productivity and Flexibility ✅

- StreamTensor outperforms FPGA LLM accelerators, Allo and DFX, on overall latency, generation speed, and TTFT (time to first token).
- As an ADL, Allo follows the traditional dataflow accelerator design paradigm, demanding manual efforts for all the dataflow components design and comprehension of FIFO sizes, etc.
- StreamTensor outperforms NVIDIA GPU on overall latency, although performs worse on TTFT, showing the advantage of dataflow accelerator on memory-bound applications (LLM generation).

[Allo] Chen, Hongzheng, et al. "Allo: A programming model for composable accelerator design." *PLDI'24*.
[DFX] Hong, Seongmin, et al. "DFX: A low-latency multi-FPGA appliance for accelerating transformer-based text generation." *MICRO'22*.