MOML @ MIT 2026 · 8th Molecular Machine Learning Conference · Extended version: EurIPS SimBioChem 2025

One nanosecond of MD, three times faster.
Kernel fusion for TensorNet

Manpreet Singh · Undergraduate student · Embedded LLM, Singapore · manpreet.singh@embeddedllm.com

embedded

QR code
or browse files

SCAN · GET THE PAPER

01 THE RESULT · ONE A100

2.82×

faster molecular dynamics. Same model. Same physics.

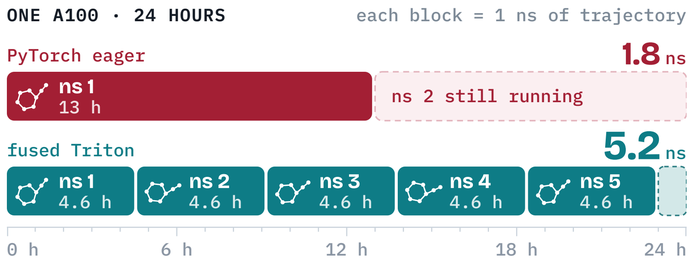
WALL TIME FOR 1 NS · 1M STEPS

PyTorch eager	13.0 h
4.6 h	fused Triton kernels

ML-driven MD does not need a new architecture to get faster. It needs fewer kernel launches.

02 WHAT IT BUYS

One GPU, one day: 1.8 ns becomes 5.2 ns



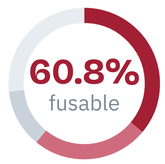
1–10 ns/day is the MLFF screening target. At **5.2 ns/day**, one A100 lands inside it.

03 THE BOTTLENECK

It's not the math. It's the memory traffic.

Neighbour lists shift every step, so gather and scatter hit memory in random order. Two memory-bound op groups hold **60.8%** of GPU time.

A100 SXM4 80GB · 4,096 atoms · 5.0 Å cutoff · 131K edges



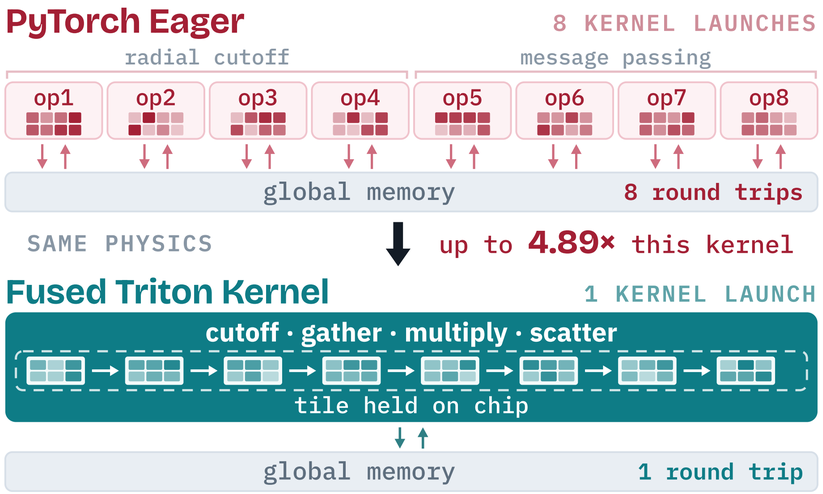
index message aggregation	36.0%
element-wise tensor algebra	24.8%
GEMM, already cuBLAS	12.4%
everything else	26.8%

torch.compile did not help. Graph-structured index ops defeat the compiler, so eager mode is the honest baseline.

04 THE FIX

Eight kernel launches become one.

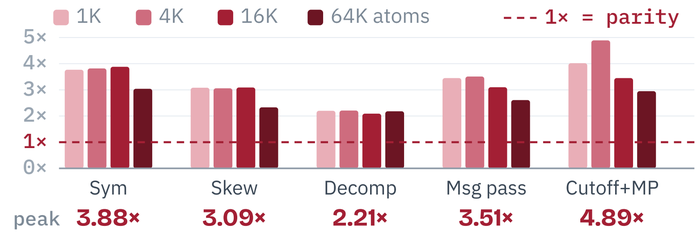
PyTorch pays a global-memory round trip for every small kernel. One fused Triton kernel keeps the tile on chip.



8 → 1	5	−88%
kernel launches, cutoff + message passing	fused Triton kernel families	launches per fused block, at best

05 IT SCALES

Faster at every size, 1K to 64K atoms

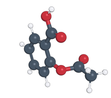


Geometric mean 3.21 / 3.38 / 3.05 / 2.60× across sizes, **3.14× overall**. Every fused kernel still beats PyTorch at 64K atoms.

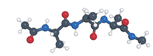
06 SAME PHYSICS

Faster, not approximate.

- ✓ Energies within 10^{-6} **kcal/mol**
- ✓ Forces within 10^{-5} **kcal/mol/Å**
- ✓ Symmetry, tracelessness within 10^{-7}
- ✓ 1,000 MD steps, **torch.allclose** at 1e-4



MD17 aspirin
21 atoms
2.85×



MD22 Ac-Ala3
42 atoms
2.96×

Reorganised computation, not approximated chemistry.

07 HONEST LIMITS

Where PyTorch still wins

1. High-contention 4D scatters: Triton up to **2.3× slower** than hand-tuned atomics.
2. Isolated ops below **~512 atoms**, where launch overhead dominates.
3. So we ship a **hybrid dispatch**: only fusion-favourable ops go to Triton.
4. End to end validated on **21–42 atom** molecules; protein scale is op-level only.