Office of Advanced Cyberinfrastructure
CSSI Elements: Award #2311372

Streaming Trajectories with MDAnalysis

Contributors

ASU
Lawson Woods
Amruthesh Thirumalaiswamy
Heekun Cho
Oliver Beckstein
Matthias Heyden

MDAnalysis
Hugo McDermott-Opeskin
Jennifer Clark
Yuxuan Zhuang

ASU School of Molecular Sciences
Arizona State University

Why Streaming of MD trajectories?

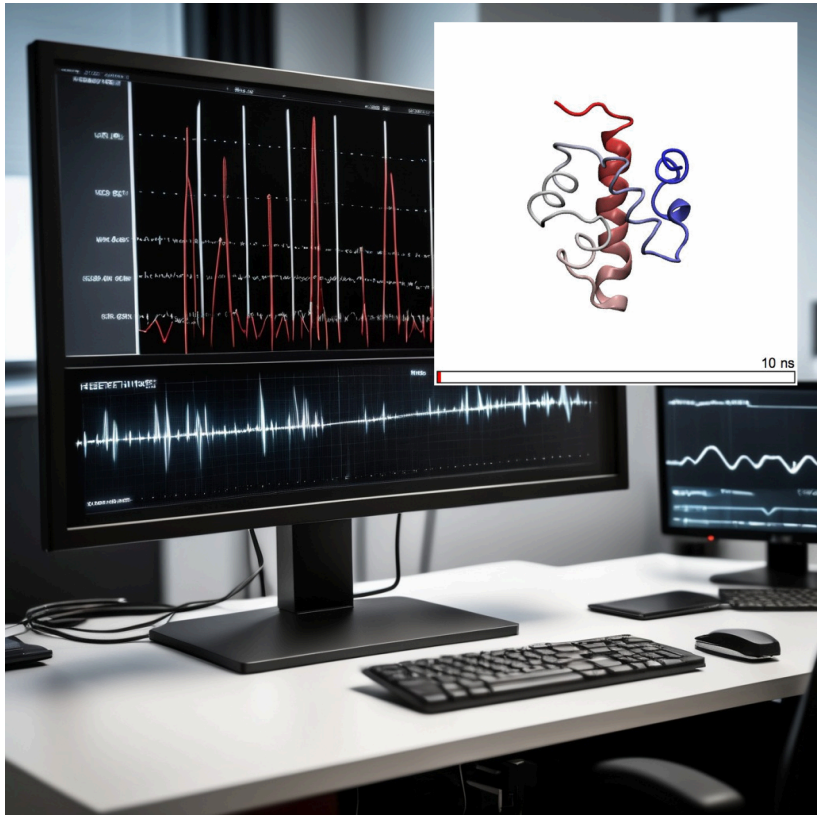
- **“On-the-fly” Analysis of Trajectories**

  → **Analyze fast** (sub-picosecond) **processes** w/o writing large trajectory output files
    - solvent dynamics (HBs, diffusion)
    - molecular vibrations
    - friction (force time correlations)

  → **Event Detection**
    - monitor arbitrary collective variables
    - weighted ensemble methods
    - water/ion flux through membrane

- **Control Output**

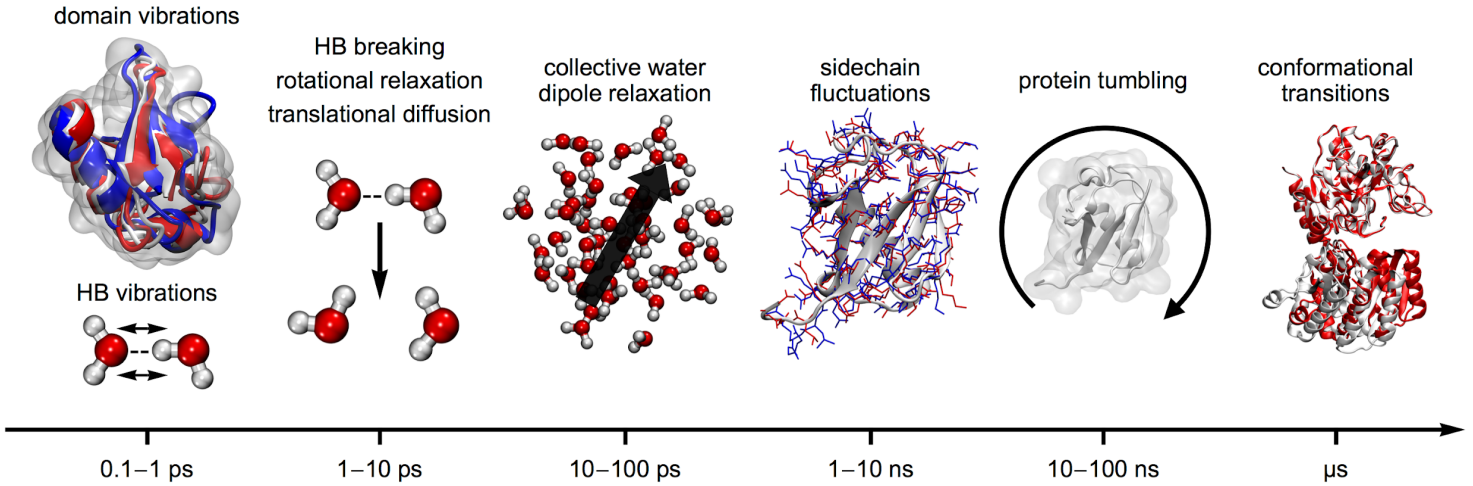
  → save disk space & post-processing
    - store only data you need (e.g., unwrapped protein coordinates)



Timescale Paradox in MD

MD simulations cover timescales over at least 9 orders of magnitude

→ **femtosecond** 10^{-15} s to **microseconds** (10^{-6} s)



Can we actually use this data?

System size: **10^5 atoms** Trajectory: 1 μs = **10^6 ps** Time resolution: 100 fs = **0.1 ps**

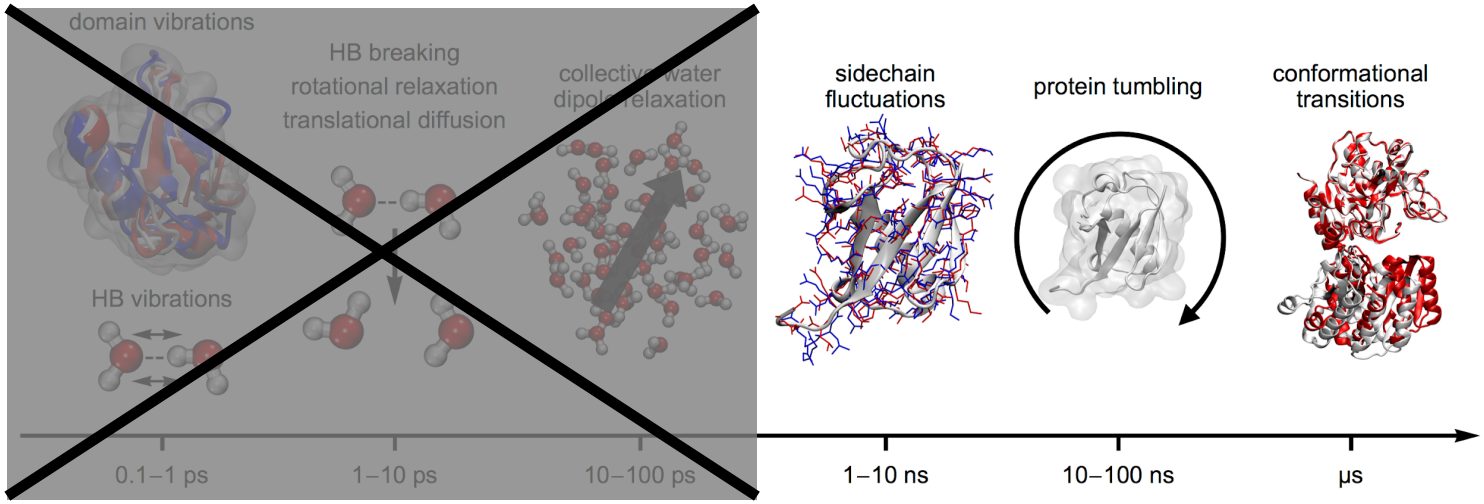
storing only coordinates = 12 bytes/atom × 10^5 atoms × 10^6 ps / 0.1 ps = **11 PetaBytes**

→ unfeasible to store this amount of data
→ 1000-fold reduction in time resolution: from 0.1 ps to 100 ps (99.9% of data ignored)

Timescale Paradox in MD

MD simulations cover timescales over at least 9 orders of magnitude

→ **femtosecond** 10^{-15} s to **microseconds** (10^{-6} s)



Can we actually use this data?

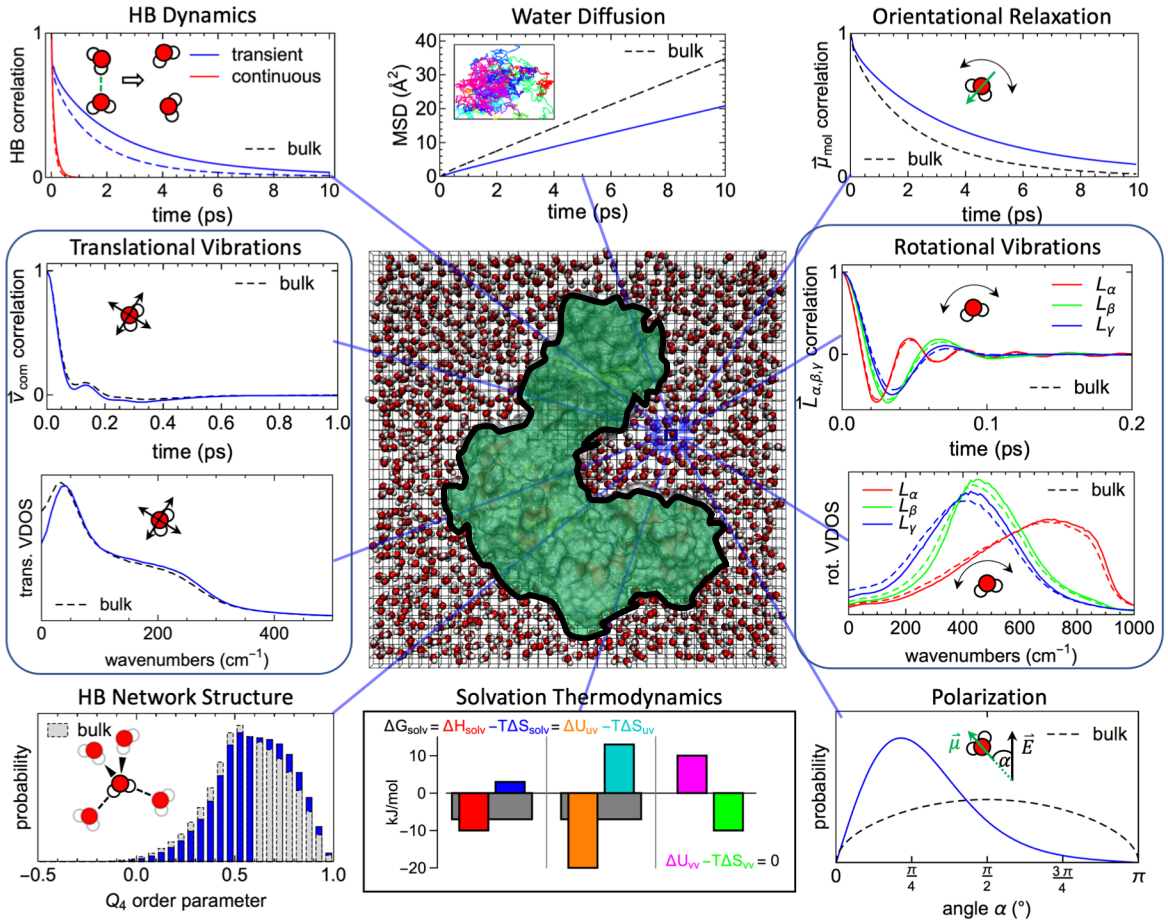
System size: **10^5 atoms** Trajectory: 1 μs = **10^6 ps** Time resolution: 100 fs = **0.1 ps**

storing only coordinates = 12 bytes/atom × 10^5 atoms × 10^6 ps / 0.1 ps = **11 PetaBytes**

→ unfeasible to store this amount of data
→ 1000-fold reduction in time resolution: from 0.1 ps to 100 ps (99.9% of data ignored)

What do we miss in such trajectories?

Example: **Solvent Dynamics**



Picosecond Dynamics in Hydration Water

- diffusion (mean squared displacements)
- HB correlation functions
- residence times
- orientational relaxation
- intermolecular vibrations (vibrational density of states)

→ thermodynamic information via 2PT

our specialty:
analyzing water dynamics with 3D resolution

Why Streaming of MD trajectories?

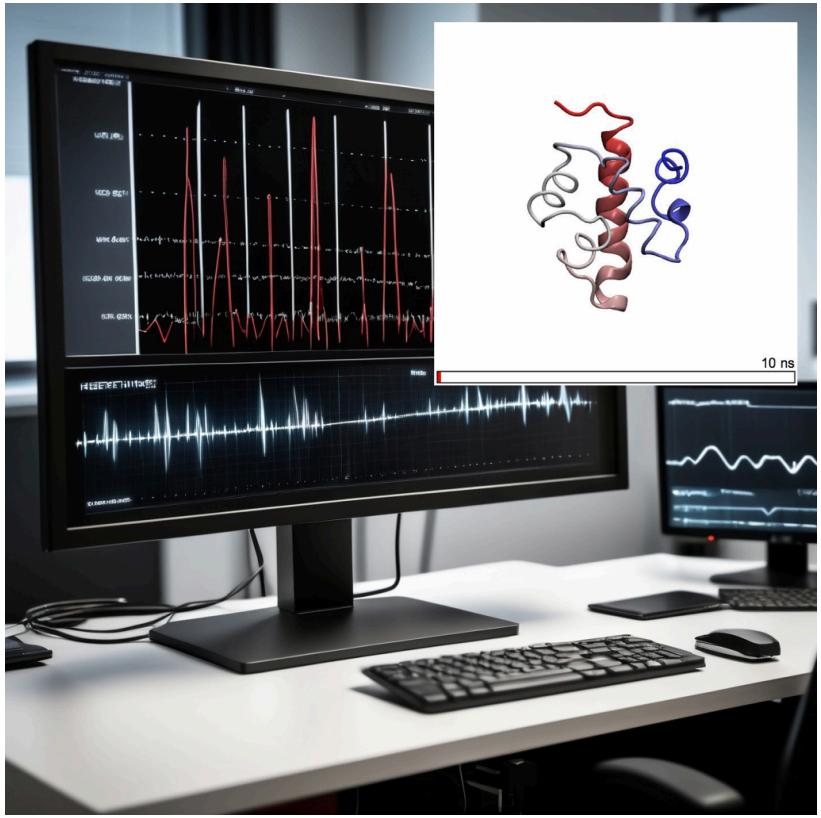
- **“On-the-fly” Analysis of Trajectories**

  → **Analyze fast** (sub-picosecond) **processes** w/o writing large trajectory output files
    - solvent dynamics (HBs, diffusion)
    - molecular vibrations
    - friction (force time correlations)

  → **Event Detection**
    - monitor arbitrary collective variables
    - weighted ensemble methods
    - water/ion flux through membrane

- **Control Output**

  → save disk space & post-processing
    - store only data you need (e.g., unwrapped protein coordinates)



Nomenclature

- **TCP/IP socket interface**

  → **Server**

  - MD Simulation software ("serves" trajectory data) ➔ producer
  - you'll need its address: localhost or IP-address

  → **Client** (here: your MDAnalysis script/notebook)

  - receives data from server ➔ consumer

  → **TCP/IP port number**

  - communication channel between **server** and **client** (→ set to same value for both)
  - 0-1023 = well-known ports (stay away, may require superuser privileges)
  - 1024-49151 = registered ports (free to use, possible conflicts)
  - 49152-65535 = dynamic, private or ephemeral ports (free to use)

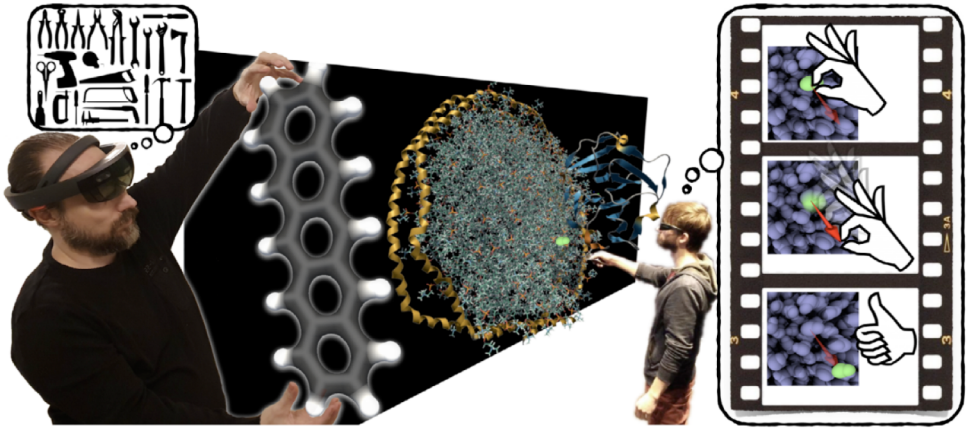
Standing on the Shoulders of Giants: Interactive Molecular Dynamics (IMD)

Original Implementation (**IMD version 2**):
John E. Stone, Justin Gullingsrud, Klaus Schulten, Paul Grayson
2001 ACM Symposium on Interactive 3D Graphics

Original: Interface NAMD & VMD

supported by GROMACS, LAMMPS, *etc.*

Main purpose:
Visualization & Interaction

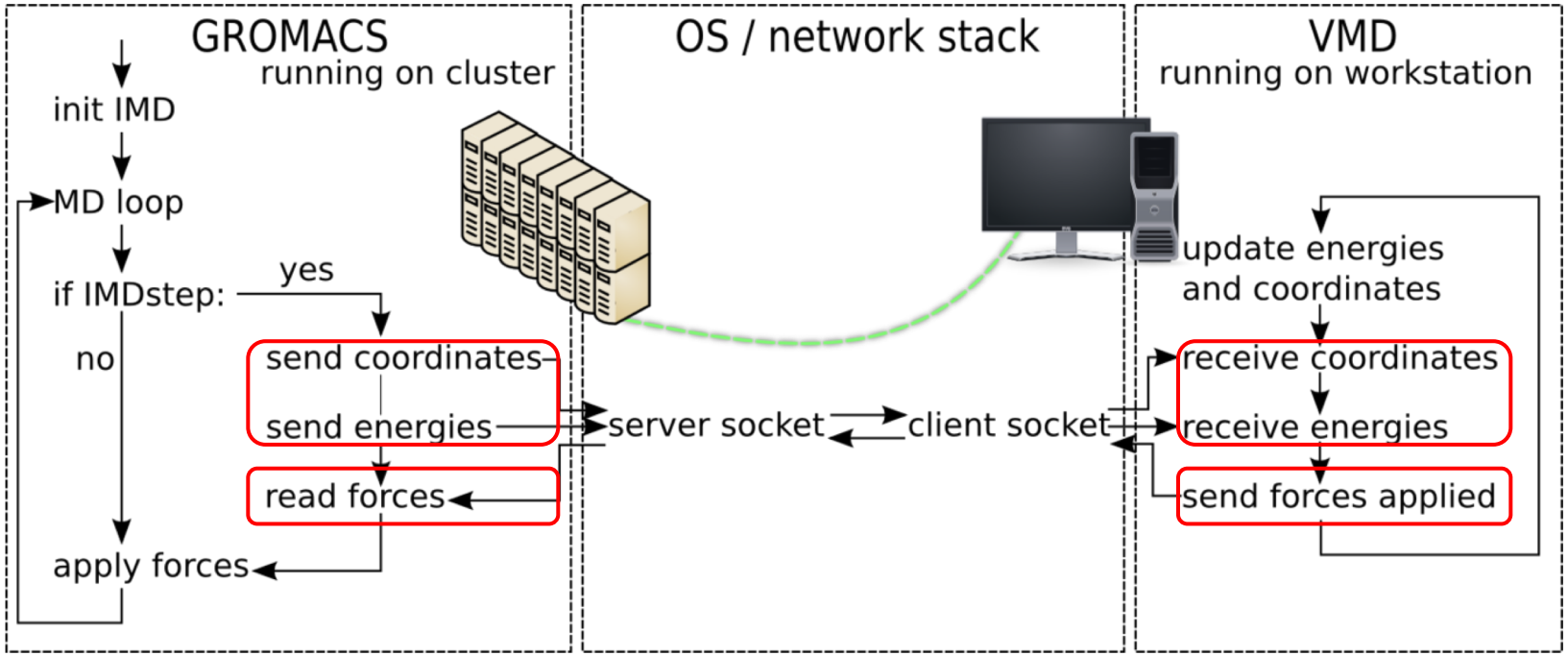


Application:
A. Lanrezac, N. Férey, M. Baaden, WIREs Comput. Mol Sci. 12, e1594 (2022).

IMD (version 2)

Functionality:

- communicate coordinates [optional: and energies] from MD simulation (server) to VMD (client) for visualization
- communicate interactive forces from VMD back to MD engine



(c) M. Hoefling & H. Grubmueller, Max Planck Institute for Multidisciplinary Sciences

Why IMD version 3?

IMD version 2

- only atom coordinates communicated to client → sufficient for visualization
- can set transmit frequency (e.g., communicate every 10 MD steps)

IMD version 3

For analysis of MD trajectories, we **need** (at least as an *option*):

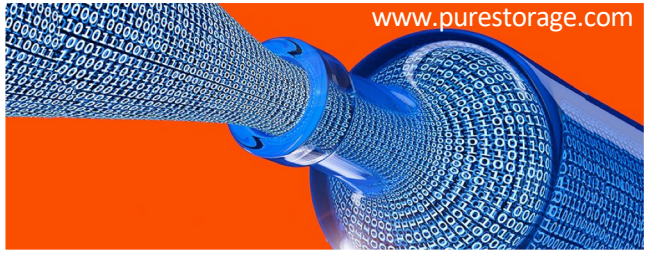
- simulation box dimensions (if not constant)

  → otherwise cannot compute distances in periodic space
- atomic velocities (e.g., for time correlation functions, molecular vibrations)
- atomic forces (e.g., to compute friction coefficients)

Nice to have:

- control wrapping/unwrapping of molecules in periodic system
- simulation time (avoids reconstruct time from integer steps)

Limits



Data Transfer Rate via TCP/IP (2018 iMac)

- on localhost:

  → >45 MBytes/s = 45 × 1,024 × 1,024 bytes/s = 47,185,920 bytes/s

  → bytes/atom = **12 (crds only)** to **36 (crds,velo,force)**

If we transmit **only coordinates**:

$$\frac{4{,}7185{,}920\ \text{bytes/s}}{\mathbf{12\ bytes/atom}} = 3.93 \times 10^6\ \text{atoms/s}$$

If we transmit **full atom info**:

$$\frac{4{,}7185{,}920\ \text{bytes/s}}{\mathbf{36\ bytes/atom}} = 1.31 \times 10^6\ \text{atoms/s}$$

with 10^5 atoms

~ 40 frames/s

~ 13 frames/s

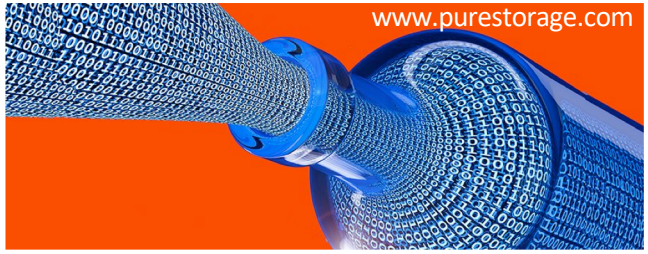
Assuming MD performance on GPU of **100 ns/day** with **2 fs** time steps

→ 578 frames/s

➔ adjust transmit frequency (time resolution in stream) or simulation will slow down

➔ stream resolution $\sim 30 - 100$ fs

Limits



Data Transfer Rate via TCP/IP (2018 iMac)

- on localhost:

→ >45 MBytes/s = 45 × 1,024 × 1,024 bytes/s = 47,185,920 bytes/s

This example is for **standard consumer hardware**!

Data buses on HPC hardware >100 MB/s

➔ stream resolution: **<10-40 fs**

Assuming MD performance on GPU of **100 ns/day** with **2 fs** time steps

→ 578 frames/s

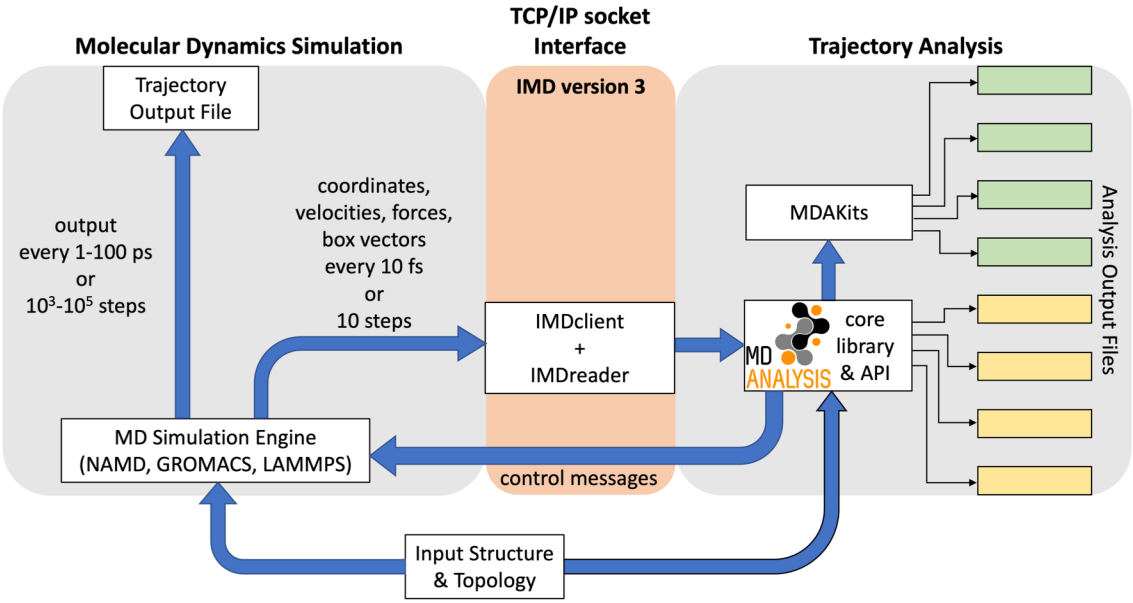
➔ adjust transmit frequency (time resolution in stream) or simulation will slow down

➔ stream resolution $\sim 30 - 100$ fs

IMD version 3 protocol
(working title)

https://imdclient.readthedocs.io/en/latest/protocol_v3.html

- new message types
- backwards compatible with IMD v2
- straightforward to expand
- implemented in: **GROMACS, NAMD3, LAMMPS** (forks of recent release)
- IMDclient & IMDreader class for **MDAnalysis**



IMD version 3 protocol

(working title)

https://imdclient.readthedocs.io/en/latest/protocol_v3.html

IMD version 2 *vs.* version 3 →

Setup in MD input file (here: GROMACS)

```
; IMD group
IMD-group                = System
IMD-version              = 3
IMD-nst                  = 10
IMD-time                 = yes
IMD-coords               = yes
IMD-vels                 = yes
IMD-forces               = yes
IMD-box                  = yes
IMD-unwrap               = yes
IMD-energies             = no
```

new in IMD version 3

Setup in MDAnalysis script

```
u = mda.Universe("topol.tpr", "imd://localhost:8888")
```

```
for ts in u.trajectory:
    DO ANYTHING YOU WANT HERE
```

Message Type	IMD version	receiver
handshake	2	server (MD)
go	2	server (MD)
disconnect	2	server (MD)
kill	2	server (MD)
md-communication	2	server (MD)
pause	2	server (MD)
transmission rate	2	server (MD)
I/O error	2	server (MD)
coordinates	2	client
sessioninfo	3	client
resume	3	server (MD)
no wait	3	server(MD)
time	3	client
velocities	3	client
forces	3	client
box	3	client