DISPATCH - aiming for exa-scale

Jon P. Ramsey
Åke Nordlund

Centre for Star and Planet Formation, and the Niels Bohr Institute, Københavns Universitet

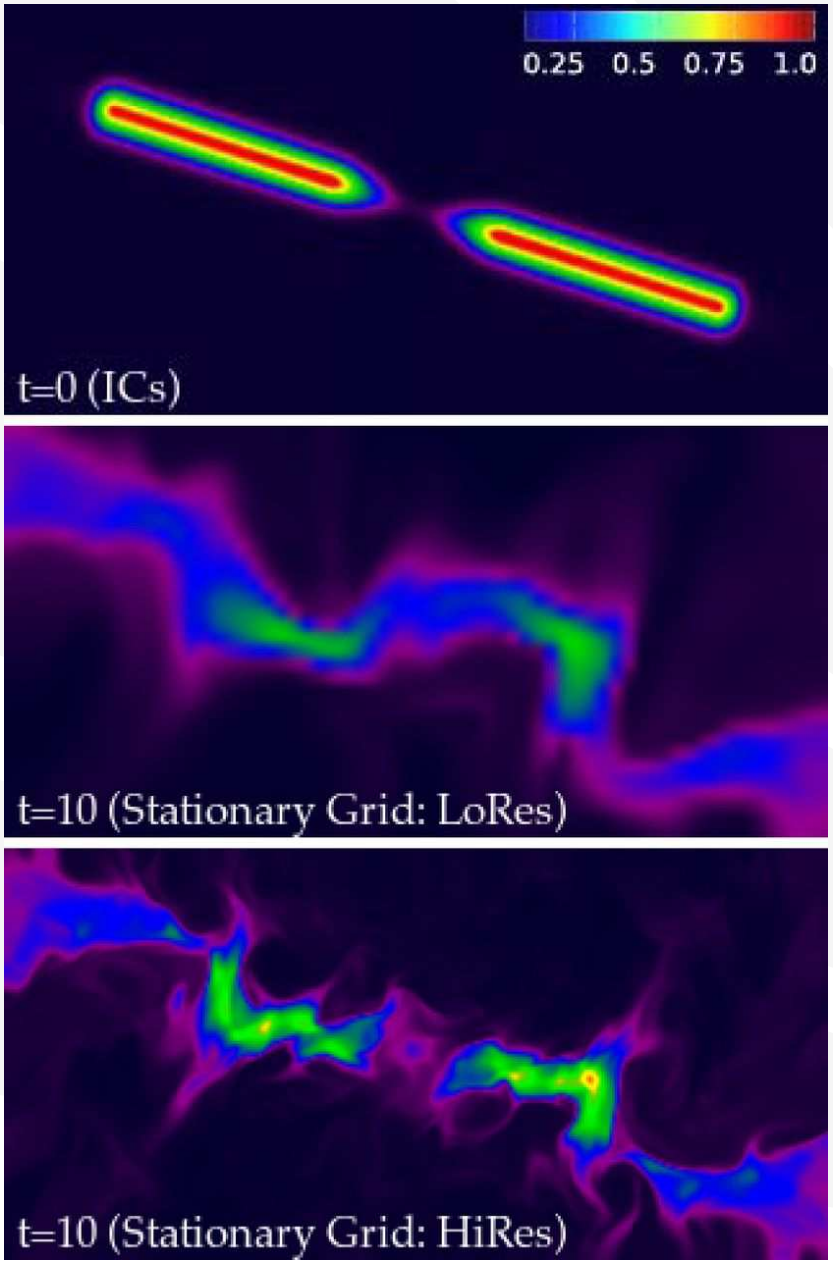
September 23, 2015

Motivation

- Eulerian (adaptive) mesh and Lagrangian smoothed particle (M)HD methods have been extremely useful tools for understanding astrophysical processes.
- However, potentially critical flaws in these methods have recently been highlighted in the literature (e.g., Springel 2010; Hopkins 2015).
- Alternative techniques are now available, but at increased complexity and computational cost (e.g., Mocz et al. 2015).
- There are scalability challenges with traditional techniques due to 'locked' time-stepping (i.e., all time-steps in a simulation are a power of 2 of the smallest time step).
- The zoom-in simulations (Nordlund et al. 2014) are pioneering and one-of-a-kind, but it is proving difficult to push the RAMSES AMR-MHD code any further in spatial contrast or temporal longevity; RAMSES also suffers from the algorithmic issues mentioned above.

Motivation

- Motivation
- A new approach
- DISPATCH
- Scaling
- Challenges
- Closing remarks



Credit: Hopkins (2015)

Motivation

- Eulerian (adaptive) mesh and Lagrangian smoothed particle (M)HD methods have been extremely useful tools for understanding astrophysical processes.
- However, potentially critical flaws in these methods have recently been highlighted in the literature (e.g., Springel 2010; Hopkins 2015).
- Alternative techniques are now available, but at increased complexity and computational cost (e.g., Mocz et al. 2015).
- There are scalability challenges with traditional techniques due to 'locked' time-stepping (i.e., all time-steps in a simulation are a power of 2 of the smallest time step).
- The zoom-in simulations (Nordlund et al. 2014) are pioneering and one-of-a-kind, but it is proving difficult to push the RAMSES AMR-MHD code any further in spatial contrast or temporal longevity; RAMSES also suffers from the algorithmic issues mentioned above.

YAMC! (but let's try something different)

A) Small, Cartesian patches that move with the local flow velocity.

- Permits a Galilean transformation within the patch.
- Good cache and vectorisation performance on modern CPUs.
- Communicating a patch to other MPI ranks takes negligible time.

B) Task-based scheduling.

- Each patch is a task.
- With (A), load balancing process is now simple and fast.

C) Asynchronous evolution of these patches in time and space.

- Exceptional time steps in a single region no longer affect the time step in the entire simulation.
- Coupled with (B), a slow-down in one task has a reduced effect on the performance of the entire simulation.

DISPATCH

- "Dispatchers are communications personnel responsible for receiving and transmitting pure and reliable messages, tracking vehicles and equipment, and recording other important information" (Source: Wikipedia article on Dispatchers).
- Or "DISconnected PATCHes".
- Recycle what we can for this new framework: use well-established, well-tested algorithms for MHD and particle integration.
- Specifically target exa-scale computing and therefore require exceptional parallel scalability.

  $\Rightarrow$ Built from the ground up to exploit hybrid MPI(v3)/OpenMP(v4).

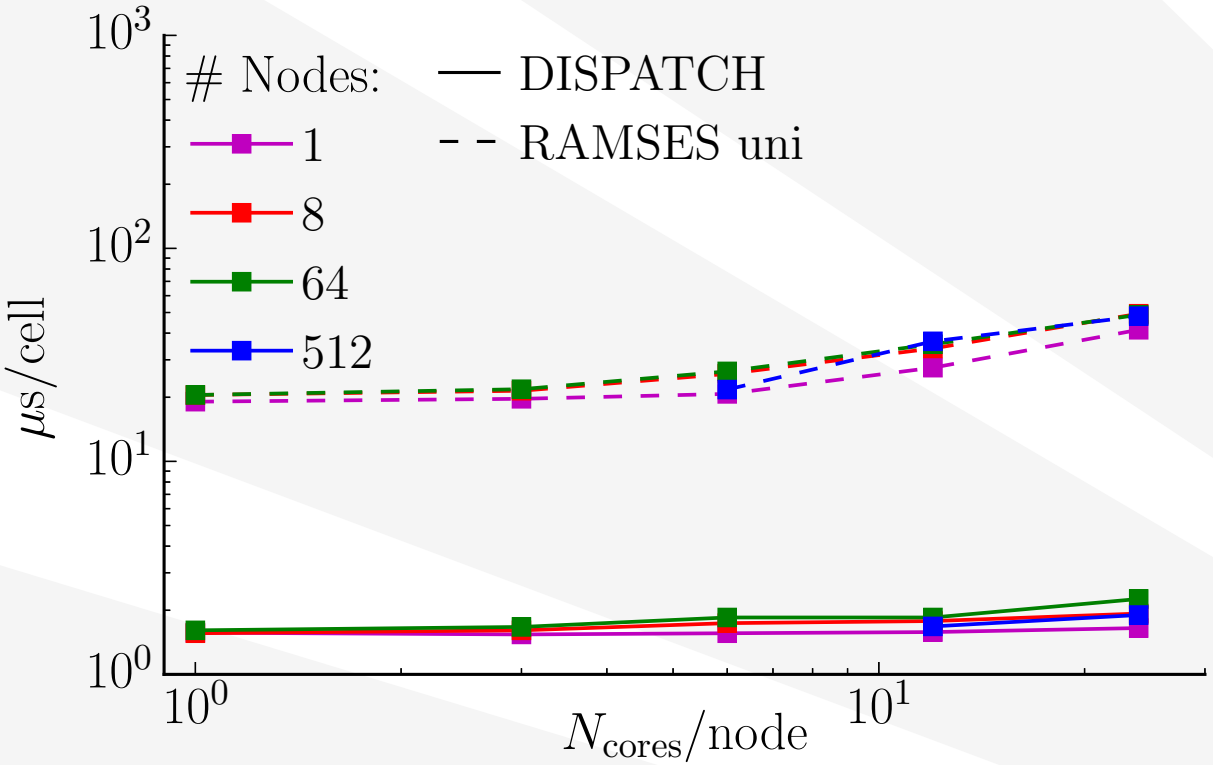
Current status

- Hybrid parallelism and efficient MPI communication already implemented.
- AZEuS and STAGGER MHD solvers already implemented.
- Simple but fast particle integrator which can handle the two-way interaction between gas and dust already implemented.
- We can already read RAMSES snapshots (in particular, from zoom-in runs of star formation).

  $\Rightarrow$ *SCIENCE!*: Gas-dust dynamics in protoplanetary disks and the early stages of planet formation.

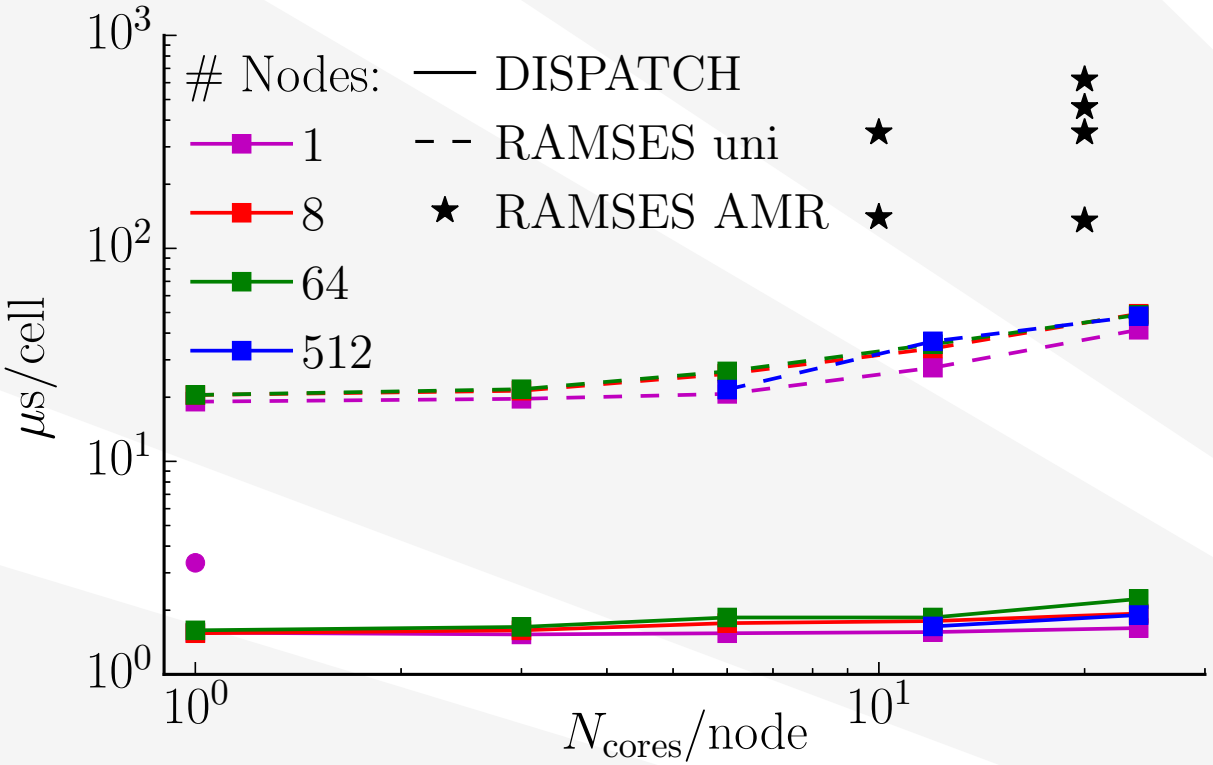
- Proof-of-concept movie: Orszag-Tang MHD vortex.

Weak Scaling



- Test problem is a uniform resolution, driven, isothermal turbulence.
- Each MPI rank is given 256^3 cells, split into 512 patches, up to a global resolution of 2048^3 cells.
- This test gives near ideal performance for RAMSES (and DISPATCH).
- Only a lack of resources is preventing us from testing to larger numbers of cores.

Weak Scaling



- Test problem is a uniform resolution, driven, isothermal turbulence.
- Each MPI rank is given 256^3 cells, split into 512 patches, up to a global resolution of 2048^3 cells.
- This test gives near ideal performance for RAMSES (and DISPATCH).
- Only a lack of resources is preventing us from testing to larger numbers of cores.

Current challenges on the Xeon Phi

1. Memory restrictions

2. DISPATCH performance

Current challenges on the Xeon Phi

1. Memory restrictions
   - The card is low-memory relative to a regular (CPU) node.
   - Running OOM can crash the card (at least on our cluster)!
2. DISPATCH performance

Current challenges on the Xeon Phi

1. Memory restrictions
   - The card is low-memory relative to a regular (CPU) node.
   - Running OOM can crash the card (at least on our cluster)!
2. DISPATCH performance
   - At the moment, 1 Xeon Phi card is roughly equal to 2 cores of a Xeon CPU (!).
   - The scaling test uses a constant 512 patches per MPI rank (i.e., 512 tasks); this is not enough work to keep 240 threads (or even 60 threads) happy.

Current challenges on the Xeon Phi

1. Memory restrictions
   - The card is low-memory relative to a regular (CPU) node.
   - Running OOM can crash the card (at least on our cluster)!
2. DISPATCH performance
   - At the moment, 1 Xeon Phi card is roughly equal to 2 cores of a Xeon CPU (!).
   - The scaling test uses a constant 512 patches per MPI rank (i.e., 512 tasks); this is not enough work to keep 240 threads (or even 60 threads) happy.
3. The compiler raises an error over alignments...
   - Only raises a warning on the CPU.
   - One must add a compiler option: `-warn noalignments`; `-align sequence`; `-align array64byte`

Closing remarks

- Even though it is still early in DISPATCH development, the code is already showing a lot of promise.
- The parallel scaling is excellent so far (for CPUs, anyway).
- We are targeting exa-scale computing and realistic *ab initio* planet formation simulations.
- It will eventually be (entirely) open-source and published online.