mini-ramses

Romain Teyssier

PRINCETON UNIVERSITY

mini-ramses on the CPU

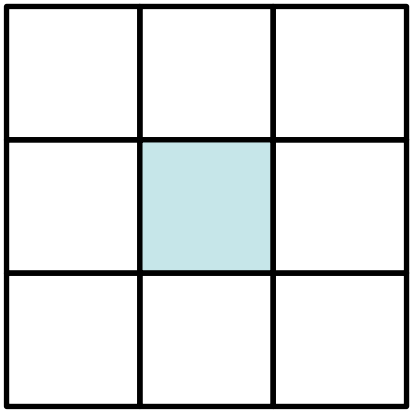
- On-the-fly clump finder and merger tree (James Sunseri)
- Fast Multipole Method (Jun-Young Lee)
- Sink particles and AGN feedback (DRAGON model, Nick Choustikov)
- Sink merging (William Groger)
- Optimization required:
  - Pre-fetch and vectorization (Romain Teyssier)
  - Particle load balancing (James Sunseri)
  - OpenMP (Tine Colman and San Han?)

mini-ramses on the GPU

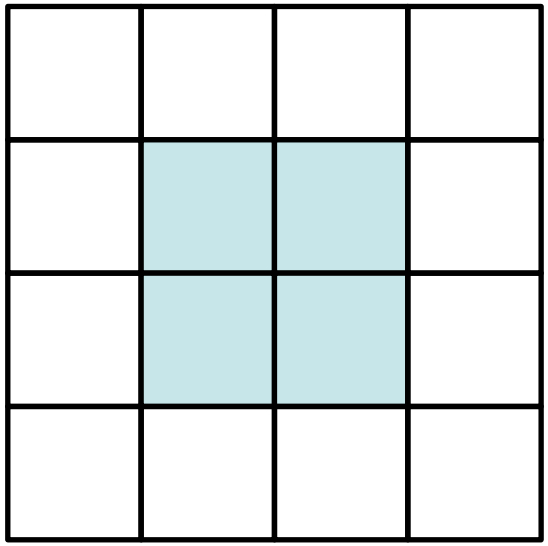
- CUDA Fortran with NVHPC Fortran compiler: `make COMPILER=NVHPC`
- Single GPU for now
- Data fully resident on the GPU
- Hash table on the GPU
  - `f1nva` hash function and fast linear probing (cuckoo hash)
- Fast sorting on the GPU with bucket sort + prefix sum
  - Use `shfl_up/shfl_down` CUDA intrinsics
- Octs are sorted in memory first per refinement level, then per Hilbert key
- Use large cache memory for ghost octs at coarse-fine boundaries
- Hydro kernels use `nbor/father` arrays for neighbors and parents connectivity

Different GPU hydro kernel configs (Bob Caddy)

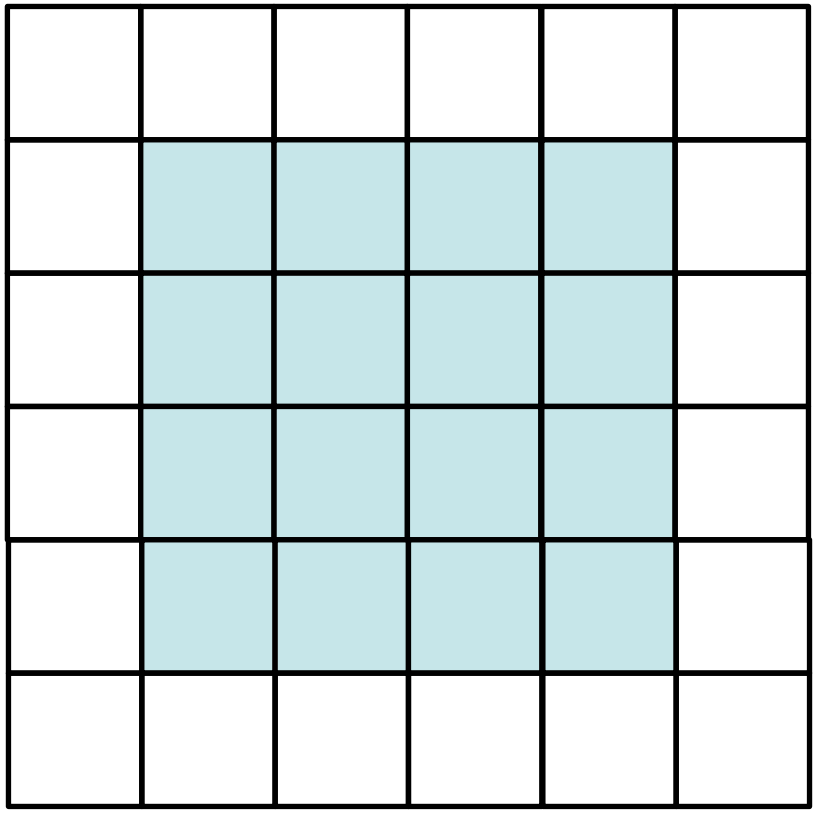
3x3x3



4x4x4



6x6x6

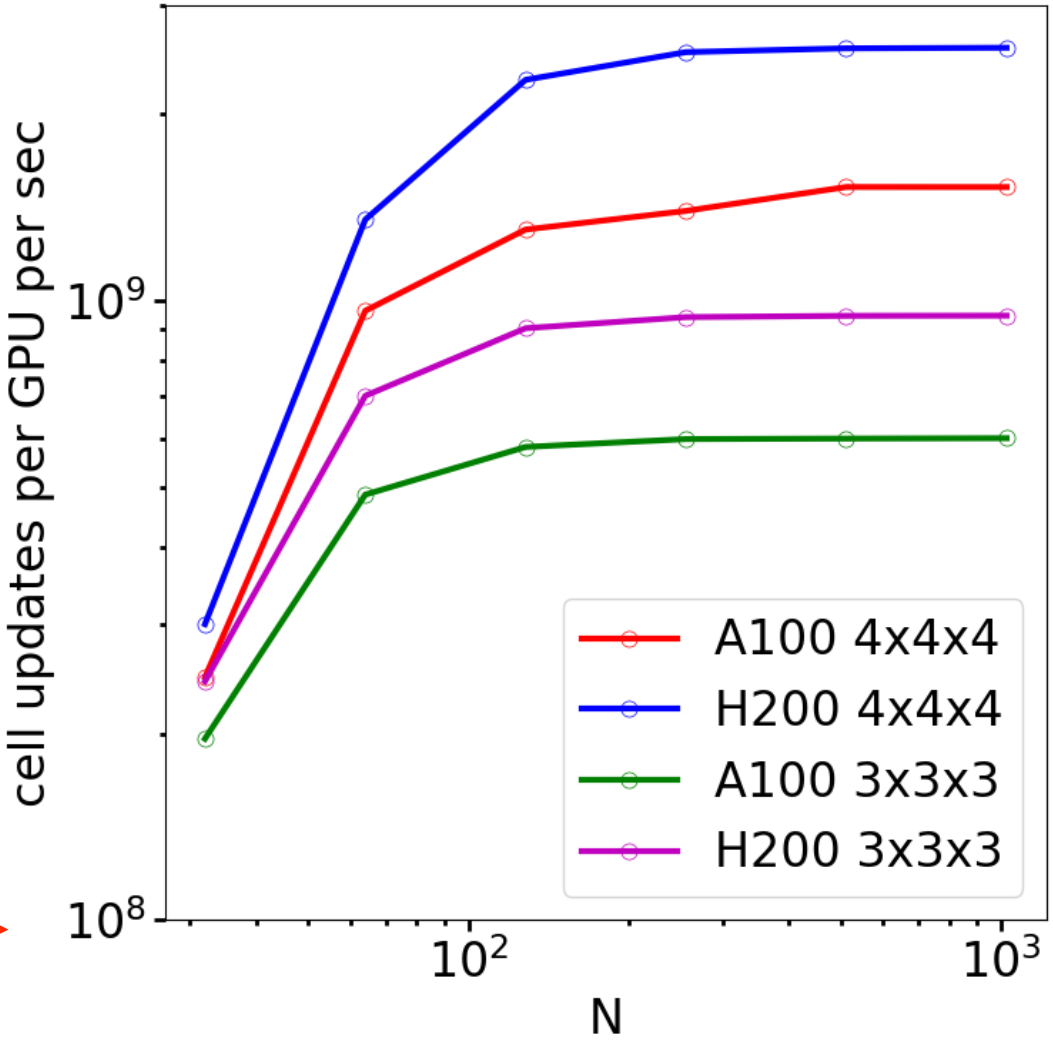


overhead=26
shared mem=12kB
block size=64

overhead=8
shared mem=41kB
block size=256

overhead=3
shared mem=210kB
block size=?

Roofline analysis



MPI 64 core

1e8 cell/node/sec

Hydro solver with LLF and N^3 cell unigrid with single precision

Sedov3d with AMR with 3x3x3 on A100

```
Main step=  5234 mcons=-2.56E-08 econs=-5.48E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5234 t= 4.99940E-02 dt= 1.444E-05 a= 1.000E+00 mem=32.6%
Time elapsed since last coarse step:     0.3332520
Used memory:  15.1 Gb
Writing output file to disk in output_00004/
Time elapsed writing to disk:     3.7615967
Mesh structure
Level  8 has     2097152 grids ( 2097152, 2097152, 2097152,)
Level  9 has     1602360 grids ( 1602360, 1602360, 1602360,)
Level 10 has     4116153 grids ( 4116153, 4116153, 4116153,)
Main step=  5235 mcons=-2.56E-08 econs=-5.50E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5235 t= 5.00085E-02 dt= 1.444E-05 a= 1.000E+00 mem=32.6%
Run completed
Total elapsed time:   697.1949959
Used memory:  15.1 Gb

    seconds          %    STEP
    116.085       16.8    refine
    149.324       21.7    cache
      9.514        1.4    time step
      5.163        0.7    hydro - set unew
    330.520       48.0    hydro - godunov
      5.059        0.7    hydro - set uold
     41.737        6.1    hydro - upload
     31.426        4.6    flag
    688.990      100.0    TOTAL
```

Sedov3d with AMR with 3x3x3 on H200

```
Main step=  5234 mcons=-2.53E-08 econs=-5.47E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5234 t= 4.99940E-02 dt= 1.444E-05 a= 1.000E+00 mem=32.6%
Time elapsed since last coarse step:     0.2007446
Used memory:  15.1 Gb
Writing output file to disk in output_00004/
Time elapsed writing to disk:     2.5839844
Mesh structure
Level  8 has     2097152 grids ( 2097152, 2097152, 2097152,)
Level  9 has     1602359 grids ( 1602359, 1602359, 1602359,)
Level 10 has     4116151 grids ( 4116151, 4116151, 4116151,)
Main step=  5235 mcons=-2.54E-08 econs=-5.49E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5235 t= 5.00085E-02 dt= 1.444E-05 a= 1.000E+00 mem=32.6%
Run completed
Total elapsed time:   422.0856476
Used memory:  15.1 Gb

    seconds          %    STEP
     66.125       15.9    refine
     76.263       18.3    cache
      6.143        1.5    time step
      2.510        0.6    hydro - set unew
    220.275       53.0    hydro - godunov
      2.520        0.6    hydro - set uold
     23.273        5.6    hydro - upload
     18.756        4.5    flag
    415.910      100.0    TOTAL
```

Sedov3d with AMR and 64 MPI cores

```
Main step=  5234 mcons=-6.66E-15 econs=-2.20E-13 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5234 t= 4.99940E-02 dt= 1.444E-05 a= 1.000E+00 mem= 4.1%
Time elapsed since last coarse step:     2.1025391
Used memory:   2.8 Gb
Writing output file to disk in output_00004/
Time elapsed writing to disk:     7.4599609
Mesh structure
Level  8 has     2097152 grids (   32768,   32768,   32768,)
Level  9 has     1602360 grids (   25036,   25037,   25036,)
Level 10 has     4116164 grids (   64315,   64316,   64315,)
Main step=  5235 mcons=-8.33E-15 econs=-2.17E-13 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5235 t= 5.00085E-02 dt= 1.444E-05 a= 1.000E+00 mem= 4.1%
Run completed
Total elapsed time:  6763.9206324
Used memory:   2.8 Gb

    seconds          %    STEP
    391.517        5.8    refine
    135.710        2.0    time step
     81.880        1.2    hydro - set unew
   3489.688       51.5    hydro - godunov
     81.292        1.2    hydro - set uold
     50.889        0.8    hydro - upload
   2550.352       37.6    flag
   6782.292      100.0    TOTAL
```

Sedov3d with AMR with 4x4x4 on A100

```
Main step=  5355 mcons=-2.57E-08 econs=-5.89E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5355 t= 4.99872E-02 dt= 1.438E-05 a= 1.000E+00 mem=36.9%
Time elapsed since last coarse step:     0.2908936
Used memory:  15.1 Gb
Writing output file to disk in output_00004/
Time elapsed writing to disk:     4.3482056
Mesh structure
Level  8 has     2097152 grids ( 2097152, 2097152, 2097152,)
Level  9 has     1813176 grids ( 1813176, 1813176, 1813176,)
Level 10 has     4958240 grids ( 4958240, 4958240, 4958240,)
Main step=  5356 mcons=-2.57E-08 econs=-5.90E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5356 t= 5.00016E-02 dt= 1.438E-05 a= 1.000E+00 mem=37.0%
Run completed
Total elapsed time:   473.4867210
Used memory:  15.1 Gb

    seconds          %    STEP
    134.862       29.1    refine
    117.793       25.4    cache
     10.257        2.2    time step
      5.751        1.2    hydro - set unew
    136.343       29.4    hydro - godunov
      5.570        1.2    hydro - set uold
     23.106        5.0    hydro - upload
     30.297        6.5    flag
    464.088      100.0    TOTAL
```

Sedov3d with AMR with 4x4x4 on H200

```
Main step=  5355 mcons=-2.55E-08 econs=-5.91E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5355 t= 4.99872E-02 dt= 1.438E-05 a= 1.000E+00 mem=36.9%
Time elapsed since last coarse step:     0.1946411
Used memory:  15.1 Gb
Writing output file to disk in output_00004/
Time elapsed writing to disk:     2.9716492
Mesh structure
Level  8 has     2097152 grids ( 2097152, 2097152, 2097152,)
Level  9 has     1813176 grids ( 1813176, 1813176, 1813176,)
Level 10 has     4958240 grids ( 4958240, 4958240, 4958240,)
Main step=  5356 mcons=-2.55E-08 econs=-5.92E-08 epot= 0.00E+00 ekin= 1.25E-01 eint= 9.79E-02
Fine step=   5356 t= 5.00016E-02 dt= 1.438E-05 a= 1.000E+00 mem=37.0%
Run completed
Total elapsed time:   295.1758232
Used memory:  15.1 Gb

    seconds          %    STEP
     77.762       27.0    refine
     78.158       27.1    cache
      6.627        2.3    time step
      2.788        1.0    hydro - set unew
     83.071       28.8    hydro - godunov
      2.792        1.0    hydro - set uold
     18.737        6.5    hydro - upload
     18.450        6.4    flag
    288.439      100.0    TOTAL
```

GPU mini-ramses: a simple example of kernel

```
recursive subroutine r_set_unew(pst,ilevel,input_size)
  use mdl_module
  use ramses_commons, only: pst_t
  use mdl_parameters
  implicit none
  type(pst_t)::pst
  integer,VALUE::input_size
  integer::ilevel

  integer::rID

  if(pst%nLower>0)then
     rID = mdl_send_request(pst%s%mdl,MDL_SET_UNEW,pst%iUpper+1,input_size,0,ilevel)
     call r_set_unew(pst%pLower,ilevel,input_size)
     call mdl_get_reply(pst%s%mdl,rID,0)
  else
#ifdef _CUDA
    call gpu_set_unew(pst%s, ilevel)
#else
     call set_unew(pst%s%r,pst%s%g,pst%s%m,ilevel)
#endif
  endif
```

GPU mini-ramses: a simple example of kernel

```
subroutine gpu_set_unew(sim, ilevel)
 use hydro_device, only: set_unew_kernel
  use ramses_commons, only: ramses_t
  type(ramses_t), intent(inout)::sim
  integer, intent(in)::ilevel

  integer(kind=4) :: head_idx, num_octs
  integer :: threadsX_per_block, threadsY_per_block, num_blocks

  head_idx = sim%m%head(ilevel)
  num_octs = sim%m%noct(ilevel)
  threadsX_per_block = 8
  threadsY_per_block = 16
  num_blocks = (num_octs + threadsY_per_block - 1) / threadsY_per_block
  call set_unew_kernel<<<num_blocks, dim3(threadsX_per_block,threadsY_per_block,1)>>>(uold, unew,
head_idx, num_octs)

end subroutine gpu_set_unew
```

GPU mini-ramses: a simple example of kernel

```
attributes(global) subroutine set_unew_kernel(uold, unew, head_idx, num_octs)
  real(kind=dp), device :: uold(1:twotondim,1:nvar,*)
  real(kind=dp), device :: unew(1:twotondim,1:nvar,*)
  integer(kind=4), intent(in), value :: head_idx
  integer(kind=4), intent(in), value :: num_octs

  integer :: block_idx ! The index of this block, corresponds to the oct to load
  integer :: cell_idx ! The index of the cell within the block
  integer :: oct_idx ! The index of the oct  within the block

  block_idx = blockIdx%x - 1

  ! Compute the oct idx
  oct_idx = block_idx * blockDim%y + threadIdx%y - 1
  if (oct_idx >= num_octs) return
  oct_idx = head_idx + oct_idx

  ! Compute the cell idx
  cell_idx = threadIdx%x

  unew(cell_idx, 1, oct_idx) = uold(cell_idx, 1, oct_idx)
  unew(cell_idx, 2, oct_idx) = uold(cell_idx, 2, oct_idx)
  unew(cell_idx, 3, oct_idx) = uold(cell_idx, 3, oct_idx)
  unew(cell_idx, 4, oct_idx) = uold(cell_idx, 4, oct_idx)
  unew(cell_idx, 5, oct_idx) = uold(cell_idx, 5, oct_idx)

end subroutine set_unew_kernel
```

GPU mini-ramses: work in progress

- Multigrid Poisson solver (in progress)
- PIC mass deposition and force interpolation (race conditions?)
- Cooling routine (warp divergence?)
  - Isothermal/barotropic EoS (done)
  - Galaxy/ISM cooling tables
  - Non-equilibrium chemistry
- Hydro solver
  - Passive scalars (shared memory?)
  - MHD (shared memory?)
  - Radiative transfer (group by group?)