

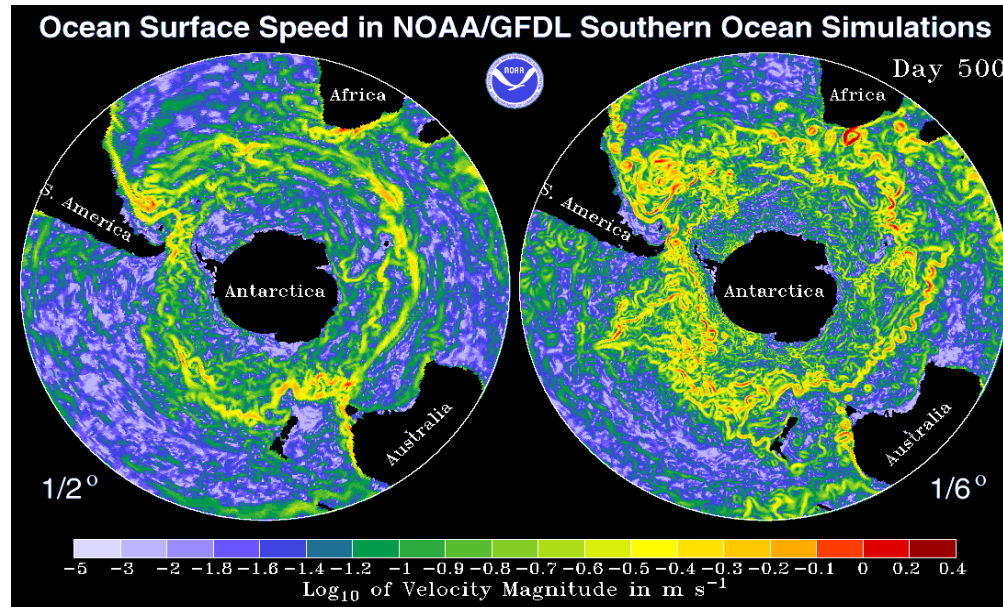
Discovering Performance Bottlenecks in Large Scale Parallel Applications

Prof. Jaswinder Pal Singh

Adrian Soviani

Princeton University, Dept. Of Computer Science

GFDL, MOM4



- Clock system design
- Measured performance, bottlenecks
- Computation model; hiding latency
- Code improvements, examples

GFDL MOM4 Project

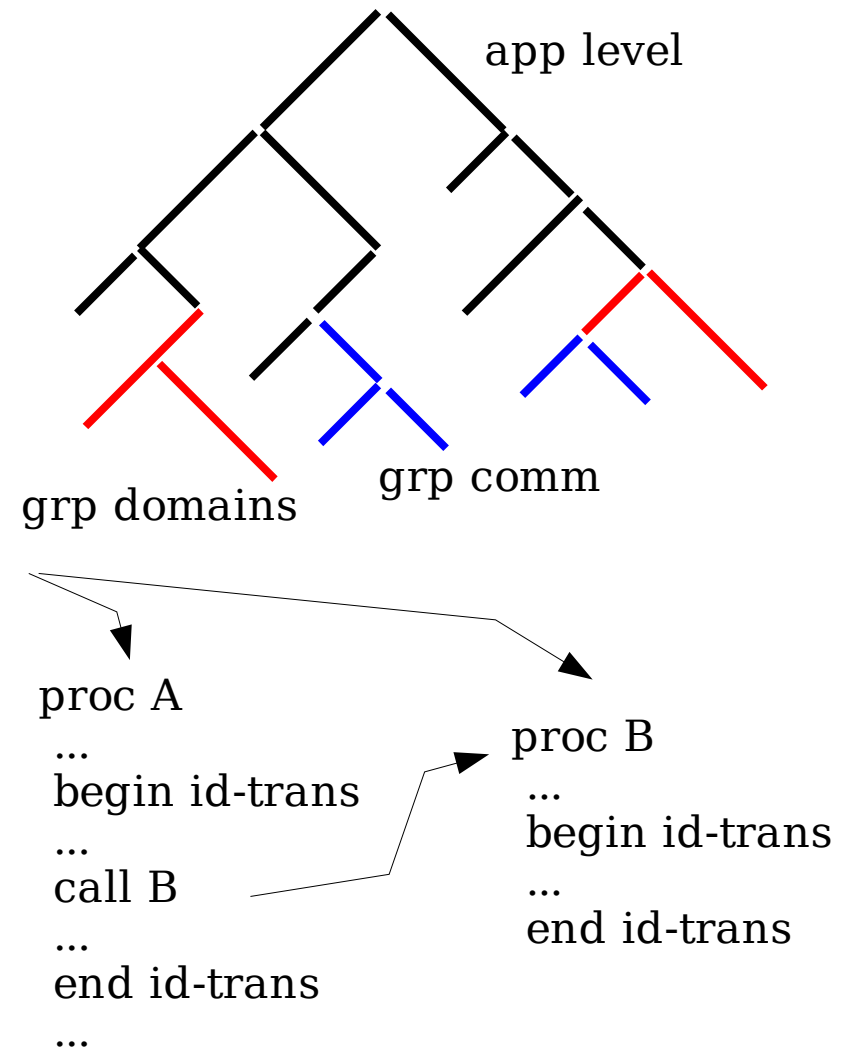
- algorithmically similar to PDE solvers
 - comm/comp growth rate
 - FFT: $1/\log(\text{size})$
 - ocean: $(n/\text{size})^{1/2}$ – highly scalable
 - different: global communication FFT, regrid, etc.
- working set size $200 \times 200 \times 50$
- scalability over 90-120 CPU an issue
- 50% spent in library layer
- 200,000 lines of code, 20 modules, multiple layers
- discover performance bottlenecks

Clocking system - main features

- trace multiple code sections under same label
- trace arbitrary code sections, not language constructs
- report call tree - abstract vs. lang hierarchy
- report custom data for each section (halo size, vector sizes, MPI msg size)
- group labels (physics, libs: [halo upd](#), [glb field](#), [MPI](#), [I/O](#))
 - turn on / off groups
 - split tree across group boundaries (explained later)
- compute unbalance across all CPU
- MOM4: about 3 mil calls / 100 sec (MPI on)
- 1-1.5% overhead; 0.2 us each caliper call

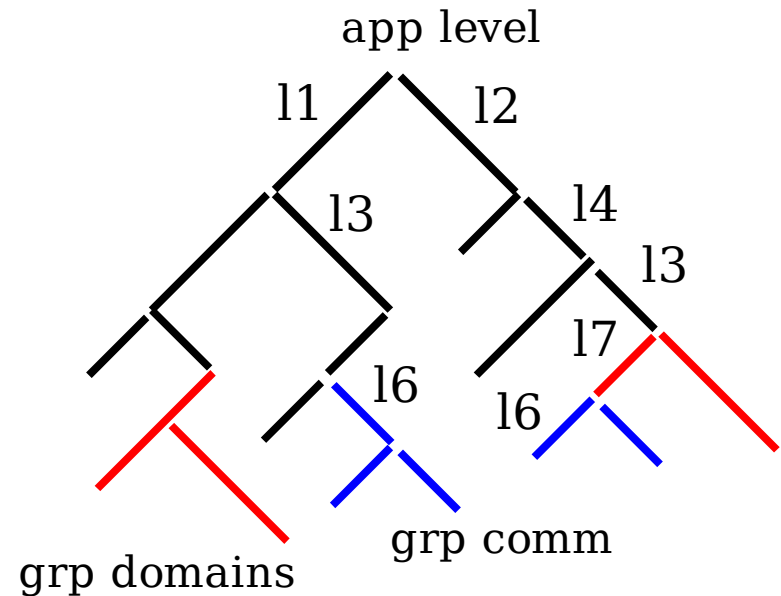
Clocking System

- caliper points
 - create group (grp)
 - create label (id, grp)
 - begin (id)
 - end (id, [aux])
- any section, cross proc boundaries
- 1 label, multiple sections
- reentrant
- create labels to generate abstract call tree (user choice)



Clocking System

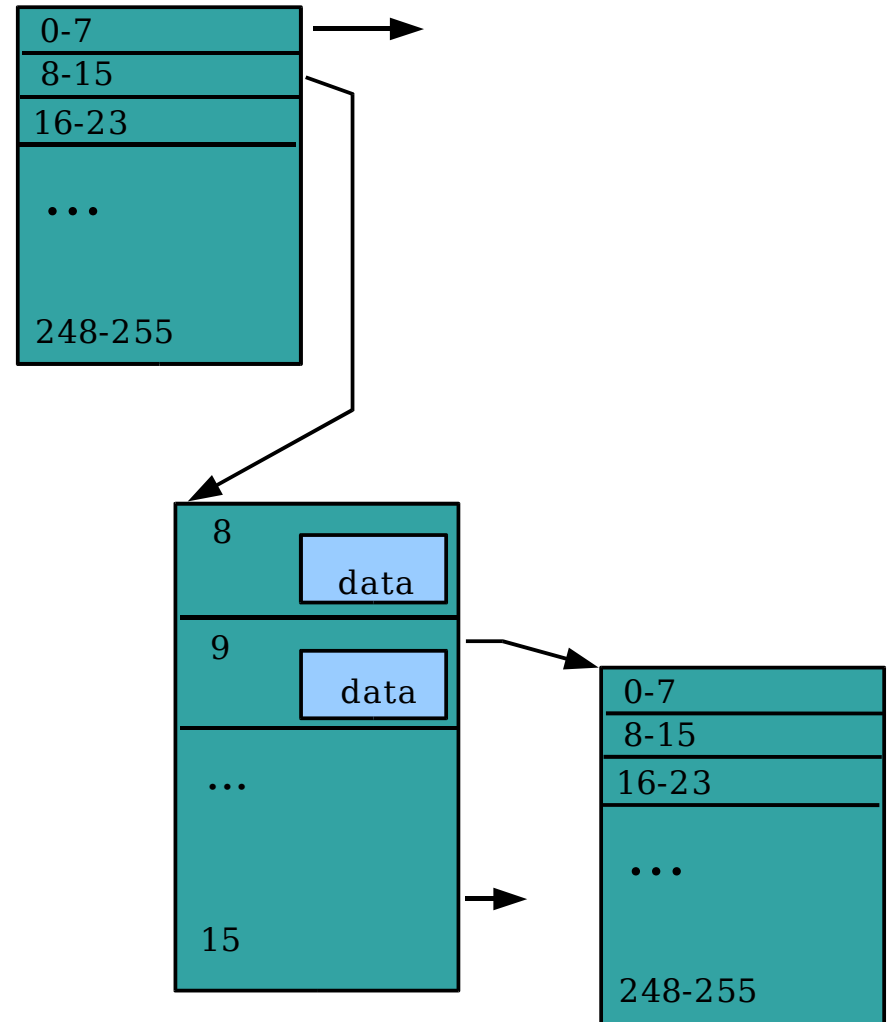
- record stats for any sequence la,lb,lc,...
 - 11,13,16
 - 12,14,13,17,16
 - halo update -> transmit
 - global field -> transmit



- for each code section sequence, each CPU
 - sum up time
 - auxiliary data
- avg, min, max, subtotals, etc – postprocessor
- custom analysis possible

Internal tree

- 2 level indexes
- 64,16 recs, 1024 clks
- m leafs = # possible call sequences
- d depth
- tree size $< 2md$
- caliper point: move one level up/down, or ignore end: update sums
- fast
- mem const (runlength)



Sample code

coupler_main.f90:

```
call mpp_init()
...
call mpp_clock_enable (.TRUE., 'MPP')
call mpp_clock_enable (.FALSE., 'MPP_DBG')
call mpp_clock_enable (.TRUE., 'MPI')
call mpp_clock_enable (.TRUE., 'DOMAINS')
call mpp_clock_enable (.TRUE., 'MPP_IO')
call mpp_clock_enable (.TRUE., 'DIAG')
...
do nc = 1, num_cpld_calls
  if (nc > 2) call mpp_clock_begin(profClock)
```

mpp.F90:

```
grp_mpp = mpp_clock_group ('MPP')
clk_mpp = mpp_clock_id ('mpp',grp=grp_mpp)
clk_transmit = mpp_clock_id ('transmit',grp=grp_mpp)
clk_reduce = mpp_clock_id ('reduce',grp=grp_mpp)
clk_broadcast = mpp_clock_id ('broadcast',grp=grp_mpp)
clk_sync = mpp_clock_id ('sync',grp=grp_mpp)

grp_mpp_dbg = mpp_clock_group ('MPP_DBG')

grp_mpi = mpp_clock_group ('MPI')
clk_mpi_send = mpp_clock_id ('mpi_send',val=.TRUE.,grp=grp_mpi)
clk_mpi_recv = mpp_clock_id ('mpi_recv',val=.TRUE.,grp=grp_mpi)
clk_mpi_broadcast = mpp_clock_id ('mpi_broadcast',val=.TRUE.,grp=grp_mpi)
clk_mpi_reduce = mpp_clock_id ('mpi_reduce',grp=grp_mpi)
clk_mpi_wait = mpp_clock_id ('mpi_wait',grp=grp_mpi)
clk_mpi_wait_self = mpp_clock_id ('mpi_wait_self',grp=grp_mpi)
clk_mpi_barrier = mpp_clock_id ('mpi_barrier',grp=grp_mpi)
```

mpp_transmit.h:

```
call mpp_clock_begin (clk_mpi_send)
call MPI_ISEND( put_data, put_len, MPI_TYPE_, to_pe, tag, MPI_COMM_WORLD, request(to_pe), error )
call mpp_clock_end_val (clk_mpi_send, put_len*(1.0*MPP_TYPE_BYTELEN_))
```

MOM4 clocks, 150 CPU

```
100 PROFILE 102.59 102.59 102.59 - 6 150
  0 sync 0.06 0.03 0.12 - 30 150
    0 mpi_barrier 0.06 0.03 0.12 - 30 150
    0 100_Total sync 0.06 0.00 0.06
  4 Ice 3.97 3.78 4.22 - 84 150
    1 domains 0.94 0.83 0.99 - 36 150
      1 transmit 0.86 0.76 0.91 - 10728 150
        0 mpi_send 0.17 0.12 0.36 - 5364 150 3840
        1 mpi_rcv 0.67 0.51 0.76 - 5364 150 3840
        1 98_Total transmit 0.84 0.02 0.86
      0 mpi_wait_self 0.01 0.01 0.01 - 5364 150
      1 92_Total domains 0.86 0.07 0.94
    2 dm_update 1.93 1.59 2.33 - 14298 150
      1 transmit 1.07 0.73 1.49 - 47114 150
        0 mpi_send 0.30 0.25 0.34 - 23557 150 807.494
        1 mpi_rcv 0.72 0.37 1.15 - 23557 150 807.494
        0 mpi_wait 0.00 0.00 0.00 - 9 18
        1 95_Total transmit 1.02 0.06 1.07
      0 mpi_wait_self 0.05 0.04 0.06 - 23617 150
      1 58_Total dm_update 1.12 0.80 1.93
    0 Send data 0.04 0.04 0.05 - 294 150
    3 73_Total Ice 2.91 1.06 3.97
90 Ocean 92.34 92.33 92.35 - 72 150
  2 (Ocean Advection Velocity) 1.57 1.40 1.90 - 72 150
    0 dm_update 0.18 0.03 0.50 - 72 150
      0 transmit 0.16 0.02 0.49 - 1103 150
        0 mpi_send 0.01 0.01 0.01 - 551 150 93.3681
        0 mpi_rcv 0.15 0.01 0.48 - 551 150 93.3681
        0 99_Total transmit 0.16 0.00 0.16
      0 mpi_wait_self 0.00 0.00 0.00 - 551 150
      0 94_Total dm_update 0.16 0.01 0.18
    1 Send data 0.52 0.43 0.63 - 144 150
    1 45_Total (Ocean Advection Velocity) 0.70 0.87 1.57
  2 (Ocean Density) 2.05 1.85 2.28 - 72 150
    1 Send data 0.98 0.83 1.20 - 288 150
    1 48_Total (Ocean Density) 0.98 1.07 2.05
  0 (Ocean rho-tilde) 0.08 0.06 0.11 - 72 150
  1 (Ocean u-rho) 0.55 0.39 0.61 - 72 150
  6 (Ocean Vertical Mixing Coeff) 5.73 5.19 6.25 - 72 150
    1 dm_update 0.82 0.27 1.84 - 144 150
      1 transmit 0.73 0.17 1.74 - 2206 150
        0 mpi_send 0.07 0.04 0.08 - 1103 150 9336.81
        1 mpi_rcv 0.66 0.09 1.68 - 1103 150 9336.81
        1 100_Total transmit 0.73 0.00 0.73
      0 mpi_wait_self 0.00 0.00 0.00 - 1103 150
      1 89_Total dm_update 0.73 0.09 0.82
    0 Send data 0.51 0.43 0.62 - 216 150
    1 23_Total (Ocean Vertical Mixing Coeff) 1.33 4.39 5.73
  21 (Ocean Neutral Physics) 21.94 21.52 22.45 - 72 150
    8 dm_update 7.95 5.97 8.96 - 10872 150
      6 transmit 5.88 3.82 7.00 - 335327 150
        2 mpi_send 1.54 0.86 1.93 - 167663 150 108.217
        4 mpi_rcv 3.94 1.72 5.83 - 167663 150 108.217
        0 mpi_wait 0.02 0.02 0.02 - 10800 1
        5 93_Total transmit 5.49 0.39 5.88
      0 mpi_wait_self 0.37 0.22 0.51 - 167591 150
      6 79_Total dm_update 6.25 1.70 7.95
    0 Send data 0.28 0.24 0.36 - 432 150
    0 (Ocean neutral density derivs) 0.39 0.28 0.44 - 72 150
    1 (Ocean neutral slopes ) 0.60 0.47 0.68 - 72 150
    2 (Ocean neutral fz-terms) 2.16 2.05 2.73 - 72 150
    3 (Ocean neutral fx-flux) 3.05 2.85 3.60 - 3600 150
    3 (Ocean neutral fy-flux) 2.97 2.82 3.52 - 3600 150
    1 (Ocean neutral fz-flux) 0.96 0.88 1.04 - 3600 150
    0 (Ocean neutral eady rate) 0.01 0.00 0.01 - 72 150
    0 (Ocean neutral baroclinicity) 0.02 0.01 0.02 - 72 150
      0 Send data 0.01 0.01 0.01 - 72 150
      0 58_Total (Ocean neutral baroclinicity) 0.01 0.01 0.02
    0 (Ocean neutral rossby radius) 0.17 0.13 0.19 - 72 150
      0 dm_update 0.04 0.02 0.08 - 72 150
        0 transmit 0.03 0.01 0.07 - 1103 150
          0 mpi_send 0.01 0.00 0.01 - 551 150 93.3681
          0 mpi_rcv 0.02 0.00 0.06 - 551 150 93.3681
          0 95_Total transmit 0.03 0.00 0.03
        0 mpi_wait_self 0.00 0.00 0.00 - 551 150
        0 79_Total dm_update 0.03 0.01 0.04
      0 24_Total (Ocean neutral rossby radius) 0.04 0.13 0.17
    0 (Ocean neutral diffusivity) 0.13 0.10 0.16 - 72 150
      0 dm_update 0.09 0.06 0.11 - 72 150
        0 transmit 0.06 0.04 0.08 - 1103 150
          0 mpi_send 0.02 0.01 0.04 - 551 150 4668.41
          0 mpi_rcv 0.04 0.02 0.06 - 551 150 4668.41
          0 97_Total transmit 0.06 0.00 0.06
        0 mpi_wait_self 0.00 0.00 0.00 - 551 150
        0 73_Total dm_update 0.06 0.02 0.09
      0 Send data 0.00 0.00 0.01 - 72 150
      0 69_Total (Ocean neutral diffusivity) 0.09 0.04 0.13
    18 85_Total (Ocean Neutral Physics) 18.68 3.25 21.94
  1 (Ocean Shortwave) 0.70 0.48 0.84 - 72 150
    1 (Ocean Shortwave Penetration) 0.53 0.33 0.62 - 1440 150
    1 75_Total (Ocean Shortwave) 0.53 0.17 0.70
  0 (Ocean sponge) 0.00 0.00 0.00 - 72 150
  1 (Ocean xlandmix) 1.41 1.30 1.52 - 72 150
    1 dm_update 0.97 0.86 1.11 - 720 150
      1 transmit 0.83 0.72 0.96 - 10693 150
        0 mpi_send 0.19 0.09 0.33 - 5346 150 3794.81
        1 mpi_rcv 0.63 0.46 0.80 - 5346 150 3794.81
        1 98_Total transmit 0.81 0.02 0.83
      0 mpi_wait_self 0.01 0.01 0.03 - 5346 150
      1 87_Total dm_update 0.84 0.13 0.97
    1 69_Total (Ocean xlandmix) 0.97 0.44 1.41
```

MOM4 clocks, 150 CPU

```
0 (Ocean rivermix) 0.35 0.18 0.55 - 72 150
1 (Ocean overflow) 0.79 0.61 1.09 - 72 150
  0 dm_update 0.25 0.06 0.55 - 216 150
    0 transmit 0.22 0.03 0.52 - 3309 150
      0 mpi_send 0.02 0.01 0.02 - 1654 150 248.982
      0 mpi_recv 0.20 0.01 0.50 - 1654 150 248.982
      0 98 Total transmit 0.21 0.00 0.22
    0 mpi_wait_self 0.00 0.00 0.01 - 1654 150
    0 89 Total_dm_update 0.22 0.03 0.25
  0 Send data 0.02 0.02 0.02 - 288 150
  0 34 Total (Ocean overflow) 0.27 0.53 0.79
0 (Ocean sigma diffusion) 0.26 0.18 0.34 - 72 150
  0 dm_update 0.19 0.11 0.28 - 216 150
    0 transmit 0.14 0.07 0.23 - 6684 150
      0 mpi_send 0.03 0.02 0.04 - 3342 150 93.1667
      0 mpi_recv 0.10 0.03 0.19 - 3342 150 93.1667
      0 mpi_wait 0.00 0.00 0.00 - 216 1
    0 94 Total transmit 0.13 0.01 0.14
    0 mpi_wait_self 0.01 0.00 0.01 - 3340 150
    0 79 Total_dm_update 0.15 0.04 0.19
  0 Send data 0.01 0.01 0.01 - 144 150
  0 74 Total (Ocean sigma diffusion) 0.20 0.07 0.26
16 (Ocean tracer update) 16.52 16.11 16.81 - 72 150
  1 Send data 1.37 1.18 1.61 - 648 150
  5 (Ocean tracer_adv horz 4th) 4.64 4.50 4.72 - 72 150
    4 dm_update 4.22 4.04 4.32 - 10800 150
      3 transmit 2.89 2.65 3.24 - 221712 150
        1 mpi_send 1.11 0.57 1.41 - 110856 150 143.667
        1 mpi_recv 1.51 1.05 2.50 - 110856 150 143.667
        0 mpi_wait 0.01 0.01 0.01 - 3600 1
      3 91 Total transmit 2.63 0.26 2.89
        0 mpi_wait_self 0.24 0.15 0.31 - 110832 150
      3 74 Total_dm_update 3.13 1.09 4.22
    4 91 Total (Ocean tracer_adv horz 4th) 4.22 0.41 4.64
0 (Ocean tracer_adv vert 4th) 0.18 0.16 0.22 - 72 150
3 (Ocean tracer_adv horz quick) 3.05 2.97 3.13 - 72 150
  2 dm_update 2.28 2.17 2.41 - 3672 150
    2 transmit 1.56 1.42 1.69 - 112511 150
      1 mpi_send 0.58 0.31 0.71 - 56255 150 187.681
      1 mpi_recv 0.85 0.56 1.26 - 56255 150 187.681
      0 mpi_wait 0.01 0.01 0.01 - 3600 1
    1 92 Total transmit 1.43 0.13 1.56
      0 mpi_wait_self 0.12 0.07 0.15 - 56231 150
    2 74 Total_dm_update 1.68 0.59 2.28
  2 75 Total (Ocean tracer_adv horz quick) 2.28 0.77 3.05
0 (Ocean tracer_adv vert quick) 0.30 0.28 0.34 - 72 150
  4 (Ocean tracer_adv MDFL-sweby) 3.59 3.48 3.67 - 72 150
    2 dm_update 2.25 2.05 2.48 - 7200 150
      2 transmit 1.61 1.40 1.86 - 110304 150
        1 mpi_send 0.61 0.32 0.81 - 55152 150 194.674
        1 mpi_recv 0.87 0.47 1.43 - 55152 150 194.674
        1 92 Total transmit 1.48 0.13 1.61
      0 mpi_wait_self 0.09 0.06 0.13 - 55152 150
      2 76 Total_dm_update 1.70 0.54 2.25
    2 63 Total (Ocean tracer_adv MDFL-sweby) 2.25 1.35 3.59
13 79 Total (Ocean tracer update) 13.13 3.39 16.52
0 (Ocean surface flux) 0.25 0.15 0.39 - 72 150
  0 dm_update 0.20 0.09 0.34 - 288 150
    0 transmit 0.17 0.06 0.31 - 4440 150
      0 mpi_send 0.03 0.01 0.03 - 2220 150 116.676
      0 mpi_recv 0.13 0.02 0.28 - 2220 150 116.704
      0 mpi_wait 0.00 0.00 0.00 - 72 8
    0 97 Total transmit 0.16 0.01 0.17
      0 mpi_wait_self 0.00 0.00 0.01 - 2216 150
    0 87 Total_dm_update 0.17 0.03 0.20
  0 Send data 0.02 0.02 0.02 - 360 150
  0 86 Total (Ocean surface flux) 0.22 0.04 0.25
0 (Ocean bottom flux) 0.12 0.03 0.56 - 72 150
  0 dm_update 0.11 0.02 0.55 - 72 150
    0 transmit 0.10 0.01 0.54 - 1131 150
      0 mpi_send 0.01 0.00 0.01 - 565 150 184.821
      0 mpi_recv 0.09 0.00 0.53 - 565 150 184.93
      0 mpi_wait 0.00 0.00 0.00 - 72 8
    0 99 Total transmit 0.10 0.00 0.10
      0 mpi_wait_self 0.00 0.00 0.00 - 562 150
    0 94 Total_dm_update 0.10 0.01 0.11
  0 93 Total (Ocean bottom flux) 0.11 0.01 0.12
4 (Ocean restoring flux) 3.80 3.61 4.14 - 72 150
  3 domains 3.59 3.17 4.08 - 144 150
    3 transmit 3.22 2.81 3.63 - 42912 150
      1 mpi_send 0.66 0.60 1.48 - 21456 150 3840
      2 mpi_recv 2.50 1.97 2.71 - 21456 150 3840
      3 98 Total transmit 3.16 0.06 3.22
    0 mpi_wait_self 0.03 0.03 0.05 - 21456 150
    3 91 Total_domains 3.25 0.34 3.59
  0 dm_update 0.18 0.02 0.89 - 72 150
    0 transmit 0.17 0.01 0.88 - 1103 150
      0 mpi_send 0.01 0.00 0.02 - 551 150 93.3681
      0 mpi_recv 0.16 0.00 0.86 - 551 150 93.3681
    0 99 Total transmit 0.17 0.00 0.17
      0 mpi_wait_self 0.00 0.00 0.00 - 551 150
    0 96 Total_dm_update 0.17 0.01 0.18
  0 Send data 0.01 0.01 0.01 - 144 150
  4 99 Total (Ocean restoring flux) 3.78 0.02 3.80
0 (Ocean TPM sbc) 0.01 0.00 0.01 - 72 150
0 (Ocean TPM source) 0.12 0.10 0.15 - 72 150
0 (Ocean TPM bbc) 0.00 0.00 0.00 - 72 150
```

MOM4 clocks, 150 CPU

```
0 (Ocean TPM tracer) 0.00 0.00 0.00 - 72 150
6 (Ocean explicit acceleration) 6.02 5.67 6.59 - 72 150
  0 Send data 0.31 0.26 0.36 - 72 150
  4 (Ocean Smag lap friction) 3.66 3.27 4.23 - 72 150
    2 dm_update 1.76 1.27 2.45 - 3600 150
      1 transmit 1.36 0.86 2.03 - 55152 150
        0 mpi_send 0.37 0.20 0.45 - 27576 150 746.945
        1 mpi_rcv 0.92 0.39 1.58 - 27576 150 746.945
        1 95 Total transmit 1.29 0.07 1.36
        0 mpi_wait_self 0.05 0.03 0.06 - 27576 150
        1 80 Total dm_update 1.41 0.35 1.76
      2 48 Total (Ocean Smag lap friction) 1.76 1.90 3.66
    4 66 Total (Ocean explicit acceleration) 3.97 2.06 6.02
  1 (Ocean implicit friction) 1.04 0.98 1.10 - 72 150
  0 (Ocean implicit Coriolis) 0.19 0.17 0.20 - 72 150
  0 (Ocean surf height tendency) 0.01 0.01 0.01 - 72 150
  0 (Ocean freesurface drag) 0.02 0.02 0.03 - 72 150
  0 (Ocean freesurface forcing) 0.18 0.12 0.28 - 72 150
    0 dm_update 0.09 0.03 0.18 - 72 150
      0 transmit 0.07 0.01 0.17 - 1131 150
        0 mpi_send 0.01 0.01 0.01 - 565 150 184.821
        0 mpi_rcv 0.06 0.00 0.16 - 565 150 184.93
        0 mpi_wait 0.00 0.00 0.00 - 72 8
        0 98 Total transmit 0.07 0.00 0.07
        0 mpi_wait_self 0.00 0.00 0.00 - 562 150
        0 84 Total dm_update 0.07 0.01 0.09
      0 49 Total (Ocean freesurface forcing) 0.09 0.09 0.18
13 (Ocean freesurface update) 13.59 13.19 14.09 - 72 150
  3 domains 3.57 3.16 4.07 - 144 150
    3 transmit 3.29 2.91 3.76 - 42912 150
      1 mpi_send 0.68 0.61 1.46 - 21456 150 3840
      2 mpi_rcv 2.54 2.07 2.73 - 21456 150 3840
      3 98 Total transmit 3.22 0.07 3.29
      0 mpi_wait_self 0.03 0.03 0.05 - 21456 150
      3 93 Total domains 3.32 0.25 3.57
    8 dm_update 8.02 7.87 8.09 - 27144 150
      5 transmit 5.42 4.94 5.97 - 419446 150
        2 mpi_send 2.12 1.12 2.69 - 209723 150 124.215
        3 mpi_rcv 2.79 1.69 4.52 - 209723 150 124.252
        0 mpi_wait 0.02 0.02 0.03 - 9000 8
        5 91 Total transmit 4.93 0.49 5.42
        0 mpi_wait_self 0.47 0.27 0.61 - 209243 150
        6 73 Total dm_update 5.89 2.13 8.02
      0 Send data 0.03 0.03 0.04 - 432 150
    11 86 Total (Ocean freesurface update) 11.62 1.97 13.59
  1 (Ocean velocity update) 0.79 0.69 0.91 - 72 150
    0 Send data 0.49 0.41 0.58 - 144 150
    0 62 Total (Ocean velocity update) 0.49 0.30 0.79
  0 (Ocean polar filter) 0.00 0.00 0.00 - 72 150
  0 (Ocean energy analysis) 0.00 0.00 0.00 - 72 150
11 (Ocean numerical diagnostics) 11.22 11.05 11.51 - 72 150
  1 Send data 1.33 1.23 1.45 - 576 150
  1 12 Total (Ocean numerical diagnostics) 1.33 9.89 11.22
  1 (Ocean Robert time filter) 0.63 0.52 0.72 - 72 150
  0 (Ocean update cell thickness) 0.01 0.00 0.01 - 72 150
  1 (Ocean update halos) 1.27 0.79 1.98 - 72 150
    1 dm_update 1.26 0.79 1.98 - 576 150
      1 transmit 1.02 0.55 1.72 - 8881 150
        0 mpi_send 0.16 0.09 0.21 - 4440 150 5833.81
        1 mpi_rcv 0.85 0.32 1.53 - 4440 150 5835.19
        0 mpi_wait 0.00 0.00 0.00 - 144 8
        1 99 Total transmit 1.01 0.01 1.02
        0 mpi_wait_self 0.01 0.01 0.02 - 4433 150
        1 82 Total dm_update 1.03 0.23 1.26
      1 100 Total (Ocean update halos) 1.26 0.00 1.27
  0 (Ocean sum ocean surface) 0.01 0.01 0.01 - 72 150
  0 (Ocean average state) 0.00 0.00 0.00 - 6 150
  1 (Ocean tracer tmask limit) 0.76 0.65 0.84 - 72 150
    0 Send data 0.25 0.21 0.31 - 72 150
    0 32 Total (Ocean tracer tmask limit) 0.25 0.51 0.76
90 100 Total Ocean 91.99 0.35 92.34
6 Land-ice-atm coupler 6.13 5.86 6.33 - 228 150
  3 SFC boundary layer 3.07 2.97 3.58 - 72 150
    1 transmit 1.37 0.97 1.78 - 29102 150
      0 mpi_send 0.22 0.05 0.70 - 14551 150 1099.23
      1 mpi_rcv 1.12 0.55 1.60 - 14551 150 1099.23
      1 97 Total transmit 1.34 0.04 1.37
      0 mpi_wait_self 0.03 0.01 0.07 - 8656 150
      1 dm_update 0.54 0.17 1.10 - 576 150
        0 transmit 0.48 0.11 1.05 - 8094 150
          0 mpi_send 0.04 0.02 0.06 - 4047 150 42.6262
          0 mpi_rcv 0.43 0.06 1.01 - 4047 150 42.6262
          0 mpi_wait 0.00 0.00 0.01 - 1099 42
          0 99 Total transmit 0.47 0.01 0.48
          0 mpi_wait_self 0.01 0.00 0.03 - 9623 150
          0 91 Total dm_update 0.49 0.05 0.54
        0 mpp_io 0.41 0.11 0.64 - 15 150
      2 76 Total SFC boundary layer 2.35 0.73 3.07
  2 Flux DN from atm 2.08 1.84 2.32 - 72 150
    1 transmit 0.94 0.74 1.05 - 19156 150
      0 mpi_send 0.16 0.02 0.74 - 10189 141 1116.71
      1 mpi_rcv 0.77 0.14 0.93 - 9578 150 1116.7
      1 98 Total transmit 0.93 0.01 0.94
      0 mpi_wait_self 0.01 0.01 0.06 - 7054 141
      0 dm_update 0.15 0.05 0.35 - 288 150
        0 transmit 0.12 0.03 0.32 - 4047 150
          0 mpi_send 0.02 0.01 0.03 - 2023 150 42.6262
          0 mpi_rcv 0.09 0.01 0.31 - 2023 150 42.6262
          0 mpi_wait 0.00 0.00 0.01 - 729 32
          0 98 Total transmit 0.11 0.00 0.12
          0 mpi_wait_self 0.01 0.00 0.02 - 4815 150
          0 85 Total dm_update 0.12 0.02 0.15
        0 mpp_io 0.25 0.11 0.35 - 18 150
      1 65 Total Flux DN from atm 1.35 0.73 2.08
```

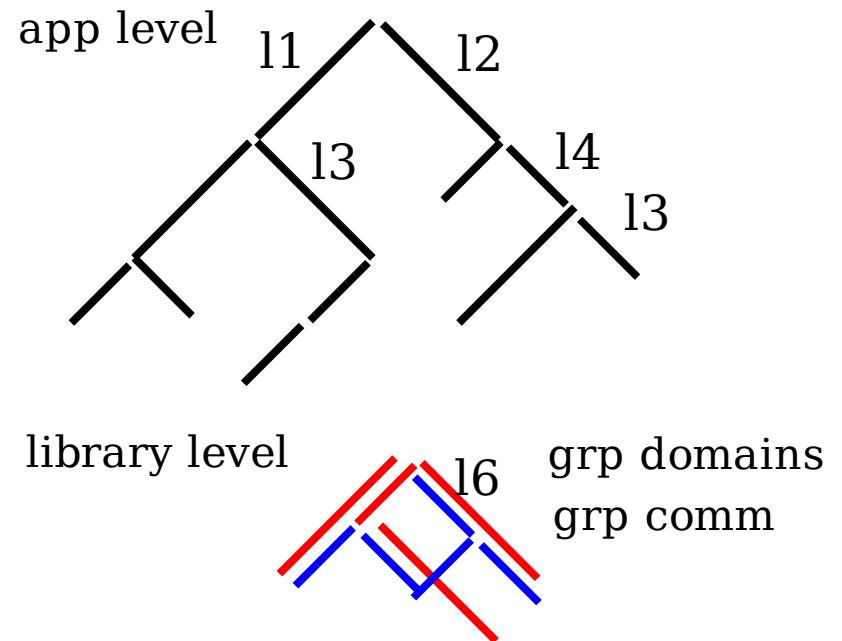
MOM4 clocks, 150 CPU

1

```
0 Flux land to ice 0.38 0.10 0.59 - 6 150
  0 transmit 0.01 0.01 0.04 - 245 150
    0 mpi_send 0.00 0.00 0.01 - 130 141 691.294
    0 mpi_rcv 0.01 0.00 0.04 - 122 150 691.294
    0 99 Total transmit 0.01 0.00 0.01
  0 mpi_wait_self 0.00 0.00 0.00 - 127 150
  0 mpp_io 0.32 0.05 0.51 - 3 150
  0 87 Total Flux land to ice 0.33 0.05 0.38
0 XGrid generation 0.01 0.00 0.01 - 6 150
1 Flux UP to atm 0.59 0.12 0.89 - 72 150
  0 transmit 0.44 0.02 0.78 - 4752 150
    0 mpi_send 0.03 0.01 0.06 - 2376 150 411.801
    0 mpi_rcv 0.40 0.00 0.77 - 2376 150 411.801
    0 99 Total transmit 0.43 0.01 0.44
  0 mpi_wait_self 0.01 0.00 0.04 - 2320 150
  0 76 Total Flux UP to atm 0.44 0.14 0.59
6 100 Total Land-ice-atm coupler 6.13 0.00 6.13
0 Ice-ocean coupler 0.07 0.07 0.08 - 12 150
0 Flux ice to ocean 0.05 0.04 0.05 - 6 150
  0 dm_redist 0.04 0.04 0.04 - 72 150
    0 transmit 0.03 0.03 0.03 - 1440 150
      0 mpi_send 0.01 0.01 0.01 - 720 150 384
      0 mpi_rcv 0.02 0.01 0.02 - 720 150 384
      0 94 Total transmit 0.03 0.00 0.03
    0 mpi_wait_self 0.00 0.00 0.00 - 720 150
    0 73 Total dm_redist 0.03 0.01 0.04
  0 89 Total Flux ice to ocean 0.04 0.01 0.05
0 Flux ocean to ice 0.03 0.02 0.03 - 6 150
  0 dm_redist 0.02 0.02 0.02 - 36 150
    0 transmit 0.01 0.01 0.02 - 720 150
      0 mpi_send 0.01 0.00 0.01 - 360 150 384
      0 mpi_rcv 0.01 0.01 0.01 - 360 150 384
      0 94 Total transmit 0.01 0.00 0.01
    0 mpi_wait_self 0.00 0.00 0.00 - 360 150
    0 68 Total dm_redist 0.02 0.01 0.02
  0 84 Total Flux ocean to ice 0.02 0.00 0.03
0 100 Total Ice-ocean coupler 0.07 0.00 0.07
100 100 Total PROFILE 102.57 0.02 102.59
```

Aggregate results

- tree still too large
- tree split across group boundaries
- subtrees are merged
- top-level – application
- merged tree – libs
- other schemes possible
 - each group independent
 - split only some groups
- libs take 50% of runtime
- IO takes 6%



MOM4 aggregated clocks, 150 CPU

```
100 PROFILE 102.59 102.59 102.59 - 6 150
  4 Ice 3.97 3.78 4.22 - 84 150
 90 Ocean 92.34 92.33 92.35 - 72 150
    2 (Ocean Advection Velocity) 1.57 1.40 1.90 - 72 150
    2 (Ocean Density) 2.05 1.85 2.28 - 72 150
    0 (Ocean rho-tilde) 0.08 0.06 0.11 - 72 150
    1 (Ocean u-rho) 0.55 0.39 0.61 - 72 150
    6 (Ocean Vertical Mixing Coeff) 5.73 5.19 6.25 - 72 150
 21 (Ocean Neutral Physics) 21.94 21.52 22.45 - 72 150
    0 (Ocean neutral density derivs) 0.39 0.28 0.44 - 72 150
    1 (Ocean neutral slopes ) 0.60 0.47 0.68 - 72 150
    2 (Ocean neutral fz-terms) 2.16 2.05 2.73 - 72 150
    3 (Ocean neutral fx-flux) 3.05 2.85 3.60 - 3600 150
    3 (Ocean neutral fy-flux) 2.97 2.82 3.52 - 3600 150
    1 (Ocean neutral fz-flux) 0.96 0.88 1.04 - 3600 150
    0 (Ocean neutral eady rate) 0.01 0.00 0.01 - 72 150
    0 (Ocean neutral baroclinicity) 0.02 0.01 0.02 - 72 150
    0 (Ocean neutral rossby radius) 0.17 0.13 0.19 - 72 150
    0 (Ocean neutral diffusivity) 0.13 0.10 0.16 - 72 150
 10 48 Total (Ocean Neutral Physics) 10.45 11.49 21.94
    1 (Ocean Shortwave) 0.70 0.48 0.84 - 72 150
      1 (Ocean Shortwave Penetration) 0.53 0.33 0.62 - 1440 150
      1 75 Total (Ocean Shortwave) 0.53 0.17 0.70
    0 (Ocean sponge) 0.00 0.00 0.00 - 72 150
    1 (Ocean xlandmix) 1.41 1.30 1.52 - 72 150
    0 (Ocean rivermix) 0.35 0.18 0.55 - 72 150
    1 (Ocean overflow) 0.79 0.61 1.09 - 72 150
    0 (Ocean sigma diffusion) 0.26 0.18 0.34 - 72 150
 16 (Ocean tracer update) 16.52 16.11 16.81 - 72 150
    5 (Ocean tracer_adv horz 4th) 4.64 4.50 4.72 - 72 150
    0 (Ocean tracer_adv vert 4th) 0.18 0.16 0.22 - 72 150
    3 (Ocean tracer_adv horz quick) 3.05 2.97 3.13 - 72 150
    0 (Ocean tracer_adv vert quick) 0.30 0.28 0.34 - 72 150
    4 (Ocean tracer_adv MDFL-sweby) 3.59 3.48 3.67 - 72 150
 11 71 Total (Ocean tracer update) 11.76 4.76 16.52
    0 (Ocean surface flux) 0.25 0.15 0.39 - 72 150
    0 (Ocean bottom flux) 0.12 0.03 0.56 - 72 150
    4 (Ocean restoring flux) 3.80 3.61 4.14 - 72 150
    0 (Ocean TPM sbc) 0.01 0.00 0.01 - 72 150
    0 (Ocean TPM source) 0.12 0.10 0.15 - 72 150
    0 (Ocean TPM bbc) 0.00 0.00 0.00 - 72 150
    0 (Ocean TPM tracer) 0.00 0.00 0.00 - 72 150
    6 (Ocean explicit acceleration) 6.02 5.67 6.59 - 72 150
      4 (Ocean Smag lap friction) 3.66 3.27 4.23 - 72 150
      4 61 Total (Ocean explicit acceleration) 3.66 2.37 6.02
    1 (Ocean implicit friction) 1.04 0.98 1.10 - 72 150
    0 (Ocean implicit Coriolis) 0.19 0.17 0.20 - 72 150
    0 (Ocean surf height tendency) 0.01 0.01 0.01 - 72 150
    0 (Ocean freesurface drag) 0.02 0.02 0.03 - 72 150
    0 (Ocean freesurface forcing) 0.18 0.12 0.28 - 72 150
 13 (Ocean freesurface update) 13.59 13.19 14.09 - 72 150
    1 (Ocean velocity update) 0.79 0.69 0.91 - 72 150
    0 (Ocean polar filter) 0.00 0.00 0.00 - 72 150
    0 (Ocean energy analysis) 0.00 0.00 0.00 - 72 150
 11 (Ocean numerical diagnostics) 11.22 11.05 11.51 - 72 150
    1 (Ocean Robert time filter) 0.63 0.52 0.72 - 72 150
    0 (Ocean update cell thickness) 0.01 0.00 0.01 - 72 150
    1 (Ocean update halos) 1.27 0.79 1.98 - 72 150
    0 (Ocean sum ocean surface) 0.01 0.01 0.01 - 72 150
    0 (Ocean average state) 0.00 0.00 0.00 - 6 150
    1 (Ocean tracer tmask limit) 0.76 0.65 0.84 - 72 150
 90 100 Total Ocean 91.99 0.35 92.34
  6 Land-ice-atm coupler 6.13 5.86 6.33 - 228 150
    3 SFC boundary layer 3.07 2.97 3.58 - 72 150
    2 Flux DN from atm 2.08 1.84 2.32 - 72 150
    0 Flux land to ice 0.38 0.10 0.59 - 6 150
    0 XGrid generation 0.01 0.00 0.01 - 6 150
    1 Flux UP to atm 0.59 0.12 0.89 - 72 150
    6 100 Total Land-ice-atm coupler 6.13 0.00 6.13
  0 Ice-ocean coupler 0.07 0.07 0.08 - 12 150
    0 Flux ice to ocean 0.05 0.04 0.05 - 6 150
    0 Flux ocean to ice 0.03 0.02 0.03 - 6 150
    0 100 Total Ice-ocean coupler 0.07 0.00 0.07
 100 100 Total PROFILE 102.51 0.08 102.59
Aggregate data
  3 transmit 2.77 1.98 3.35 - 53256 150
    0 mpi_send 0.39 0.06 1.48 - 26628 150 1042.3
    2 mpi_rcv 2.31 1.22 3.15 - 26628 150 1042.3
    3 98 Total transmit 2.70 0.07 2.77
  0 sync 0.06 0.03 0.12 - 30 150
    0 mpi_barrier 0.06 0.03 0.12 - 30 150
    0 100 Total sync 0.06 0.00 0.06
  0 mpi_wait_self 0.05 0.01 0.14 - 17736 150
  8 domains 8.09 7.31 9.14 - 324 150
    7 transmit 7.36 6.62 8.28 - 96552 150
      1 mpi_send 1.51 1.36 3.11 - 48276 150 3840
      6 mpi_rcv 5.71 4.77 6.14 - 48276 150 3840
      7 98 Total transmit 7.22 0.15 7.36
    0 mpi_wait_self 0.07 0.06 0.12 - 48276 150
    7 92 Total domains 7.44 0.66 8.09
 33 dm_update 33.47 30.07 36.58 - 81042 150
 23 transmit 24.09 20.62 27.04 - 135660 150
    7 mpi_send 7.22 4.02 8.62 - 678300 150 267.893
   15 mpi_rcv 15.23 11.39 21.51 - 678300 150 267.914
    0 mpi_wait 0.01 0.00 0.06 - 2959 55
   22 93 Total transmit 22.46 1.64 24.09
    1 mpi_wait_self 1.47 0.92 1.79 - 686107 150
   25 76 Total dm_update 25.56 7.91 33.47
  0 dm_redist 0.06 0.06 0.07 - 108 150
    0 transmit 0.04 0.04 0.05 - 2160 150
      0 mpi_send 0.02 0.01 0.02 - 1080 150 384
      0 mpi_rcv 0.02 0.02 0.03 - 1080 150 384
      0 94 Total transmit 0.04 0.00 0.04
    0 mpi_wait_self 0.00 0.00 0.00 - 1080 150
    0 71 Total dm_redist 0.04 0.02 0.06
  1 mpp_io 0.97 0.31 1.46 - 36 150
  6 Send data 6.19 5.44 7.16 - 4398 150
 50 50 Total PROFILE 51.66 50.93 102.59
```

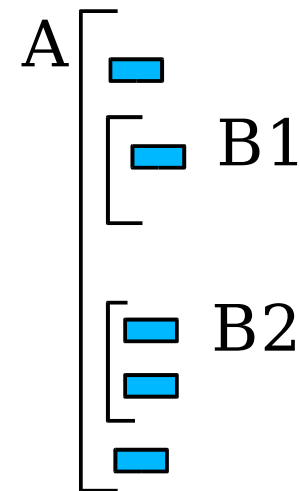
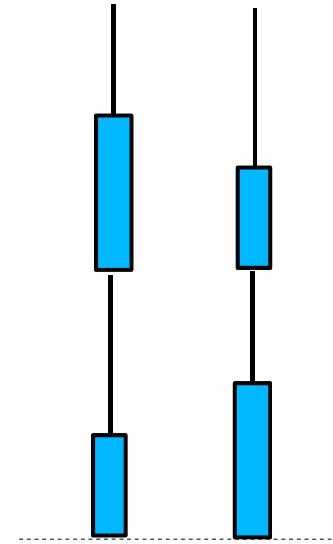
opt

??

>>

Computation unbalance

- comm, comp steps
- same sections across CPUs
- for given code section A
 $U(A) = \max(\text{comp}) - \text{avg}(\text{comp})$
- $U(A)$ is lower bound
- $\text{comp} = \text{time} - \text{MPI, I/O time}$
- A has subtrees B_i
- $U(A) = \text{sum } U(B_i) + M \cdot \text{avg} \{ \text{time}(A) - \text{MPI} - \text{sum } \text{time}(B_i) \}$
- # subsections matters (finer granularity)



comp unbalance, 150 CPU

```
100 PROFILE 102.59 102.59 102.59 - 6 150
  4 Ice 3.97 3.78 4.22 - 84 150
    0 7 Unbalance Ice 0.26 0.22 3.97
90 Ocean 92.34 92.33 92.35 - 72 150
  2 (Ocean Advection Velocity) 1.57 1.40 1.90 - 72 150
    0 10 Unbalance (Ocean Advection Velocity) 0.16 0.05 1.57
  2 (Ocean Density) 2.05 1.85 2.28 - 72 150
    0 14 Unbalance (Ocean Density) 0.29 0.08 2.05
  0 (Ocean rho-tilde) 0.08 0.06 0.11 - 72 150
    0 29 Unbalance (Ocean rho-tilde) 0.03 0.03 0.08
  1 (Ocean u-rho) 0.55 0.39 0.61 - 72 150
    0 11 Unbalance (Ocean u-rho) 0.06 0.06 0.55
  6 (Ocean Vertical Mixing Coeff) 5.73 5.19 6.25 - 72 150
    1 19 Unbalance (Ocean Vertical Mixing Coeff) 1.09 0.95 5.73
21 (Ocean Neutral Physics) 21.94 21.52 22.45 - 72 150
  0 (Ocean neutral density derivs) 0.39 0.28 0.44 - 72 150
    0 13 Unbalance (Ocean neutral density derivs) 0.05 0.05 0.39
  1 (Ocean neutral slopes ) 0.60 0.47 0.68 - 72 150
    0 14 Unbalance (Ocean neutral slopes ) 0.08 0.08 0.60
  2 (Ocean neutral fz-terms) 2.16 2.05 2.73 - 72 150
    1 26 Unbalance (Ocean neutral fz-terms) 0.57 0.57 2.16
  3 (Ocean neutral fx-flux) 3.05 2.85 3.60 - 3600 150
    1 18 Unbalance (Ocean neutral fx-flux) 0.55 0.55 3.05
  3 (Ocean neutral fy-flux) 2.97 2.82 3.52 - 3600 150
    1 18 Unbalance (Ocean neutral fy-flux) 0.55 0.55 2.97
  1 (Ocean neutral fz-flux) 0.96 0.88 1.04 - 3600 150
    0 9 Unbalance (Ocean neutral fz-flux) 0.08 0.08 0.96
  0 (Ocean neutral eady rate) 0.01 0.00 0.01 - 72 150
  0 (Ocean neutral baroclinicity) 0.02 0.01 0.02 - 72 150
  0 (Ocean neutral rossby radius) 0.17 0.13 0.19 - 72 150
    0 27 Unbalance (Ocean neutral rossby radius) 0.05 0.04 0.17
  0 (Ocean neutral diffusivity) 0.13 0.10 0.16 - 72 150
    0 16 Unbalance (Ocean neutral diffusivity) 0.02 0.01 0.13
10 48 Total (Ocean Neutral Physics) 10.45 11.49 21.94
  2 11 Unbalance (Ocean Neutral Physics) 2.35 0.15 21.94
  1 (Ocean Shortwave) 0.70 0.48 0.84 - 72 150
    1 (Ocean Shortwave Penetration) 0.53 0.33 0.62 - 1440 150
      0 19 Unbalance (Ocean Shortwave Penetration) 0.10 0.10 0.53
    1 75 Total (Ocean Shortwave) 0.53 0.17 0.70
      0 21 Unbalance (Ocean Shortwave) 0.15 0.05 0.70
  0 (Ocean sponge) 0.00 0.00 0.00 - 72 150
  1 (Ocean xlandmix) 1.41 1.30 1.52 - 72 150
    0 5 Unbalance (Ocean xlandmix) 0.07 0.03 1.41
  0 (Ocean rivermix) 0.35 0.18 0.55 - 72 150
    0 57 Unbalance (Ocean rivermix) 0.20 0.20 0.35
  1 (Ocean overflow) 0.79 0.61 1.09 - 72 150
    0 13 Unbalance (Ocean overflow) 0.10 0.09 0.79
  0 (Ocean sigma diffusion) 0.26 0.18 0.34 - 72 150
    0 12 Unbalance (Ocean sigma diffusion) 0.03 0.02 0.26
16 (Ocean tracer update) 16.52 16.11 16.81 - 72 150
  5 (Ocean tracer_adv horz 4th) 4.64 4.50 4.72 - 72 150
    0 3 Unbalance (Ocean tracer_adv horz 4th) 0.13 0.06 4.64
  0 (Ocean tracer_adv vert 4th) 0.18 0.16 0.22 - 72 150
    0 20 Unbalance (Ocean tracer_adv vert 4th) 0.04 0.04 0.18
  3 (Ocean tracer_adv horz quick) 3.05 2.97 3.13 - 72 150
    0 4 Unbalance (Ocean tracer_adv horz quick) 0.11 0.05 3.05
  0 (Ocean tracer_adv vert quick) 0.30 0.28 0.34 - 72 150
    0 15 Unbalance (Ocean tracer_adv vert quick) 0.04 0.04 0.30
  4 (Ocean tracer_adv MDFL-sweby) 3.59 3.48 3.67 - 72 150
    0 5 Unbalance (Ocean tracer_adv MDFL-sweby) 0.17 0.12 3.59
11 71 Total (Ocean tracer update) 11.76 4.76 16.52
  1 8 Unbalance (Ocean tracer update) 1.27 0.54 16.52
  0 (Ocean surface flux) 0.25 0.15 0.39 - 72 150
    0 5 Unbalance (Ocean surface flux) 0.01 0.01 0.25
  0 (Ocean bottom flux) 0.12 0.03 0.56 - 72 150
  4 (Ocean restoring flux) 3.80 3.61 4.14 - 72 150
    0 6 Unbalance (Ocean restoring flux) 0.23 0.00 3.80
  0 (Ocean TPM sbc) 0.01 0.00 0.01 - 72 150
  0 (Ocean TPM source) 0.12 0.10 0.15 - 72 150
    0 24 Unbalance (Ocean TPM source) 0.03 0.03 0.12
  0 (Ocean TPM bbc) 0.00 0.00 0.00 - 72 150
  0 (Ocean TPM tracer) 0.00 0.00 0.00 - 72 150
  6 (Ocean explicit acceleration) 6.02 5.67 6.59 - 72 150
    4 (Ocean Smag lap friction) 3.66 3.27 4.23 - 72 150
      0 4 Unbalance (Ocean Smag lap friction) 0.15 0.12 3.66
    4 61 Total (Ocean explicit acceleration) 3.66 2.37 6.02
      0 6 Unbalance (Ocean explicit acceleration) 0.34 0.14 6.02
  1 (Ocean implicit friction) 1.04 0.98 1.10 - 72 150
    0 6 Unbalance (Ocean implicit friction) 0.06 0.06 1.04
  0 (Ocean implicit Coriolis) 0.19 0.17 0.20 - 72 150
    0 7 Unbalance (Ocean implicit Coriolis) 0.01 0.01 0.19
  0 (Ocean surf height tendency) 0.01 0.01 0.01 - 72 150
  0 (Ocean freesurface drag) 0.02 0.02 0.03 - 72 150
  0 (Ocean freesurface forcing) 0.18 0.12 0.28 - 72 150
    0 6 Unbalance (Ocean freesurface forcing) 0.01 0.00 0.18
13 (Ocean freesurface update) 13.59 13.19 14.09 - 72 150
  1 5 Unbalance (Ocean freesurface update) 0.68 0.14 13.59
  1 (Ocean velocity update) 0.79 0.69 0.91 - 72 150
    0 18 Unbalance (Ocean velocity update) 0.14 0.05 0.79
  0 (Ocean polar filter) 0.00 0.00 0.00 - 72 150
  0 (Ocean energy analysis) 0.00 0.00 0.00 - 72 150
11 (Ocean numerical diagnostics) 11.22 11.05 11.51 - 72 150
  0 4 Unbalance (Ocean numerical diagnostics) 0.42 0.30 11.22
  1 (Ocean Robert time filter) 0.63 0.52 0.72 - 72 150
    0 14 Unbalance (Ocean Robert time filter) 0.09 0.09 0.63
  0 (Ocean update cell thickness) 0.01 0.00 0.01 - 72 150
  1 (Ocean update halos) 1.27 0.79 1.98 - 72 150
    0 2 Unbalance (Ocean update halos) 0.03 0.00 1.27
  0 (Ocean sum ocean surface) 0.01 0.01 0.01 - 72 150
  0 (Ocean average state) 0.00 0.00 0.00 - 6 150
  1 (Ocean tracer tmask limit) 0.76 0.65 0.84 - 72 150
    0 20 Unbalance (Ocean tracer tmask limit) 0.15 0.09 0.76
90 100 Total Ocean 91.99 0.35 92.34
  8 9 Unbalance Ocean 8.06 0.03 92.34
```

comp unbalance, 150 CPU

```
6 Land-ice-atm coupler 6.13 5.86 6.33 - 228 150
  3 SFC boundary layer 3.07 2.97 3.58 - 72 150
    1 25 Unbalance SFC boundary layer 0.78 0.48 3.07
  2 Flux DN from atm 2.08 1.84 2.32 - 72 150
    0 15 Unbalance Flux DN from atm 0.32 0.17 2.08
  0 Flux land to ice 0.38 0.10 0.59 - 6 150
    0 56 Unbalance Flux land to ice 0.21 0.02 0.38
  0 XGrid generation 0.01 0.00 0.01 - 6 150
  1 Flux UP to atm 0.59 0.12 0.89 - 72 150
    0 21 Unbalance Flux UP to atm 0.13 0.12 0.59
  6 100 Total Land-ice-atm coupler 6.13 0.00 6.13
  1 24 Unbalance Land-ice-atm coupler 1.44 0.00 6.13
0 Ice-ocean coupler 0.07 0.07 0.08 - 12 150
  0 Flux ice to ocean 0.05 0.04 0.05 - 6 150
  0 Flux ocean to ice 0.03 0.02 0.03 - 6 150
  0 100 Total Ice-ocean coupler 0.07 0.00 0.07
100 100 Total PROFILE 102.51 0.08 102.59
10 10 Unbalance PROFILE 9.77 0.00 102.59
```

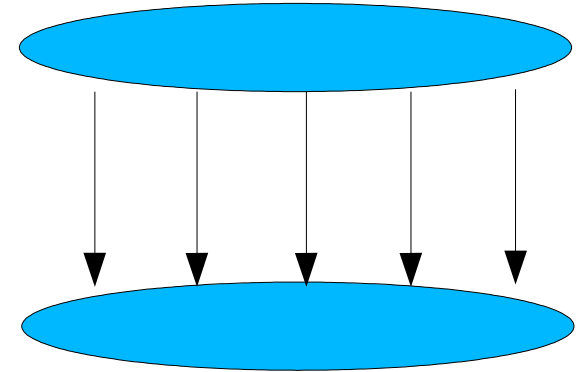
- halo upd - regular, coupler - irregular
- unbalance 10% overall, 3% libs
- 44% comm
 - 10% total unbalance
 - 34% MPI, lib
- 33% halo update
 - 8% not MPI, 3% unbalance

Aggregate data

```
3 transmit 2.77 1.98 3.35 - 53256 150
  0 mpi_send 0.39 0.06 1.48 - 26628 150 1042.3
  2 mpi_recv 2.31 1.22 3.15 - 26628 150 1042.3
  3 98 Total transmit 2.70 0.07 2.77
  0 4 Unbalance transmit 0.11 0.11 2.77
0 sync 0.06 0.03 0.12 - 30 150
  0 mpi_barrier 0.06 0.03 0.12 - 30 150
  0 100 Total sync 0.06 0.00 0.06
0 mpi_wait_self 0.05 0.01 0.14 - 17736 150
8 domains 8.09 7.31 9.14 - 324 150
  7 transmit 7.36 6.62 8.28 - 96552 150
    1 mpi_send 1.51 1.36 3.11 - 48276 150 3840
    6 mpi_recv 5.71 4.77 6.14 - 48276 150 3840
    7 98 Total transmit 7.22 0.15 7.36
    0 3 Unbalance transmit 0.24 0.24 7.36
    0 mpi_wait_self 0.07 0.06 0.12 - 48276 150
    7 92 Total domains 7.44 0.66 8.09
    0 6 Unbalance domains 0.45 0.21 8.09
33 dm_update 33.47 30.07 36.58 - 81042 150
  23 transmit 24.09 20.62 27.04 - 1356600 150
    7 mpi_send 7.22 4.02 8.62 - 678300 150 267.893
    15 mpi_recv 15.23 11.39 21.51 - 678300 150 267.914
    0 mpi_wait 0.01 0.00 0.06 - 2959 55
    22 93 Total transmit 22.46 1.64 24.09
    0 1 Unbalance transmit 0.22 0.22 24.09
    1 mpi_wait_self 1.47 0.92 1.79 - 686107 150
  25 76 Total dm_update 25.56 7.91 33.47
  1 3 Unbalance dm_update 0.85 0.63 33.47
0 dm_redist 0.06 0.06 0.07 - 108 150
  0 transmit 0.04 0.04 0.05 - 2160 150
    0 mpi_send 0.02 0.01 0.02 - 1080 150 384
    0 mpi_recv 0.02 0.02 0.03 - 1080 150 384
    0 94 Total transmit 0.04 0.00 0.04
    0 mpi_wait_self 0.00 0.00 0.00 - 1080 150
    0 71 Total dm_redist 0.04 0.02 0.06
1 mpp_io 0.97 0.31 1.46 - 36 150
  1 54 Unbalance mpp_io 0.53 0.53 0.97
6 Send data 6.19 5.44 7.16 - 4398 150
  1 18 Unbalance Send data 1.12 1.12 6.19
50 50 Total PROFILE 51.66 50.93 102.59
  3 3 Unbalance PROFILE 3.06 2.59 102.59
```

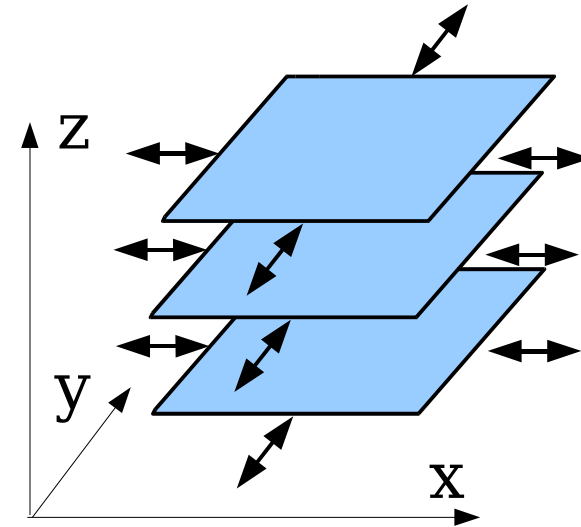
Bottlenecks

- Application structure
 - comp step
 - comm step
 - ...
- Too many short comp steps, small msgs
- **Latency hurts:** $nT(s) = n(K+ks)$, $T(ns) = K+nks$
- Hide latency – techniques
 - **aggregate multiple msgs**
 - overlap comm with comp
 - trade comp for comm
 - find better algorithms – less comm



Halo updates

- Several variables mapped to 3D grid
- Processing & halo update done for each k layer
- restructure application to work on 3D cells (50x)



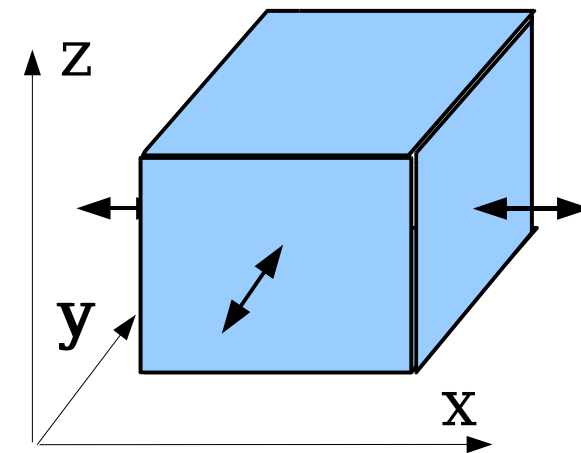
```

21 (Ocean Neutral Physics) 21.94 21.52 22.45 - 72 150
  8 dm_update 7.95 5.97 8.96 - 10872 150
    6 transmit 5.88 3.82 7.00 - 335327 150
      2 mpi_send 1.54 0.86 1.93 - 167663 150 108.217
      4 mpi_rcv 3.94 1.72 5.83 - 167663 150 108.217
      0 mpi_wait 0.02 0.02 0.02 - 10800 1
      5 93 Total transmit 5.49 0.39 5.88
    0 mpi_wait_self 0.37 0.22 0.51 - 167591 150
    6 79 Total dm_update 6.25 1.70 7.95
  0 Send data 0.28 0.24 0.36 - 432 150
  
```

z=50, #tr

```

16 (Ocean tracer update) 16.52 16.11 16.81 - 72 150
  1 Send data 1.37 1.18 1.61 - 648 150
  5 (Ocean tracer_adv horz 4th) 4.64 4.50 4.72 - 72 150
    4 dm_update 4.22 4.04 4.32 - 10800 150
      3 transmit 2.89 2.65 3.24 - 221712 150
        1 mpi_send 1.11 0.57 1.41 - 110856 150 143.667
        1 mpi_rcv 1.51 1.05 2.50 - 110856 150 143.667
        0 mpi_wait 0.01 0.01 0.01 - 3600 1
        3 91 Total transmit 2.63 0.26 2.89
      0 mpi_wait_self 0.24 0.15 0.31 - 110832 150
      3 74 Total dm_update 3.13 1.09 4.22
    4 91 Total (Ocean tracer_adv horz 4th) 4.22 0.41 4.64
  
```



neutral physics

before

```
subroutine neutral_physics (rho_prev, pressure_at_depth, T_prog)

  real, intent(in), dimension(isd:ied,jsd:jed,nk)      :: rho_prev,
  pressure_at_depth
  type(ocean_prog_tracer_type), intent(inout)          :: T_prog(:)
  real, dimension(isd:ied,jsd:jed)                    :: tchg,
  tmp_flux

  ...

  do k=1,nk

    call fz_flux(T_prog,k)

    do nn=1,num_prog_tracers
      fx(nn)%field(:, :) = agm_array(:, :, k)*dTdx(nn)%field(:, :, k)
      *FMX(Grd%dht(:, :, k)*Grd%tmask(:, :, k))
      fy(nn)%field(:, :) = agm_array(:, :, k)*dTdy(nn)%field(:, :, k)
      *FMY(Grd%dht(:, :, k)*Grd%tmask(:, :, k))

      if (Grd%tripolar) then
        call mpp_update_domains(fx(nn)%field(:, :), fy(nn)%field
          (:, :), Dom_flux%domain2d, gridtype=CGRID_NE) !for fold redundancies
      endif

      tchg(:, :) = (BDX_ET(fx(nn)%field(:, :)) + BDY_NT(fy(nn)%field
        (:, :)))
      T_prog(nn)%wrk1(isc:iec, jsc:jec, k) = &
        Grd%tmask(isc:iec, jsc:jec, k)*( tchg(isc:iec, jsc:jec) +
        (fz1(nn)%field(isc:iec, jsc:jec) - fz2(nn)%field(isc:iec, jsc:jec)))
      T_prog(nn)%source(isc:iec, jsc:jec, k) = T_prog(nn)%source
        (isc:iec, jsc:jec, k) + &
        T_prog(nn)%wrk1(isc:iec, jsc:jec, k) ! add neutral
        contribution to source term
      fz1(nn)%field(isc:iec, jsc:jec) = fz2(nn)%field
        (isc:iec, jsc:jec)
      flux_x(nn)%field(:, :, k) = Grd%dyte(:, :)*fx(nn)%field(:, :
        )
      flux_y(nn)%field(:, :, k) = Grd%dxtn(:, :)*fy(nn)%field(:, :
        )
    enddo
  enddo

  ...
```

after

```
subroutine neutral_physics (rho_prev, pressure_at_depth, T_prog)

  real, intent(in), dimension(isd:ied,jsd:jed,nk)      :: rho_prev,
  pressure_at_depth
  type(ocean_prog_tracer_type), intent(inout)          :: T_prog(:)
  real, dimension(isd:ied,jsd:jed)                    :: tchg,
  tmp_flux
  real, dimension(isd:ied,jsd:jed,nk,num_prog_tracers) :: txz, tyz

  ...

  do k=1,nk
    call fz_flux (T_prog,k) ! for all nn, single k
    do nn=1,num_prog_tracers
      txz(:, :, k, nn) = agm_array(:, :, k)*dTdx(nn)%field(:, :, k)*FMX
        (Grd%dht(:, :, k)*Grd%tmask(:, :, k))
      tyz(:, :, k, nn) = agm_array(:, :, k)*dTdy(nn)%field(:, :, k)*FMY
        (Grd%dht(:, :, k)*Grd%tmask(:, :, k))
    enddo
  enddo

  if (Grd%tripolar) then
    call mpp_update_domains(txz, tyz, Dom_flux%domain2d, gridtype=
      CGRID_NE) !for fold redundancies
  endif

  do k=1,nk
    do nn=1,num_prog_tracers
      fx(nn)%field(:, :) = txz(:, :, k, nn);
      fy(nn)%field(:, :) = tyz(:, :, k, nn);
    enddo

    do nn=1,num_prog_tracers
      tchg(:, :) = (BDX_ET(fx(nn)%field(:, :)) + BDY_NT(fy(nn)%field
        (:, :)))
      T_prog(nn)%wrk1(isc:iec, jsc:jec, k) = &
        Grd%tmask(isc:iec, jsc:jec, k)*( tchg(isc:iec, jsc:jec) +
        (fz1(nn)%field(isc:iec, jsc:jec) - fz2(nn)%field(isc:iec, jsc:jec)))
      T_prog(nn)%source(isc:iec, jsc:jec, k) = T_prog(nn)%source
        (isc:iec, jsc:jec, k) + &
        T_prog(nn)%wrk1(isc:iec, jsc:jec, k) ! add neutral
        contribution to source term
      fz1(nn)%field(isc:iec, jsc:jec) = fz2(nn)%field
        (isc:iec, jsc:jec)
      flux_x(nn)%field(:, :, k) = Grd%dyte(:, :)*fx(nn)%field(:, :
        )
      flux_y(nn)%field(:, :, k) = Grd%dxtn(:, :)*fy(nn)%field(:, :
        )
    enddo
  enddo

  ...
```

neutral physics

before

```
21 (Ocean Neutral Physics) 21.94 21.52 22.45 - 72 150
 8 dm_update 7.95 5.97 8.96 - 10872 150
 6 transmit 5.88 3.82 7.00 - 335327 150
 2 mpi_send 1.54 0.86 1.93 - 167663 150 108.217
 4 mpi_recv 3.94 1.72 5.83 - 167663 150 108.217
 0 mpi_wait 0.02 0.02 0.02 - 10800 1
 5 93 Total transmit 5.49 0.39 5.88
 0 mpi_wait_self 0.37 0.22 0.51 - 167591 150
 6 79 Total_dm_update 6.25 1.70 7.95
 0 Send data 0.28 0.24 0.36 - 432 150
```

after

```
19 (Ocean Neutral Physics) 16.34 15.76 16.80 - 72 150
 0 dm_update 0.30 0.08 0.74 - 72 150 4800
 0 transmit 0.26 0.05 0.70 - 1103 150
 0 mpi_send 0.02 0.01 0.03 - 551 150 4668.41
 0 mpi_recv 0.24 0.02 0.68 - 551 150 4668.41
 0 100 Total transmit 0.26 0.00 0.26
 0 mpi_wait_self 0.00 0.00 0.00 - 551 150
 0 88 Total_dm_update 0.26 0.04 0.30
 1 dm_updatev 0.95 0.32 2.06 - 72 150
 0 transmit 0.14 0.01 0.48 - 220 15
 0 mpi_send 0.01 0.00 0.01 - 110 15 10956.5
 0 mpi_recv 0.14 0.00 0.47 - 110 15 10956.5
 0 mpi_wait 0.00 0.00 0.00 - 72 1
 0 100 Total transmit 0.14 -0.00 0.14
 0 mpi_wait_self 0.00 0.00 0.00 - 105 15
 1 dm_update 0.93 0.32 1.98 - 144 150 14400
 1 transmit 0.80 0.20 1.83 - 2206 150
 0 mpi_send 0.08 0.05 0.10 - 1103 150 14005.2
 1 mpi_recv 0.72 0.10 1.75 - 1103 150 14005.2
 1 100 Total transmit 0.80 0.00 0.80
 0 mpi_wait_self 0.00 0.00 0.01 - 1103 150
 1 87 Total_dm_update 0.81 0.12 0.93
 1 113 Total dm_updatev 1.07 -0.12 0.95
```

- decreased #call 50 times, increased msgs size
- result unchanged (checksums)
- 7.95 sec vs. 1.25 sec
- tracer update:

Global Fields

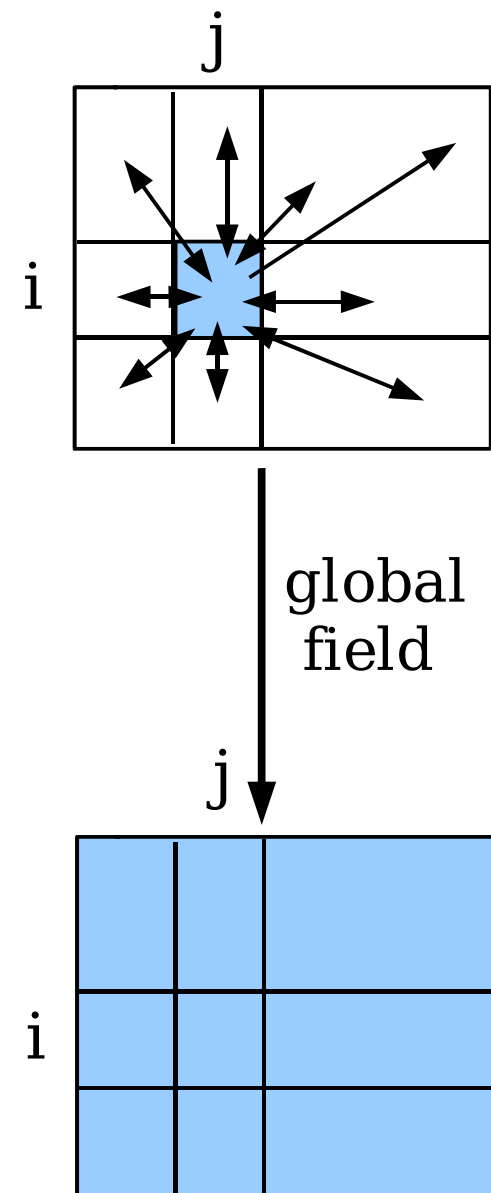
- each CPU stores a cell
- global field sends the cell to all CPUs, and receives their cell
- $n(n-1)$ msgs, D/n size
- full domain can be used
- huge perf penalty (#CPU)

```
14 (Ocean freesurface update) 11.60 11.44 11.72 - 72 150
```

```
...
3 dm_global_field 2.69 2.54 2.80 - 144 150 480
  3 transmit 2.44 2.27 2.52 - 42912 150
    1 mpi_send 0.65 0.60 0.72 - 21456 150 3840
    2 mpi_recv 1.76 1.62 1.87 - 21456 150 3840
    3 99 Total transmit 2.41 0.03 2.44
  0 mpi_wait_self 0.02 0.02 0.04 - 21456 150
  3 92 Total dm_global_field 2.47 0.23 2.69
```

```
4 (Ocean restoring flux) 3.80 3.61 4.14 - 72 150
```

```
...
3 domains 3.59 3.17 4.08 - 144 150
  3 transmit 3.22 2.81 3.63 - 42912 150
    1 mpi_send 0.66 0.60 1.48 - 21456 150 3840
    2 mpi_recv 2.50 1.97 2.71 - 21456 150 3840
    3 98 Total transmit 3.16 0.06 3.22
  0 mpi_wait_self 0.03 0.03 0.05 - 21456 150
  3 91 Total domains 3.25 0.34 3.59
```

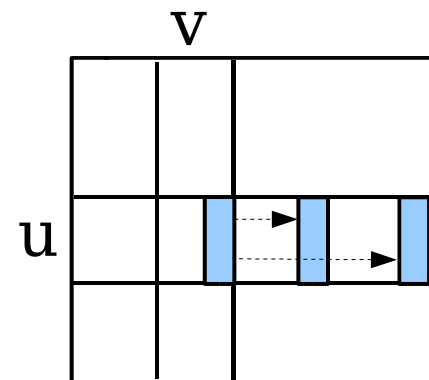
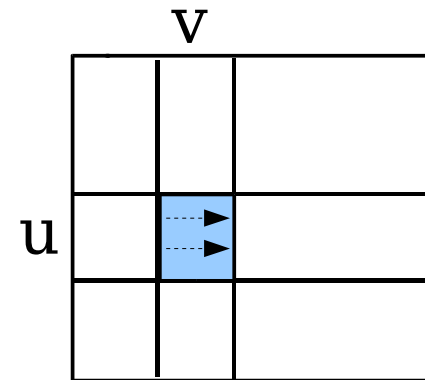
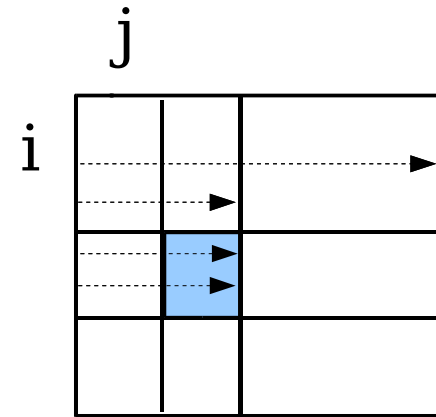


Freesurface update

- For each i,j $s_{ij} = \sum_{k=1}^j a_{ik}$
- global field generates full matrix, sums are computed
- better aggregation scheme:

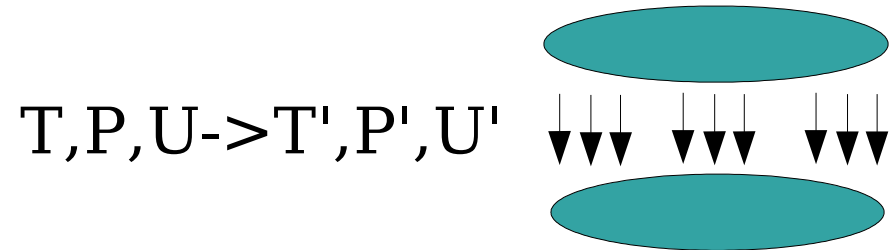
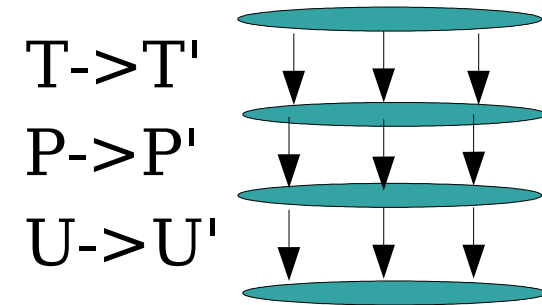
$$p_{iv} = \sum_{k=xm}^{(v+1)m-1} a_{ik}$$

- m is cell size, (u,v) are cell coord
- cell (u,v) sends p_{iv} to all (u,x) cells, where $x > v$
- s_{ij} can be reconstructed
- $\frac{1}{2}n\sqrt{n}$ messages of size $\sqrt{\frac{D}{n}}$
- checksums changed
- same for vertical



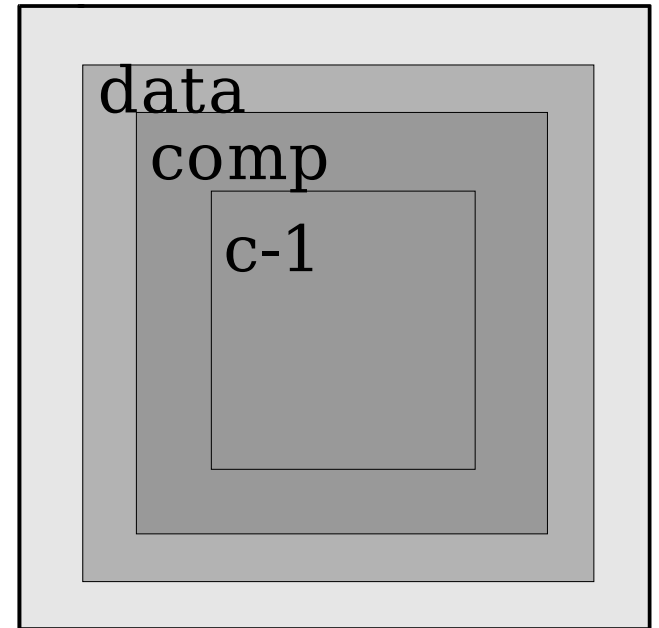
Grouping halo updates

- update T
- $T \rightarrow T'$
- update P
- $P \rightarrow P'$
- update U
- $U \rightarrow U'$
- merge T,P,U \rightarrow one vector, higher dim
- if P depends on T', can we use T instead?



Merging 2 halo updates

- trade comp for comm
- merge 2 comm/comp steps into 1
- send/recv halo 2x size
- 1st step: compute data domain
- 2nd step: compute comp domain
- should work across seams
- overlap comp, comm
- send halo
- compute c-1
- recv halo
- compute comp \ c-1



Code improvements

- domain update aggregation
 - 24% total runtime, speedup factor 4-5 (18%)
- global field elimination
 - 7% total runtime, speedup factor 10 (6%)
- static vs. dynamic arrays
 - 12% total runtime, speedup factor 2 (6%)
- A few code sections have been recoded

CPU	ver	libs	comp	total	lib %	total %
90	orig	53.64	87.20	140.85	100	100
	new	41.24	88.18	129.42	77	92
150	orig	51.66	50.93	102.59	100	100
	new	34.27	51.32	85.59	66	83

Improvements: Global field

- global field elimination
 - restoring flux: 3%
 - freesurface update: 3% (prev slides)
 - ice: 1% (sum, cut_check – send/recv to sym CPU)
 - 7% total runtime, save 6%

```
real, dimension(1 :im, 1 :jm) :: g_x

call mpp_global_field(Domain0, x, g_x)
g_sum = sum(g_x)

call mpp_global_field ( Domain, uv, global_uv )

do i=1,im/2-1
  if (global_uv(i,jm)/=-global_uv(im-i,jm)) then
    cut_error =.true.
!   if (mpp_pe()==0) &
!     print *, mesg, i, im-i, global_uv(i,jm), global_uv(im-i,jm)
    fix = (global_uv(i,jm)-global_uv(im-i,jm))/2
    global_uv(i ,jm) = fix
    global_uv(im-i,jm) = -fix
  end if
end do

uv = global_uv(is:ie,js:je)
```

Domain update

- k-level halo aggregation
 - neutral physics: 8%
 - tracer update - horz 4th, horz quick, sweby: 5%, 3%, 4%
 - ice_sis: 2%
 - ocean smag lap friction: 2%
 - 24% total runtime, speedup factor 4-5, save 18%
- multiple-vector aggregation, merge steps
 - freesurface update: 8%, speedup 2-3, save 5%

```
if(have_abc) then
  call mpp_update_domains (eta_t_fs(:,:,fstaup1), Dom%domain2d)
  call ocean_abc_freesurf(eta_t_fs, fstaum1, fstau, fstaup1, tdt)
endif

if(eta_filter) then
  call mpp_update_domains (eta_t_fs(:,:,fstaup1), Dom%domain2d)
  eta_t_fs(:,:,fstaup1) = eta_t_fs(:,:,fstaup1) + tdt*LAP_eta(eta_t_fs(:,:,fstaup1),eta_mix(:,:))
endif

call mpp_update_domains (eta_t_fs(:,:,fstaup1), Dom%domain2d)
...
if (Grd%tripolar) then
  call mpp_update_domains (Ext_mode%grad_ps(:,:,1),Ext_mode%grad_ps(:,:,2), Dom%domain2d, gridtype=BGRID_NE)
endif
```

Code improvements

- static vs. dynamic arrays
 - Ocean neutral fx-flux, fy-flux, fz-terms, ...
 - 12% total runtime, speedup factor 2, save 6%

```
slope = -Grd%tmask(i+ip,j,kkpkr)&  
        *(drhodT(i+ip,j,kk)*dTdx(index_temp)%field(i,j,kk)   +&  
          drhodS(i+ip,j,kk)*dTdx(index_salt)%field(i,j,kk)) &  
        /((drhodT(i+ip,j,kk)*dTdz(index_temp)%field(i+ip,j,kk-1+kr) +&  
          drhodS(i+ip,j,kk)*dTdz(index_salt)%field(i+ip,j,kk-1+kr) - epsln)
```

	Dynamic	Static
loop	3.35	2.95
loop-struct	4.56	4.46
array	3.02	2.99
array-struct	3.20	3.30

Notes

- Latency hurts: $nT(s) = n(K+ks) \gg T(ns) = K+nks$
- Hide latency
 - merge comm/comp steps
 - k level
 - multiple vectors
 - time steps
 - overlap comm/comp
- Avoid comm by changing algorithm
- Optimize comp where needed
- target: 0.9 efficiency for ocean kernel
- SGI Origin 3000 not a limitation