a break through TM5 limits

"the TM6 project"

Ph. Le Sager, KNMI

Royal Netherlands Meteorological Institute (KNMI)
The Netherlands

2012-10-15 Mon

Outline

Outline

TM5 Limitations

Fast but not enough

1. EC-Earth : couple of decades max, no ensemble run
2. (very) Hi-Res slower than real time!
3. MPI Processor starvation > **27** or **1**

TM5 Limitations

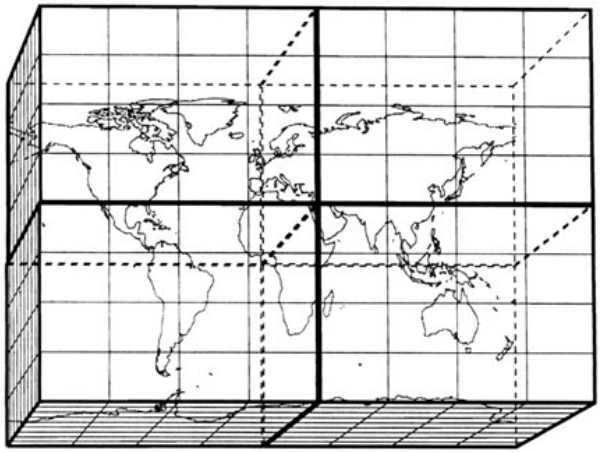
Fast but not enough

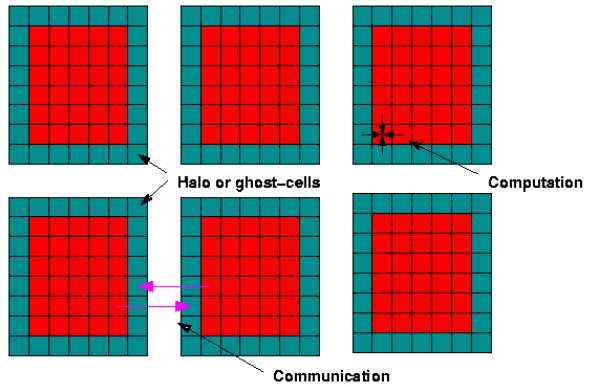
1. EC-Earth : couple of decades max, no ensemble run
2. (very) Hi-Res slower than real time!
3. MPI Processor starvation > **27** or **1**

Technically

1. Two MPI decompositions (levels, tracers)
   - add complexity: which_par, lmloc, duplicate code
   - add MPI comm: switching (>3 %)
2. Tracer decomp => meteo is **not** decomposed
   - heavy MPI communication: **half** runtime is in MPI_Bcast
   - large memory requirements (1x1: 10 Gb)

TM6 strategy : Revised domain decomposition





	TM5	TM6
processor starvation	27	30x22 = 660 (@6x4) 60x45 = 2700 (@3x2) 180*90=16200 (@1x1)
meteo communication	broadcast all	halo update (snd/rcv)

Outline

Structure

- INFRA
  - MPI domains defined
  - communications:
    - point-to-point (fill halo)
    - collective (gather, scatter)
    - semi-collective (eg scatter meridional data)
- SUPRA
  - test suite (TDD) for bitwise comparison of restart/output

Restart & Meteo

RESTART OPTIONS

- implemented: 1, 2, 30, 31, 33, 4, 5, 9
- tested : 33 (w/ read-write restart in parallel)

DECOMPOSED METEO, but

- read on 1 proc, then scattered
- works with all formats/source

Processes

All done!

- **advection**
- **convection**
- **diffusion**
- **wet dep**
- **dry dep**
- **chemistry**
  - **emissions**
  - **photolysis**
  - **M7**, incl. online dust [not tested]
- **sedimentation**
- **strat. boundary**

Outputs

Half done

- From BASE
  - **mmix**
  - **budgets** (incl. extra ‘Box’ fluxes)
- From PROJ / USER_OUTPUT
  - time-series (pdump)
  - **station** [not tested]
  - **mix** [not tested]
  - aerocom
  - **settings**
  - **planeflight** [not tested]
  - noaa

ToDo list

Test

- M7 & outputs: mix, station, planeflight
- debug : “1x8” case, “-qflttrap=enable:inv” required

ToDo list

Test

- M7 & outputs: mix, station, planeflight
- debug : “1x8” case, “-qfltrap=enable:inv” required

Code & test

- chunk reading of meteo in netCDF-4
- aerocom & time-series outputs
- EC-Earth proj
- updated chem emissions (edgar 4.2 + GFED3)

ToDo list

Test

- M7 & outputs: mix, station, planeflight
- debug : “1x8” case, “-qfilttrap=enable:inv” required

Code & test

- chunk reading of meteo in netCDF-4
- aerocom & time-series outputs
- EC-Earth proj
- updated chem emissions (edgar 4.2 + GFED3)

Missing features

reduced grid ; zoom regions

Outline

Model Set Up

Full chemistry (w/o M7)

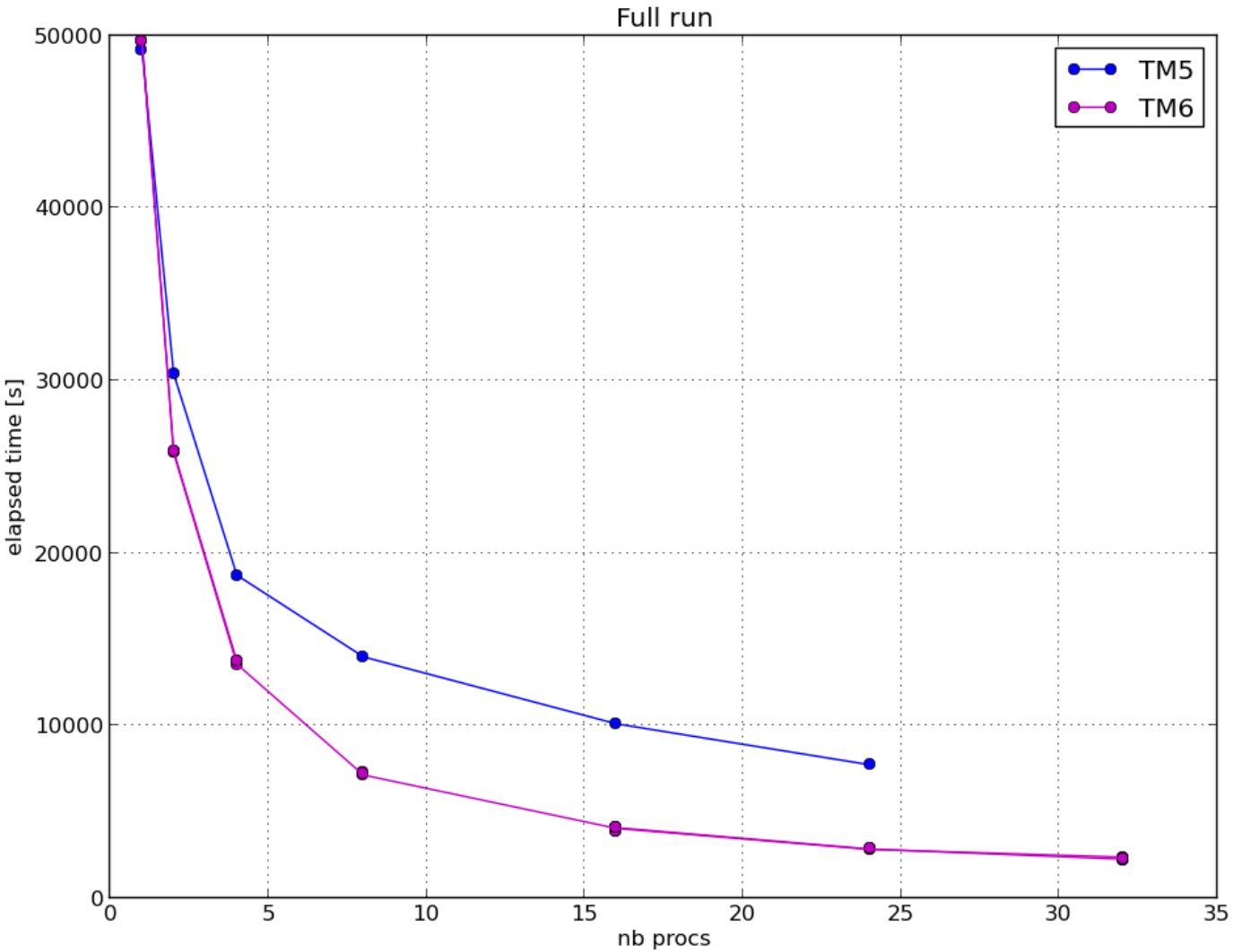
- summer 2012 trunk version
  - Edgar 4.1, AR5 (BB), new photolysis (no GFED3, Edgar4.2)
- output : mmix + profile + with_budgets
- everything on (**no** without_*)

- 3x2, 34-levels
- meteo : ei, glb100x100, tm5-nc

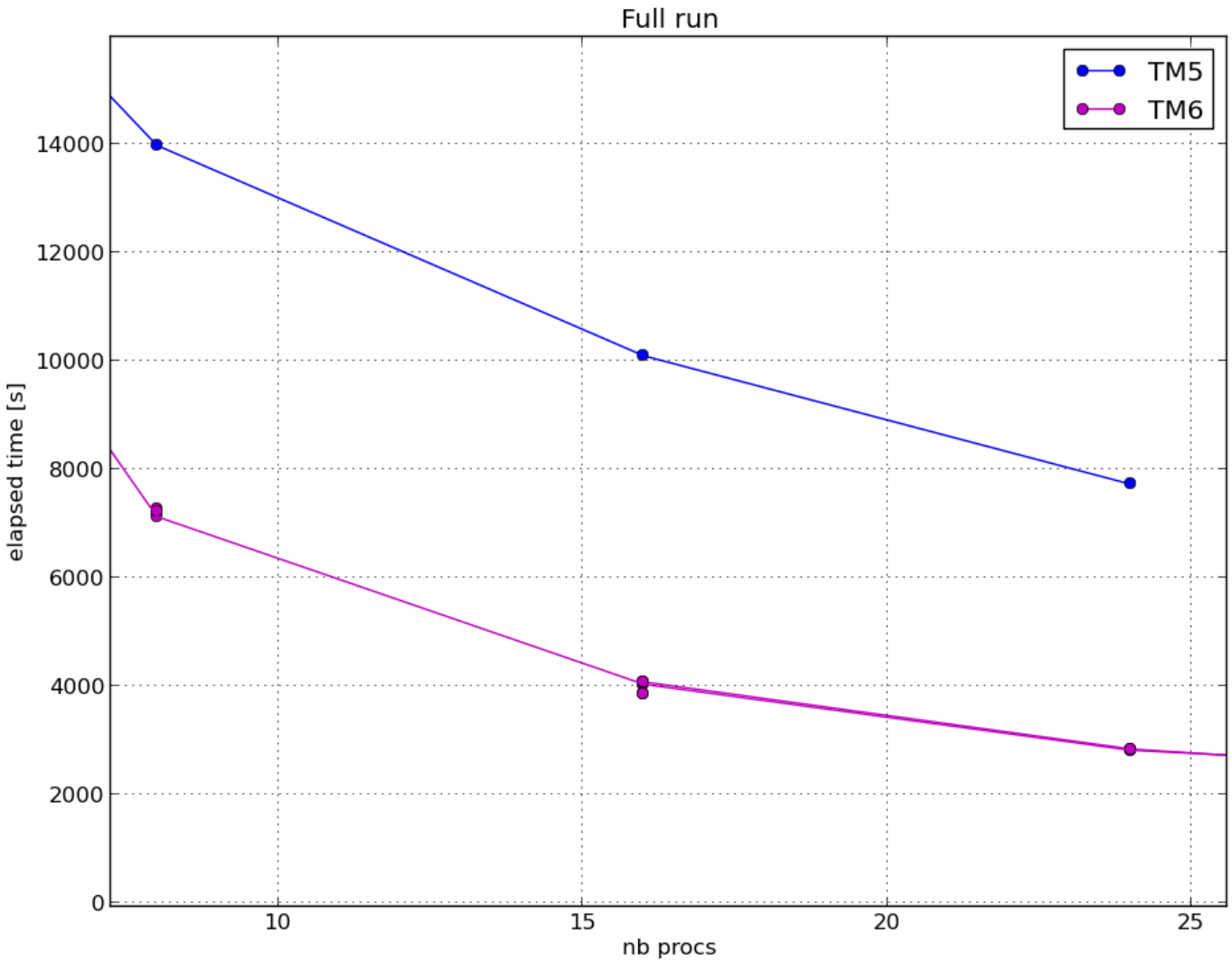
Runs Set Up

- 4-days runs
- all combinations from
  - **1, 2, 4, 8, 16, 24, 32** procs along Lon./Lat.
  - limited to 32
- TM5 => 7 runs (1 failed: 32)
- TM6 => 23 runs (1 failed: 1x8)

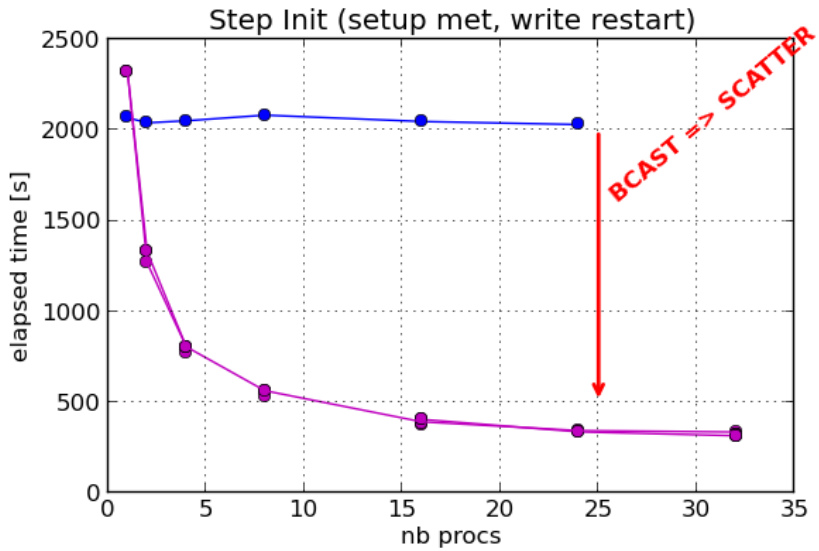
Overall perf

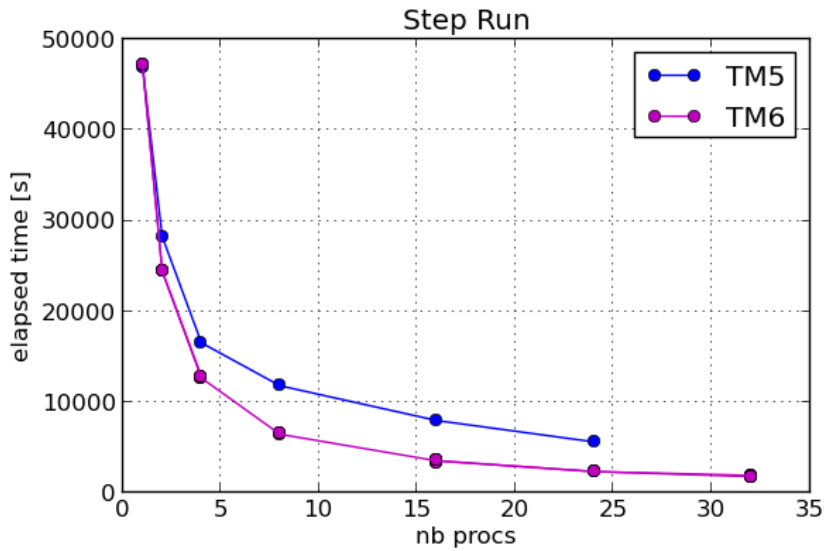


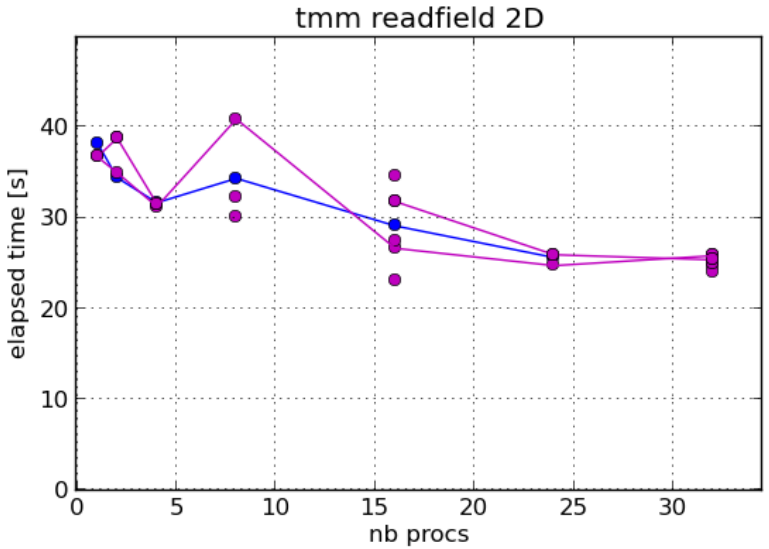
Overall perf (zoom)

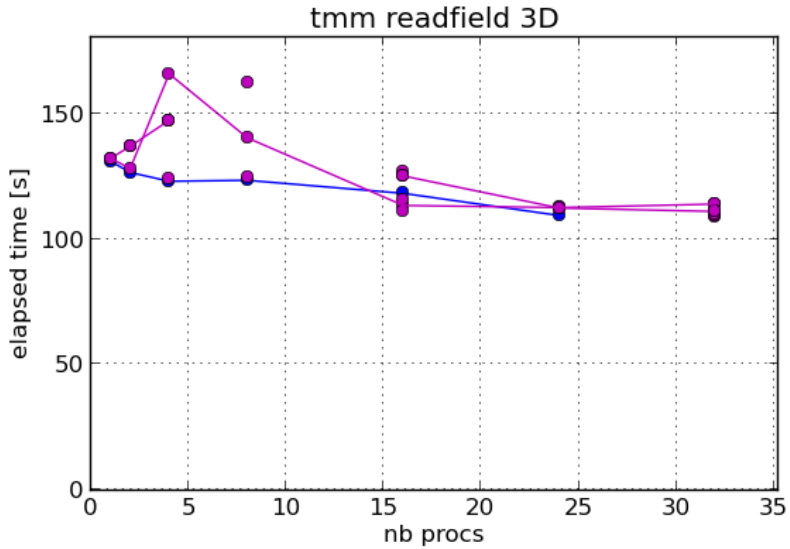


Overall detailed

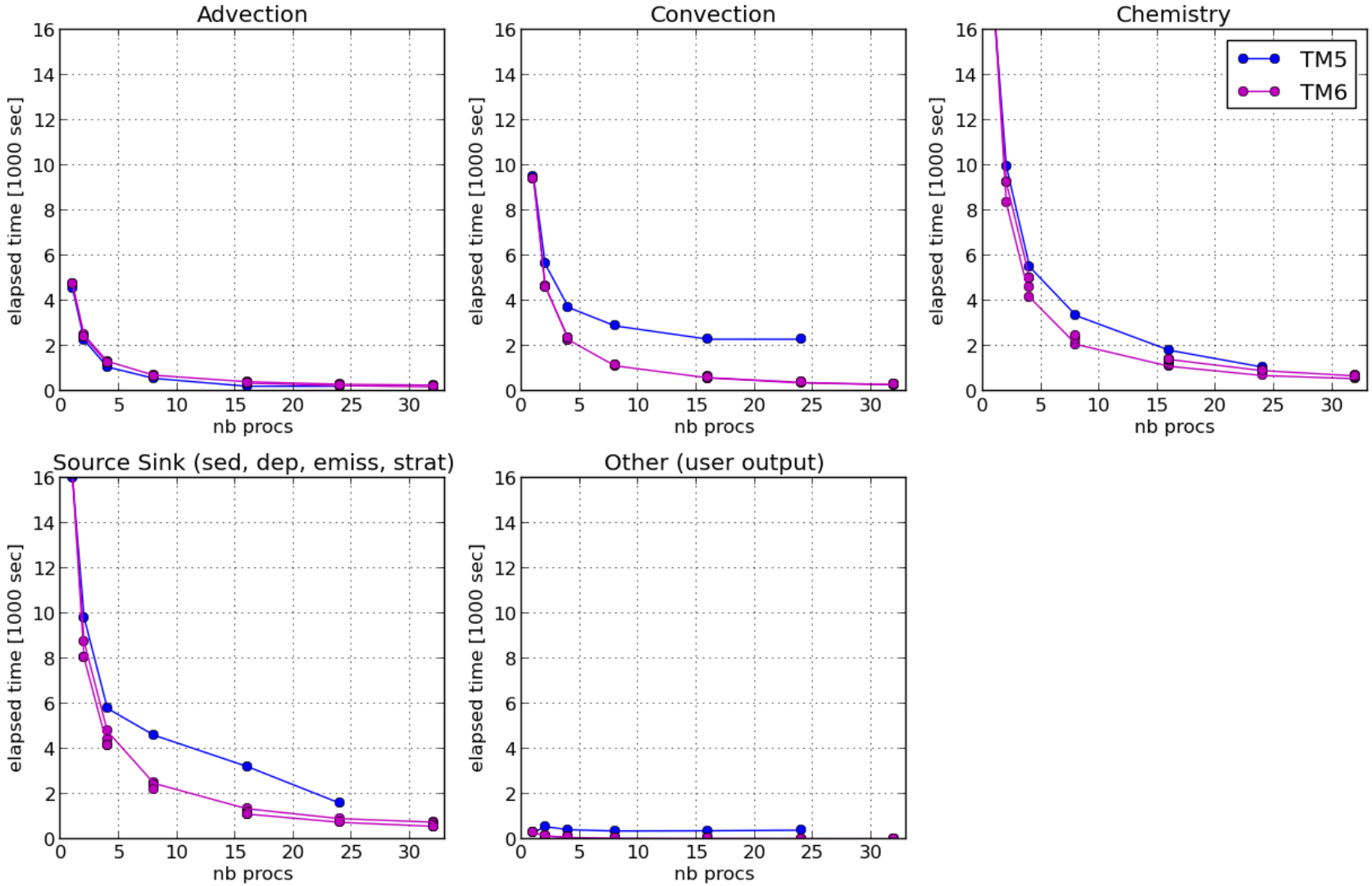




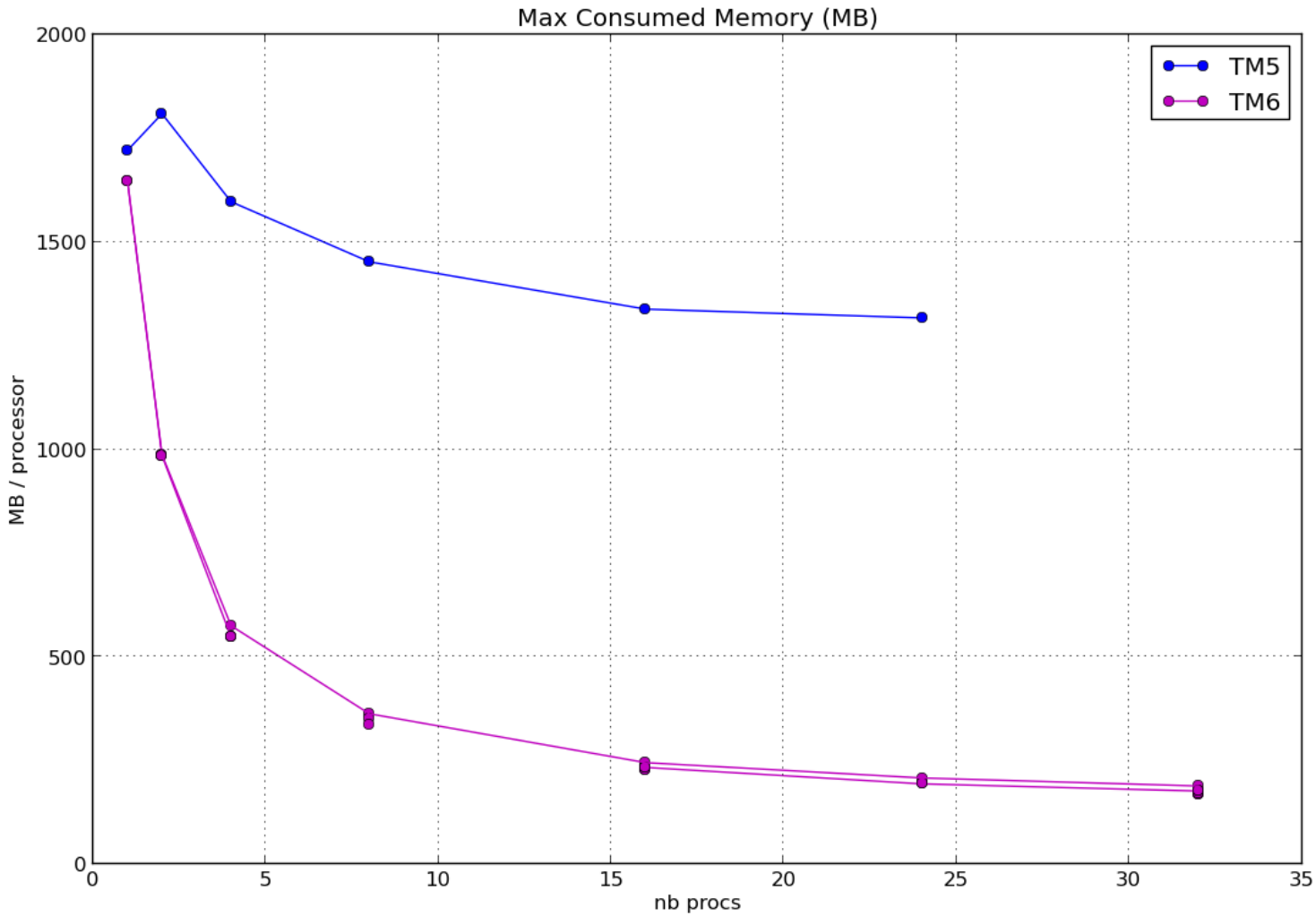




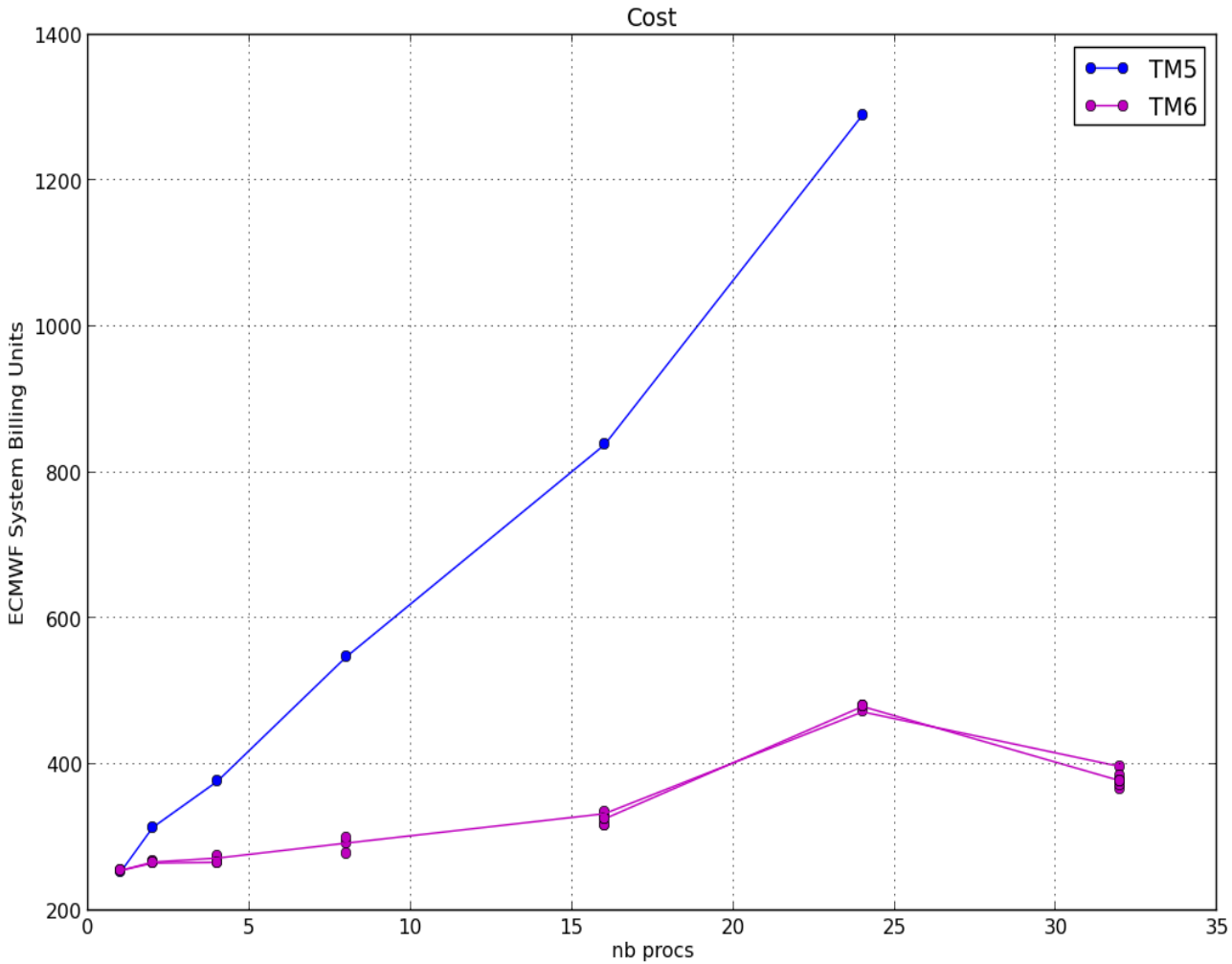
Inside Step run



Memory

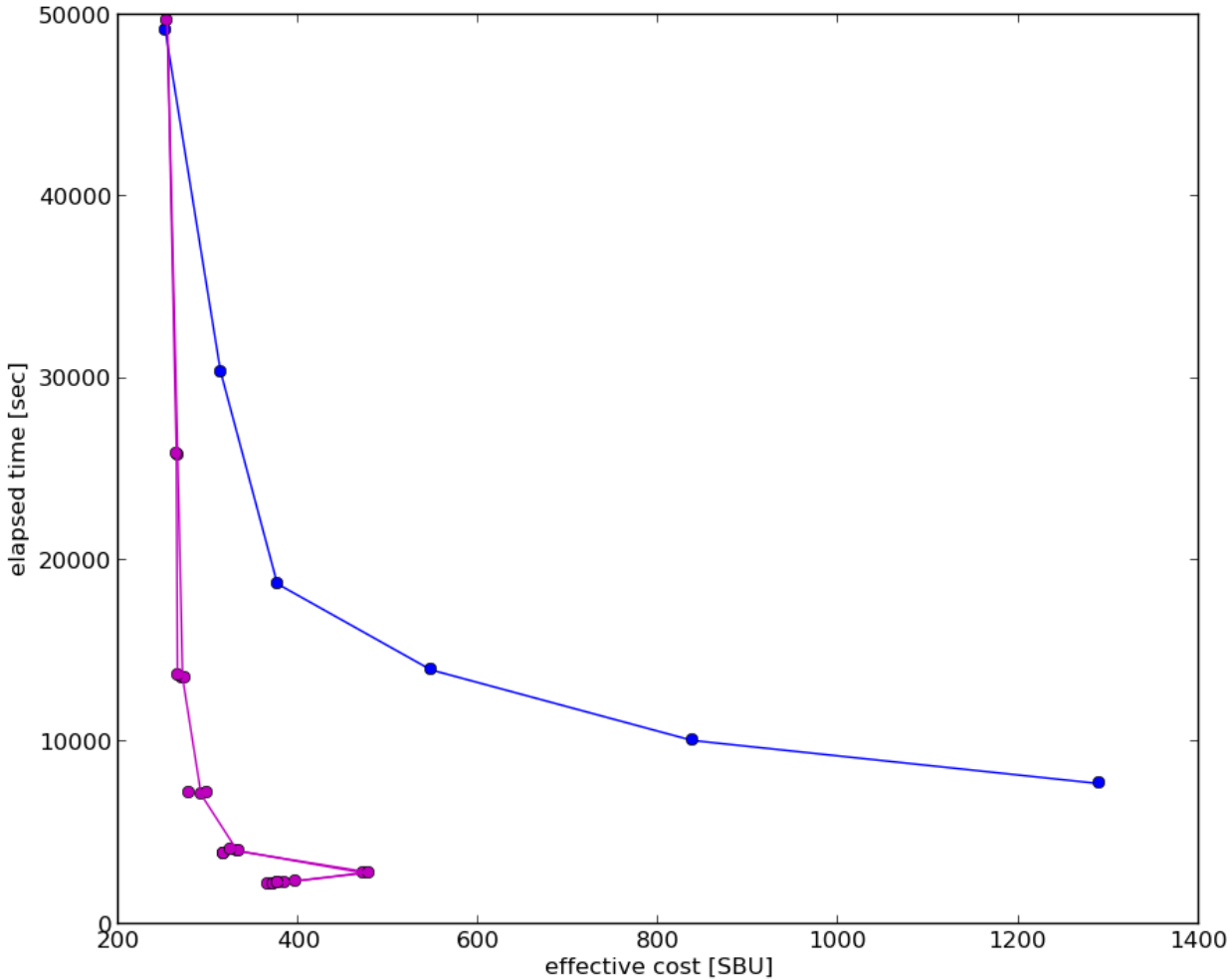

Max Consumed Memory (MB)
TM5
TM6
MB / processor
nb procs
0
500
1000
1500
2000
0
5
10
15
20
25
30
35

Cost



Same cost
TM6 32 cpu
TM5 4 cpu

Overall Perf #2



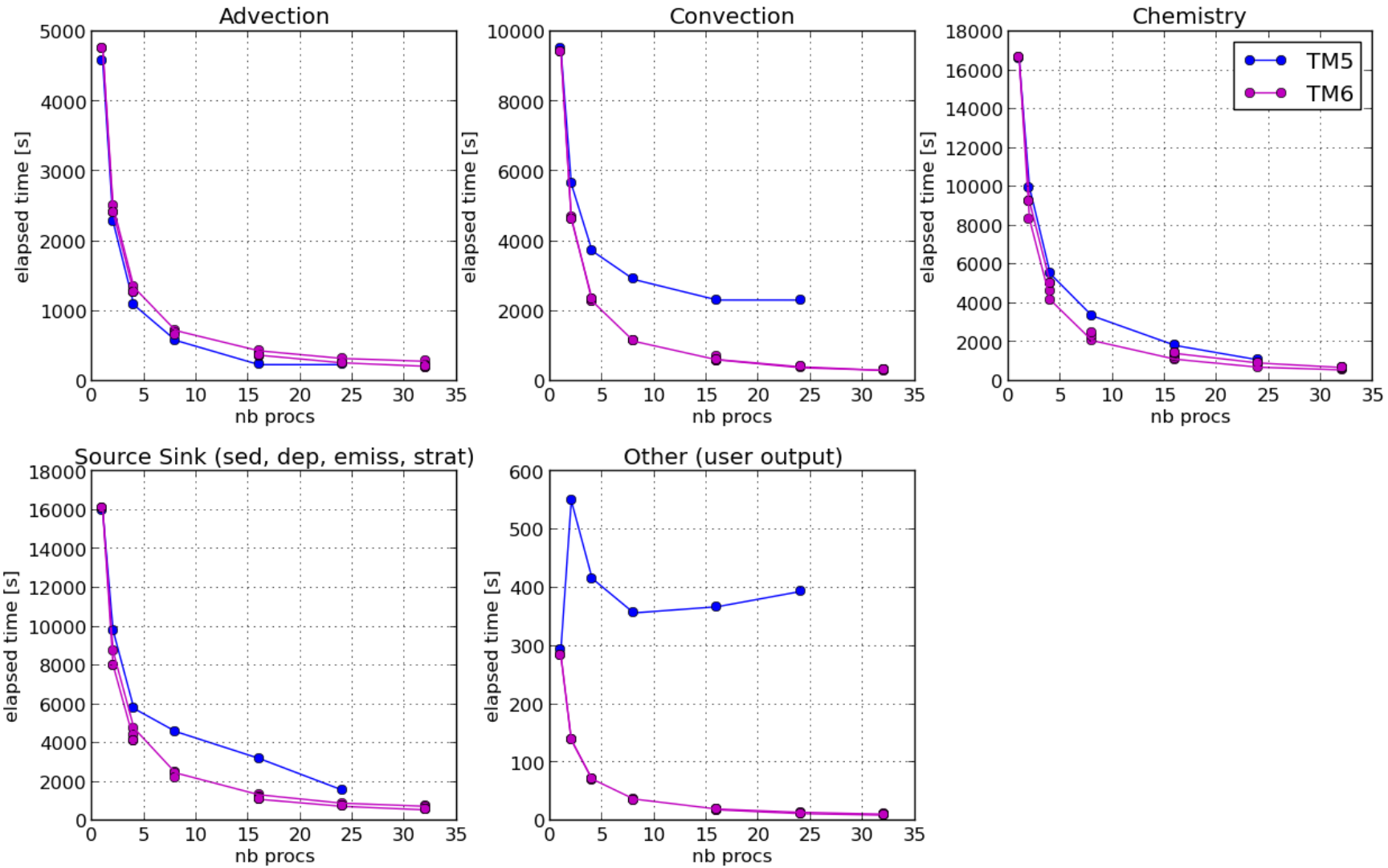
8x faster !
same price!

CONCLUSION

- huge gain
  - 7x less memory
  - **faster** meteo setup, convection, mmix
- w/r/t procs
  - 2.5 x **faster**
  - 60% speed up
- w/r/t ressources
  - 8 x **faster**
  - 87% speed up
  - HIGHER LIMITS. . . more procs, higher-res

Outline

Step run - shows perf of output_mmix_step

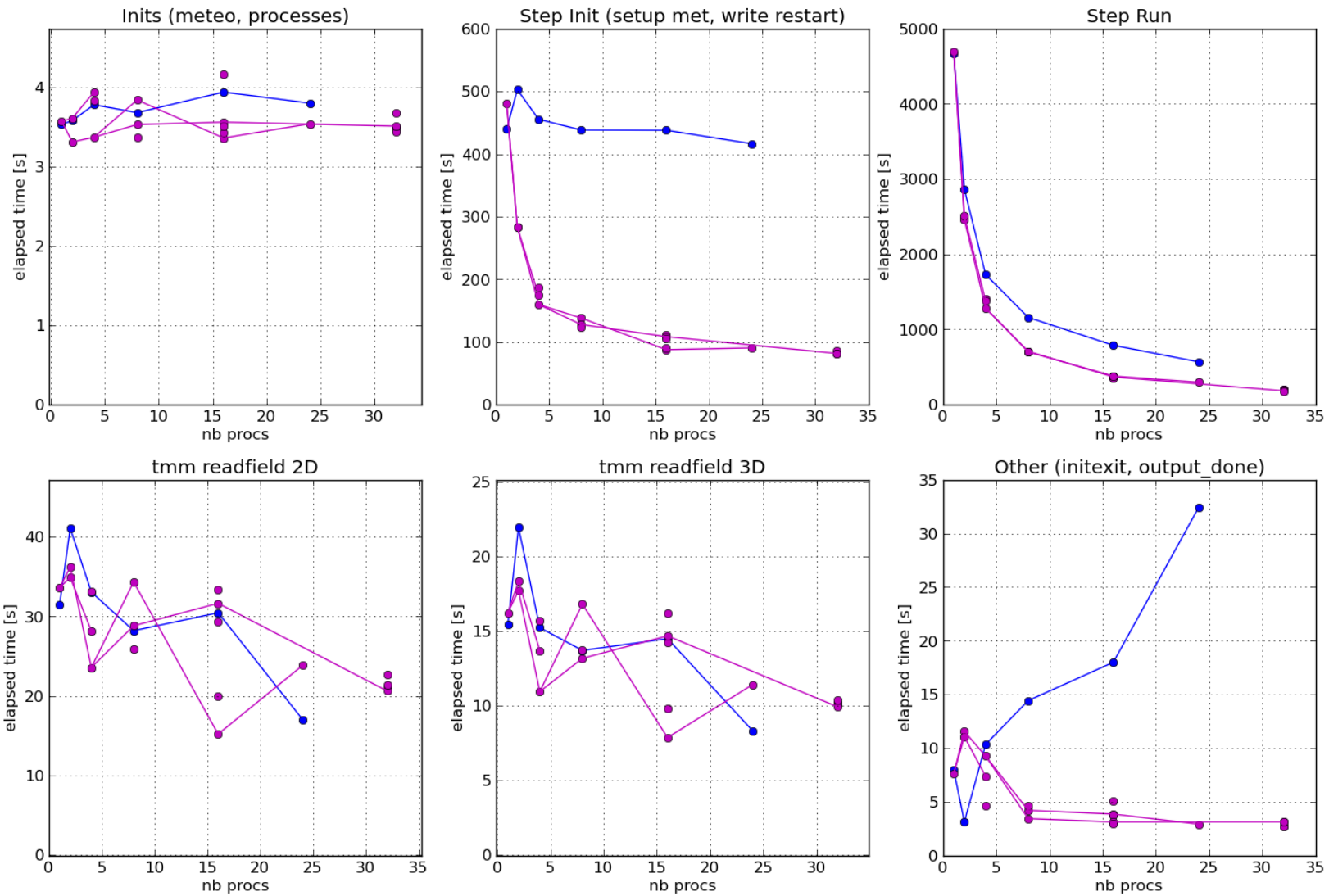


Experiment #2 - 6x4, coarsened meteo

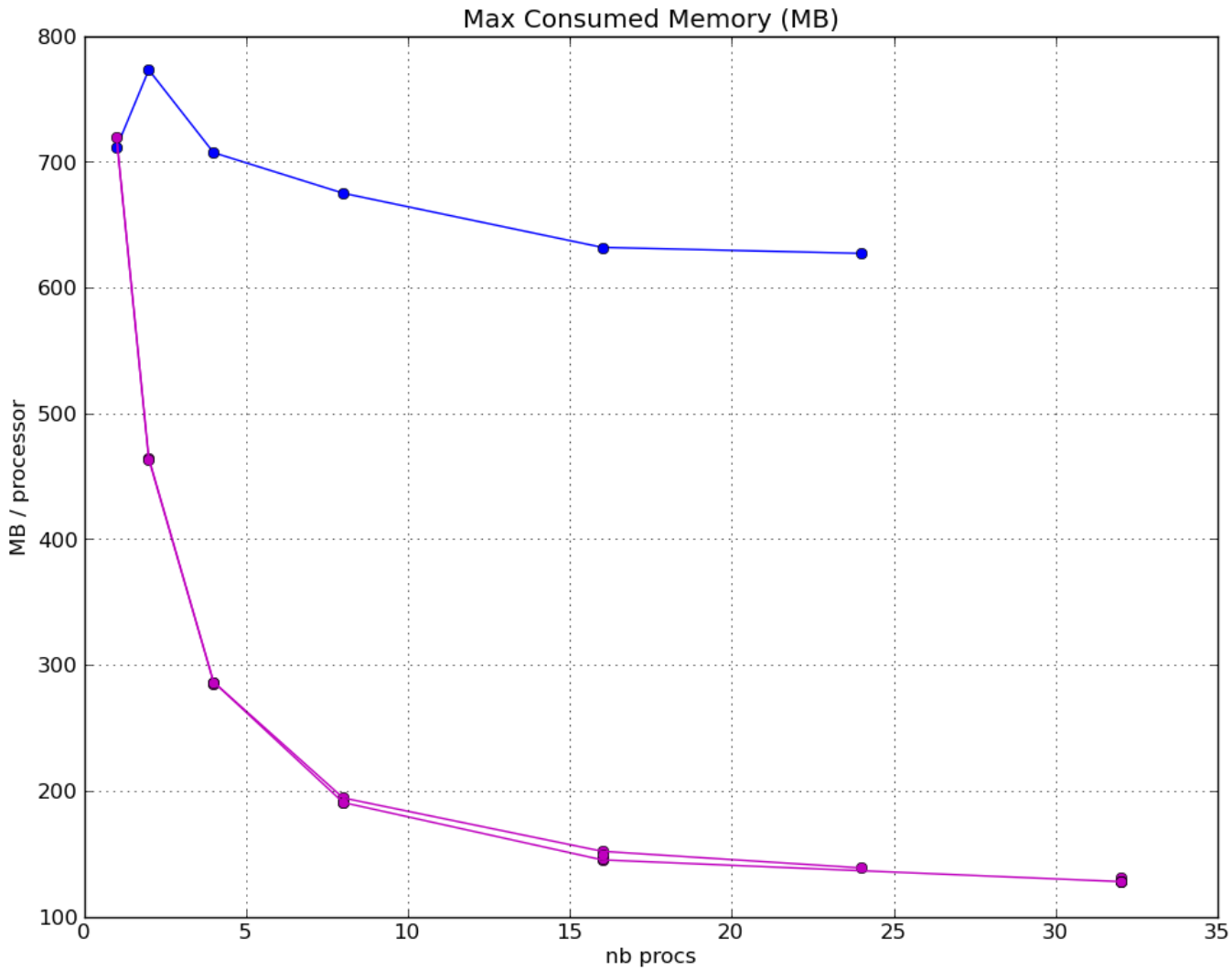
full chemistry (w/o m7)

- same as experiment #1, except:
  - **6x4** instead of 3x2 res.
  - meteo : **coarsened** instead of glb100x100
- TM5 => 7 runs (1 failed: 32)
- TM6 => 23 runs (2 failed: 1x32, 1x8)

Exp. #2 - Overall detailed



Exp. #2 - Memory


Max Consumed Memory (MB)
MB / processor
nb procs
800
700
600
500
400
300
200
100
0
5
10
15
20
25
30
35