Status of TM6

Philippe Le Sager, KNMI

2013-04-25

Outline

Context

Previously in TM6

TM6 Status & Perf

Extra

Outline

Context

Previously in TM6

TM6 Status & Perf

Extra

Performance issue

speed

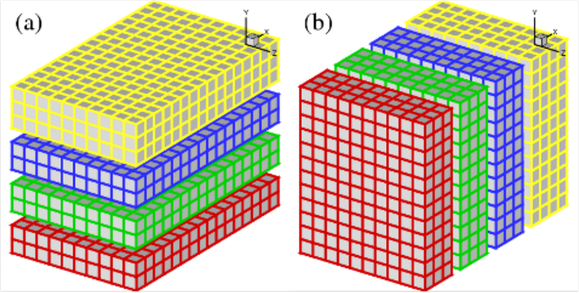
- fast as standalone, but slow for CGCM (EC-Earth)
  - decade/day wanted
  - BUT max nb processors = nb Tracers (**27**, **1**,..)

resolution

- high resolution
  - very demanding in memory (10 Gb/proc @ 1x1)

Basic idea of MPI: domain decomposition

- Arrays are split across processors, along any dimension.



- TM5 **4D MASS** arrays are distributed along either LEVELS or TRACERS.

The bottleneck

meteo fields are NOT distributed… but COPIED !!!!

- every 3h => **FREQUENT** communication
- 50+ met fields
  - **HUGE** memory requirement
  - **HEAVY** communication

MPI profiling

TM5-chem v3, 2-days run, 4 MPI tasks

	% of elapsed time
switching decomposition	3 %
broadcasting meteo	**50 %**
other MPI comm	2 %
total MPI comm	**55 %**

Outline

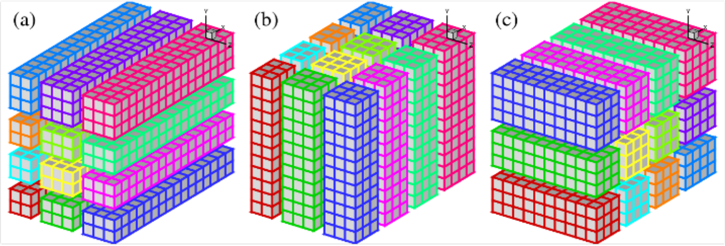
Context

Previously in TM6

TM6 Status & Perf

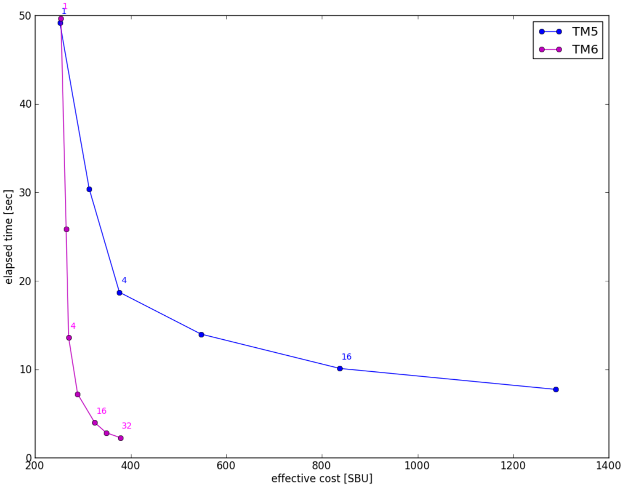
Extra

Revised domain decomposition = (b)



	TM5	TM6
max #processor	27	30x22 = 660 (@6x4) 60x45 = 2700 (@3x2) 180*90=16200 (@1x1)
meteo communication	broadcast	scatter

Performance TM-chemistry



8x faster
same price

ToDo list as of last Crete meeting

To test/fix

- **M7**, ~~online dust~~ & outputs: ~~mix~~, station, planeflight
- ~~debug : “1x8” case, “-qflttrap=enable:inv” required (EBI)~~

To 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

Context

Previously in TM6

TM6 Status & Perf

Extra

Porting to ECMWF/c2a (IBM/AIX power7)

Fixed:

- Pure MPI-2 :
  - MPI_GET_EXTENT –> MPI_TYPE_GET_EXTENT
  - MPI_TYPE_HVECTOR –> MPI_TYPE_CREATE_HVECTOR
- libs
- totalview requires ssh

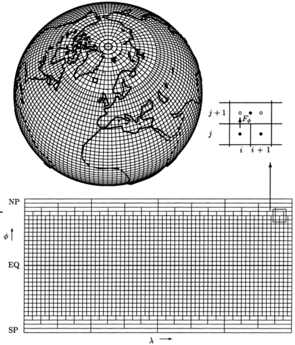
but still issues

- unexplained frozen runs
- M7 : crashes w/ 5+ cpus, sedimentation bug

Reduced grid - implementation

first question at every talk!

- implement case of '*no decomposition along longitudes*'

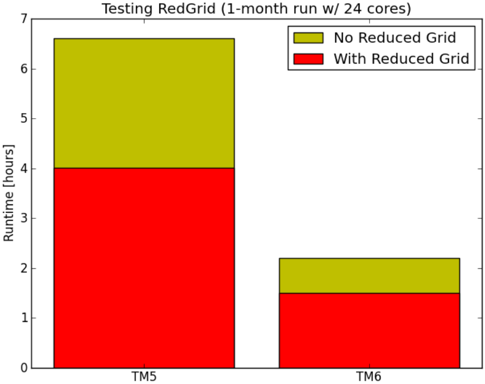


TM5	TM6	TM6 w/ redgrid
27	660 (@6x4)	22 (@6x4)
	2700 (@3x2)	45 (@3x2)
	16200 (@1x1)	90 (@1x1)

Max #processors

Reduced grid // 1-month runs // chemistry w/o M7

- 3x2 w/ reduced grid
  - 7 bands (90-74) at each pole
  - merging [40, 8, 8, 4, 4, 4, 2] cells



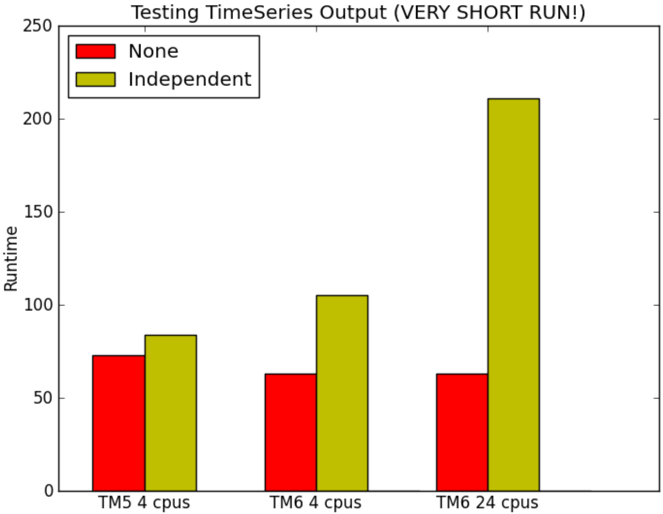
- TM5 —> TM6 : 60-70 % speed-up
- Reduced grid : 30-40 % speed-up

The I/O experiment (1) - Time-Series Output

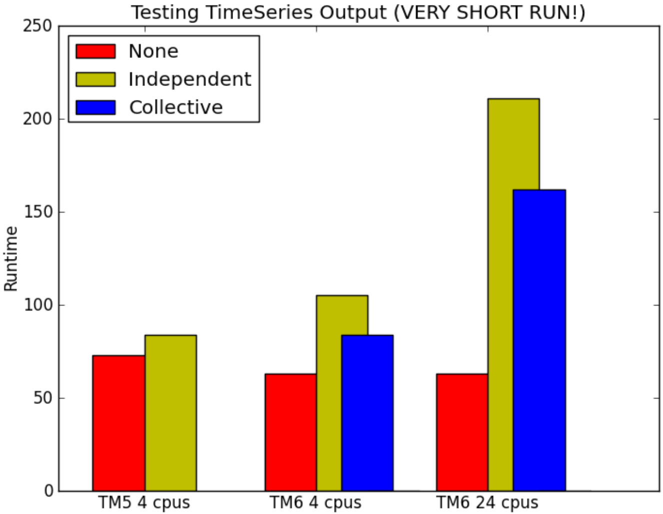
(former RETRO output)

- TM5
  - pnetcdf –> netcdf4 (MDF)
  - INDEPENDENT access mode <= **unlimited dimensions**
- TM6
  - case 1: stick to INDEPENDENT
  - case 2: switch to COLLECTIVE access mode
  - case 3: write once a day (not every time step!)

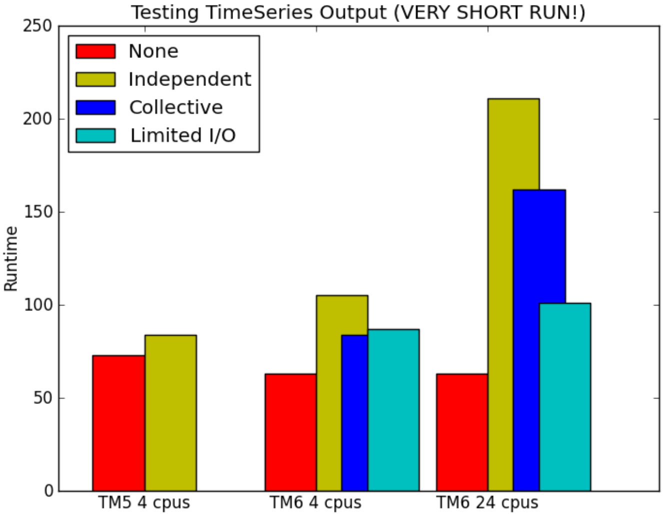
The I/O experiment (1) - Time-Series Output



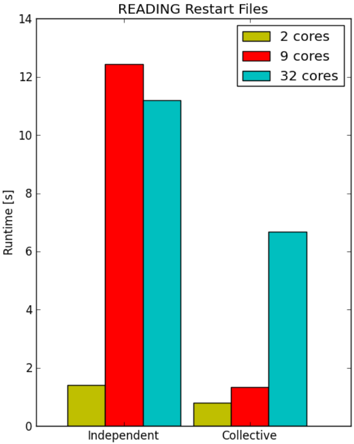
The I/O experiment (1) - Time-Series Output



The I/O experiment (1) - Time-Series Output



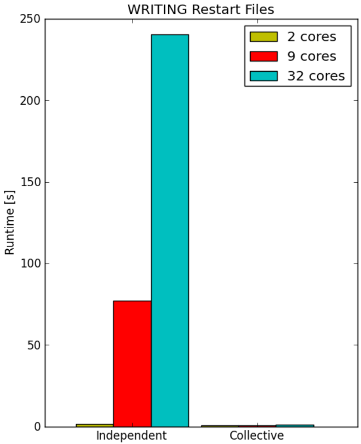
The I/O experiment (2) - Does netCDF4 w/ parallel I/O scale?



READING restart

- collective **faster** than independent (1.5-9x)
- **time increases w/ nb cores**
  - impact for meteo (**must** account for the scatter time saved)

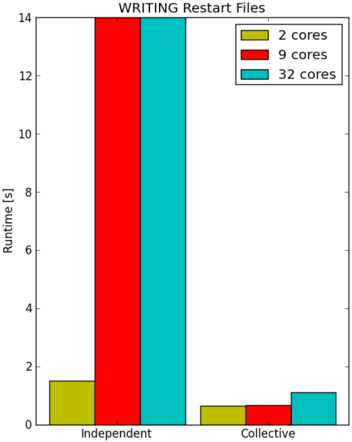
The I/O experiment (2) - Does netCDF4 w/ parallel I/O scale?



WRITING restart

- collective **really faster** (2.3-110x)

The I/O experiment (2) - Does netCDF4 w/ parallel I/O scale?



WRITING restart

- collective **really faster** (2.3-110x)
- writing time : no increase with nb cores

I/O next steps - Optimize

- Time-Series Output
  - test w/ longer runs
  - one file /month & /tracer instead of /day & all tracers?
  - **quilting** : asynchronized I/O for MPI (eg, WRF)?
  - file splitting?
- Read/Write restart
  - file splitting
  - quilting
- Meteo Input
  - switch to parallel reading

"As-fast-as-you-can" experiment @1x1

one node (50 Gb, 32 procs max), one day sim, no reduced grid

Model	**Regions**	**Resources**	**Runtime**	**Cost** (SBU)
TM6	global 1x1	32 procs	1h 4mn	643
TM5	global 1x1	6 procs	11h 25mn	6850
TM5	global 3x2 + euro 1x1	*broken*	-	

- zooming broken in TM5 in 3 places:
  - when nudging of CH4 emissions
  - in photolysis
    - latitudinal decomposition
    - solar zenith angle

"As-fast-as-you-can" experiment @1x1

one node (50 Gb, 32 procs max), one day sim, no reduced grid

Model	**Regions**	**Resources**	**Runtime**	**Cost** (SBU)
TM6	global 1x1	32 procs	1h 4mn	643
TM5	global 1x1	6 procs	11h 25mn	6850
TM5	global 3x2 + euro 1x1	*broken*	-	

SUCCESS!

- **10.6 x cheaper**
- **10.6 x faster!**
- **90.6% speedup**

NEXT

- fix M7
- couple to EC-Earth

- optimize reduced grid
- optimize time-series
- read netCFD meteo in parallel

Outline

Context

Previously in TM6

TM6 Status & Perf

Extra

Reduced grid - 1-month runs by numbers

- Chemistry w/o M7
  - optimized (-O3 -qstrict)
  - full chemistry (but w/o m7)
  - w/o time-series output
  - 3x2 w/ reduced grid:
    - 7 bands (90-74) at each pole
    - merging [40, 8, 8, 4, 4, 4, 2] cells

	w/o redgrid	w/ redgrid	speed-up
TM5	23799	14422	**39%**
TM6	7909	5401	**32%**
speed-up	**67%**	**63%**	77%

The I/O experiment (2) - by numbers

- Reading restart

	8x4 cpus	3x3 cpus	2x1 cpus
coll.	6.35	1.70	0.84
	7.00	0.99	0.75
ind.	10.99	12.30	1.33
	11.40	12.57	1.51

- Writing restart

	8x4 cpus	3x3 cpus	2x1 cpus
coll.	1.01	0.75	0.64
	1.23	0.59	0.65
ind.	237.19	73.28	1.51
	243.57	81.26	1.50