CONSORTIUM FOR SMALL SCALE MODELING

COSMO

Deutscher Wetterdienst

Wetter und Klima aus einer Hand

DWD

Scalability and Performance of COSMO-Model 5.1 on Cray XC30

Ulrich Schättler

Source Code Administrator

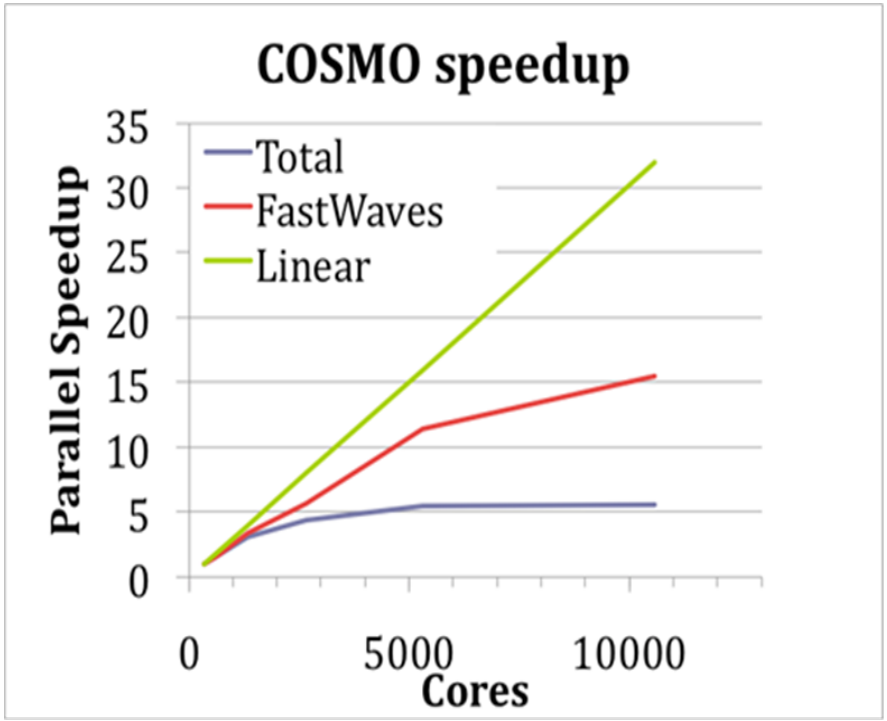
COSMO-Model

Contents

- Old Scalability Results
- Latest Changes
- Scalability Tests with COSMO-DE65 (651 x 715 x 65)
- First Conclusions (for NWP)
- What about the CCLM
- Performance Counters

Old Scalability Results

- From HP2C Report: „Performance Analysis and Prototyping of the COSMO Regional Atmospheric Model“ (Matthew Cordery, et al.)
- „We note the poor parallel scaling characteristics of COSMO beyond 1000 cores“
- Parallel speedup of COSMO for a 1-hour simulation on 1 km grid
- which domain size was used?



Latest Changes

- The old tests were done using COSMO_RAPS_4.10 version, the benchmark version based on COSMO-Model Version 4.10 (from 11th September 2009).
- Since then, quite a few things happened:
  - new strong conservative fast waves solver
  - more stable advection schemes (Strang-splitting)
  - new COSMO-ICON microphysics with new 2D (blocked) data structure and copy to/from block structure.
  - (at DWD) use RTTOV10 to compute synthetic satellite images
  - modified module mpe_io2.f90 for asynchronous GRIB I/O (now including prefetching of data – but not yet tested)
  - new module netcdf_io.f90 for asynchronous and parallel NetCDF I/O

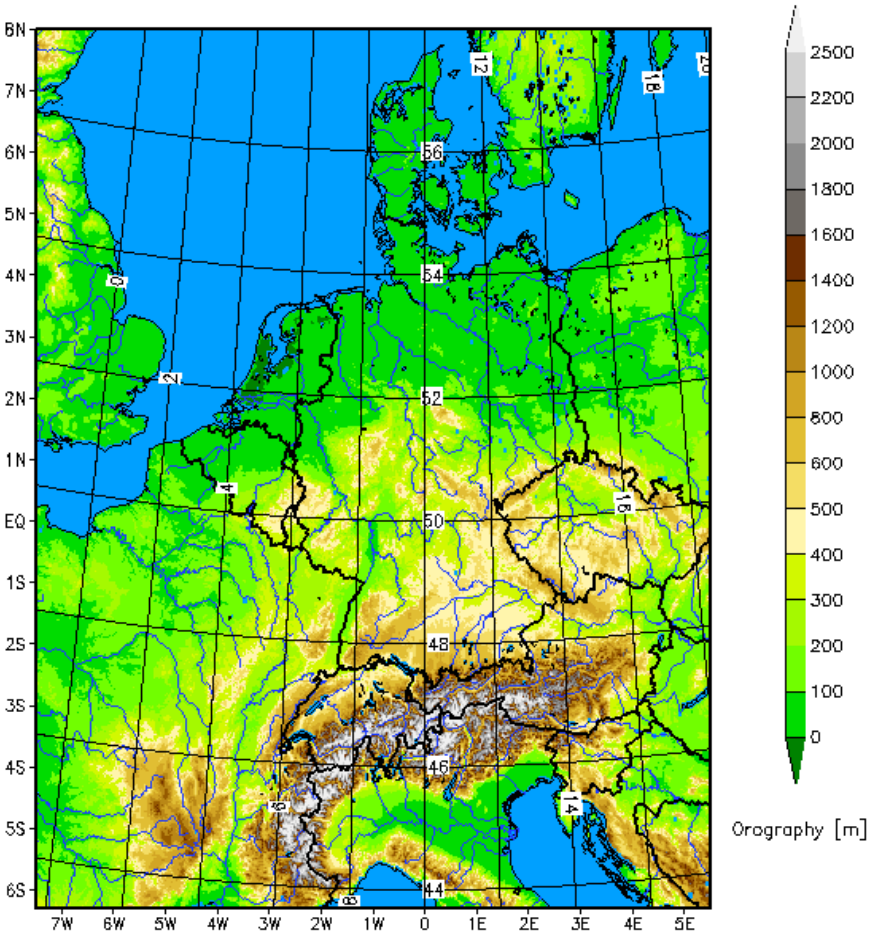
Latest Changes

- DWD now runs a Cray XC30 and no more NEC SX-9: for the second time in history we are running a machine with more than 1000 cores:
  - 2003/2004: IBM Power3 with nearly 2000 processors (but no application really used more than a few hundred processors)
  - since December 2013 (Phase 0): Cray XC30 with 364 nodes, each having 2 Intel Ivy-Bridge CPUs with 10 cores: 7280 cores
  - from December 2014 (Phase 1): extension to 784 nodes, but with mixed Ivy-Bridge and Haswell CPUs; in total: 17488 cores
  - Performance grows by a factor of 3: 15 members of a big COSMO-DE are running a 12 hour forecast in 1200 seconds on Phase 0 machine, 45 members will run on Phase 1 machine.

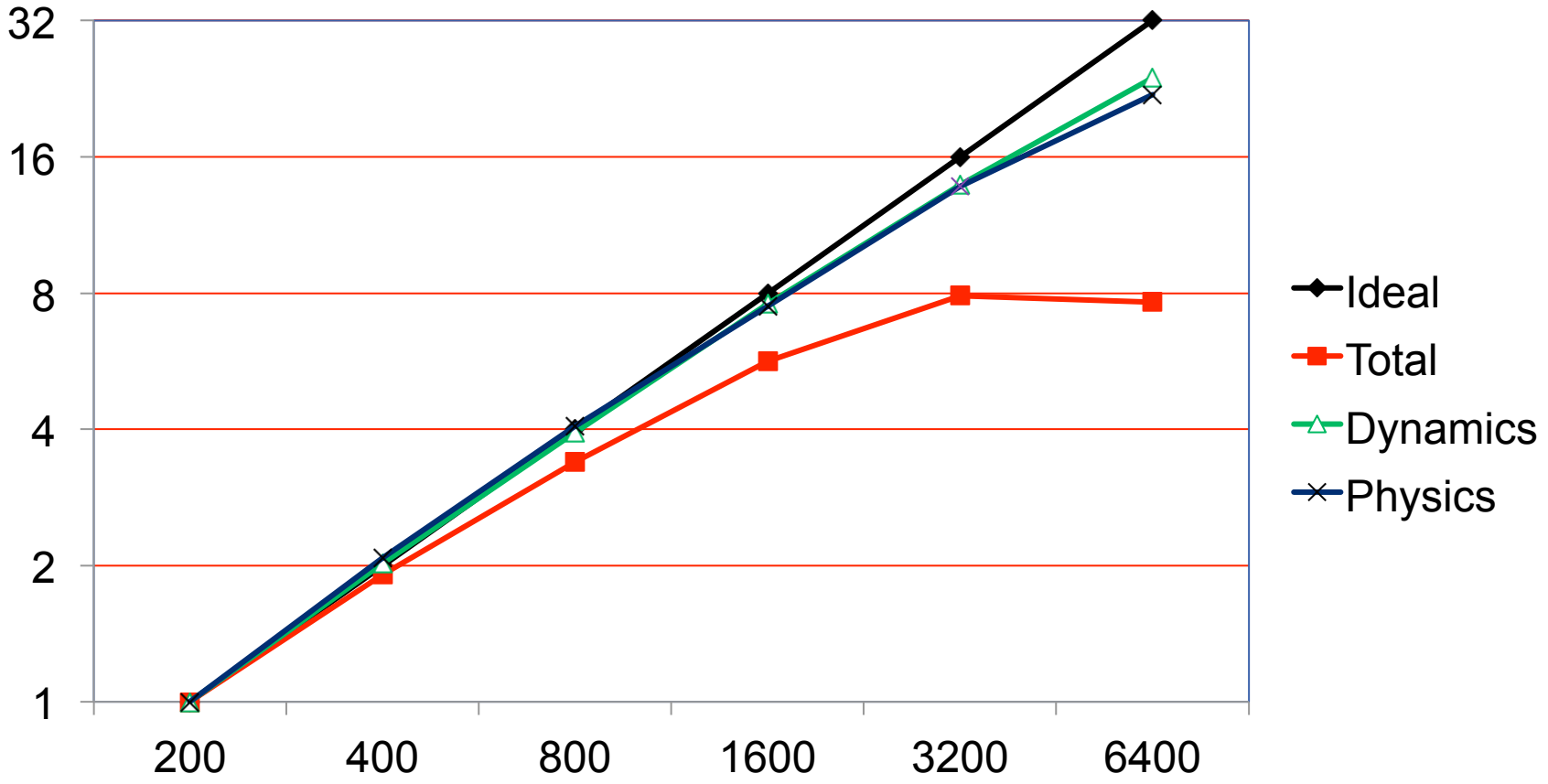
- how much cores do we need for COSMO?

The new COSMO-DE65

- In 2015, DWD will upgrade the COSMO-DE to a larger domain and 65 vertical levels: $651 \times 716 \times 65$ grid points
- Test Characteristics
  - 12 hour forecast should run
    - in ≤ 1200 s in ensemble mode
    - in ≤ 400 s in deterministic mode
  - „nudgecast“ run: nudging and latent heat nudging in the first 3h
  - SynSat pictures every 15 minutes
  - amount of output data per hour: 1.6 GByte: asynchronous output is used with 4 or 5 output cores



Scalability of COSMO-Model 5.1 for COSMO-DE65



Timings for COSMO-DE65

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Dynamics	1848.19	913.37	469.68	244.10	132.93	77,22
Dyn. Comm.	259.57	137.55	90.41	49.38	30.79	21.19
Physics	326.02	156.66	80.08	43.51	23.68	14.82
Phy. Comm.	17.08	9.92	5.41	3.44	2.52	1.89
Copying	19.26	9.21	4.71	2.25	1.03	0.47
Nudging	43.05	25.00	15.22	11.92	14.79	38.66
Nud. Comm.	27.92	34.48	29.73	35.48	49.62	77.72
Add. Comp.	726.64	400.47	216.37	117.67	54.84	27.87
Input	22.49	21.60	29.34	31.91	36.40	47.18
Output	33.75	25.22	24.06	29.33	47.62	94.40
Total	3333.62	1744.74	982.72	589.99	422.14	436.68

Timings for COSMO-DE65

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Dynamics	1848.19	913.37	469.68	244.10	132.93	77,22
Dyn. Comm.	259.57	137.55	90.41	49.38	30.79	21.19
Physics	326.02	156.66	80.08	43.51	23.68	14.82
Phy. Comm.	17.08	9.92	5.41	3.44	2.52	1.89
Copying	19.26	9.21	4.71	2.25	1.03	0.47
Nudging	43.05	25.00	15.22	11.92	14.79	38.66
Nud. Comm.	27.92	34.48	29.73	35.48	49.62	77.72
Add. Comp.	726.64	400.47	216.37	117.67	54.84	27.87
Input	22.49	21.60	29.34	31.91	36.40	47.18
Output	33.75	25.22	24.06	29.33	47.62	94.40
Total	3333.62	1744.74	982.72	589.99	422.14	436.68

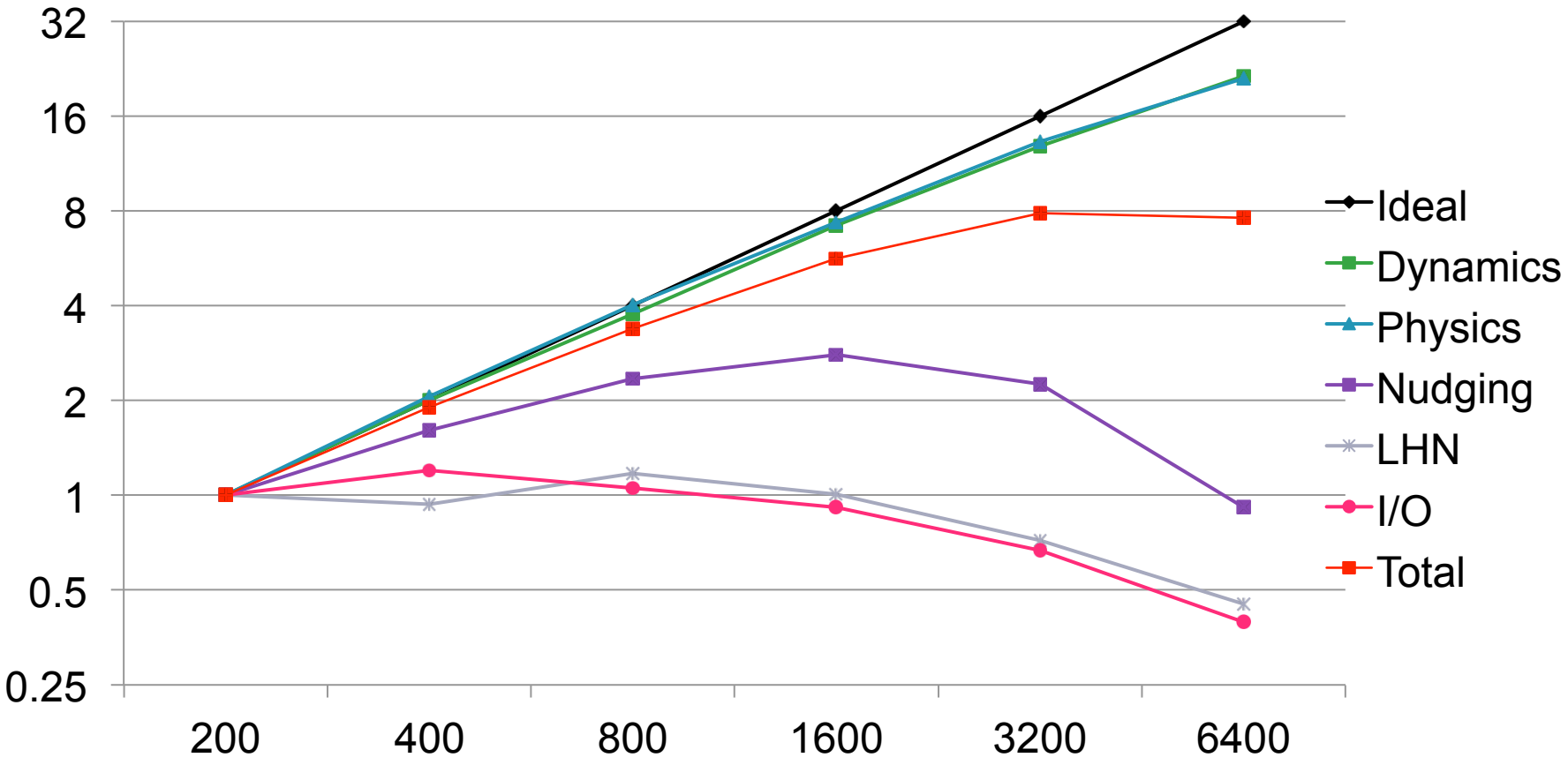
Timings for COSMO-DE65

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Nudging	26.81	16.22	9.24	6.50	5.83	15.13
Nud. Comm.	3.37	2.45	2.60	2.36	3.36	9.08
Nud. Barrier	5.57	3.61	3.36	3.92	6.66	14.92
Latent Heat N.	16.24	8.78	5.98	5.42	8.96	23.53
LHN Comm.	18.65	28.42	23.77	29.20	39.60	53.72
Nud. Total	70.64	59.48	44.95	47.40	64.41	116.38
Add. Comp.	726.64	400.47	216.37	117.67	54.84	27.87
Total	3333.62	1744.74	982.72	589.99	422.14	436.68
~ #gp /core	2330	1165	582	291	145	72

Timings for COSMO-DE65

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Input	22.49	21.60	29.34	31.91	36.40	47.18
read data	8.57	7.59	14.03	18.31	21.74	22.10
meta data	6.29	5.52	6.24	3.52	2.07	1.02
compute Input	0.72	2.34	9.07	10.08	12.58	24.06
distribute data	6.91	6.14	0.00	0.00	0.00	0.00
Output	33.75	25.22	24.06	29.33	47.62	94.90
compute Output	16.58	9.67	6.68	9.05	22.77	62.35
meta data	0.66	0.33	0.16	0.12	0.06	0.03
write data	0.14	0.08	0.04	0.02	0.01	0.01
gather data	16.35	15.12	17.17	20.13	24.77	32.50
Total	3333.62	1744.74	982.72	589.99	422.14	436.68

Scalability of COSMO Components (incl. Comm.)



First Conclusions

- Scalability of COSMO-Model for COSMO-DE65 domain size is reasonably well up to 1600 cores. Dynamics and Physics also scale beyond up to 6400 cores.
- Meeting the operational requirements:
  - for ensemble mode about 650 cores would be necessary to run a 12 hour forecast in less than 1200 seconds. But then 40 members will not fit in Phase 1 machine
  - for deterministic mode, it is not possible to run in less than 400 seconds.
- This is not a problem of the scalability, but of some expensive components!

First Conclusions (II)

- Expensive Components:
  - New fast-waves solver is more expensive than old one (40-50% of dynamics time; but not investigated further up to now)
  - Communication in the Latent Heat Nudging
  - Additional Computations: is almost only in RTTOV10
    - factor of about 10-15 compared to RTTOV7
    - very imbalanced computations, perhaps due to cloud characteristics
    - much effort for some diagnostic pictures
- Tests were done on a „usual crowded“ machine and really reflect the operational setups (no tricks, no cheating, no beautifying)

What about the CCLM

- During climate simulations
  - You do not compute nudging or latent heat nudging
  - You do not compute the synthetic satellite images: timings for additional computations will drop down to about 10% of timings shown before
  - You will do less output: only about 60% of output amount from NWP output

- How long would a simulation for 150 years take? (150 years are about 54790 forecast days)

Estimations for CCLM using COSMO-DE65 Size

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Dynamics	2107.76	1050.92	560.09	293.48	163.72	98.41
Physics	362.36	175.79	90.20	49.20	27.23	17.18
Add. Comp.	72.66	40.05	21.64	11.77	5.48	2.79
Input	22.49	21.60	29.34	31.91	36.40	47.18
Output	20.25	15.13	14.44	17.60	28.57	56.64
Total	2585.52	1303.49	715.71	403.96	261.40	222.20
Forecast days per day	16.7	33.14	60.36	106.94	165.26	194.42
Days for 150 year simulation	3281 9 years	1654	908	512	332	282 9 months

Conclusions for CCLM

- Convection permitting climate simulations are still rather expensive, but not „out of sight“ on today‘s HPC platforms.
- Times for „Additional Computations“ and „Output“ are only estimated, not measured, in the table before. All other timings taken from NWP tests.

Performance Counters

- We made also some runs on our small test machine (128 cores) with Intel Sandy Bridge processors (still with working hardware counters)
- Domain and decomposition were chosen in a way that the subdomains are as big as for COSMO-DE65, when running on about 400 cores: 320x260x65 grid points on 7x9+1 tasks
- From pat_report:
  - HW FP Ops / User Time 1973.205M/sec 264939705337 ops
    9.5% peak(DP)
  - MFLOPS (aggregate) 126285.11M/sec
  - which corresponds to 166 GFlop/s per processor (*8 = 1.32 TFlop/s: 9.5 % are: 126.16 GFlop/s)
- Are these measurements ok? We thought to get a much smaller percentage out of peak performance?


Thank you
very much
for your
attention