Performance profiling of COSMO on three generations of Cray supercomputers

Jean-Guillaume Piccinali, CSCS

Anne Roches, C2SM

Goal

Analyse performances:

- of main configurations of COSMO (running currently on CSCS machines)
- with different domain decompositions
- on several platforms

In order to:

- understand the code and its bottlenecks
- identify potential quick wins

Method (I)

Systematic procedure:

- Performance overview
- Computation scaling
- Communication scaling
- Memory usage
- I/O

Method (II)

TESTS 1 (no tools)	# of repet	CONFIG	DECOMP	# of runs
DOLE	2	3	1	6
ROSA	2	3	3	18
PALU	2	3	3	18

TESTS 2 (craypat)	Sampling + Tracing	CONFIG	DECOMP	# of runs
DOLE	2	3	1	6
ROSA	2	3	3	18
PALU	2	3	3	18

TESTS 3 (Tau+ Scalasca)	Sampling + Tracing	CONFIG	DECOMP	# of runs
PALU	2*2	3	1	12

TESTS 4	Nodes Occupatio n (6,12,24)	CONFIGS	DECOMP	# of runs
PALU	3	3	1	9

COSMO configurations

- Based on RAPS5.0
  - ➔ All runs : 24 h simulation

- 2 operational NWP versions:
  - ➔ COSMO2
  - ➔ COSMO7

- 1 climate (RCM) version:
  - ➔ "IPCC"

Configurations: grid, time

Space and time characteristics:

	COSMO2	COSMO7	“IPCC”
nx	520	393	441
ny	350	338	429
nz	60	60	40
Δx, Δy (°)	0.02	0.06	0.11
dt (s)	20	60	100

Configurations: obs

Use of observations:

	COSMO2	COSMO7	"IPCC"
luseobs	T	T	F
Use of satellite data	T	T	F
Nudging	T	T	F

Domain decompositions

- Usual decomposition for each configuration
  + 1 “small” decomposition
  + 1 “big” decomposition

	COSMO2	COSMO7	“IPCC”
usual	28 x 35 + 4 (984)	28 x 35 + 4 (984)	30 x 30 (900)
small	20 x 24 + 4 (484)	20 x 24 + 4 (484)	24 x 24 (576)
big	48 x 60 + 4 (2884)	48 x 60 + 4 (2884)	53 x 53 (2809)

(Number of cores)

=> Keep aspect ratio ~ cst

Decompositions

Aspect ratios:

	COSMO2	COSMO7	“IPCC”
usual	0.8	0.8	1
small	0.83	0.83	1
big	0.8	0.8	1

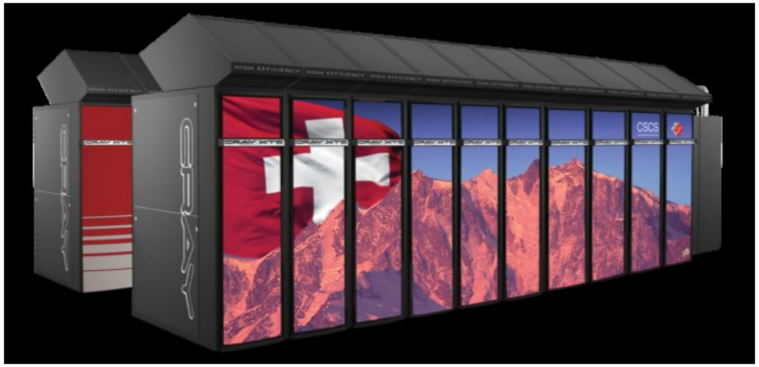
Grid points per core:

	COSMO2	COSMO7	“IPCC”
usual	~185	~135	~210
small	~376	~274	~328
big	~63	~46	~67

Platforms

- 3 generations of Cray machines
  - Dole: Cray XT4

  - Rosa: Cray XT5



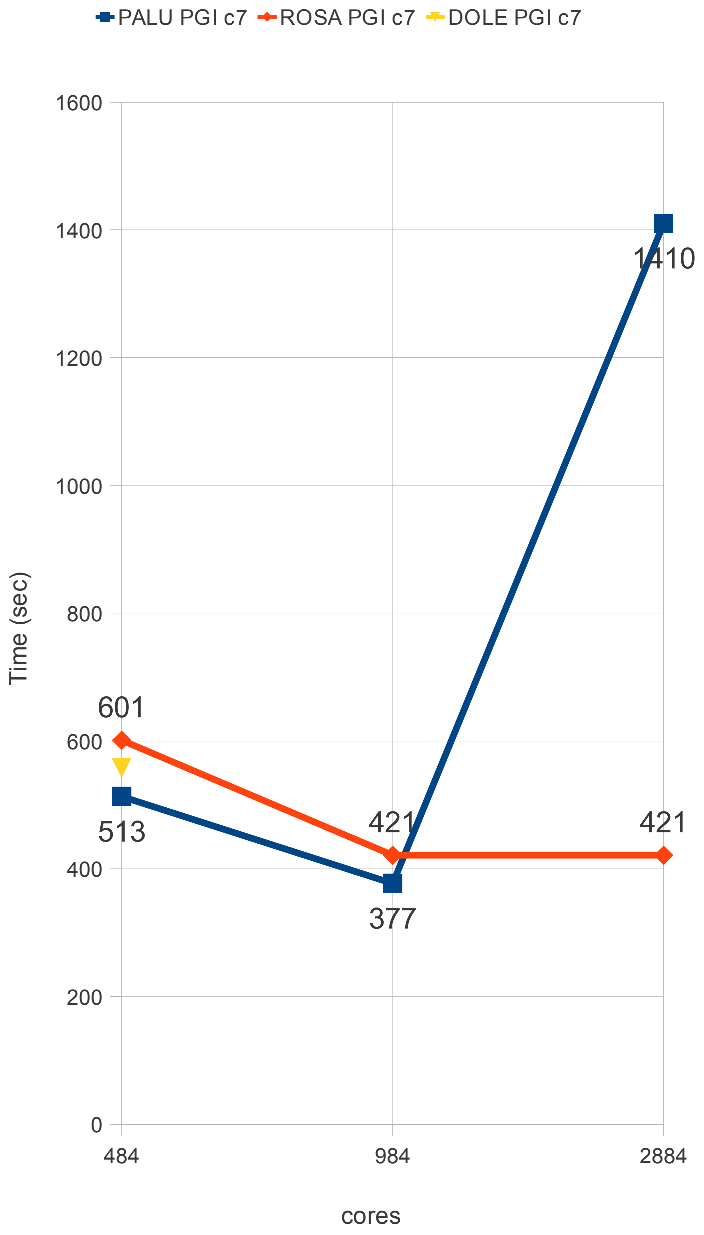
  - Palu: Cray XE6

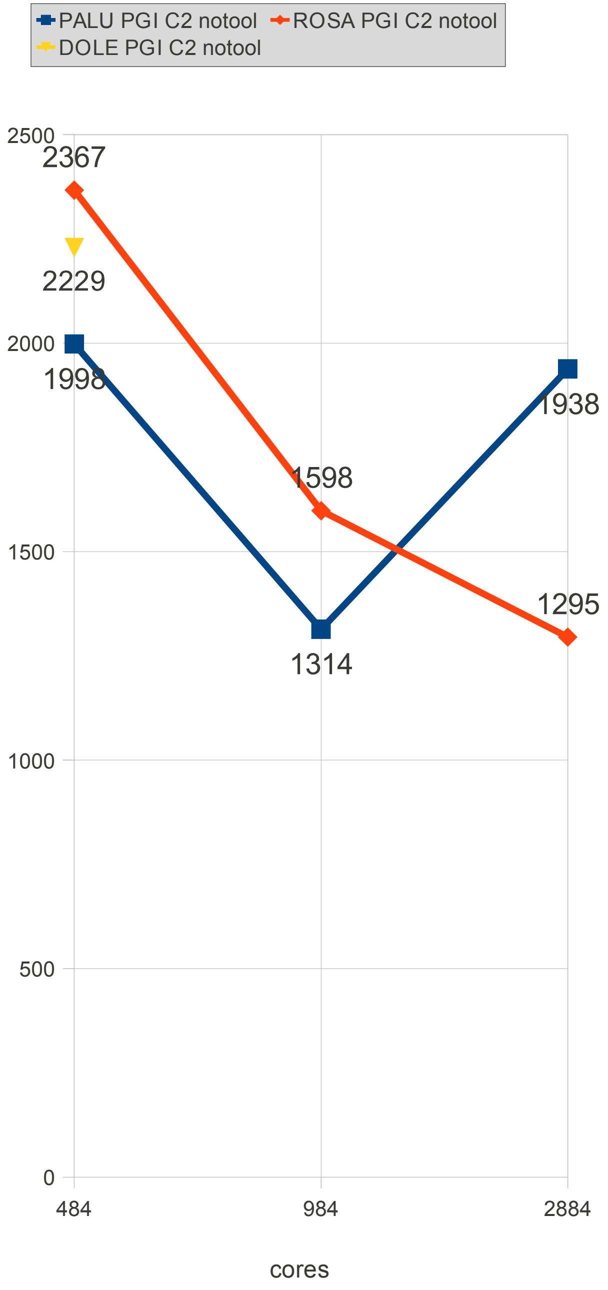
Platforms

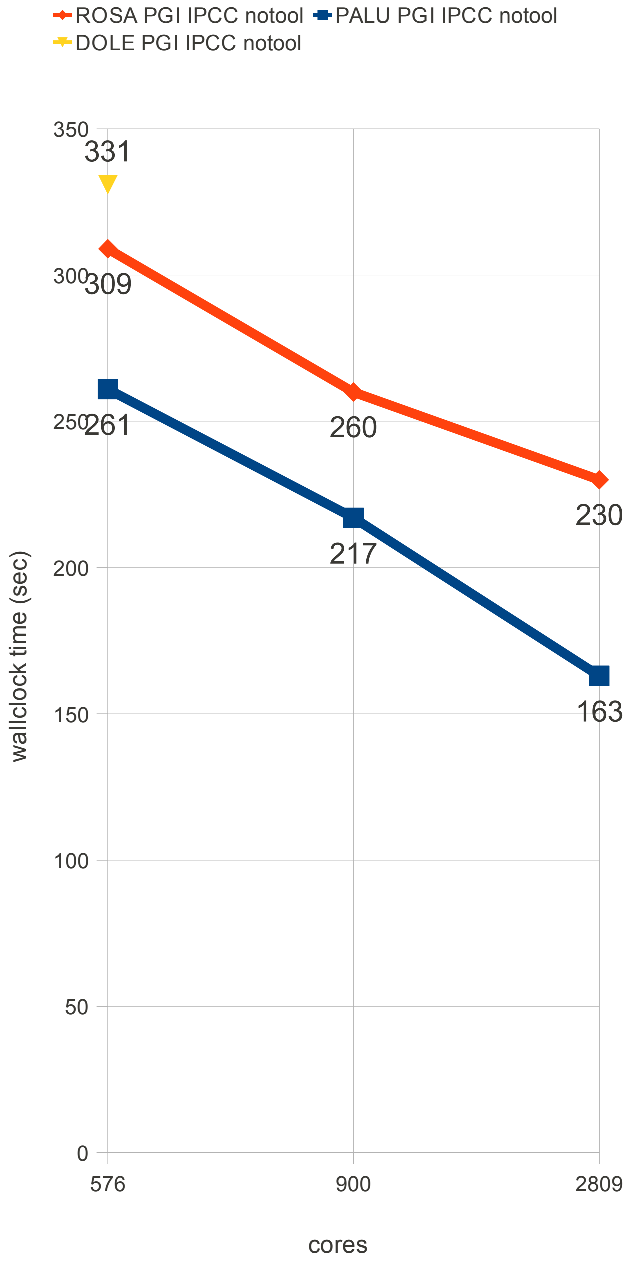
	Dole (XT4)	Rosa (XT5)	Palu (XE6)
Pes (nodes * cores/node)	172 x 4 = 688	1844 x 12 = 22128	176 x 24 = 4224
CPU Frequency	2.3 GHz	2.4 GHz	2.1 GHz
CPU type (AMD Opteron)	Barcelona	Istanbul	Magny-Cours
Interconnect injection bandwidth	2 GB/s	2 GB/s	5 GB/s
Memory per node	8 GB	16 GB	32 GB
L1 cache	4 × 128 KB	6 × 128 KB	12 × 128 KB
L2 cache	4 × 512 KB	6 × 512 KB	12 × 512 KB
L3 cache	2 MB	6 MB	2 × 6 MB
Total memory	5.5 TB	354 TB	135 TB
Peak performance	6 Tflops	212 TFlops	35 TFlops

http://fr.wikipedia.org/wiki/AMD_Opteron

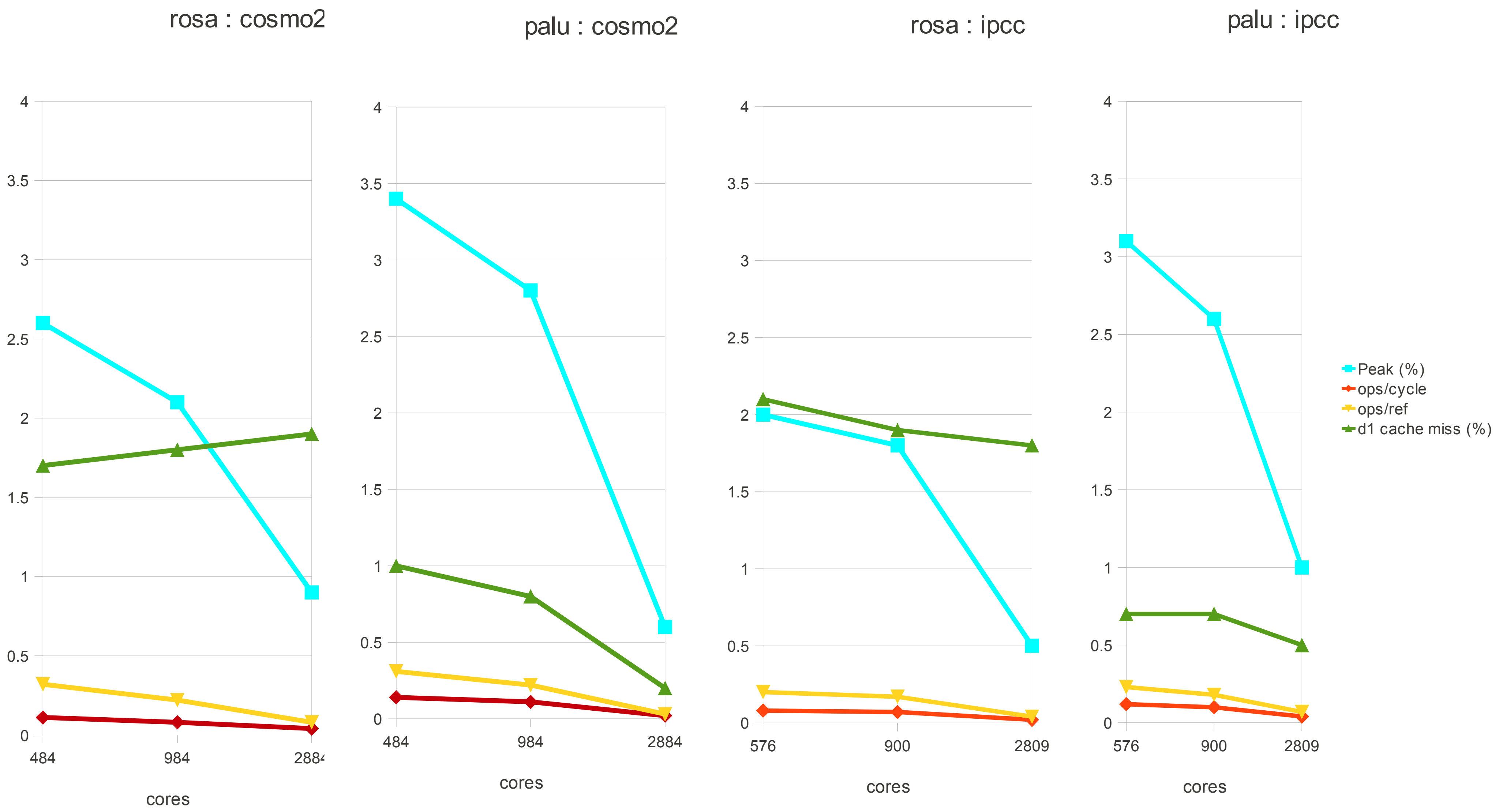
Timings (1 day)







Computational intensity (xpat)



rosa : cosmo2

palu : cosmo2

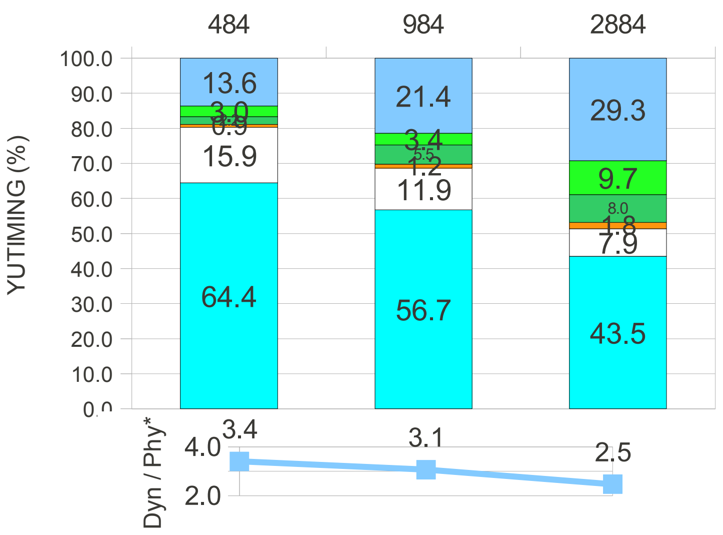
rosa : ipcc

palu : ipcc

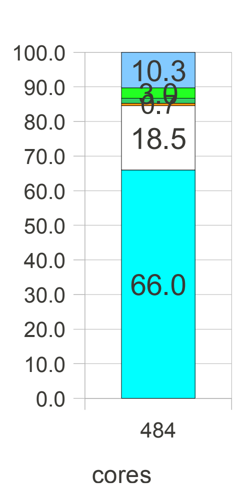
- Computational intensity (ops/ref) should be >= 1
- D1 cache miss ratio should be ~= 0%

Profile by regions (%)

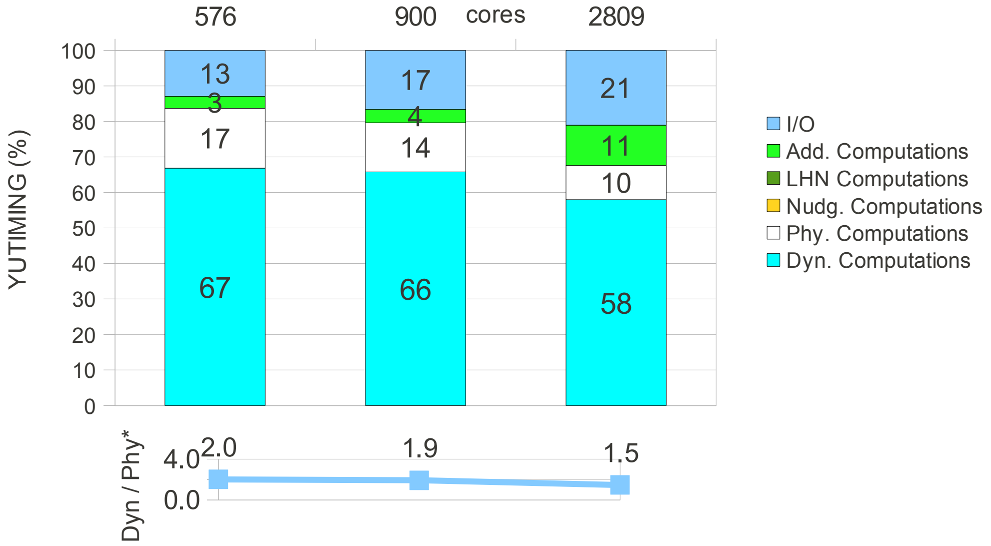
rosa : cosmo2



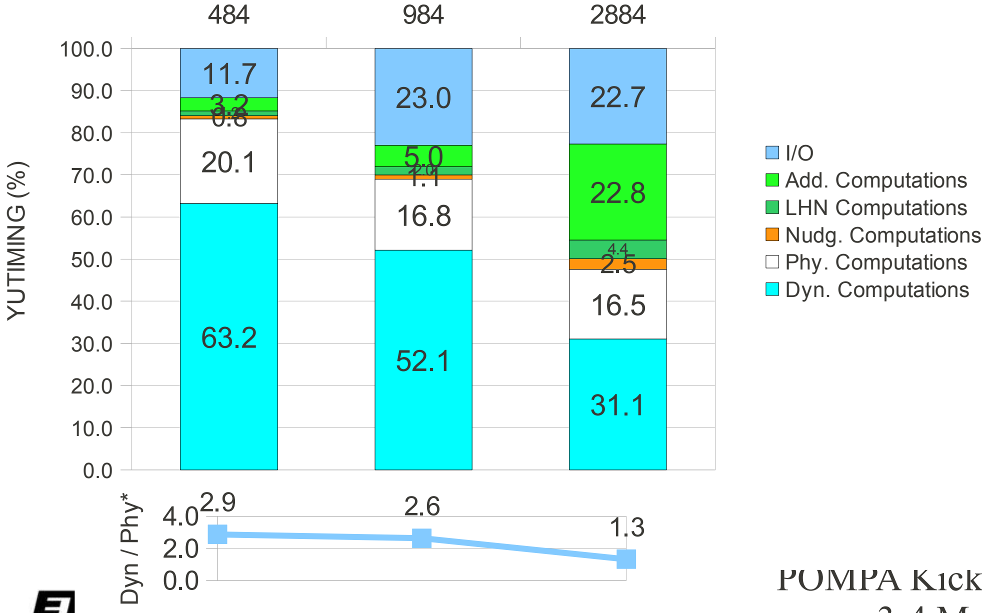
dole : cosmo2



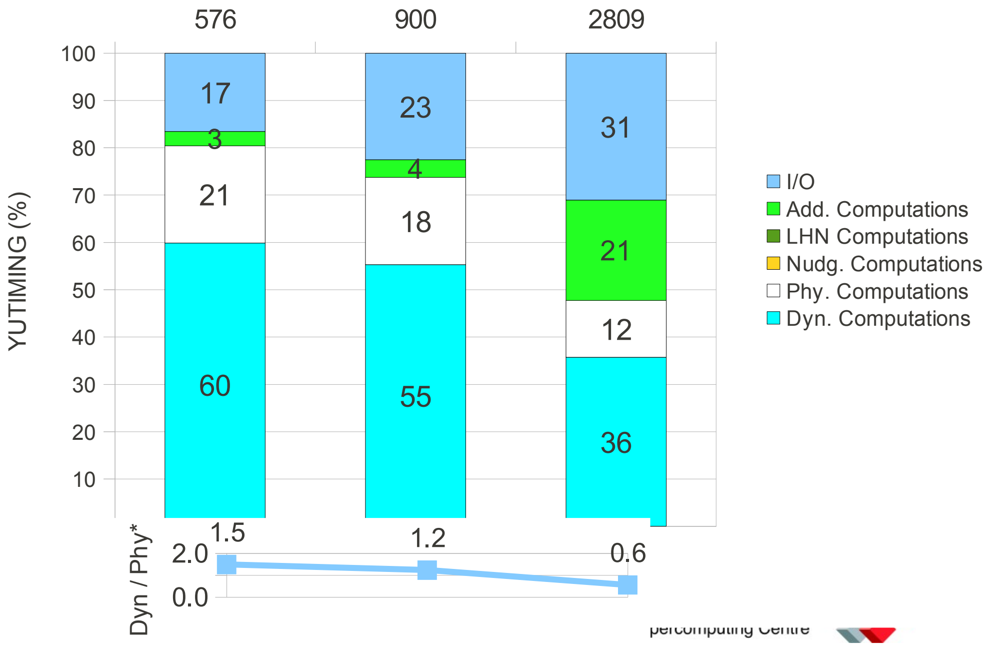
rosa : ipcc



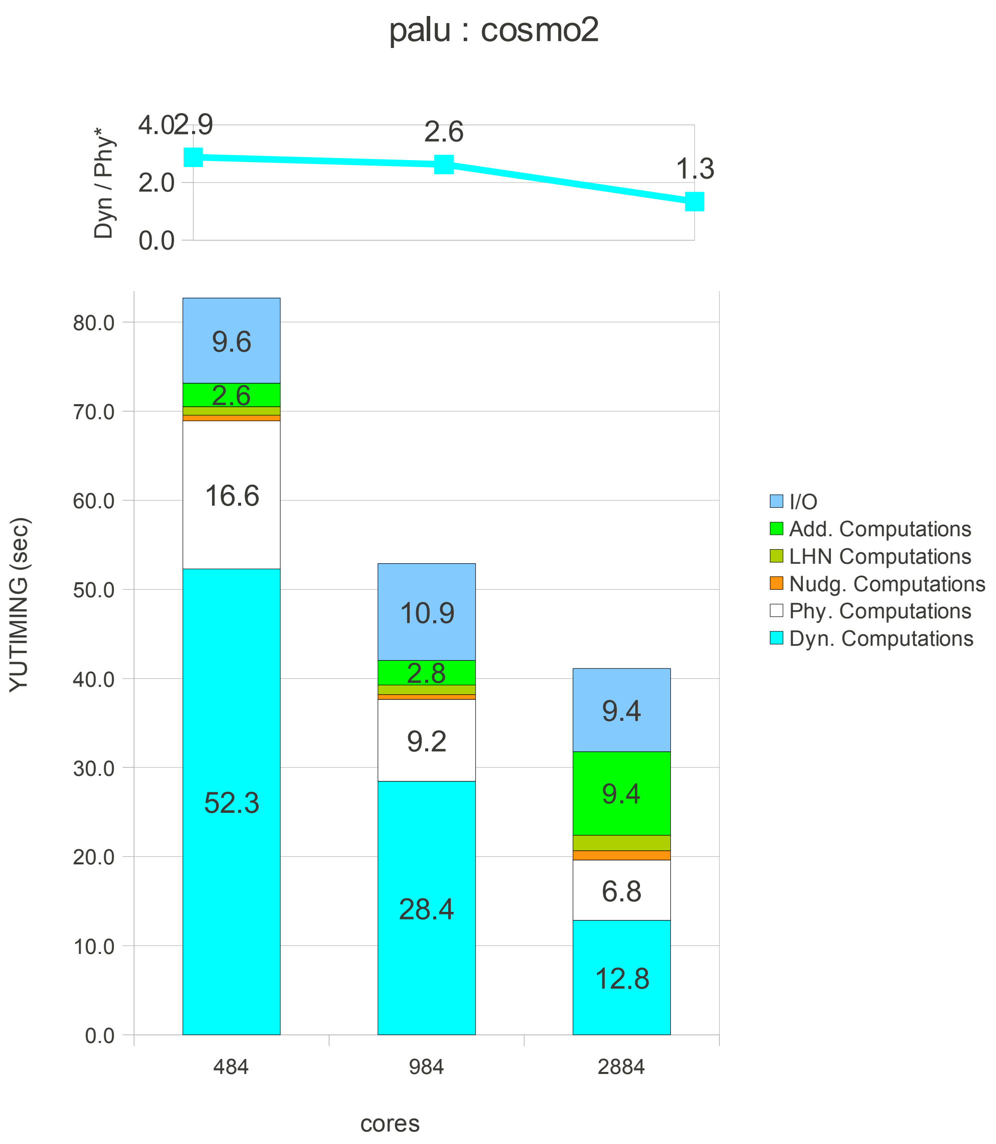
palu : cosmo2

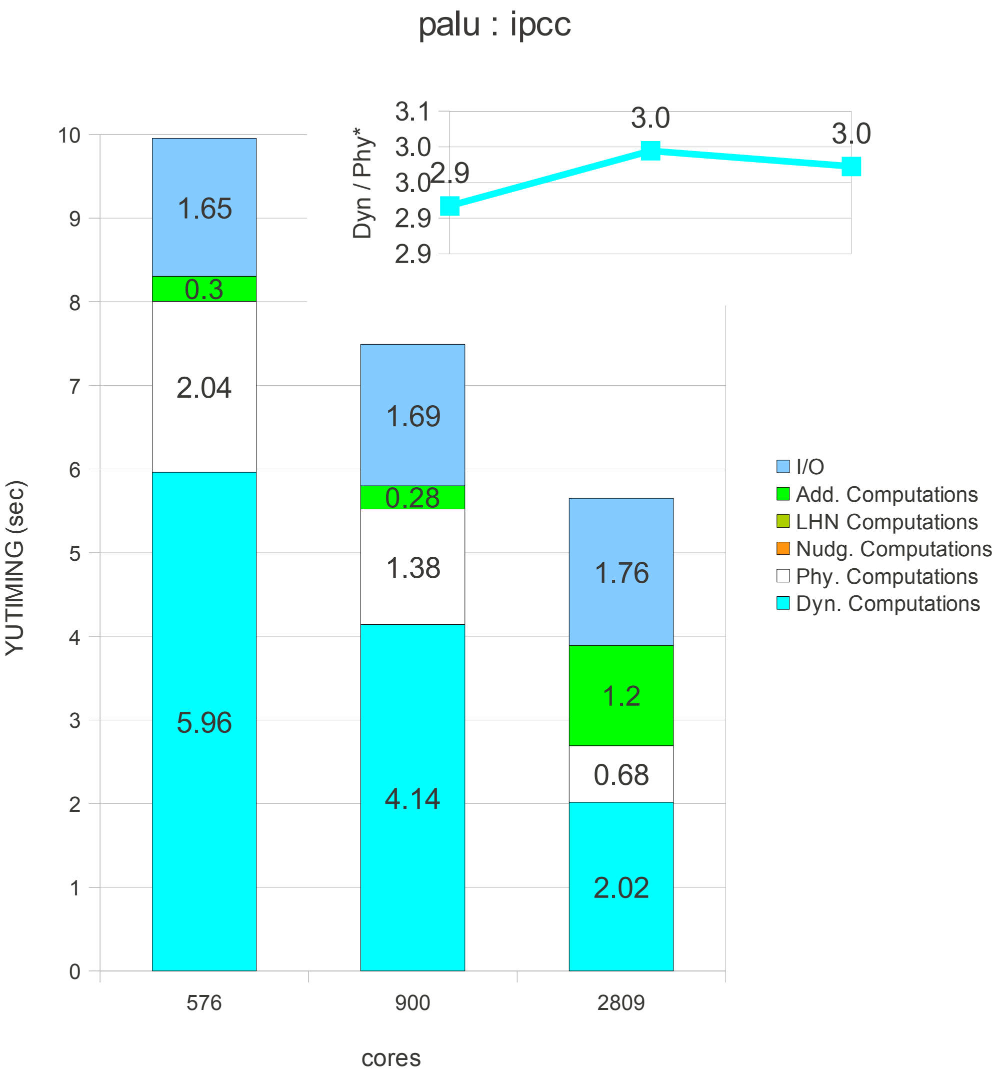


palu : ipcc

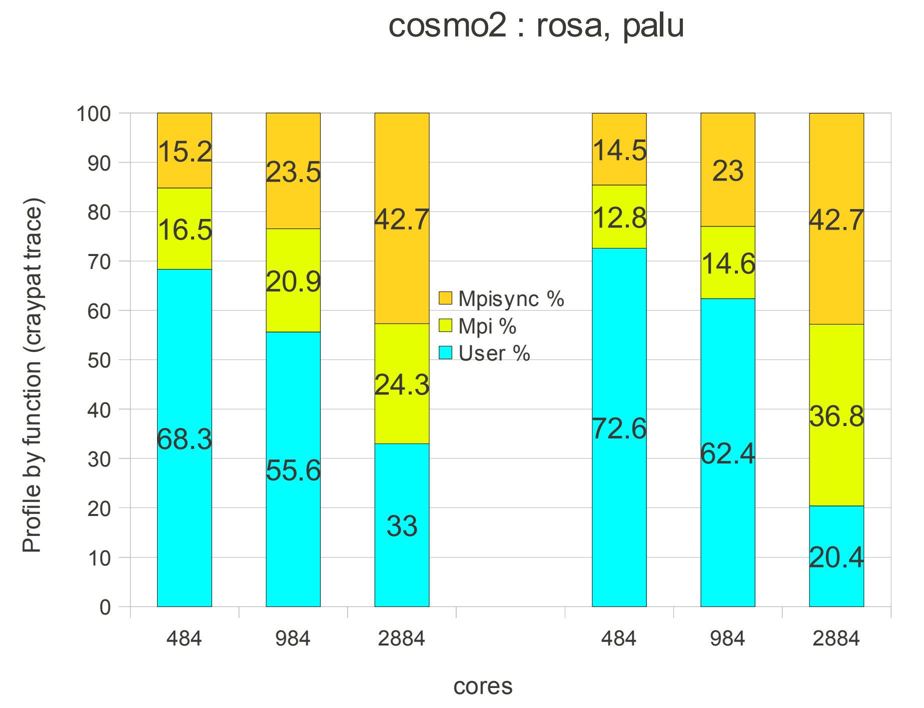


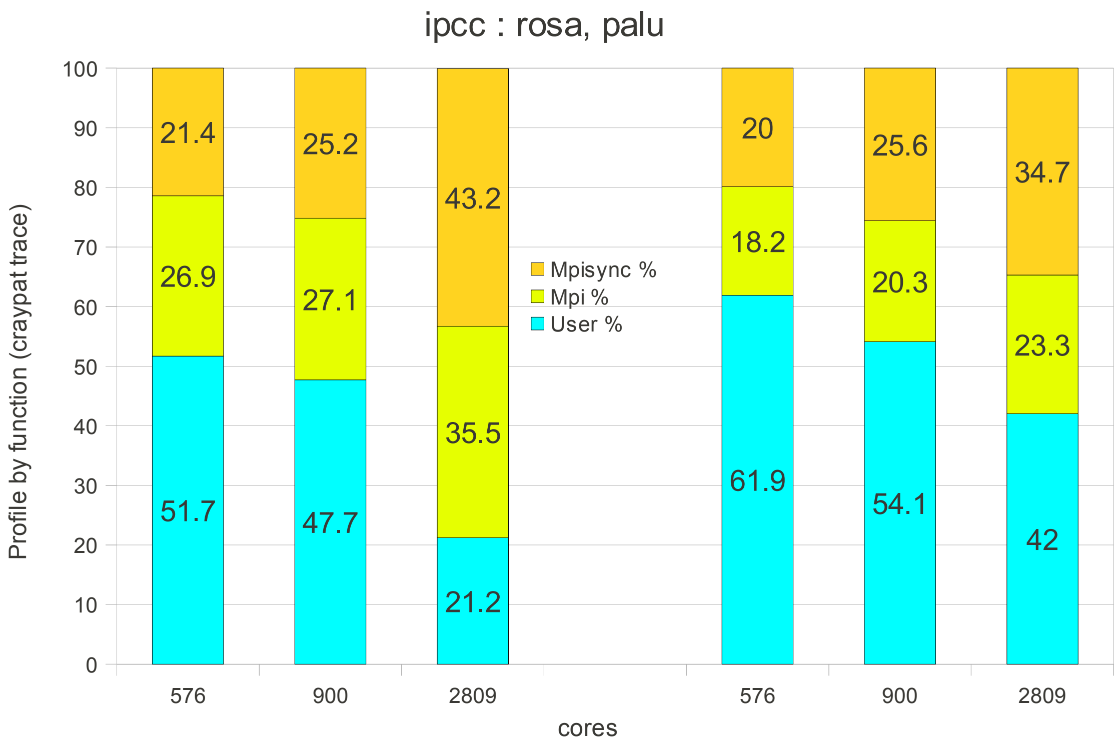
Profile by regions (sec)

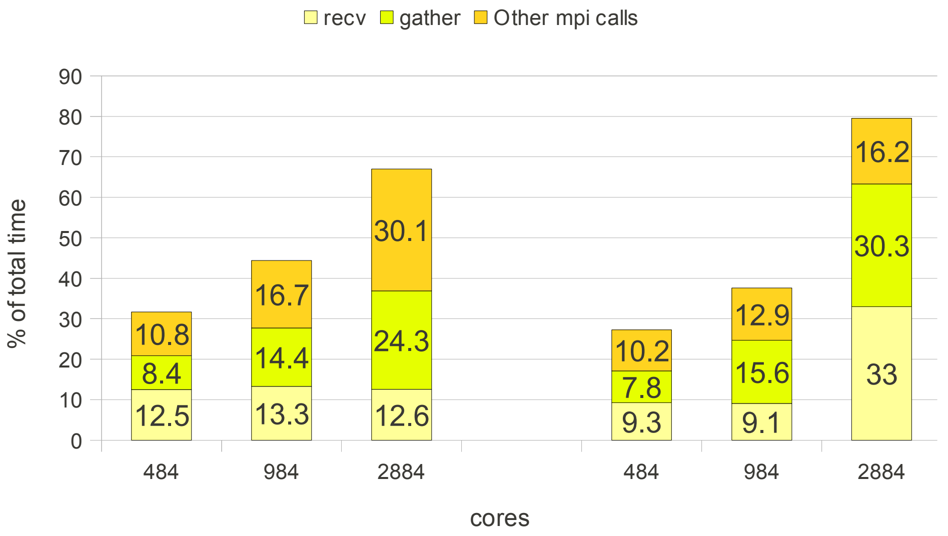


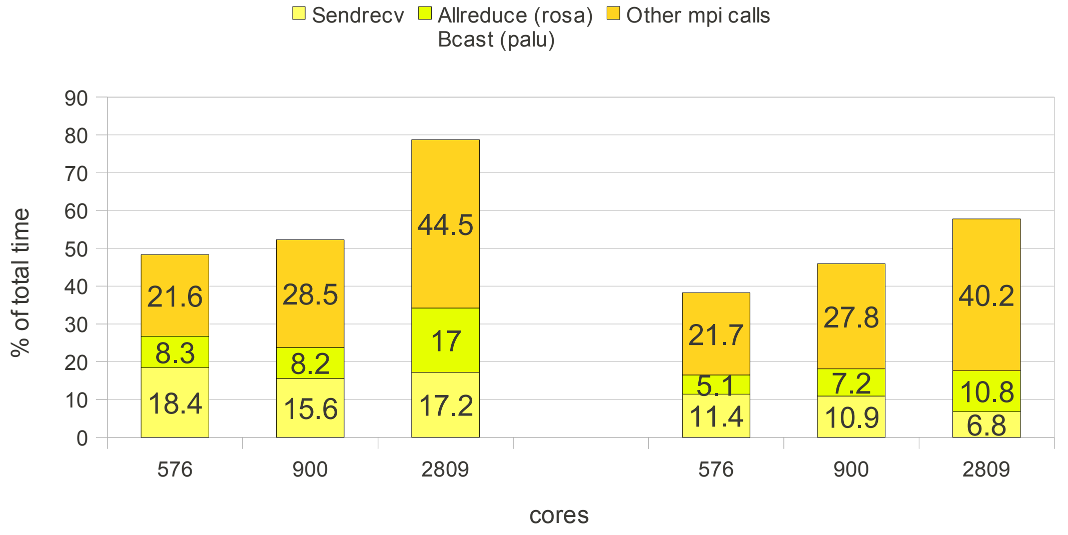


Profile by MPI calls (ipcc, xpat)

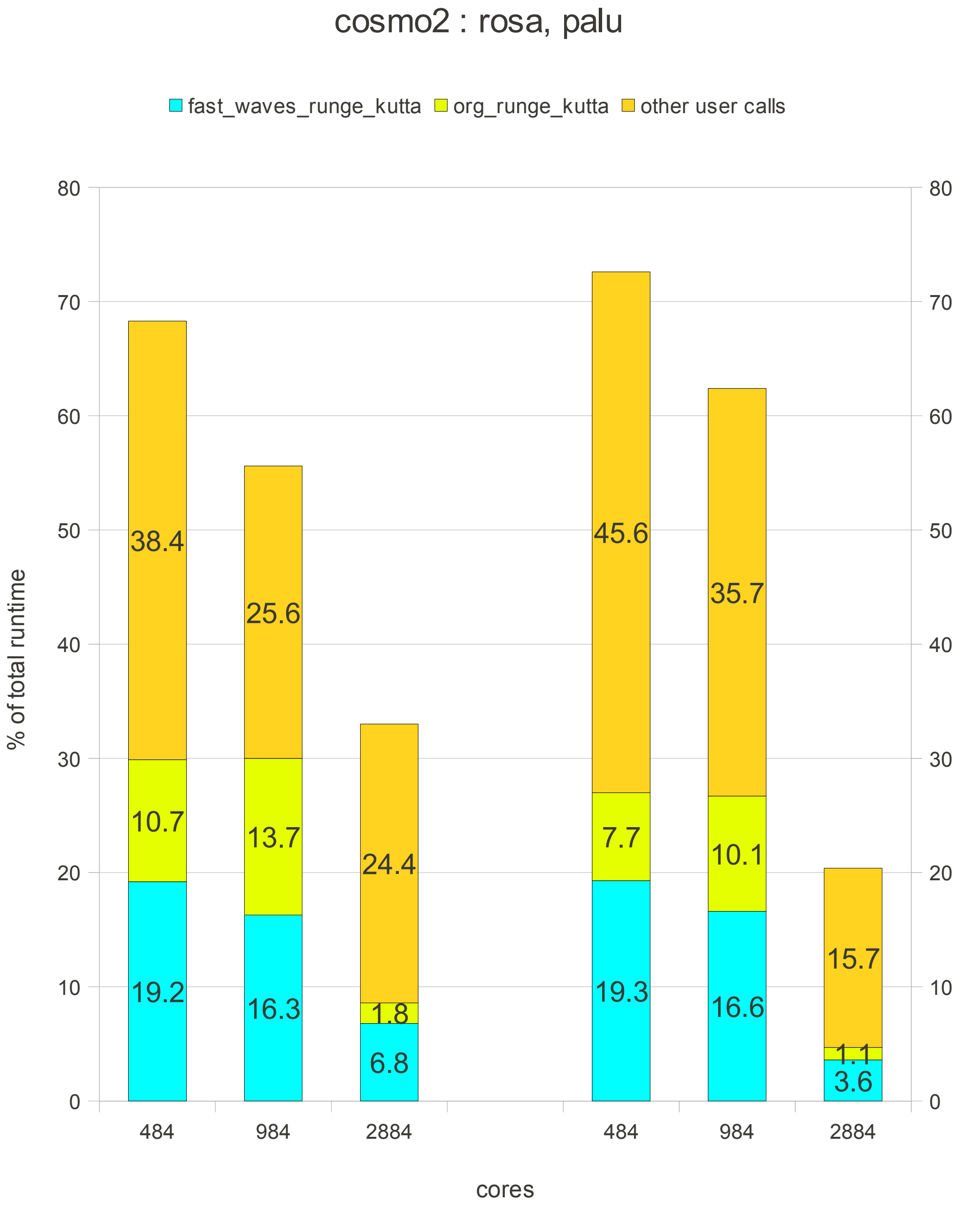




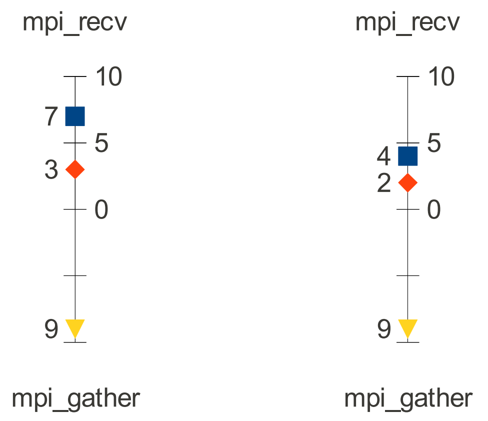




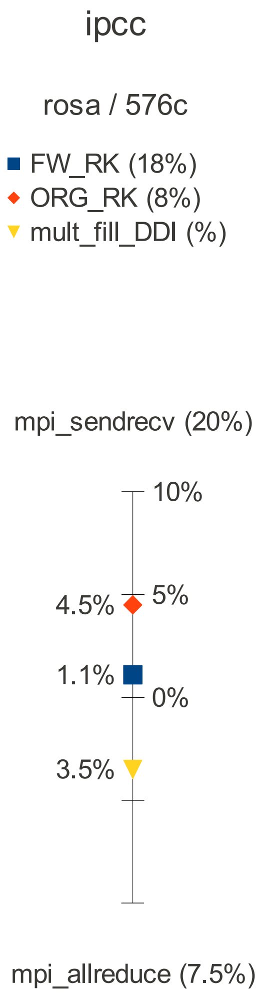
Profile by user functions (c2)

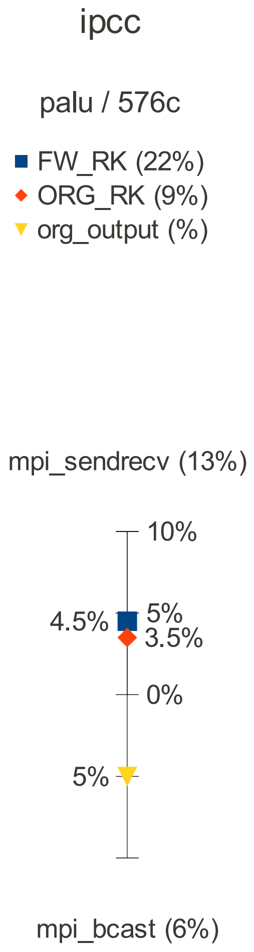


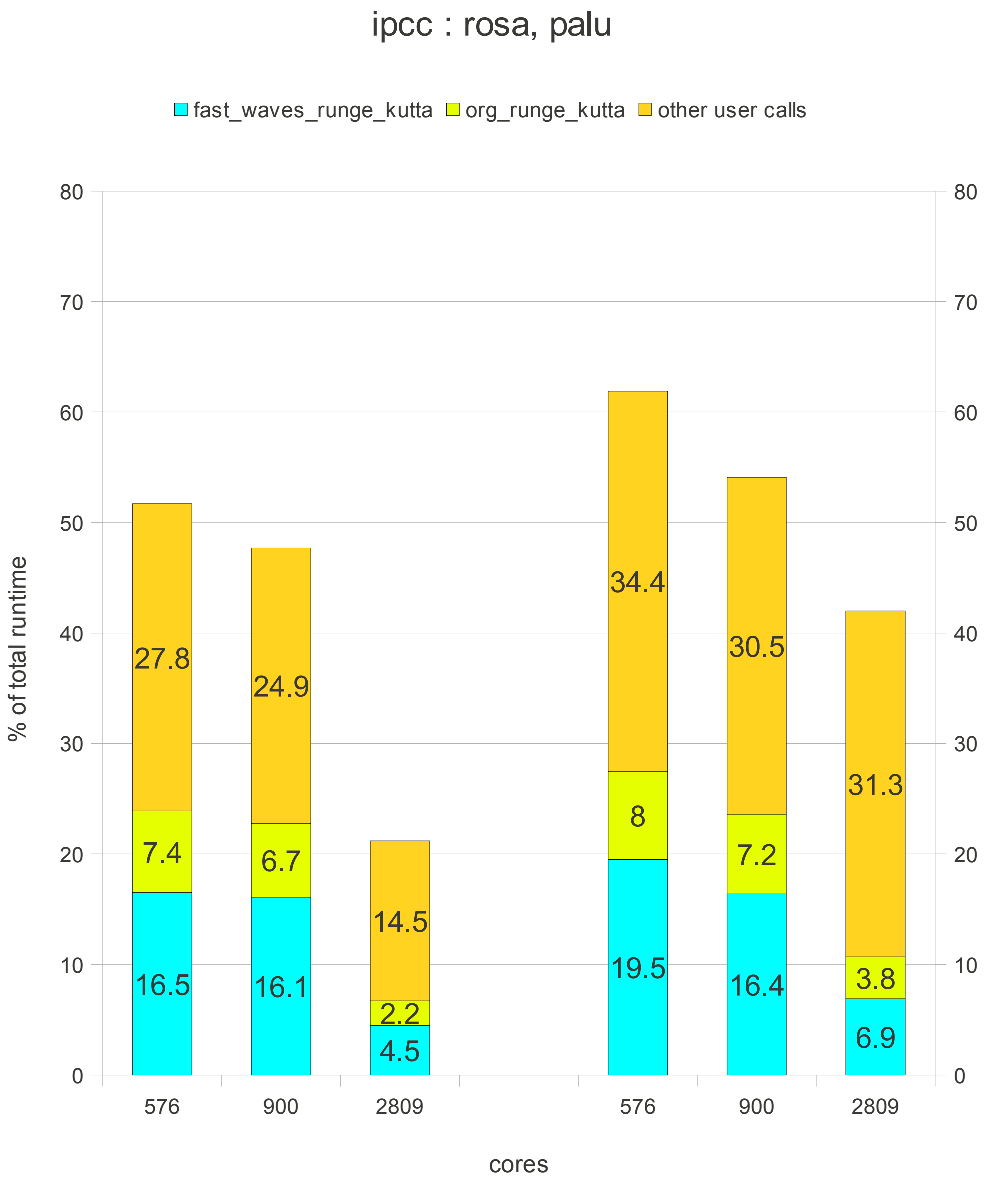
cosmo2	cosmo2
rosa / 484c	palu / 484c
■ FW_RK (21%)	■ FW_RK (22%)
◆ ORG_RK (12%)	◆ ORG_RK (9%)
▼ GATHER2D	▼ GATHER2D



Profile by user functions (ipcc)

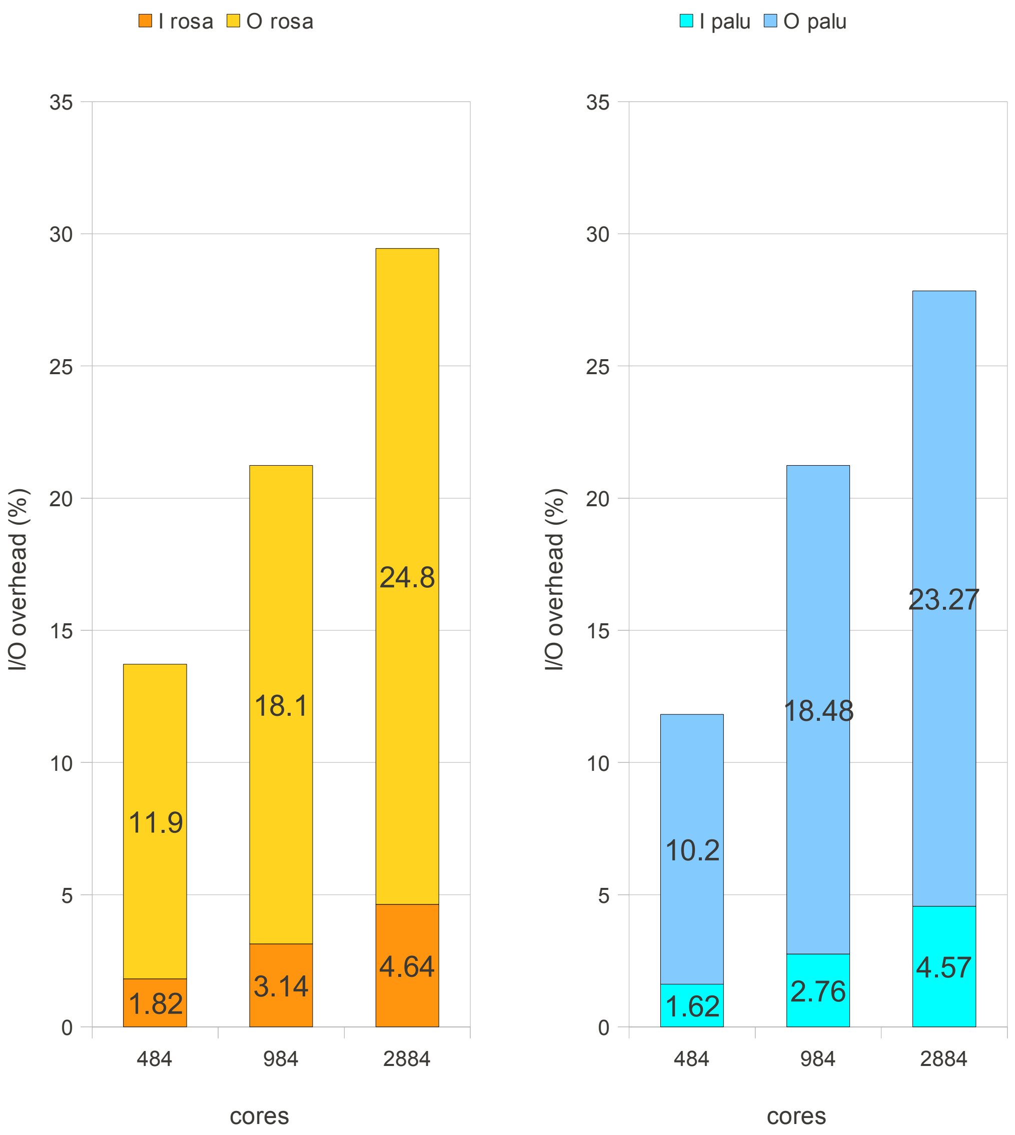




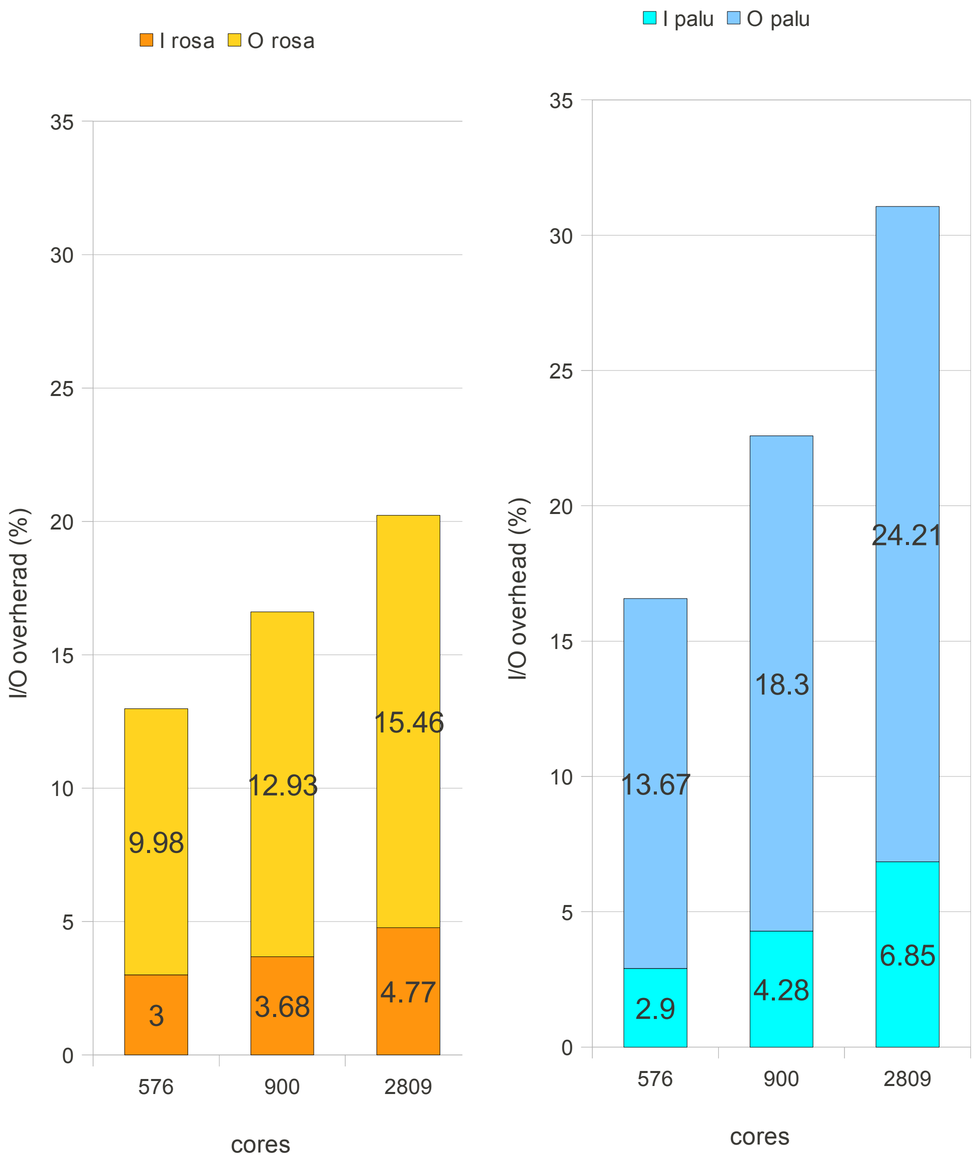


I/O overhead (yutiming)

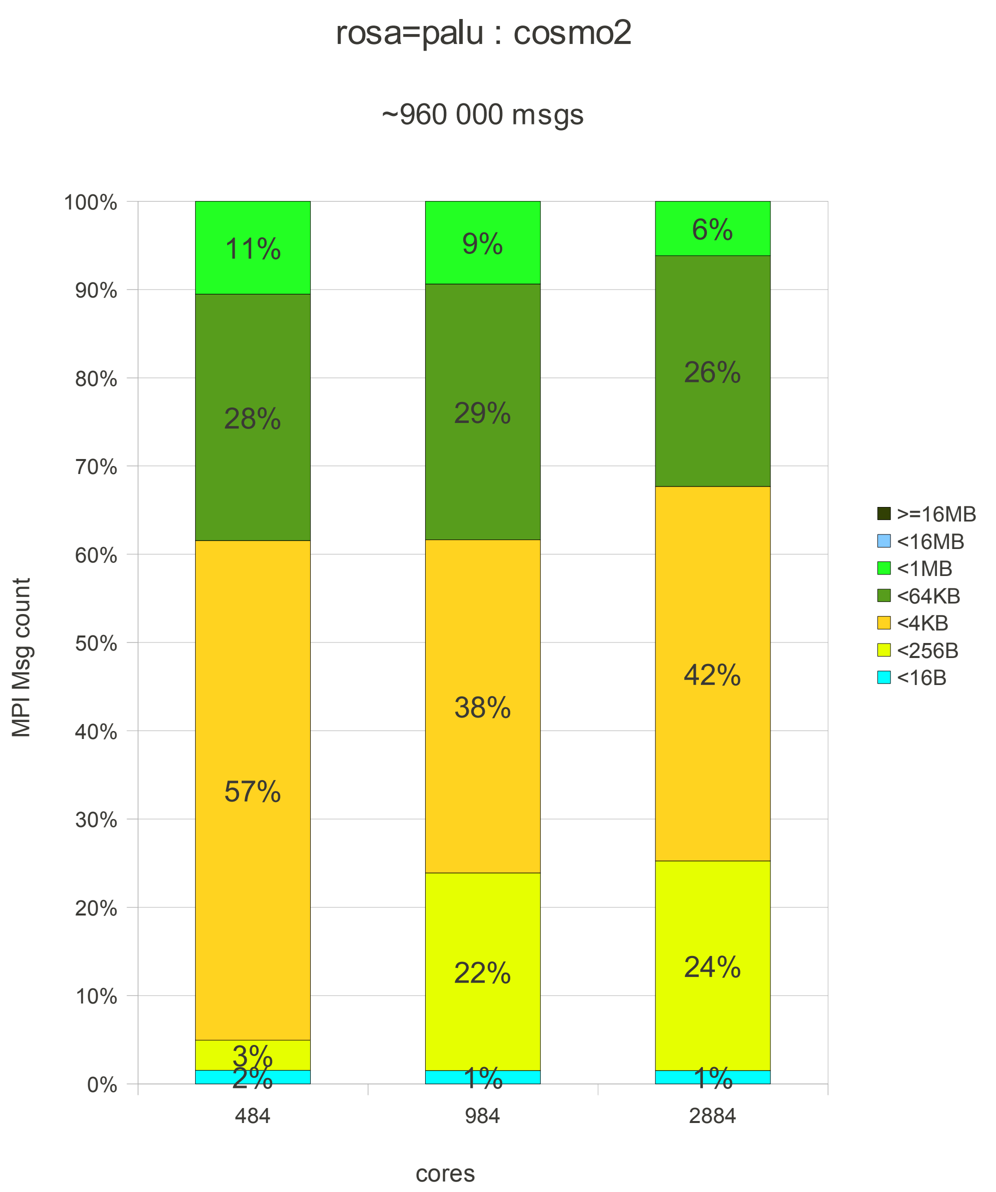
cosmo2

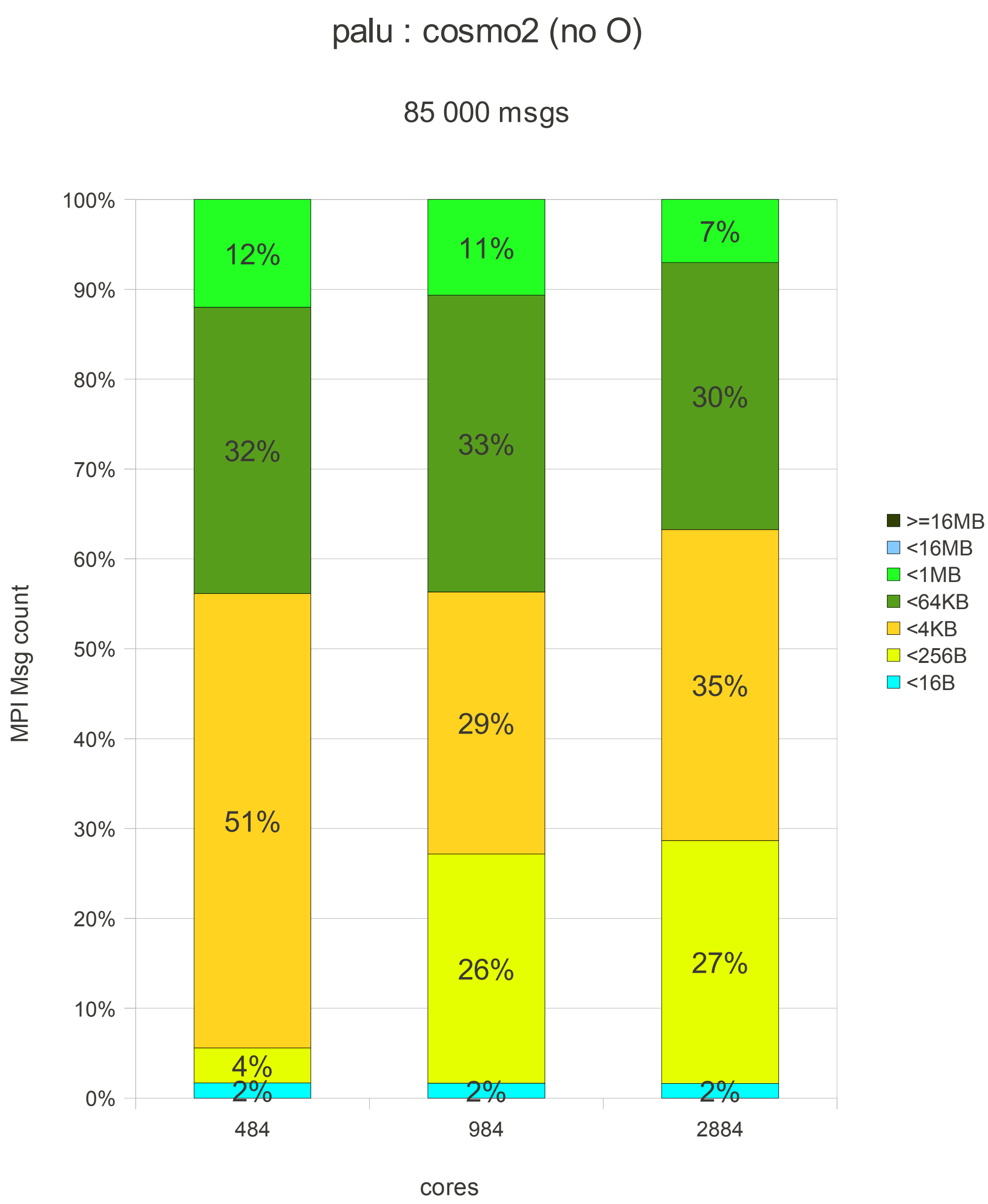


ipcc



MPI message stats (cosmo2, xpat)





Configurations: I/O

I/O:

		COSMO2	COSMO7	“IPCC”
I/O	asynchronous	T	T	F
I/O	format	Grib1	Grib1	NetCDF
I/O	nincwait,nmaxwait	>0	>0	0

Configurations: I/O

I:

		COSMO2	COSMO7	“IPCC”
I	hincbound	1	3	6
I	hnewbcdt	3	3	0
I	lana_qr_qs	T	F	F
I	llb_qr_qs	T	F	F
I	lana_rho_snow	T	F	F
I	lan_rho_snow	T	T	F

4 more 2D fields read for COSMO2 as for IPCC

Configurations: I/O

O:

		COSMO2	COSMO7	"IPCC"
o	# output blocks	2	1	7
o	on model levels	109 x 2D and 13 x 3D @ 1 h freq ------------------ 19 x 2D and 7 x 3D @ 30min freq on a reduced domain	99 x 2D and 12 x 3D @ 1 h freq	5 x 2D @ 1 h freq -------------------- 30 x 2D and 7 x 3D @ 3 h freq --------------------- 17 x 2D and 6 x 3D @ 6 h freq --------------------- 1 x 3D @ 24 h freq
o	on pressure levels	7 x 3D @ 1 h freq	8 x 3D @ 1 h freq	-
o	on z-levels	1 x 2D and 8 x 3D @ 1 h freq	7 x 3D @ 1 h freq	-
o	On sat. levels	1 x 3D @ 1 h freq	-	-

Conclusion and outlook (I)

- Systematic profiling runs performed
- Lot of profiling data
- First overview of the results
- Need more analysis in order to reach the goals
  - Understand the code
  - Identify quick wins

Conclusion and outlook (II)

- Other potential next steps:
  - Test external Lustre
  - Perform weak scaling
  - Try different compilers
  - Analyse additional configurations (ART, M7, …)
  - (Other platforms → done by the colleagues)

THANK YOU