AROME scalability

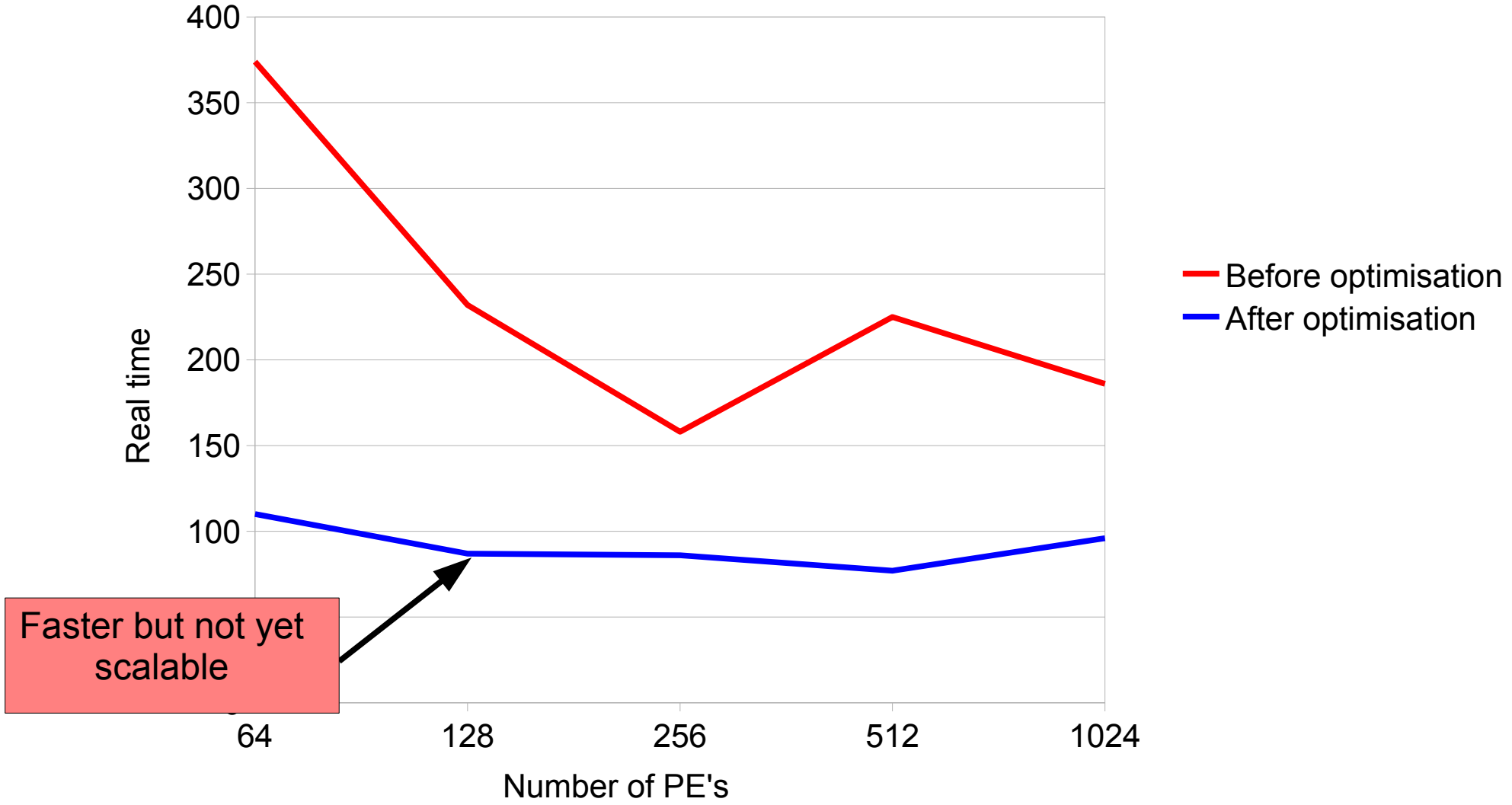
Philippe Marguinaud, Ryad El Khatib,
Eric Sevault (CNRM/GMAP), Eric Maisonnave (CERFACS)

- Development realized in 2010
  - SURFEX setup optimizations
  - OpenMP enhancements
- AROME and the scalability of the physics
- AROME and the scalability of the dynamics
- I/Os studies
- Conclusion

SURFEX setup optimizations

- **Before** :
  - SURFEX initial condition file was read by each NPROMA packet of each MPI task
- **Now** :
  - SURFEX fields are read once and stored in a memory cache
  - Reduction of the namelists accesses
  - Reduction of static allocations
- **To do** :
  - OpenMP support in the setup
  - Optimize the writing out of Surfex files

AROME setup on ECMWF c1a

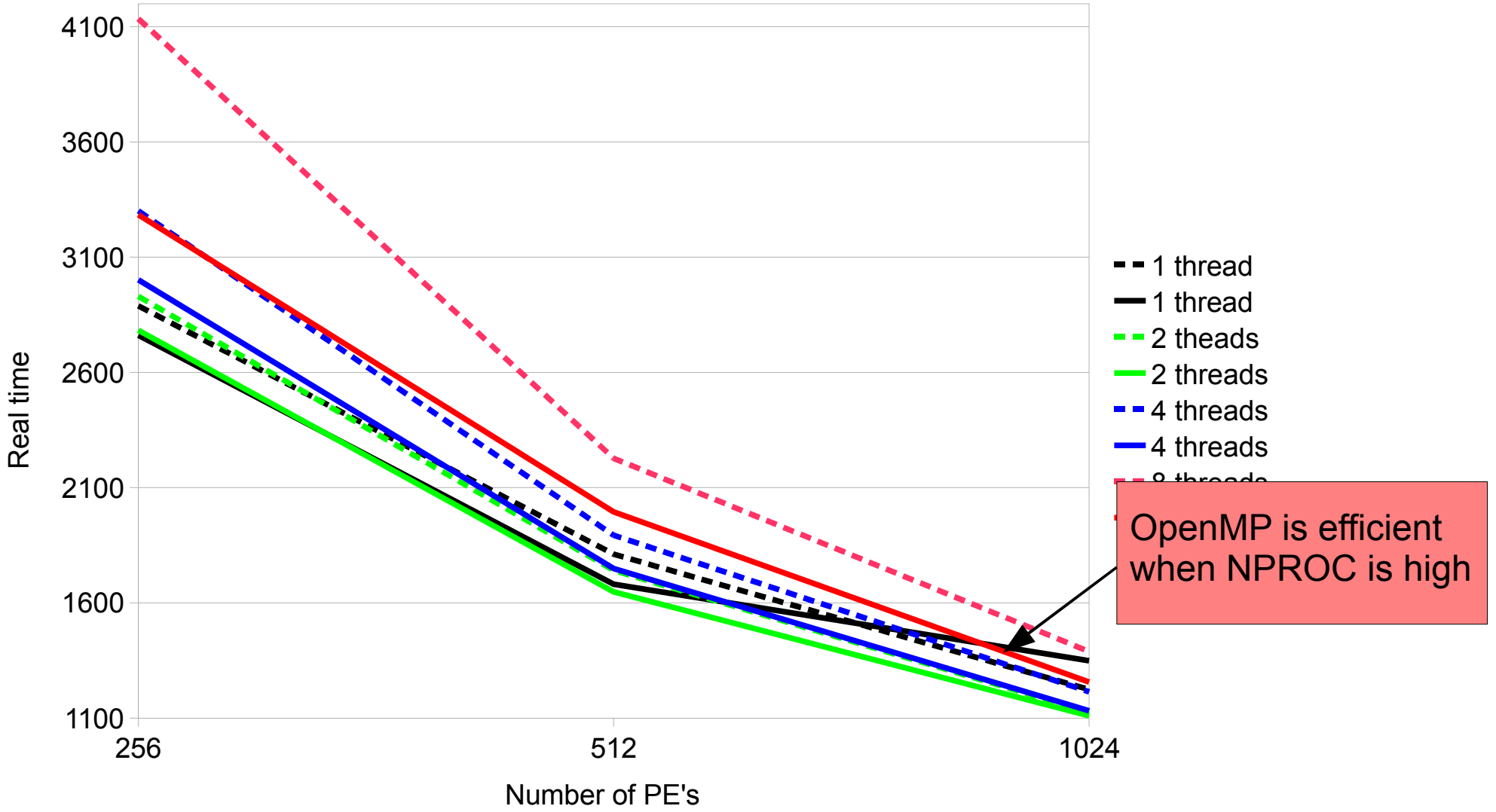


OpenMP & SURFEX

(with the participation of CERFACS)

- SURFEX V6 : not thread-safe, no OpenMP support
- The solution to make SURFEX compatible with OpenMP has been defined with the SURFEX team by the mean of « THREADPRIVATE » directives on the global variables (about 2000 directives !)
- SURFEX V7 : still not thread-safe, but works when used inside an Open-MP environment
- However :
  - No support for another parallelisation scheme
  - Probably difficult to maintain
  - Probably not OOPS-compliant

OpenMP on ECMWF c1a
AROME 30h forecast



OpenMP – Conclusion

- Memory usage reduction
- Better load balancing
- Performances :
  - IBM : better when much processors
  - NEC : detrimental because of numerous dynamic allocations (reduction under progress)
  - PC : good improvment (20%)

The performance depends on the machine and the open-mp implementation

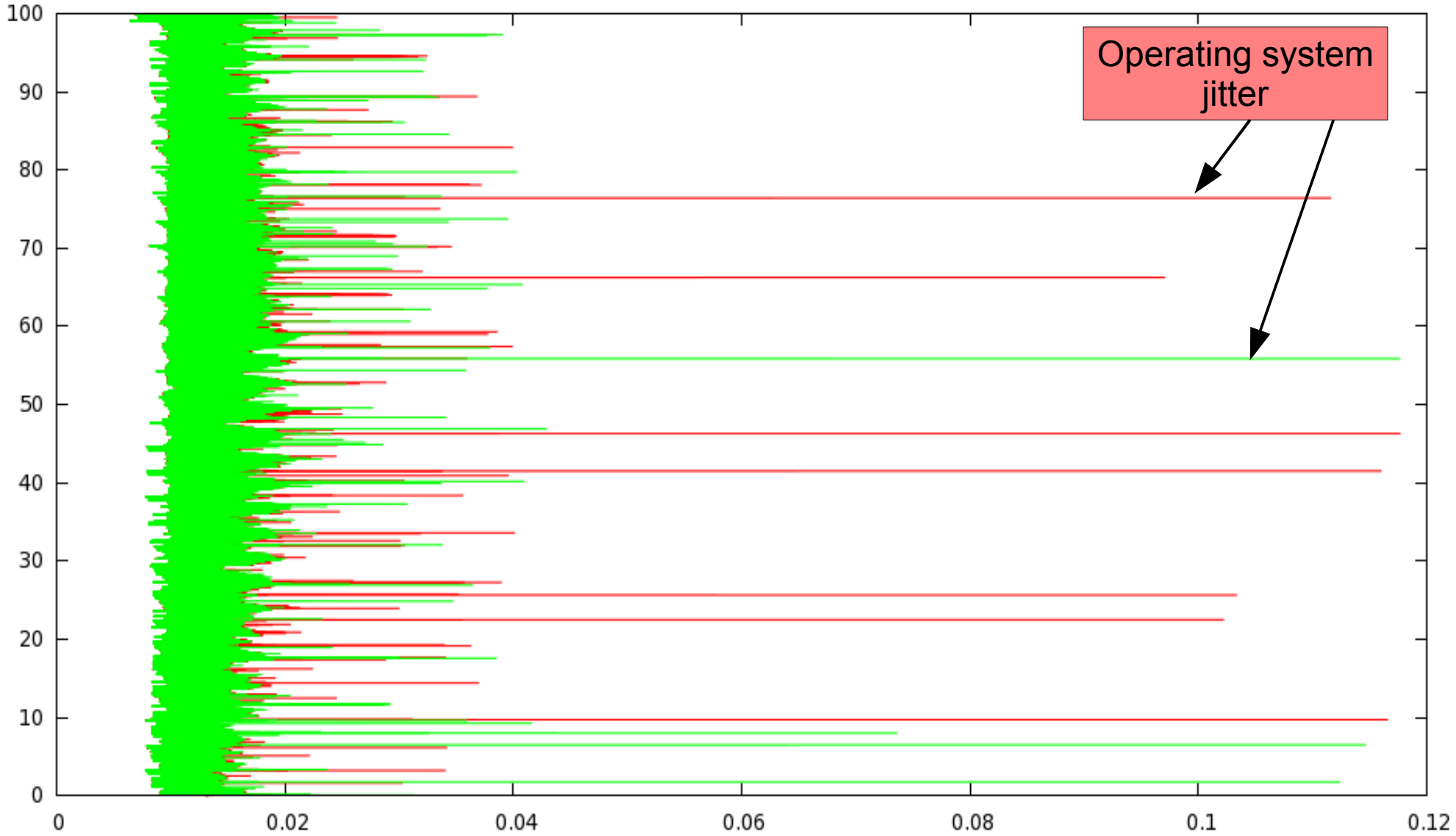
Scalability of the physics

- Independent atmosphere columns
- Computation per groups of NPROMA gridpoints (50 sur IBM, 3582 sur NEC)

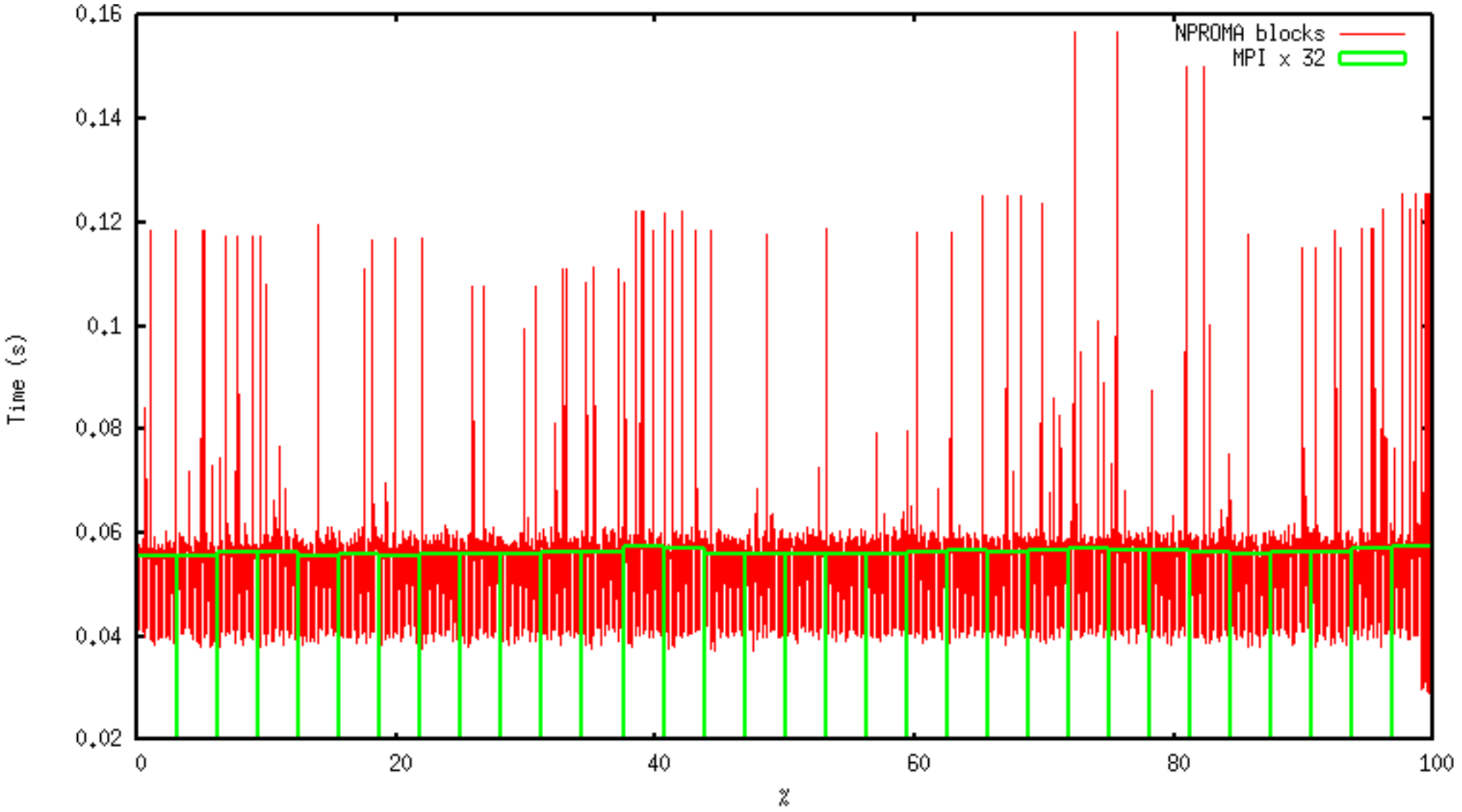
→ optimal use of memory cache or vector registers

The scalability is driven by the computation imbalances

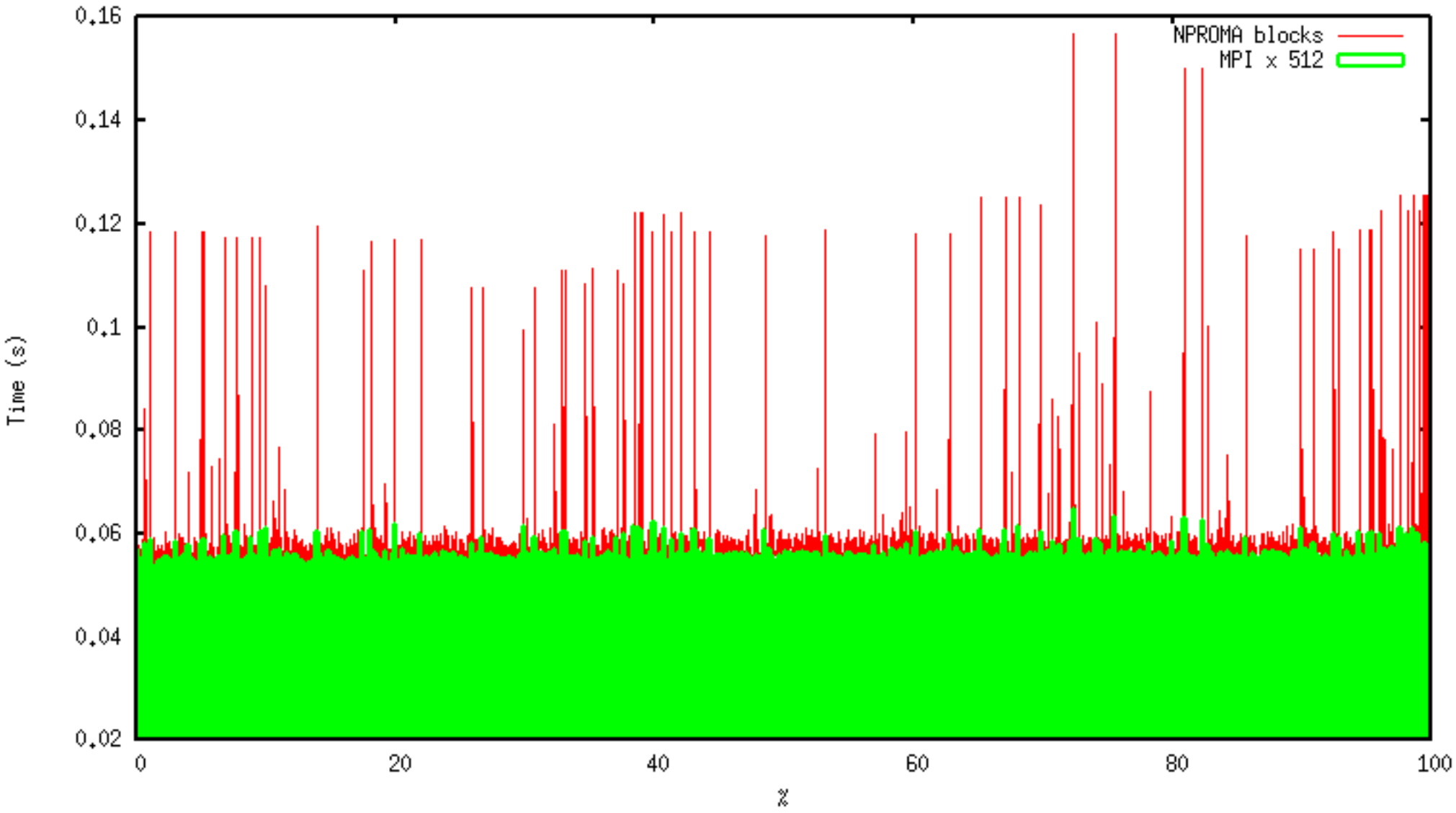
Time for each NPROMA packet



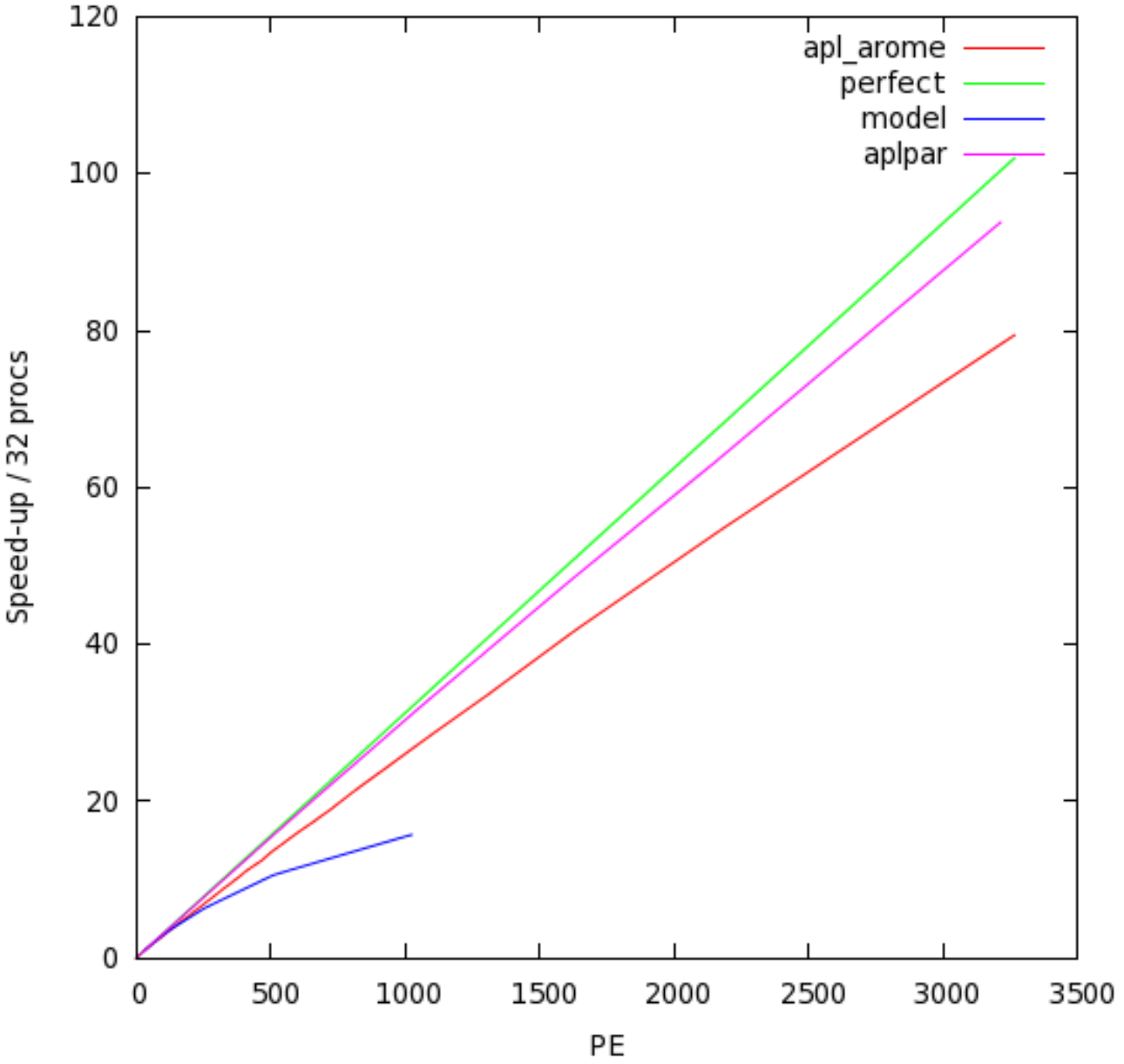
NPROMA packets for 32 MPI tasks



NPROMA packets for 512 MPI tasks

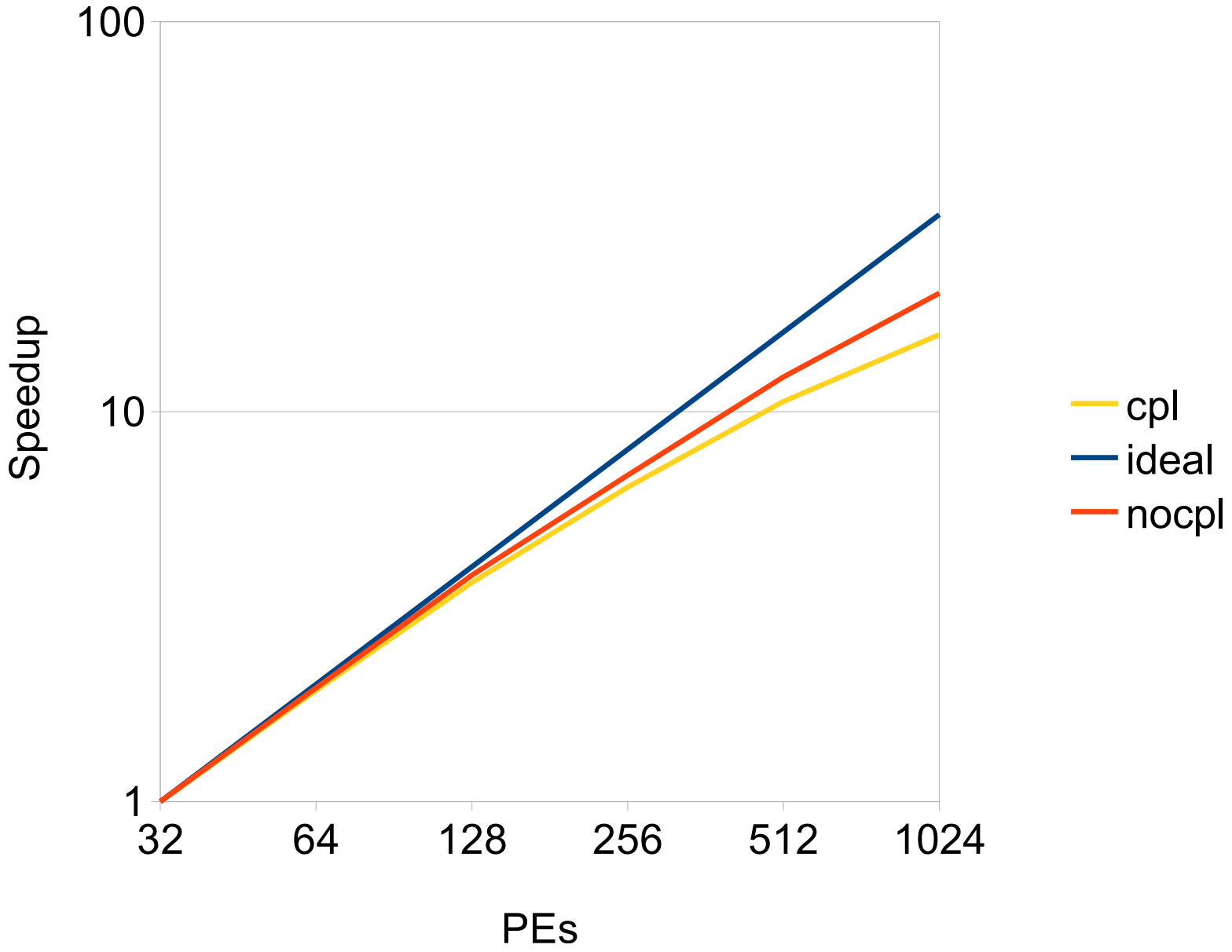


Theoretical scalability of the physics



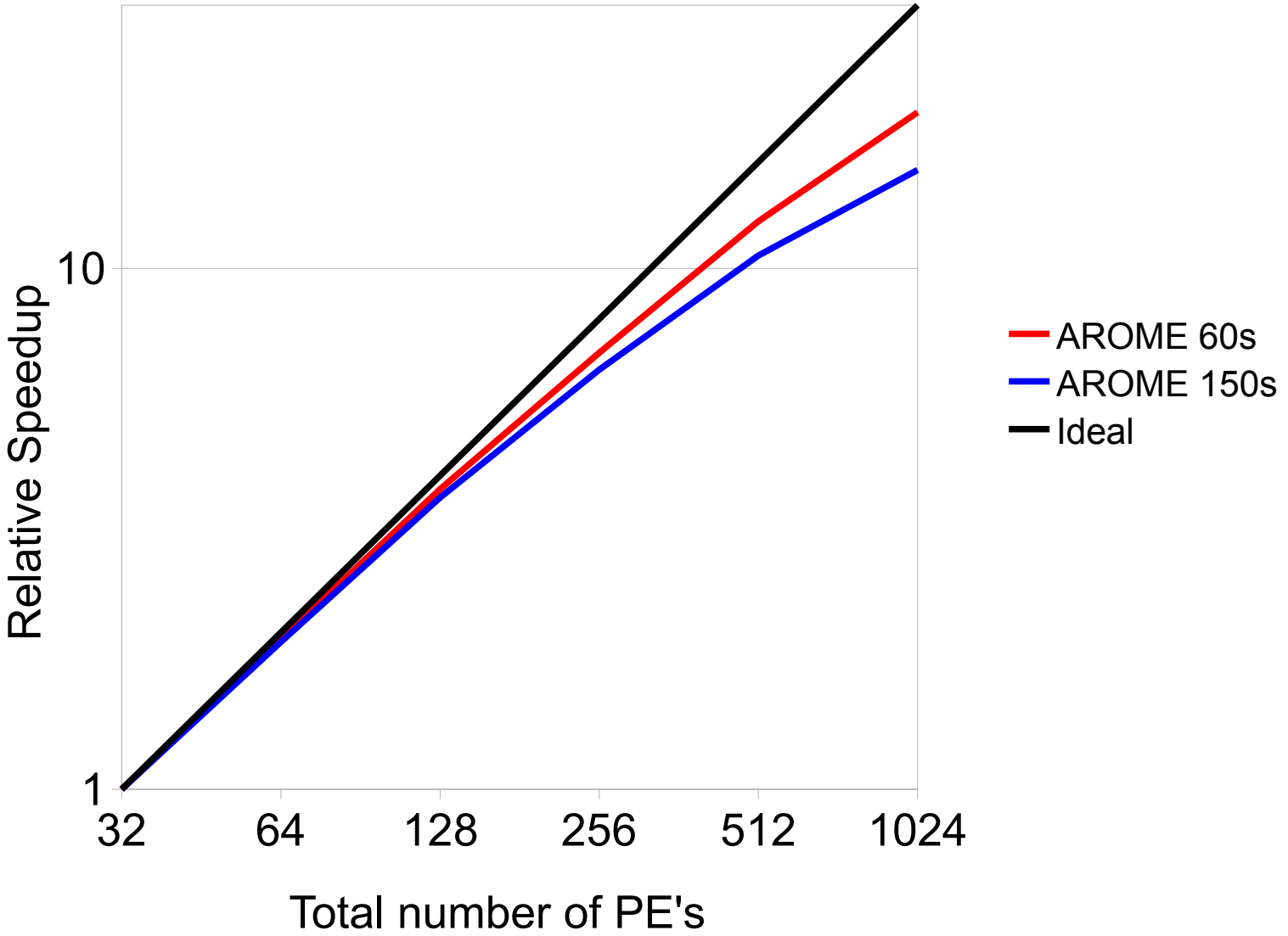
Effect of reading coupling data

AROME, with and without coupling



AROME With different time-steps

Useless computation makes scalable models ;-)



Miscellaneous issues related to the scalability of dynamics

- We can't setup up to 1 gridpoint per MPI task because of an obscure abort in the setup of SL distribution – related to the E-zone presence

  Open-MP is the only way out (...for now)

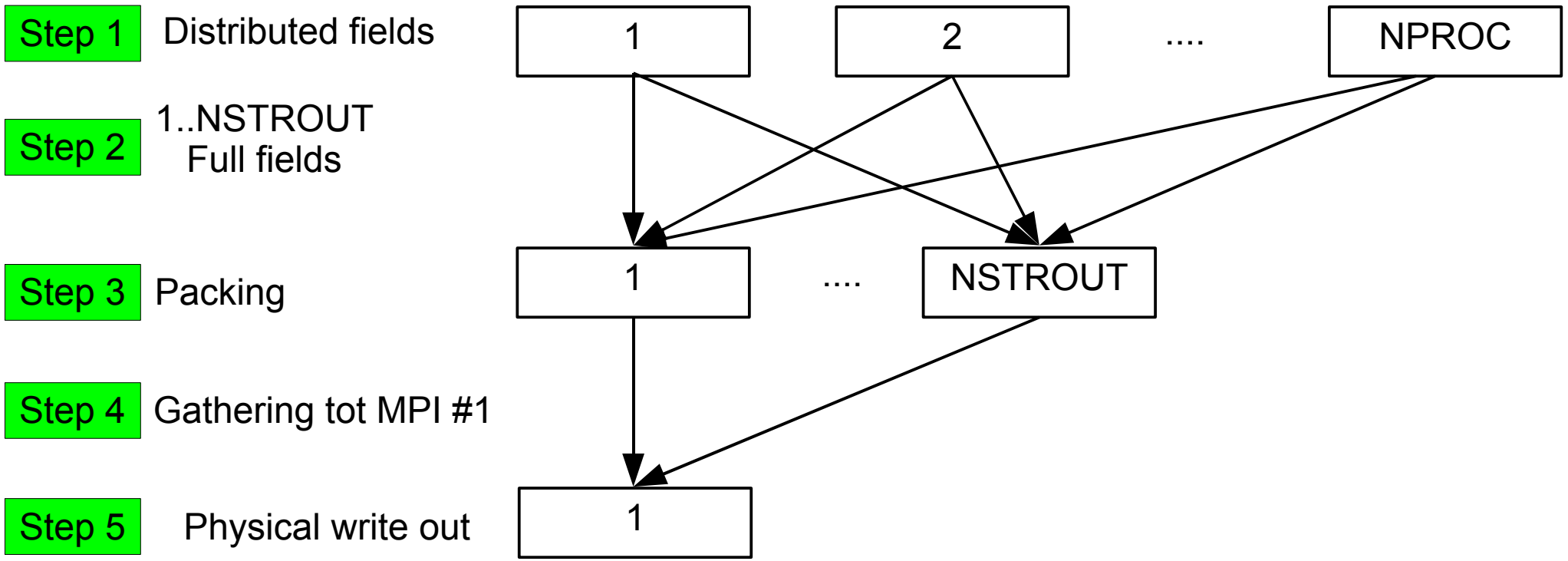
- The first limit that will be reached is in spectral space

  (½) * Number of waves * levels for ≈ optimal load balancing

- Mean wind communications in the middle of the spectral transforms reduces the Open-MP scalability

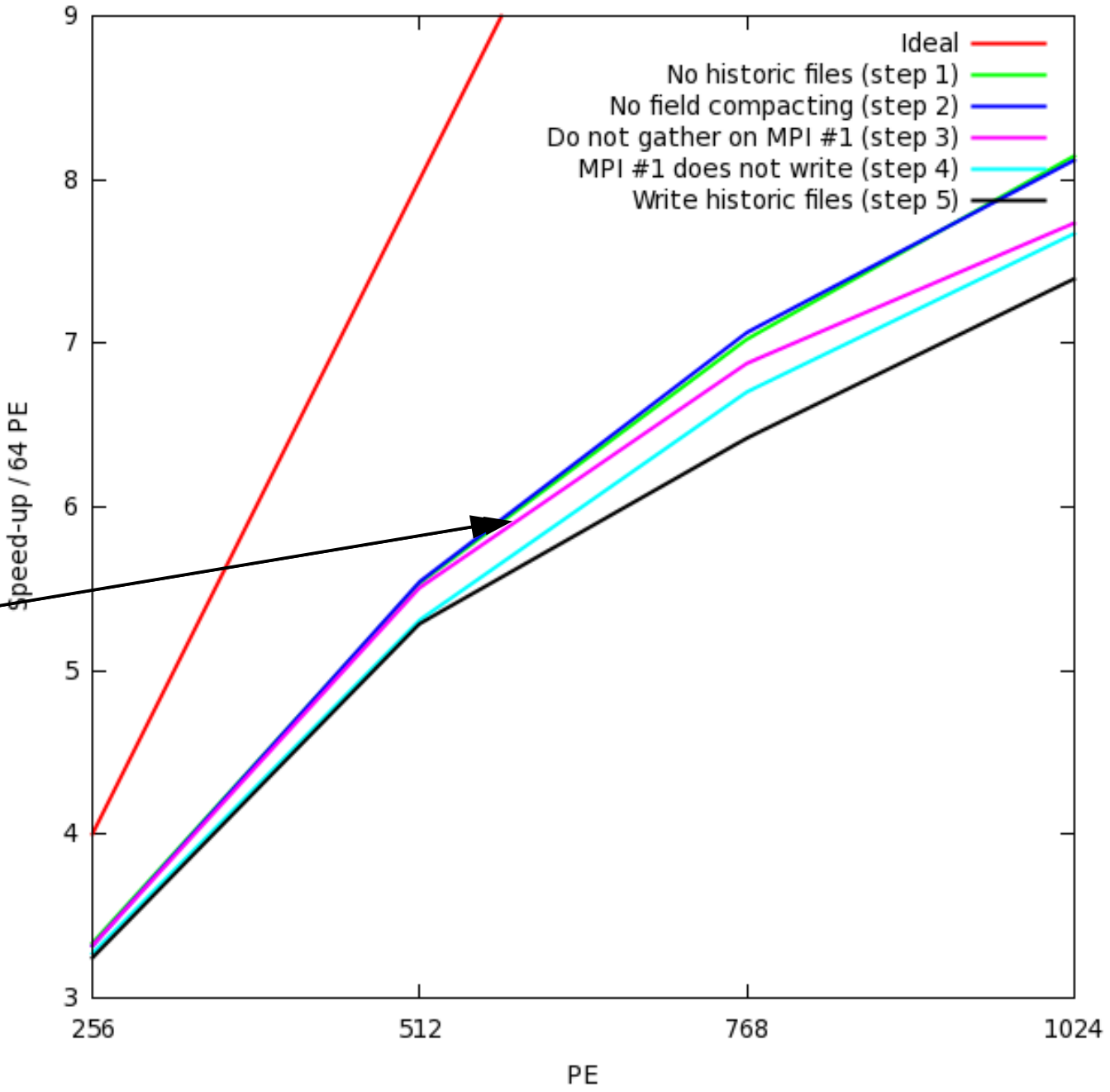
  A solution is under investigation

I/Os in ARPEGE/AROME

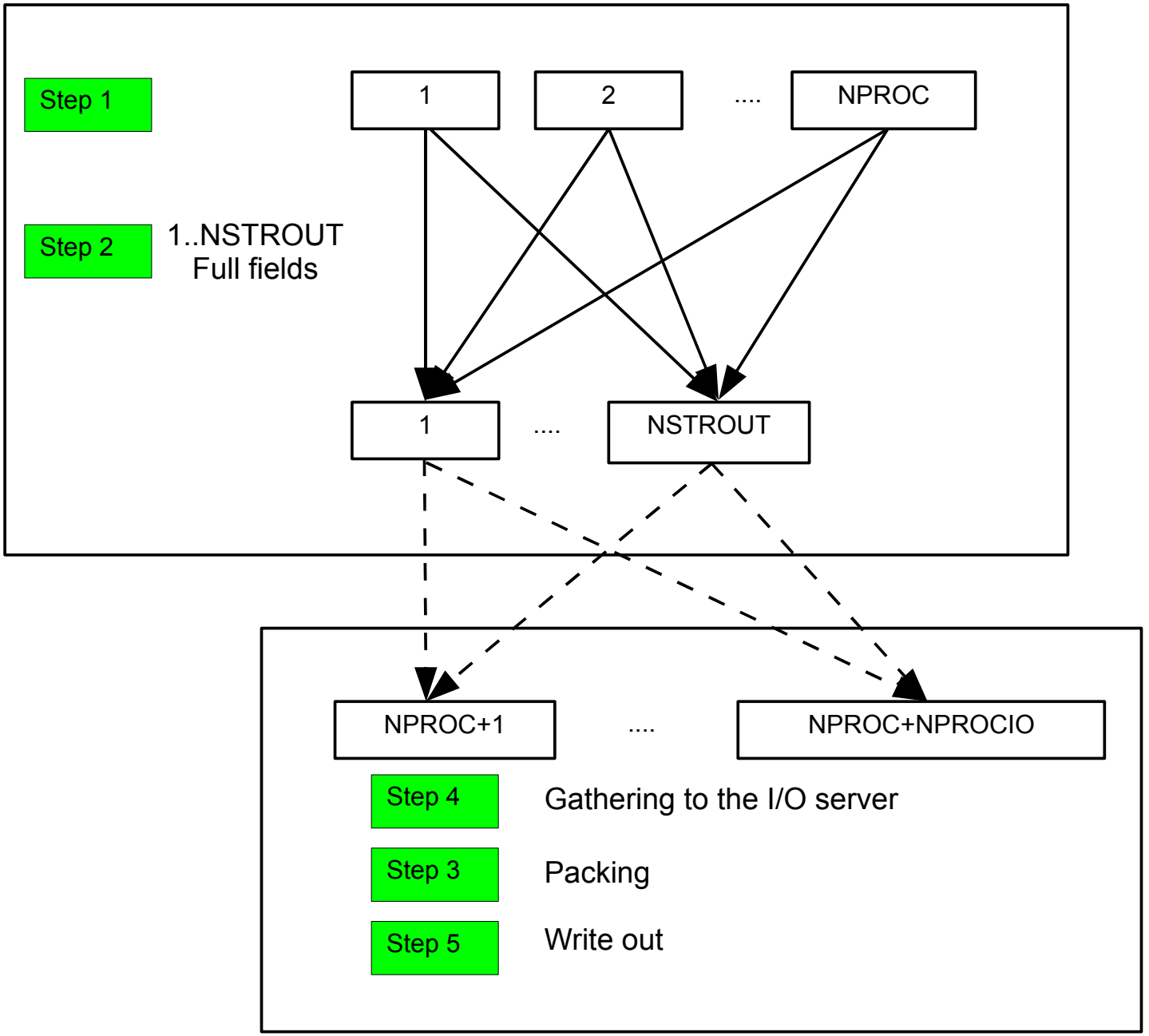


I/Os scalability

≈ 600 fields
Packing is
no more scalable



An I/O server prototype



I/O server testbed

'RAPS' configuration AROME FRANCE E700L87

1440x1350L87 = 169 128 000 points

- 6 hours forecast, 720 time steps of 30 seconds
- Historical files produced every 5 time steps
- Field packing (FA)
- I/O server size = 32 or 64 processors
- IBM/AIX c1a at ECMWF
- Traditional I/Os : 144 files ; 648 Gb = 144 x 4.5 Gb (= 1.6Tb without packing)
- I/O server : ***1 unclosed raw data file per I/O server processor***
- Takes 13 minutes without I/Os and 4096 processors

I/O server - results

Number of procesors	1024	2048	3072	4096
without IO	2120	1197	879	762
with IO	3587	2699	2664	3253
Traditional I/O cost	1467	1502	1785	2491
With I/O server	2355	1385	1114	1013
I/O server cost	235	188	235	251
I/O savings	83,00%	87,00%	86,00%	89,00%

I/O server – conclusion

- Not very difficult to set : 1000 to 2000 lines of code
- More d'I/Os ↔ More processors to dedicate to the I/O server

I/O server is helpful when :

- They are relatively numerous compared to the model computations (nowcasting or massive I/Os of a huge sized model)
- The number of is much exceeding the number of fields (no parallelism on packing)

Conclusions

- AROME has been proven to run properly over thousands of processors
- There is a solution for the I/O scalability problem
- The next computer procurements could take advantage of the I/O server prototype

- Imbalance in physics to be further investigated
- Enhancements in dynamics under progress
- Still work needed on SURFEX