www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Predicting applications performance far far away

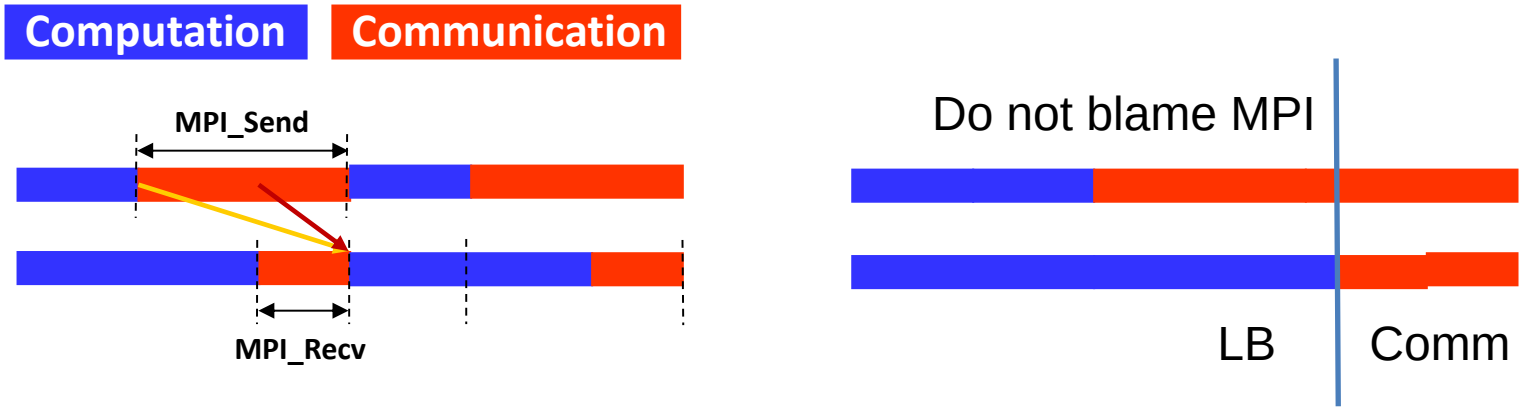
Judit Gimenez (judit@bsc.es)

Scalable Tools Workshop (Tahoe)
August 2015

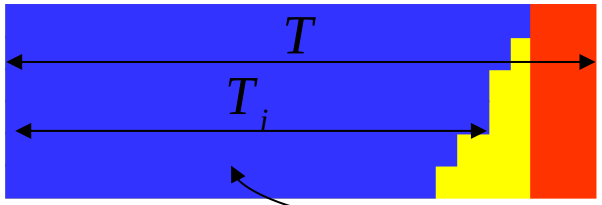
Our believe

Traces for small core counts express the symptoms that would produce performance degradation at larger scale even they have no impact

Parallel efficiency model



- Parallel efficiency = LB eff * Comm eff



$$eff_i = \frac{T_i}{T}$$

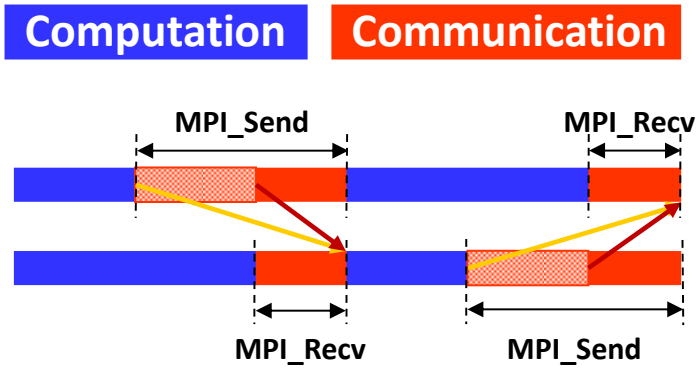
$$CommEff = \max(eff_i)$$

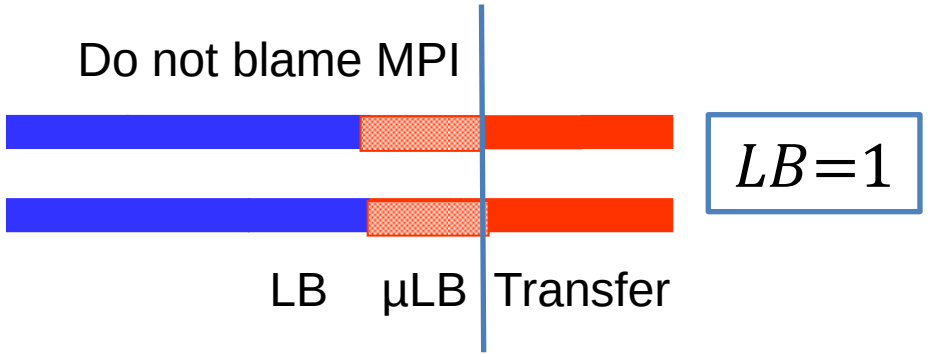
$$LB = \frac{\sum_{i=1}^{P} eff_i}{P * \max(eff_i)}$$

IPC

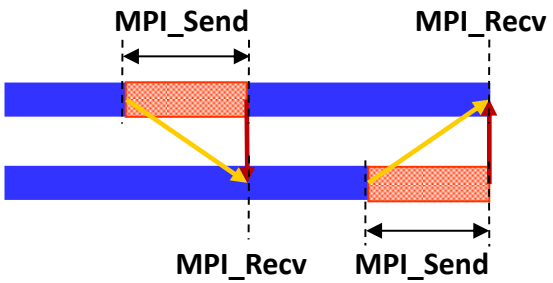
instr

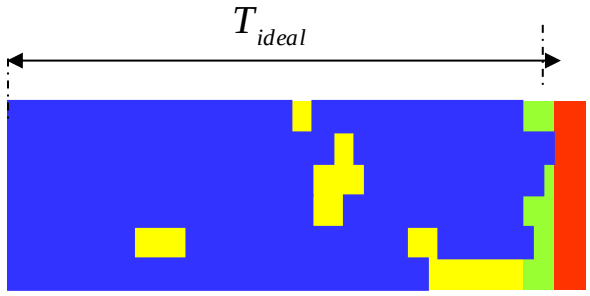
Parallel efficiency refinement: LB * µLB * Transfer





- Serializations / dependences (µLB)
- Dimemas ideal network → Transfer efficiency = 1





Migrating/local load imbalance
Serialization

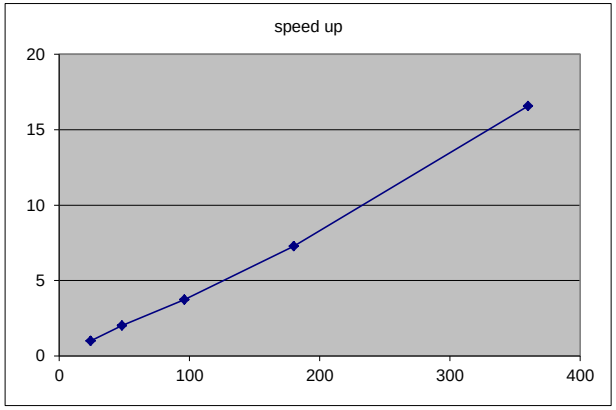
$$\mu LB = \frac{\max(T_i)}{T_{ideal}} \qquad Transfer = \frac{T_{ideal}}{T}$$

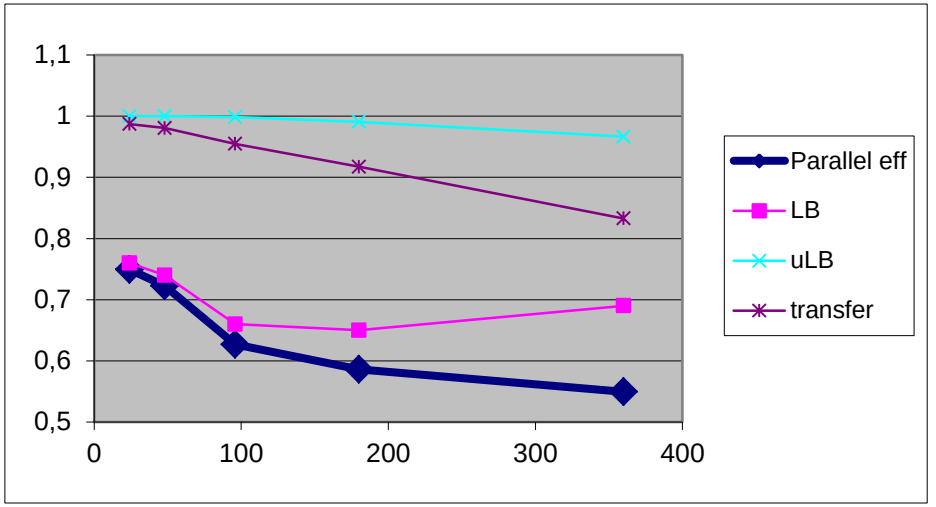
Why scaling?

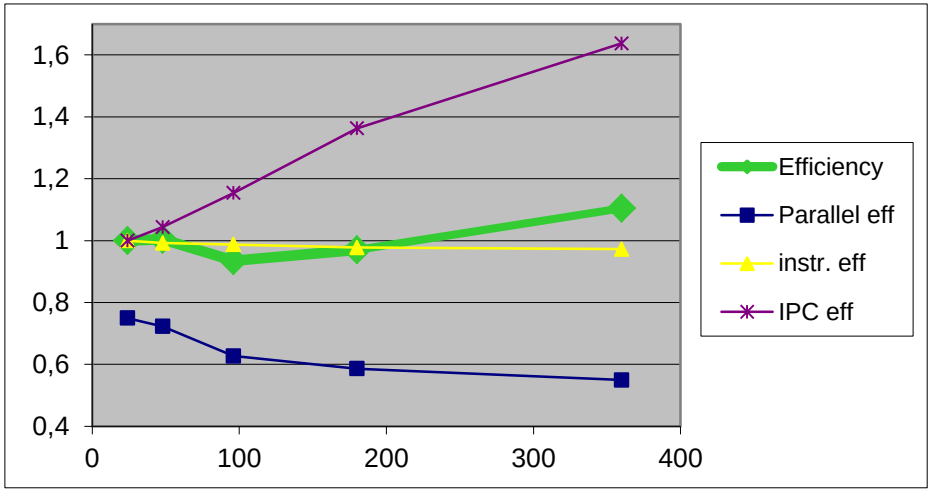
$$\eta_{\parallel} = LB * Ser * Trf$$

CG-POP mpi2s1D - 180x120

Good scalability !!
Should we be happy?



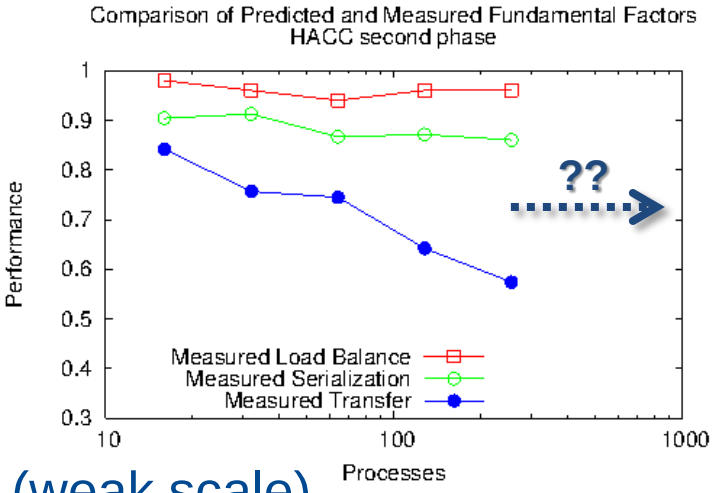




$$\eta = \eta_{\parallel} * \eta_{instr} * \eta_{IPC}$$

Scalability prediction

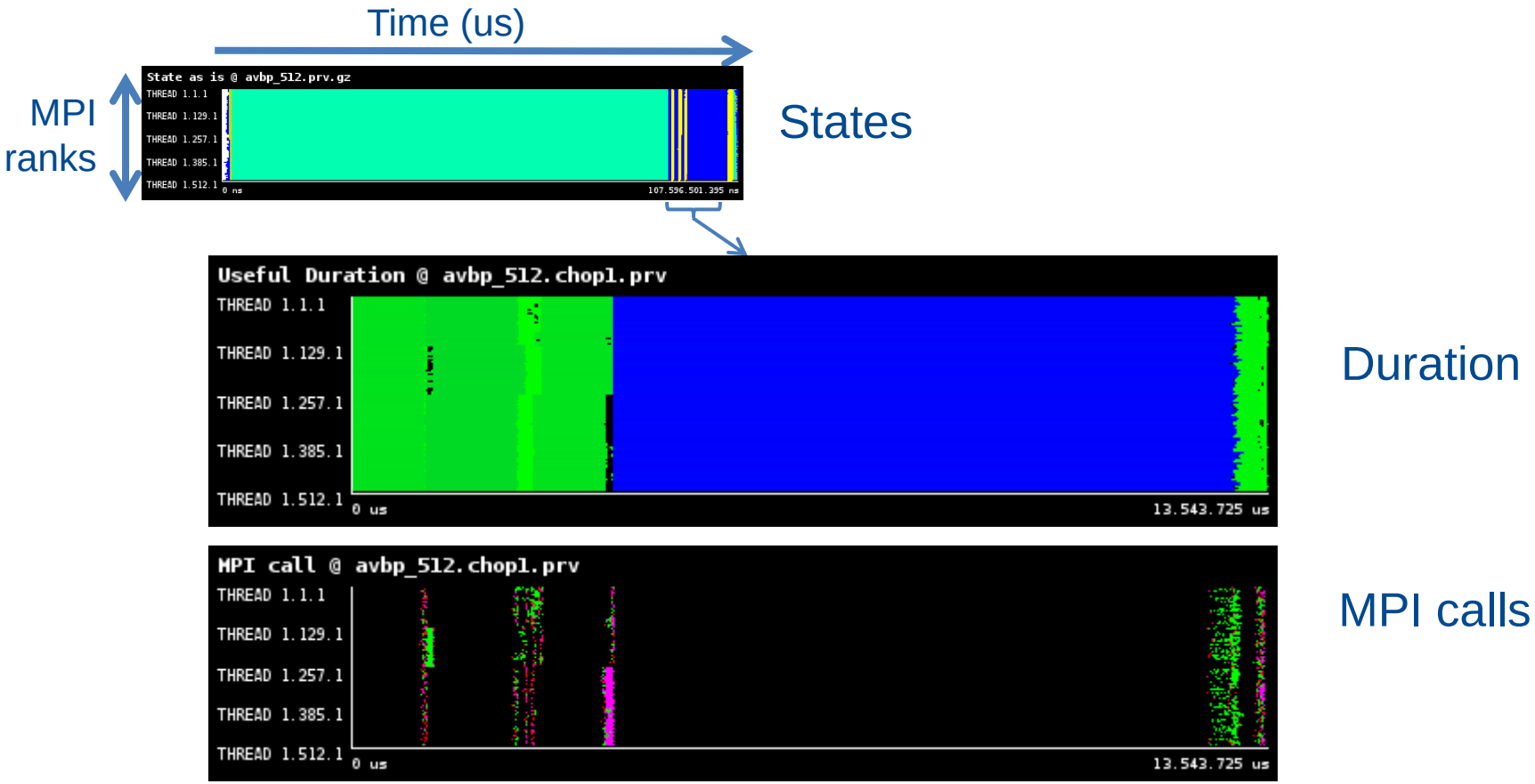
- Efficiency extrapolation
  - From few executions @ low core counts
- Fit based on
  - Reasonable fundamental behavioral models (e.g. Amdahl)
  - Guided by observed internal application behavior (tools)



- Examples from DEEP project
  - AVBP – CFD code (strong scale)
  - TURBORVB – Montecarlo simulation (weak scale)

AVBP structure

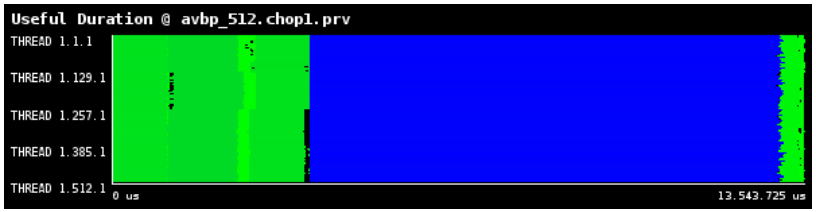
Application structure



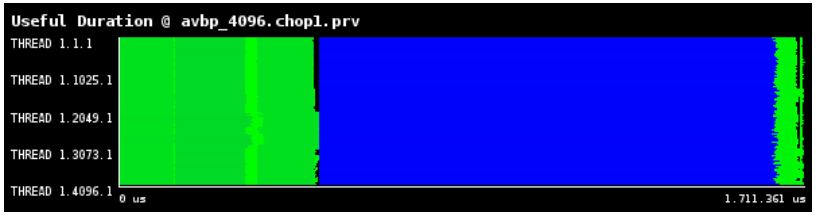
AVBP scalability

- Input
  - BG/Q runs 512, 1024, 2048, 4096
  - Pure MPI executions, strong scale
  - CFD simulations
- Analysis
  - Load Balance: aprox. constant

512



4096



Computations histogram

512

4096

AVBP scalability

- Analysis (cont.)
  - Communications – similar pattern
    ↑ avg #point 2 point calls

512

MPI call profile @ avbp_512.chop1.prv

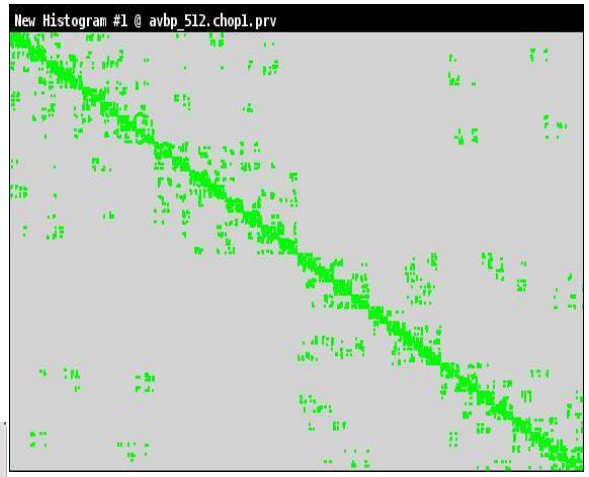
	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Allreduce
THREAD 1.502.1	88	88	11	2
THREAD 1.503.1	154	154	11	2
THREAD 1.504.1	165	165	11	2
THREAD 1.505.1	110	110	11	
THREAD 1.506.1	143	143	11	
THREAD 1.507.1	77	77	11	
THREAD 1.508.1	165	165	11	
THREAD 1.509.1	176	176	11	
THREAD 1.510.1	176	176	11	
THREAD 1.511.1	165	165	11	
THREAD 1.512.1	154	154	11	
Total	66.308	66.308	5.632	
Average	129,507812	129,507812	11	
Maximum	231	231	11	
Minimum	11	11	11	
StDev	38,135244	38,135244	0	
Avg/Max	0,560640	0,560640	1	

4096

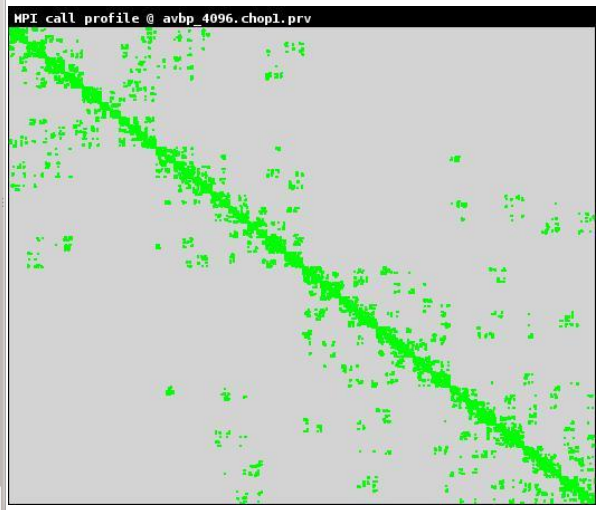
MPI call profile @ avbp_4096.chop1.prv

	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Allreduce
THREAD 1.4086.1	88	88	11	2
THREAD 1.4087.1	220	220	11	2
THREAD 1.4088.1	132	132	11	2
THREAD 1.4089.1	154	154	11	2
THREAD 1.4090.1	154	154	11	2
THREAD 1.4091.1	165	165	11	2
THREAD 1.4092.1	176	176	11	2
THREAD 1.4093.1	165	165	11	2
THREAD 1.4094.1	187	187	11	2
THREAD 1.4095.1	187	187	11	2
THREAD 1.4096.1	198	198	11	2
Total	614.768	614.768	45.056	8.192
Average	150,089844	150,089844	11	2
Maximum	231	231	11	2
Minimum	55	55	11	2
StDev	33,075305	33,075305	0	0
Avg/Max	0,649740	0,649740	1	1

512

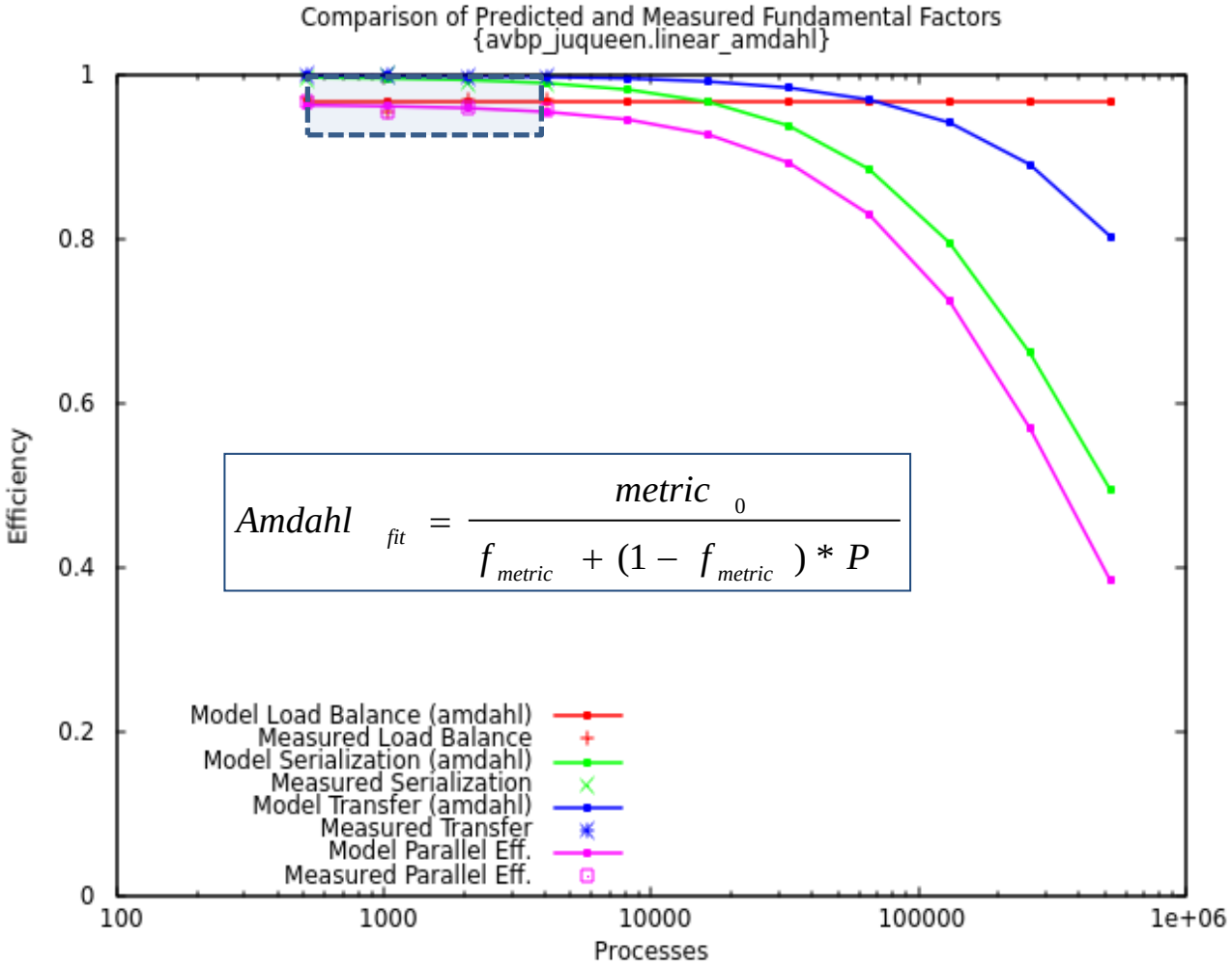


4096



AVBP extrapolation

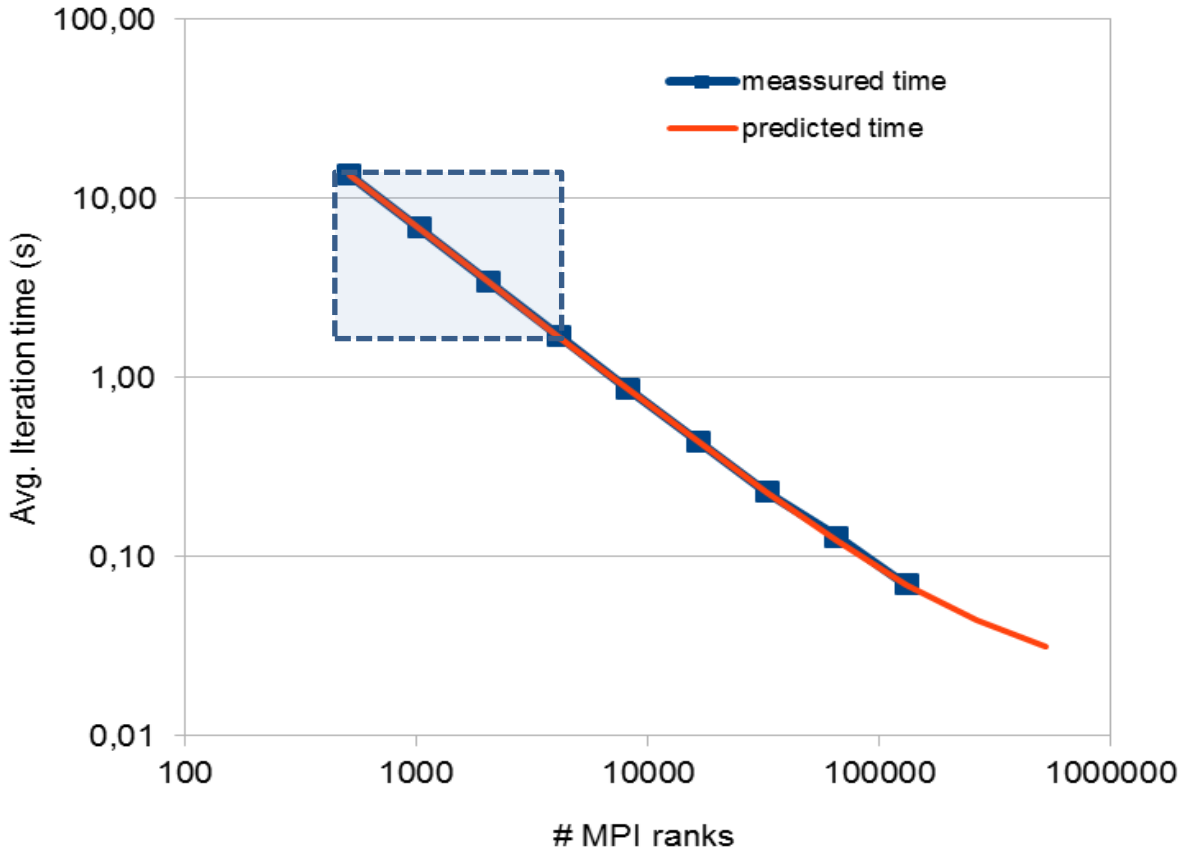
- Input: runs 512 – 4096



Strong scale → small computations betweeen MPI calls become more and more important

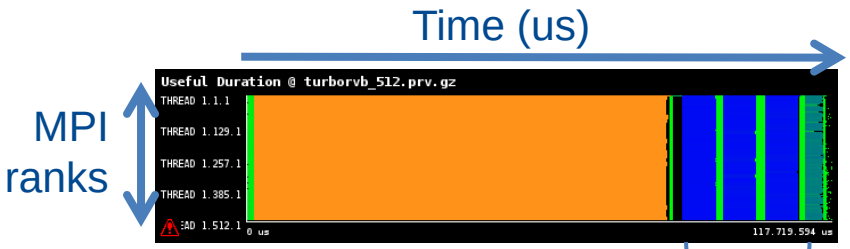
AVBP validation

- Average iteration time (seconds)
  - From Parallel efficiency (no instr., no IPC)

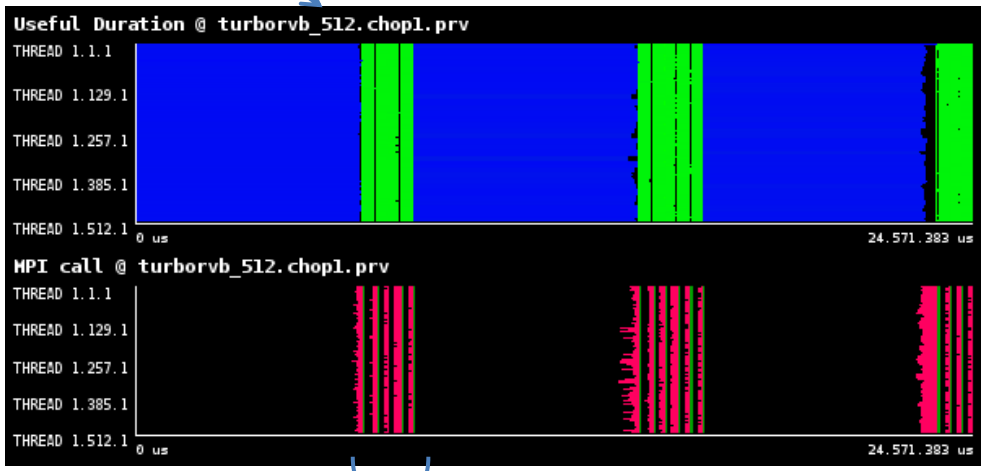


TURBORVB structure

Application structure

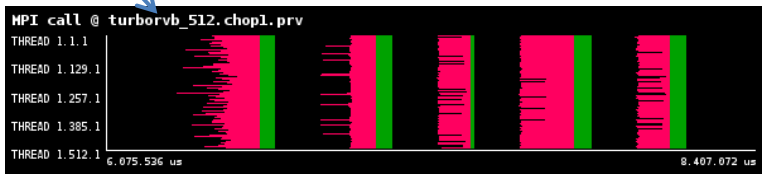


Duration of computations between MPI calls



Duration

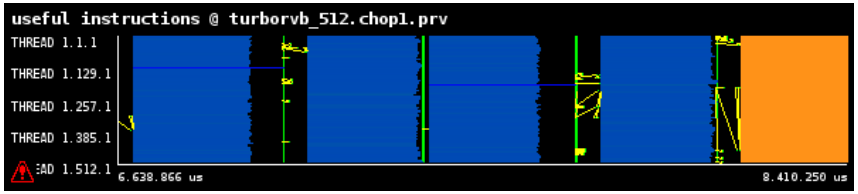
MPI calls



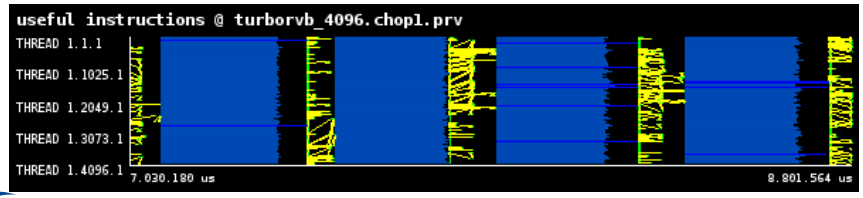
TURBORVB scalability

- Input
  - runs 512, 1024, 2048, 4096
  - Pure MPI executions
  - Montecarlo simulations
- Modeling
  - Load Balance: Small random unbalance

512

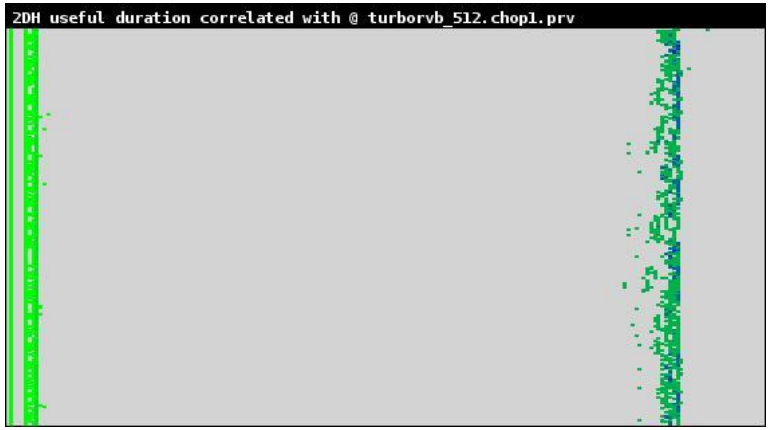


4096

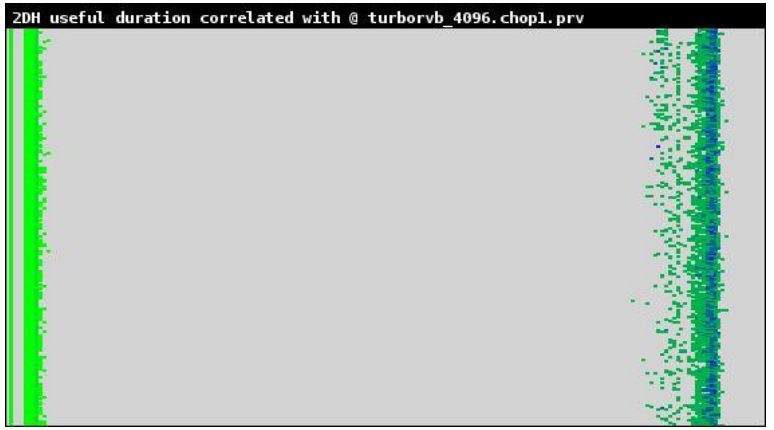


Computations histogram

512



4096

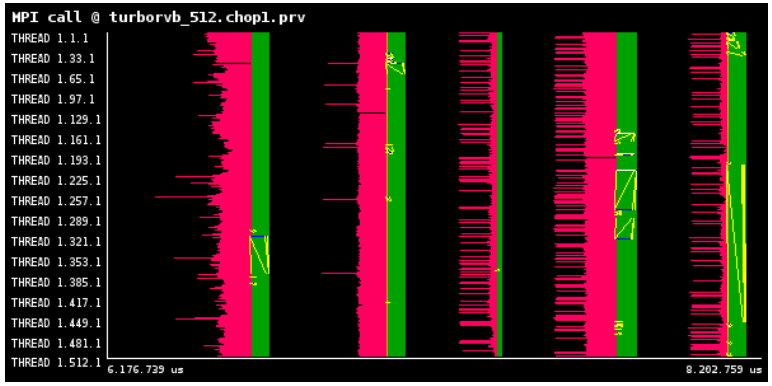


TURBORVB scalability

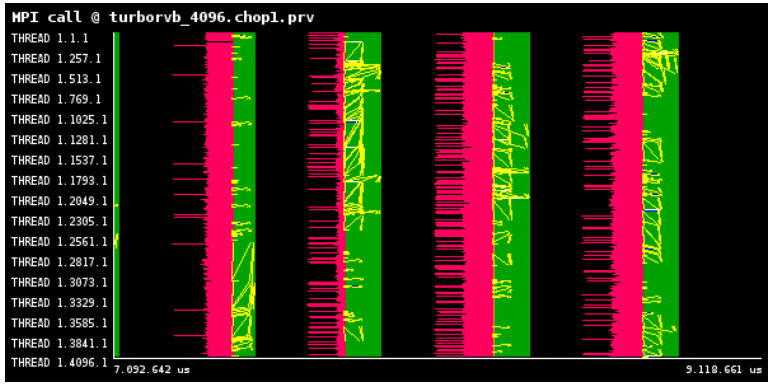
- Modeling (cont.)
  - Communications (Serialization, Transfer)
    - Collectives: Constant (size, number)
    - Point to point
      - Few processes do chains of 7 communications
      - Average calls per iteration per process constant (0.25)
      - Probability to generate contention on node network devices smoothly increases with the number of cores

$$Amdahl_{fit} = \frac{metric_0}{f_{metric} + (1 - f_{metric}) * P^{1/3}}$$

512

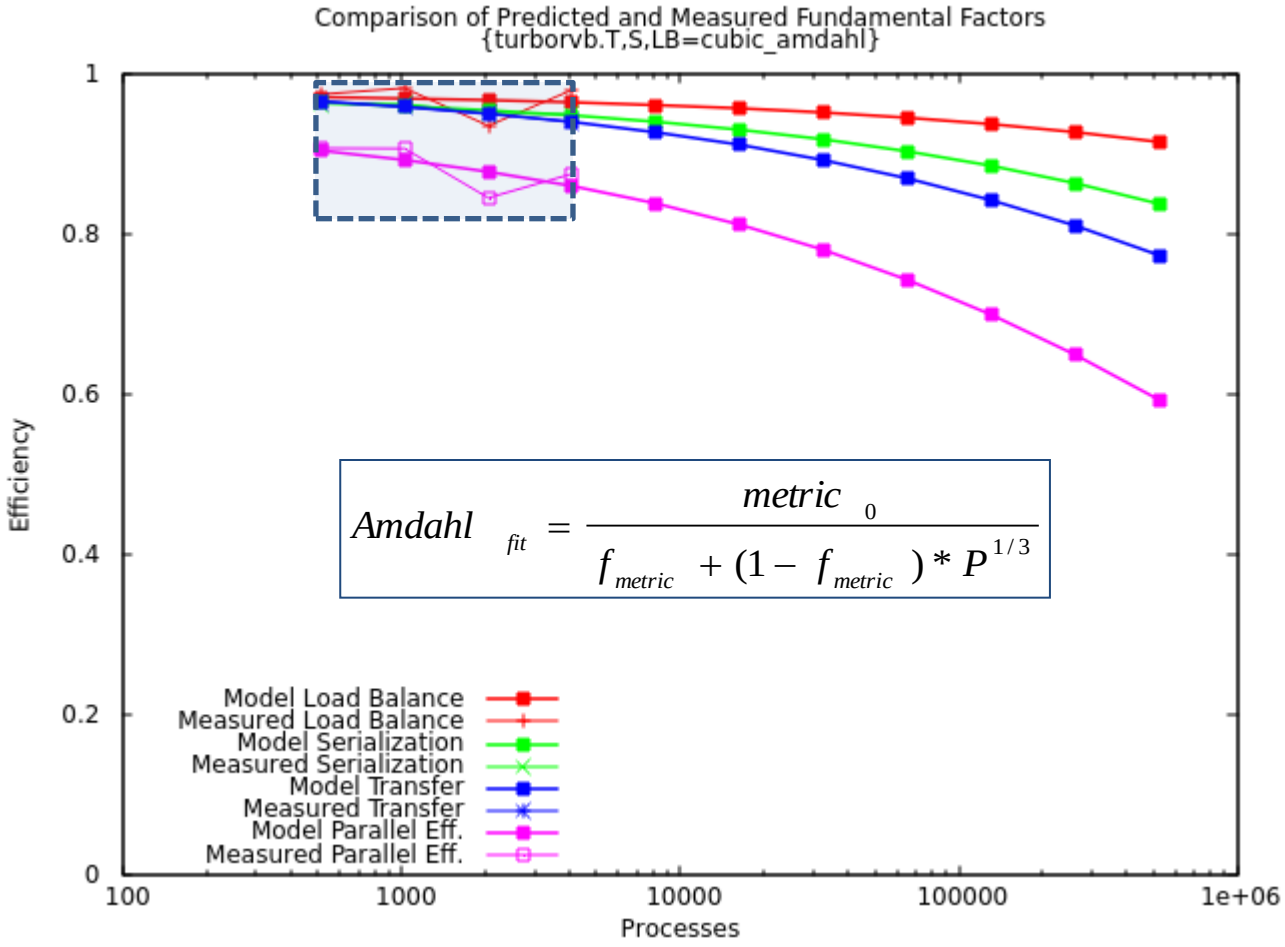


4096



TURBORVB extrapolation

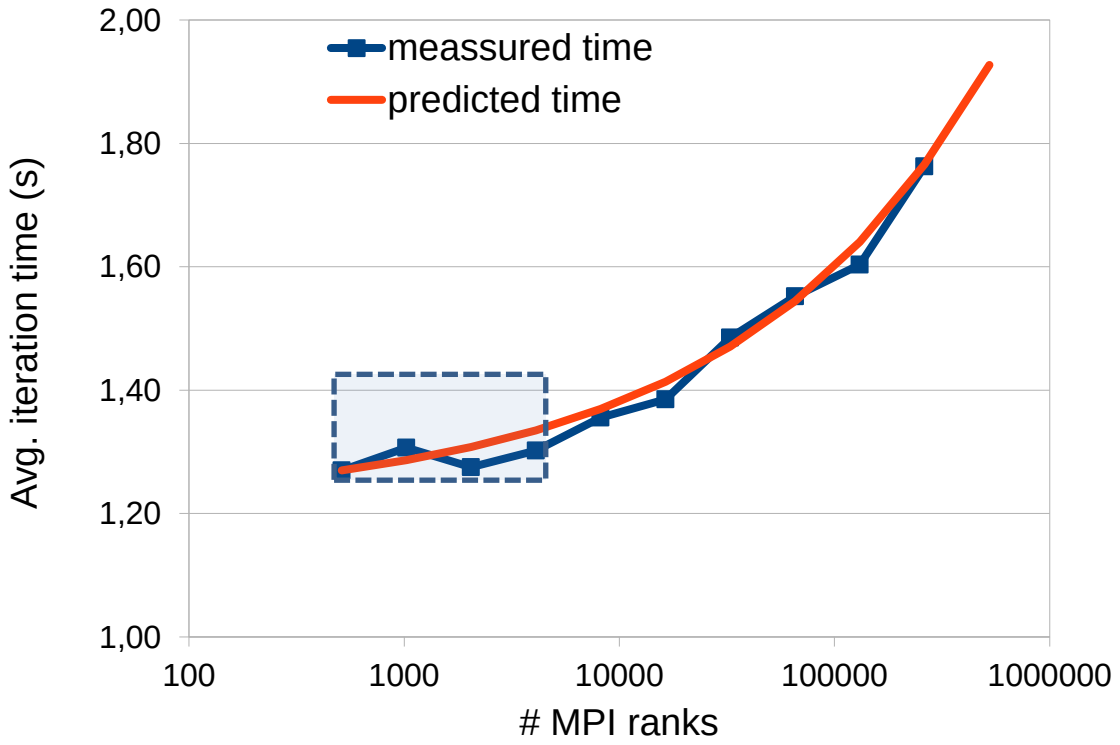
- Input: runs 512 – 4096



Network contention can be reduced balancing / limiting the random selection within a node

TURBORVB validation

- Average iteration time (seconds)
  - From Parallel efficiency (no instr., no IPC)



Conclusions

- 4 measurements can predict performance at 100x
- A lot of insight is obtained digging into traces for small core counts
- Important side effect: identify relevance of different efficiency factors (where to improve)