Reproducibility Issues and Publication Evolutions (in HPC)

Arnaud Legrand

52nd ORAP Forum, Paris
March 2024

Classical Computational Science Challenges

```
my_code --cfg=magical_param:0.94572 '*.dat' --output foo.csv
```

Tracking parameters and data

- `*.dat`? Ooh, you ran this in `data/2091293-AJXQ37`?
- Wasn't `mymap.dat` updated since then?
- That was for `foo.csv`. What about `bar.csv`? Is it reproducible?

Classical Computational Science Challenges

```
my_code --cfg=magical_param:0.94572 '*.dat' --output foo.csv
```

Tracking parameters and data

- `*.dat`? Ooh, you ran this in `data/2091293-AJXQ37`?
- Wasn't `mymap.dat` updated since then?
- That was for `foo.csv`. What about `bar.csv`? Is it reproducible?

Tracking the process (on short/long term)

- Why did I run this? What did I learn from it? When did I do this?

Classical Computational Science Challenges

```
my_code --cfg=magical_param:0.94572 '*.dat' --output foo.csv
```

Tracking parameters and data

- `*.dat`? Ooh, you ran this in `data/2091293-AJXQ37`?
- Wasn't `mymap.dat` updated since then?
- That was for `foo.csv`. What about `bar.csv`? Is it reproducible?

Tracking the process (on short/long term)

- Why did I run this? What did I learn from it? When did I do this?

Tracking code version

- `my_code` is revision `21b95ecfa0911d6ca87668482b11ab9498edd8f3`

Classical Computational Science Challenges

```
my_code --cfg=magical_param:0.94572 '*.dat' --output foo.csv
```

Tracking parameters and data

- `*.dat`? Ooh, you ran this in `data/2091293-AJXQ37`?
- Wasn't `mymap.dat` updated since then?
- That was for `foo.csv`. What about `bar.csv`? Is it reproducible?

Tracking the process (on short/long term)

- Why did I run this? What did I learn from it? When did I do this?

Tracking code version

- `my_code` is revision `21b95ecfa0911d6ca87668482b11ab9498edd8f3`

Tracking software environment

- `my_code` depends on a dozen of libraries, which depend on dozens of libraries
- `my_code` was compiled with `clang 1:9.0-49.1` and `-O3 -funroll-loops -fno-strict-aliasing -finline-functions ...`

Classical Computational Science Challenges

```
my_code --cfg=magical_param:0.94572 '*.dat' --output foo.csv
```

Tracking parameters and data

- `*.dat`? Ooh, you ran this in `data/2091293-AJXQ37`?
- Wasn't `mymap.dat` updated since then?
- That was for `foo.csv`. What about `bar.csv`? Is it reproducible?

Tracking the process (on short/long term)

- Why did I run this? What did I learn from it? When did I do this?

Tracking code version

- `my_code` is revision `21b95ecfa0911d6ca87668482b11ab9498edd8f3`

Tracking software environment

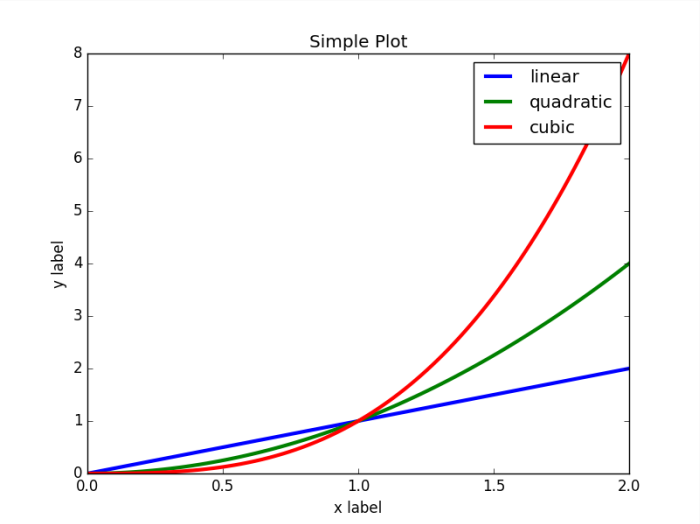
- `my_code` depends on a dozen of libraries, which depend on dozens of libraries
- `my_code` was compiled with `clang 1:9.0-49.1` and `-O3 -funroll-loops -fno-strict-aliasing -finline-functions ...`

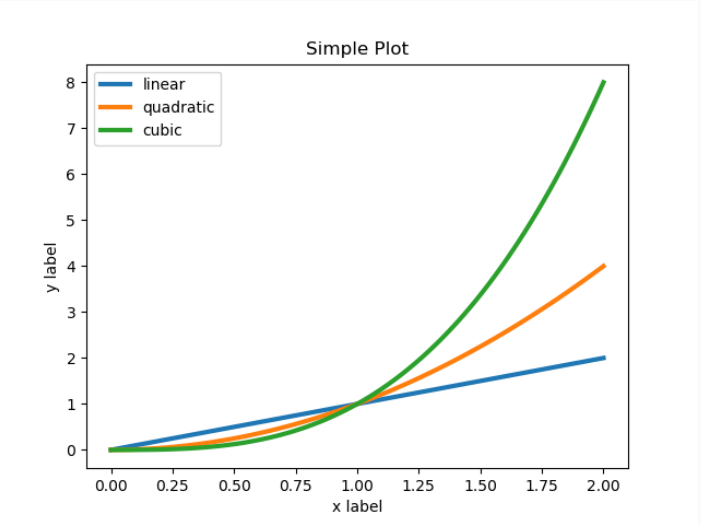
Handle complex sequences and reuse results leverage cloud/supercomputers and their ecosystem (workflow, RJMS, data warehouse)

Software dependencies: horror stories

Software dependencies: horror stories

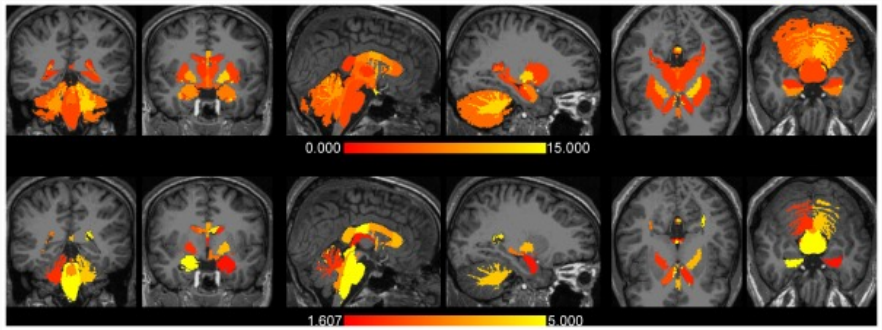
- Software environment evolution





Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity

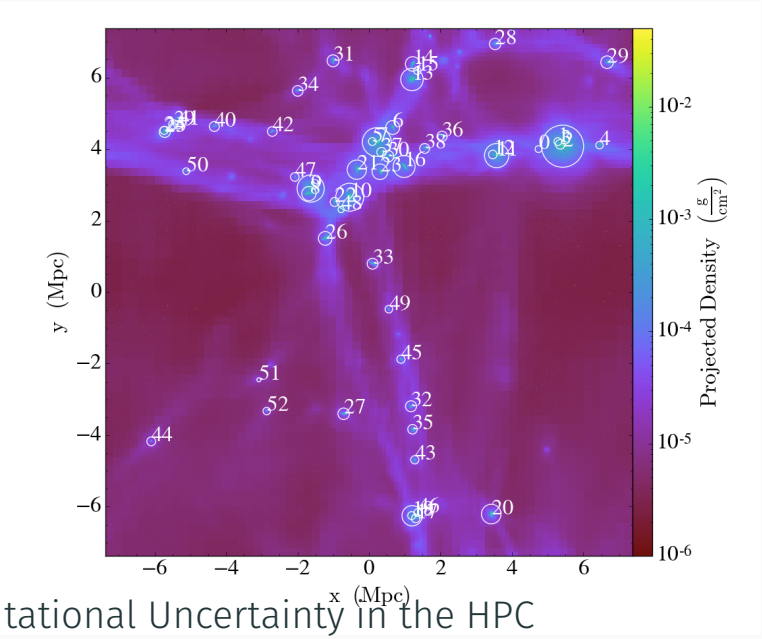


The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements (PLOS ONE, 2012)

> *Significant differences in volume and cortical thickness were revealed across FreeSurfer versions:*
>
> - *volume:* *8.8 ± 6.6% (range 1.3-64.0%)*
> - *cortical thickness:* *2.8 ± 1.3% (range 1.1-7.7%)*
>
> *About a factor two smaller differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.*
>
> *In the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system.*
>
> *Formal assessment of the accuracy of FreeSurfer is desirable.*

Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity
- Impact of the compiler

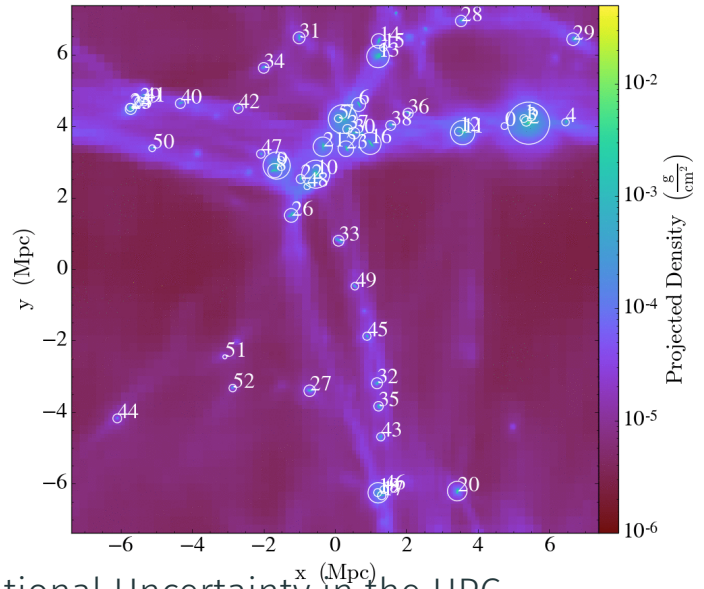


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h

Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity
- Impact of the compiler
  - Not even mentioning the lack of determinism stemming from threads, MPI collective communic and non-commutative operations

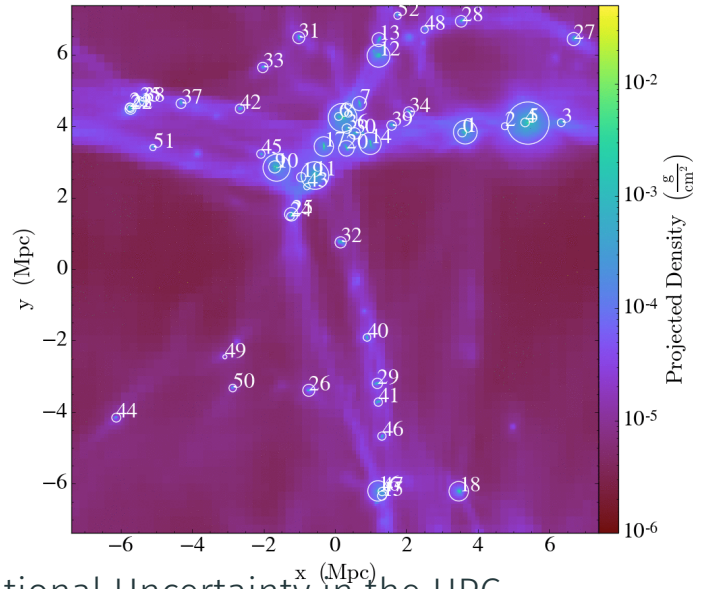


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h

Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity
- Impact of the compiler
  - Not even mentioning the lack of determinism stemming from threads, MPI collective communic and non-commutative operations

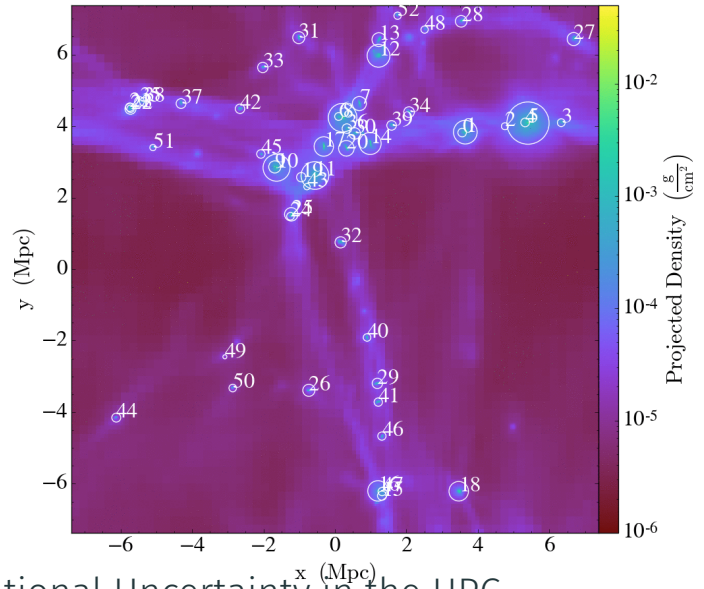


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h

Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity
- Impact of the compiler
  - Not even mentioning the lack of determinism stemming from threads, MPI collective communic and non-commutative operations

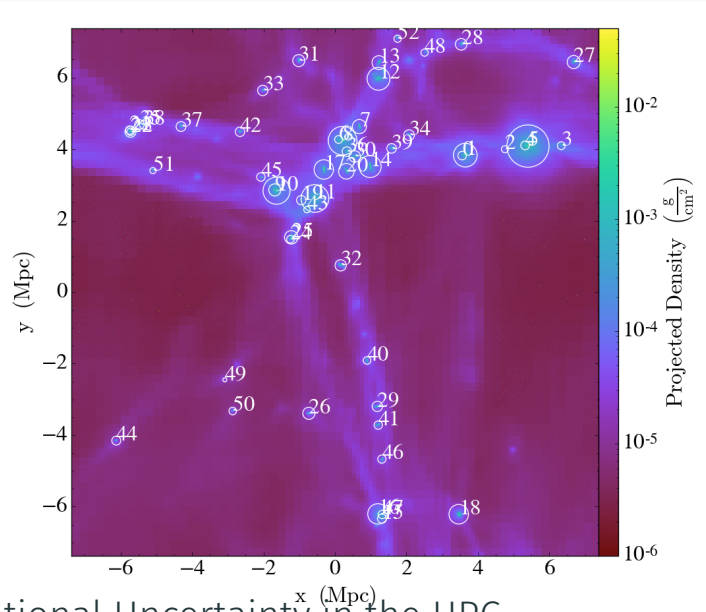


Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h
gcc@6.2.0	Normal	2.266E 46	1.218E 44	10h
gcc@6.2.0	High	2.275E 46	1.199E 44	9h

Software dependencies: horror stories

- Software environment evolution
- OS heterogeneity
- Impact of the compiler
  - Not even mentioning the lack of determinism stemming from threads, MPI collective communic and non-commutative operations



Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context (ResCuE-HPC, 2018)

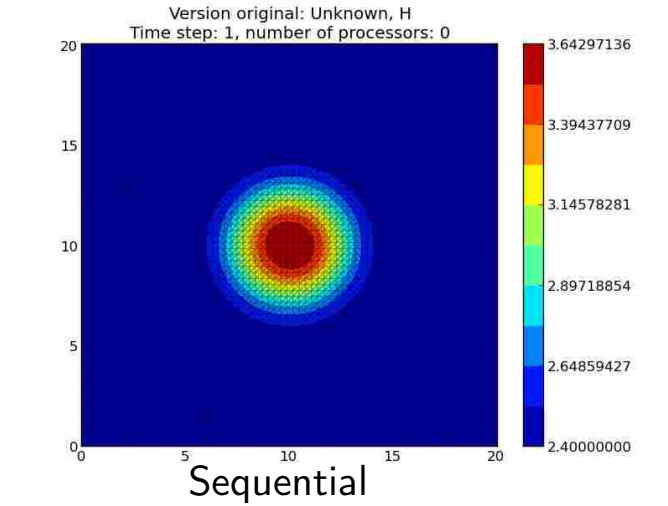
Compiler	Optim.	Largest Halo Avg Mass.	Std. Err	Walltime
gcc@6.2.0	None	2.273E 46	1.069E 44	22h
gcc@6.2.0	Normal	2.266E 46	1.218E 44	10h
gcc@6.2.0	High	2.275E 46	1.199E 44	9h
intel@16.0.3	None	22.71 E 46	1.587E 44	39h
intel@16.0.3	Normal	43.30 E 46	1.248E 44	7h
intel@16.0.3	High	2.268E 46	1.414E 44	6h
cce@8.5.5	Low	43.11 E 46	1.353E 44	16h
cce@8.5.5	Normal	2.271E 46	1.261E 44	6h
cce@8.5.5	High	2.272E 46	1.341E 44	5h

Telemac2D: the simplest gouttedo simulation

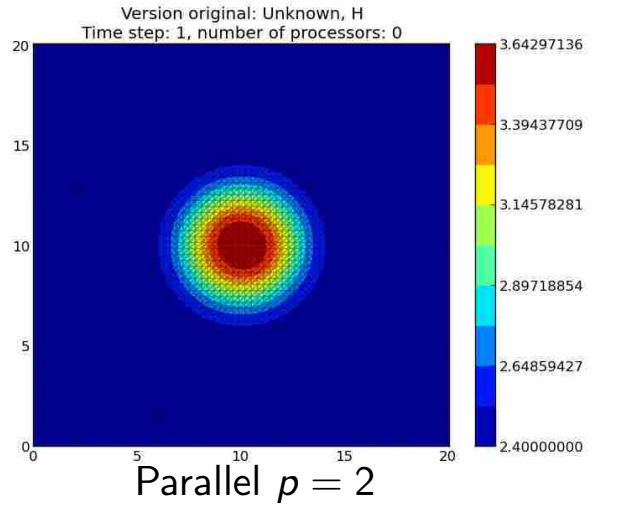
The gouttedo test case

- 2D-simulation of a water drop fall in a square bassin
- Unknown: water depth for a 0.2 sec time step
- Triangular mesh: 8978 elements and 4624 nodes

Expected numerical reproducibility (time step = 1, 2, ...)



Sequential



Parallel $p = 2$

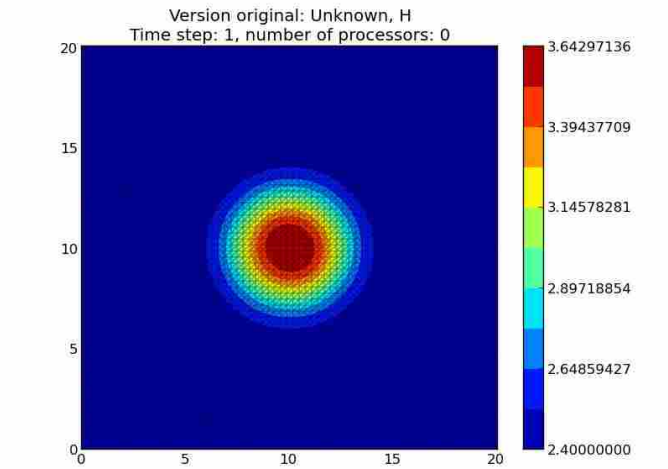
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

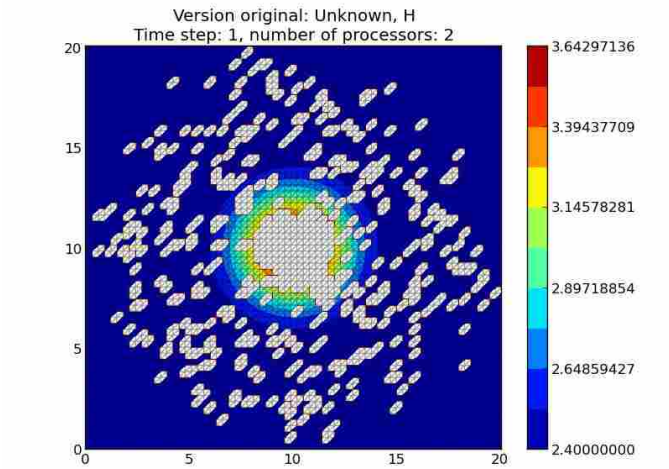
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 1



Sequential



Parallel $p = 2$

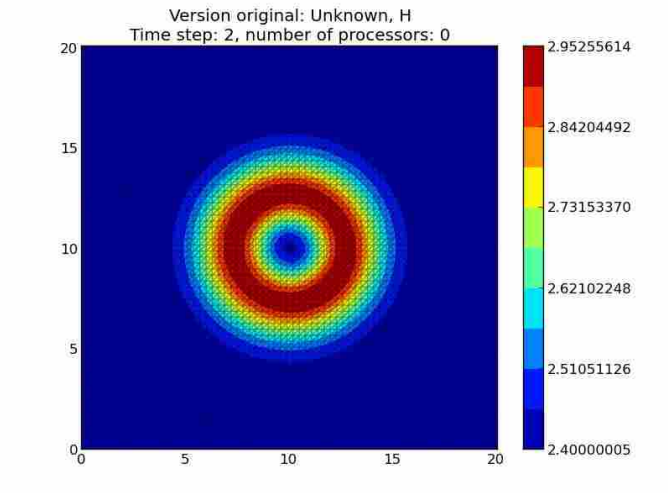
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

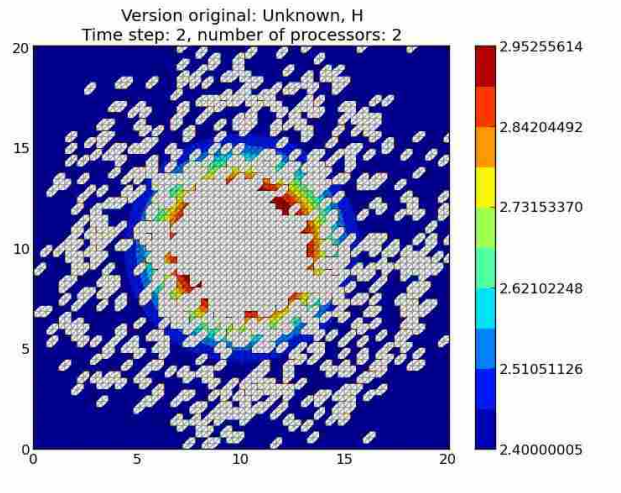
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 2



Sequential



Parallel $p = 2$

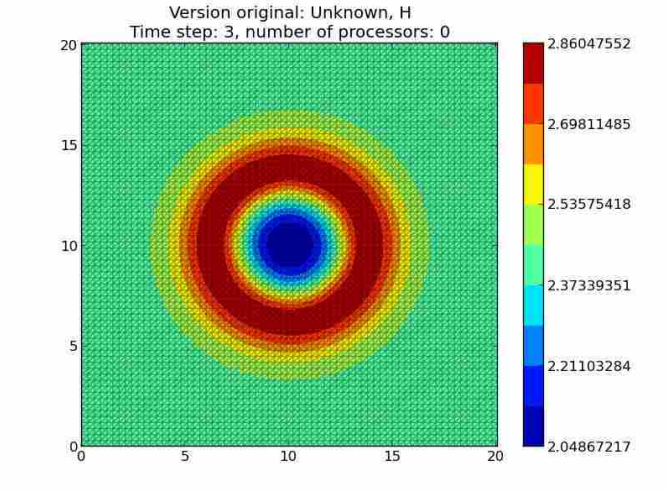
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

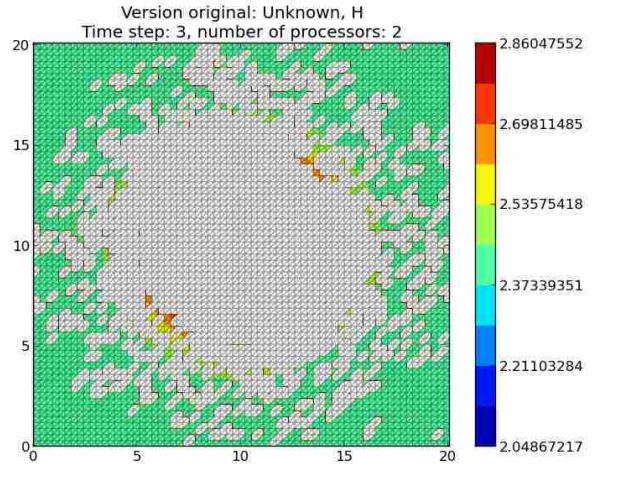
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 3



Sequential



Parallel $p = 2$

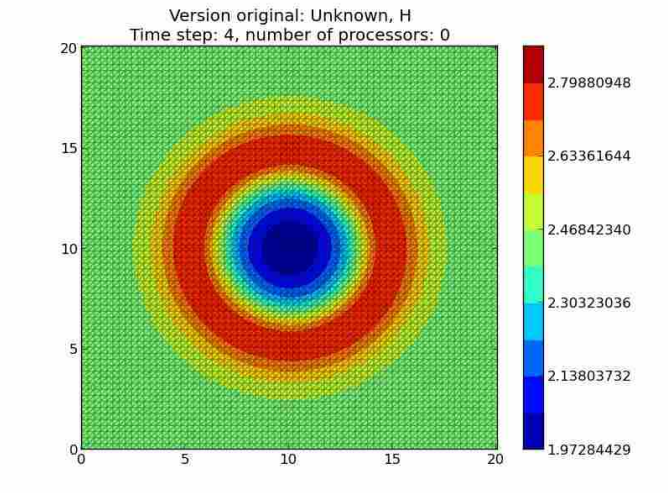
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

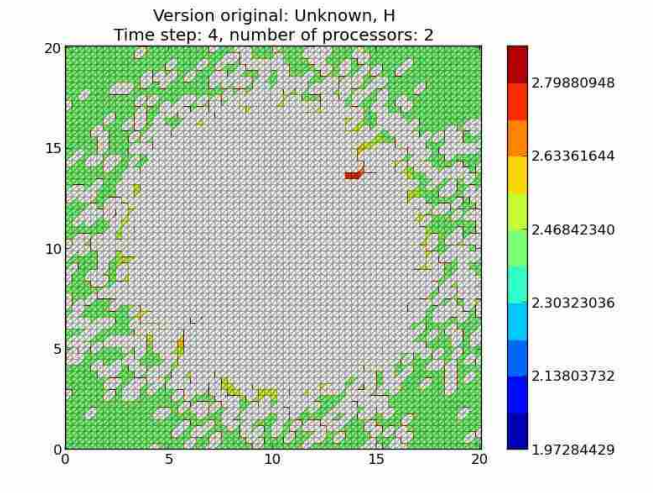
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 4



Sequential



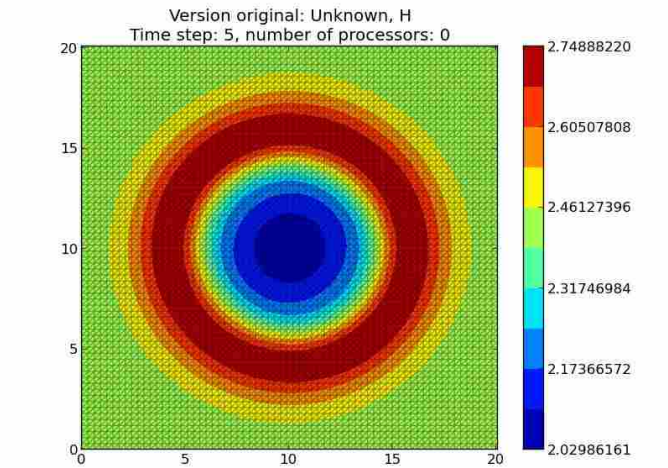
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

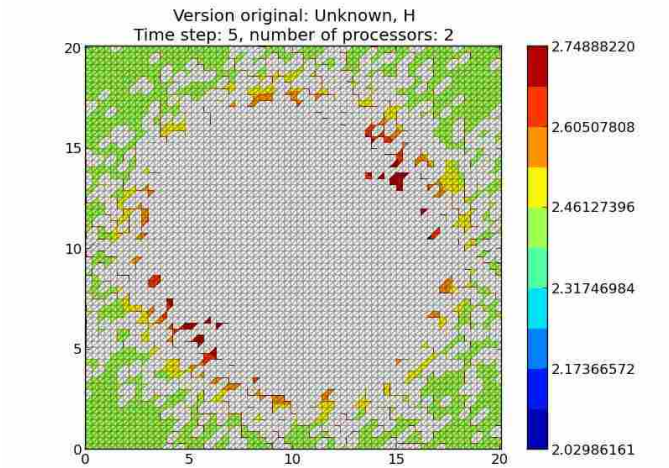
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 5



Sequential



Parallel $p = 2$

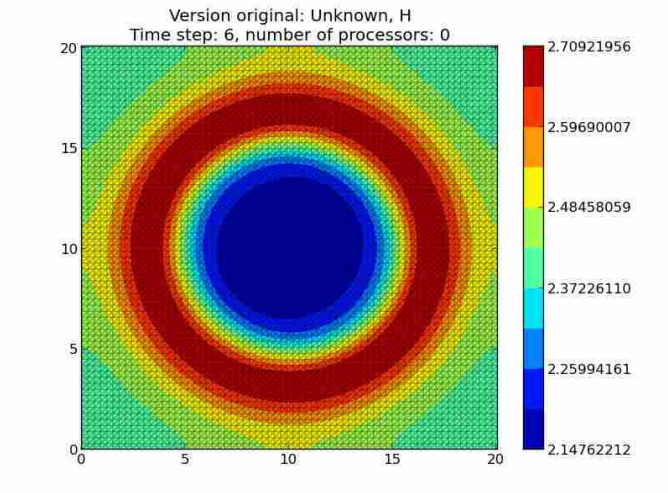
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

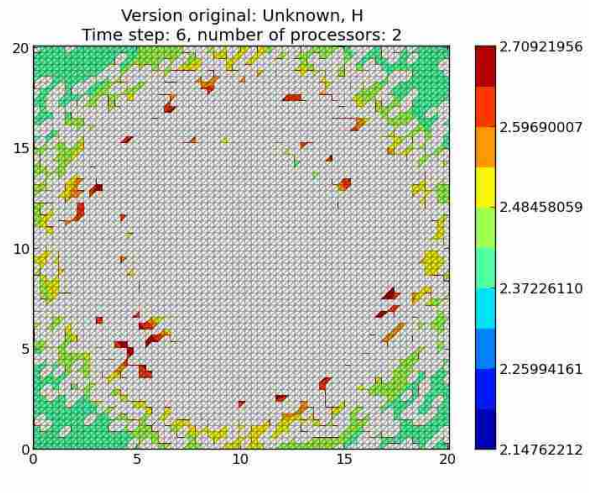
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 6



Sequential



Parallel $p = 2$

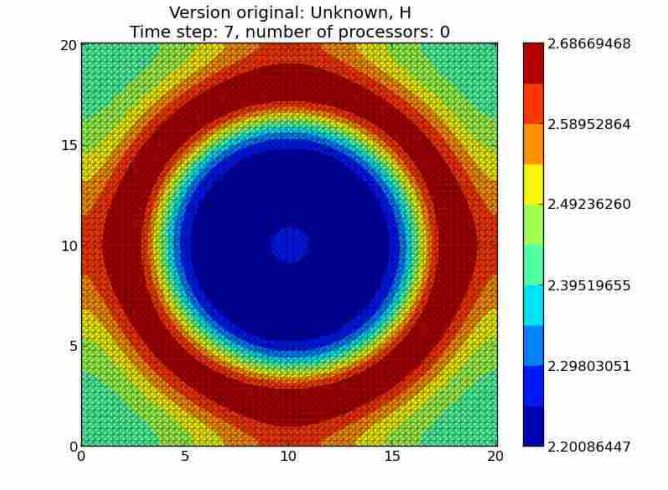
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

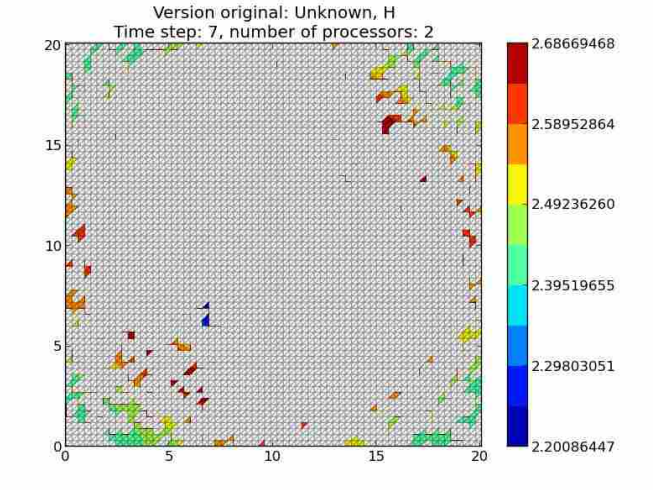
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 7



Sequential



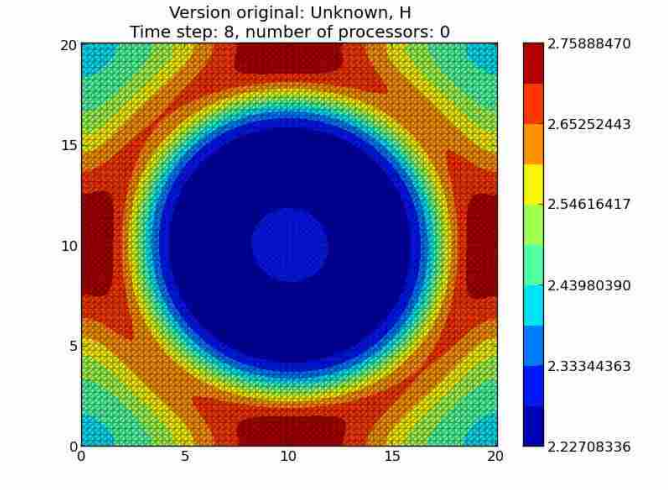
Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

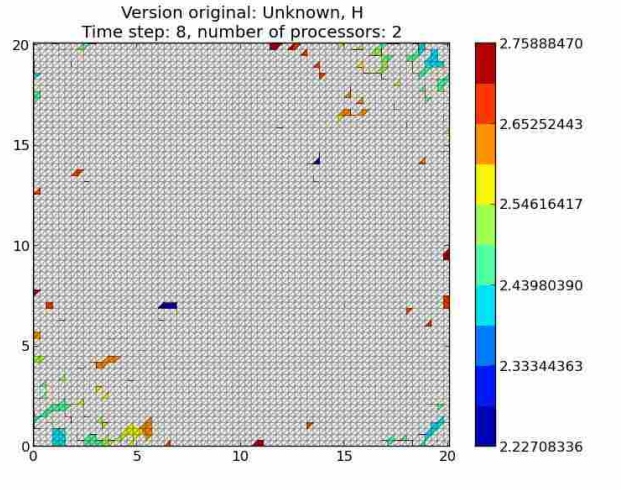
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 8



Sequential



Parallel $p = 2$

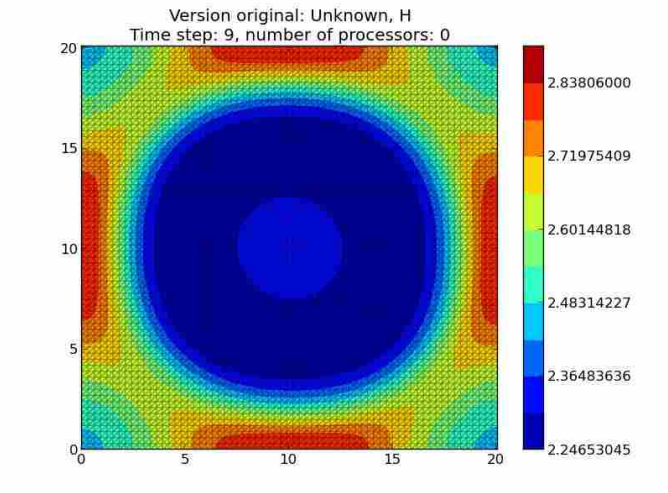
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

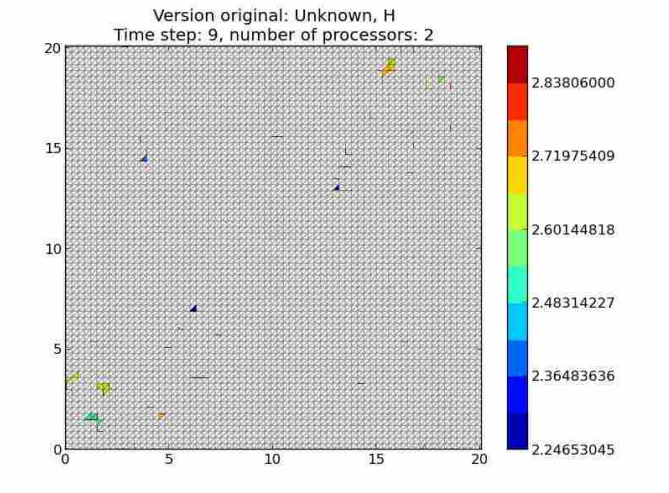
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 9



Sequential



Parallel $p = 2$

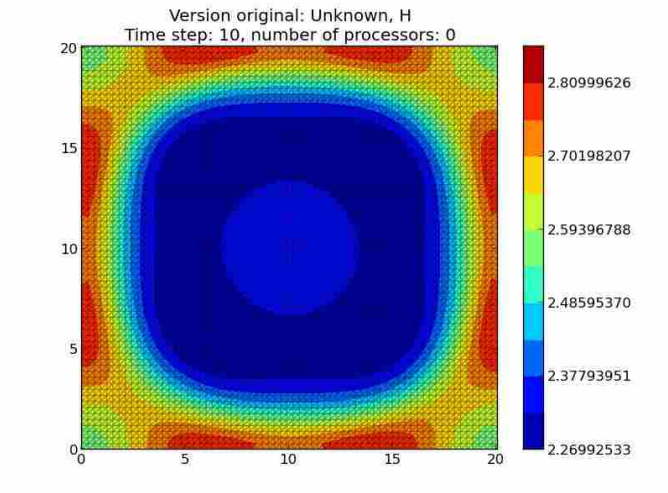
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

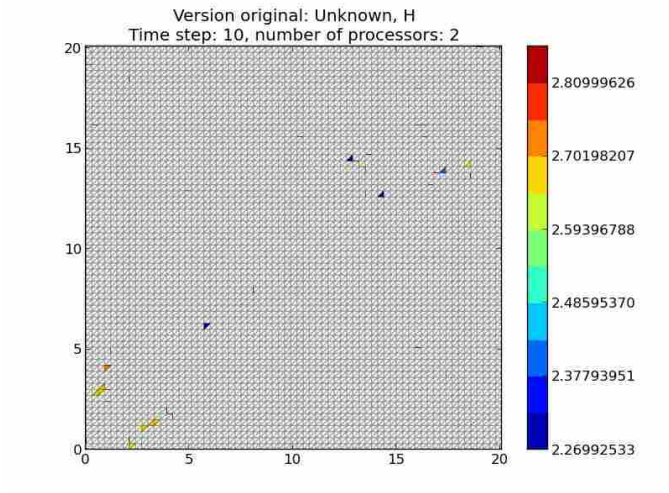
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 10



Sequential



Parallel $p = 2$

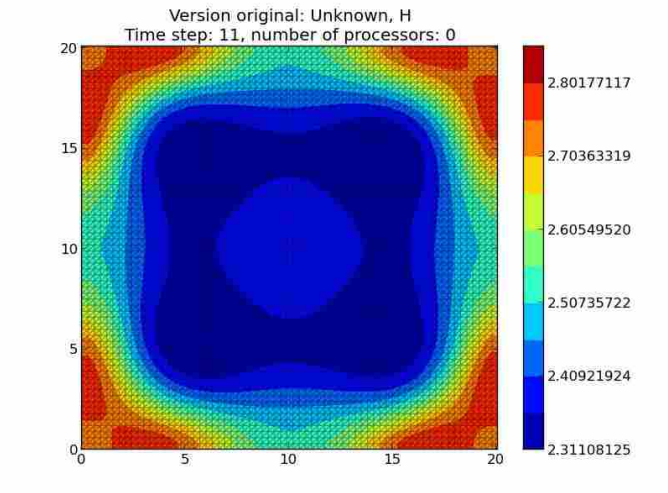
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

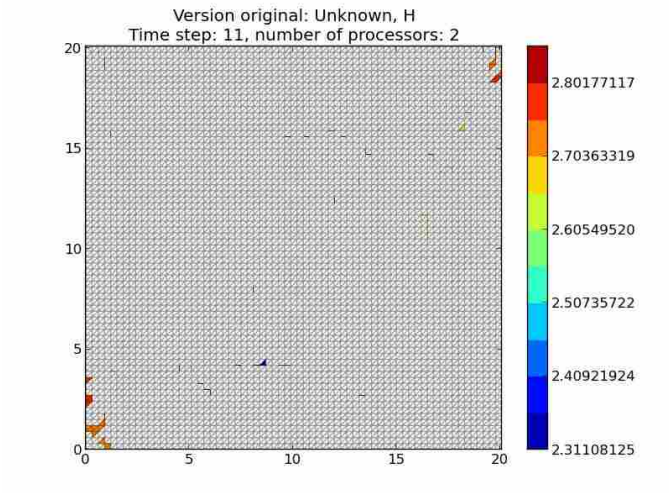
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 11



Sequential



Parallel $p = 2$

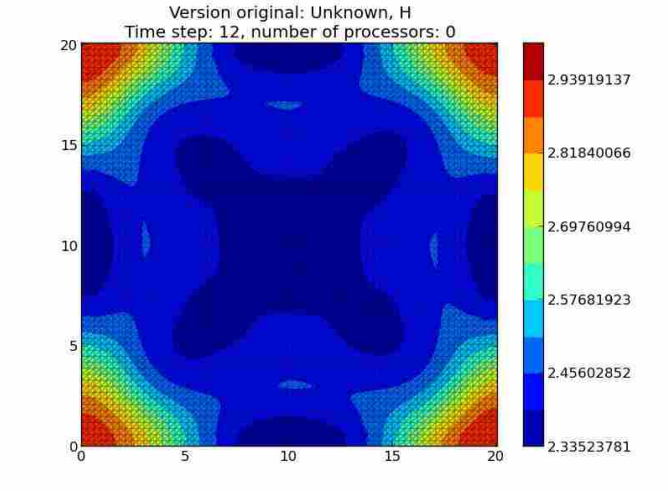
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

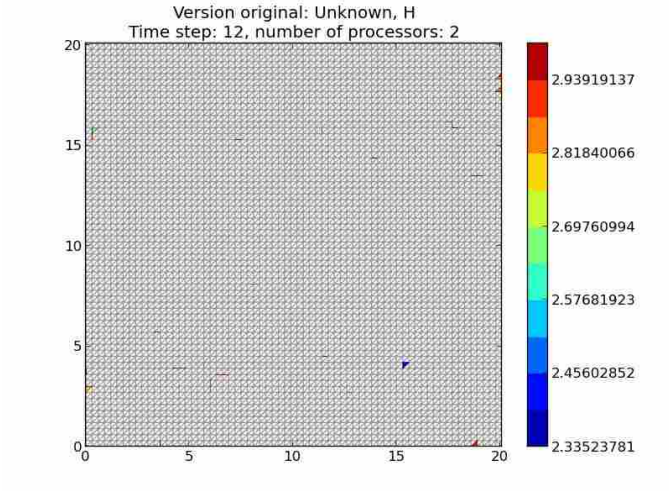
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 12



Sequential



Parallel $p = 2$

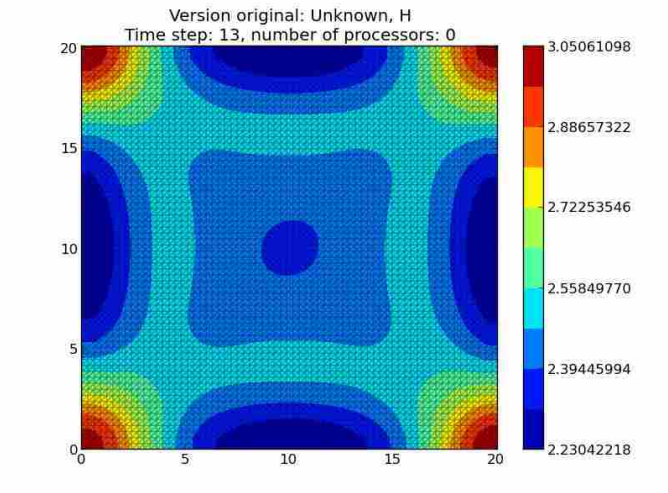
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

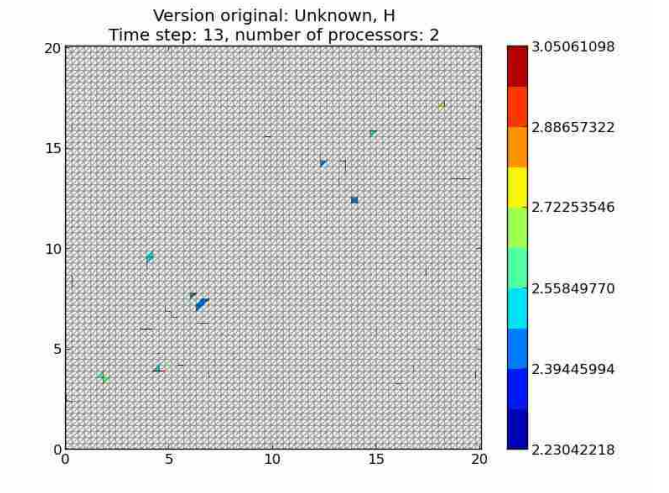
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 13



Sequential



Parallel $p = 2$

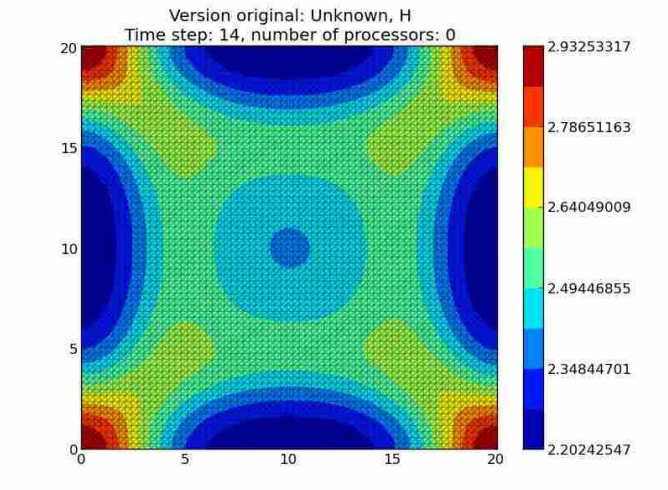
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

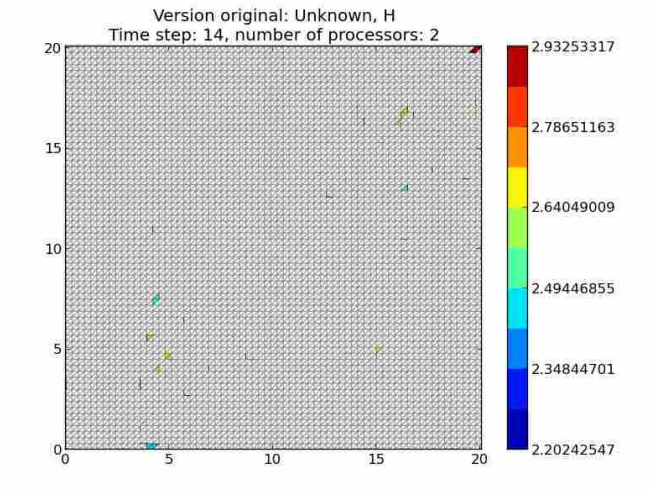
A white plot displays a non-reproducible value

Numerical reproducibility?

time step = 14



Sequential



Parallel $p = 2$

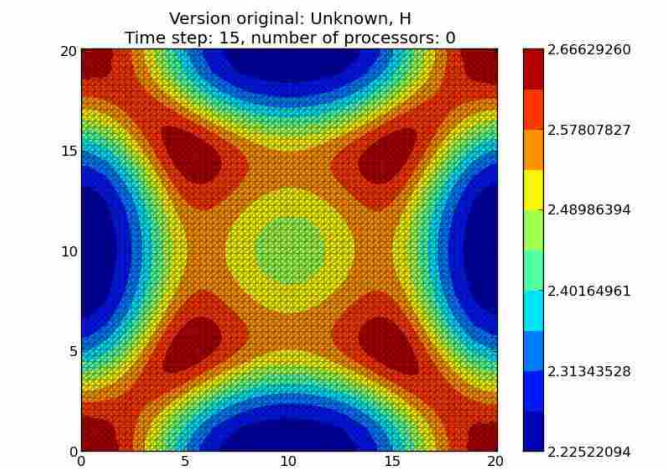
Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

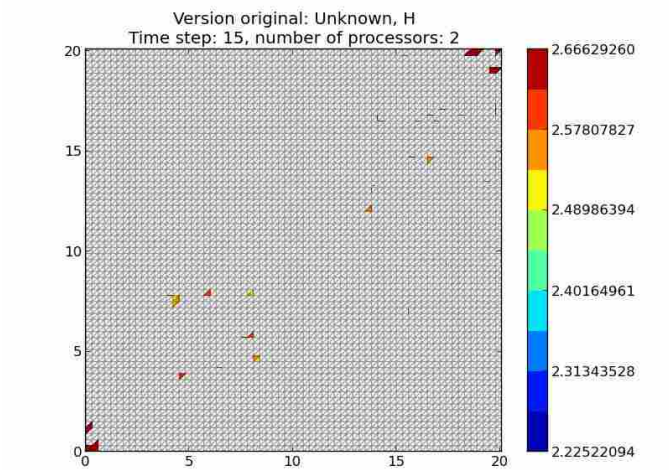
A white plot displays a non-reproducible value

NO numerical reproducibility!

time step = 15



Sequential



Parallel $p = 2$

Courtesy of P. Langlois and R. Nheili

Did I mention we have parallel machines nowadays?

round(round$(a+b)+c) \neq$
round$(a+$round$(b+c))$.
These numerical issues can become quite harmful in real use cases.

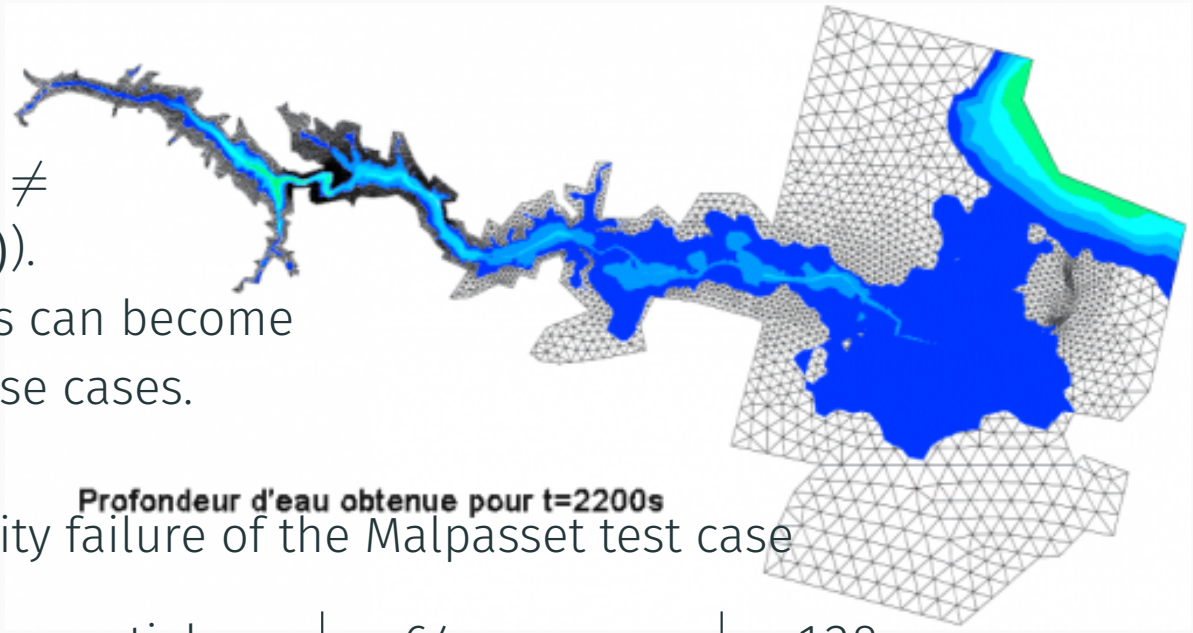


TABLE 1.1: Reproducibility failure of the Malpasset test case

	The sequential run	a 64 procs run	a 128 procs run
depth H	0.3500122E-01	0.2748817E-01	0.1327634E-01
velocity U	0.4029747E-02	0.4935279E-02	0.4512116E-02
velocity V	0.7570773E-02	0.3422730E-02	0.7545233E-02

Numerical reproducibility?: Approximations in the model, in the algorithm, in its implementation, in its execution.

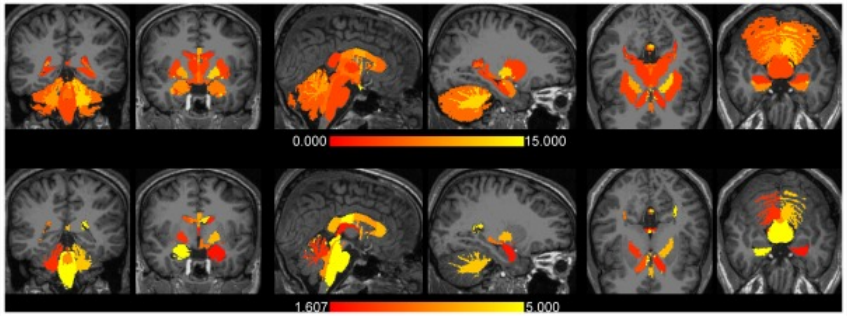
The whole chain needs to be revisited.

Courtesy of P. Langlois and R. Nheili

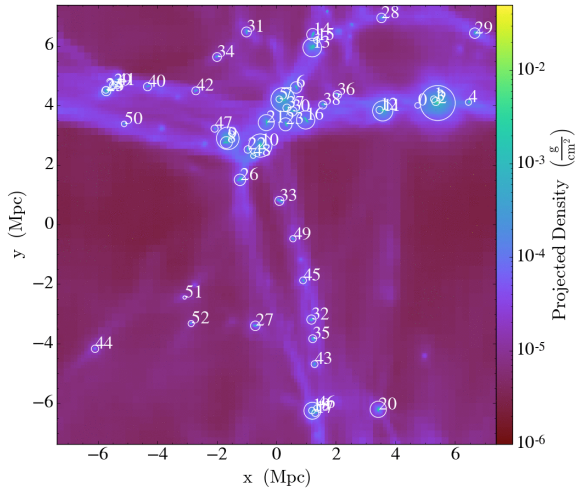
Complexity ⇝ reproducibility issues

The slightest difference **may** have significant consequences

FreeSurfer (PLOS ONE, 2012)



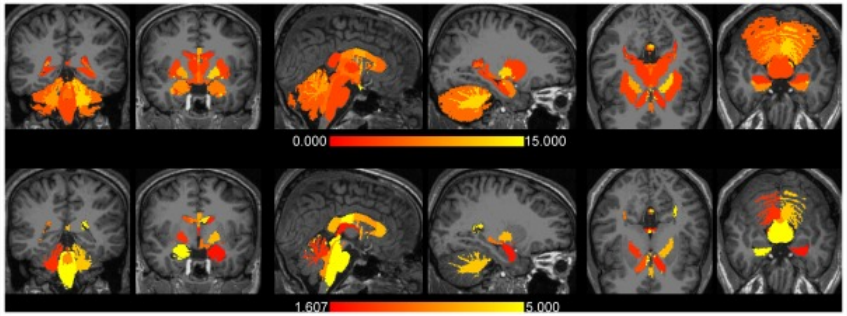
Astrophysics (ResCuE-HPC, 2018)



Complexity ⇝ reproducibility issues

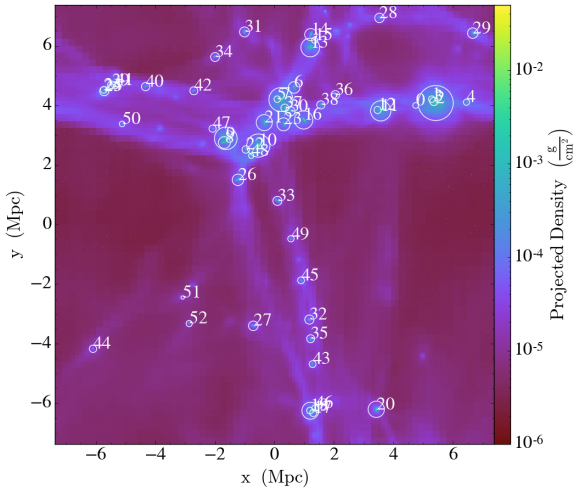
The slightest difference **may** have significant consequences

FreeSurfer (PLOS ONE, 2012)



Software environment problem
(permeability + versions differ)

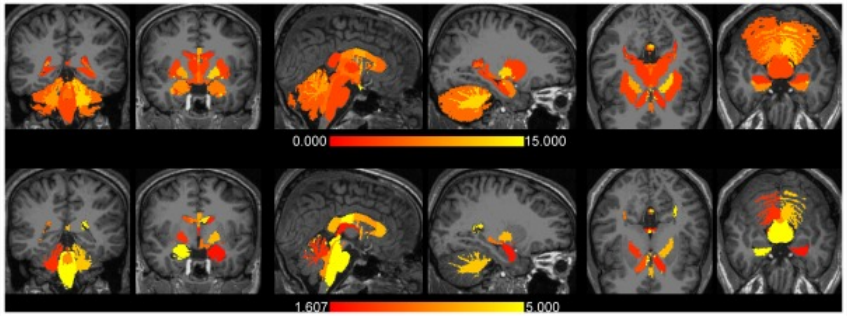
Astrophysics (ResCuE-HPC, 2018)



Complexity ⇝ reproducibility issues

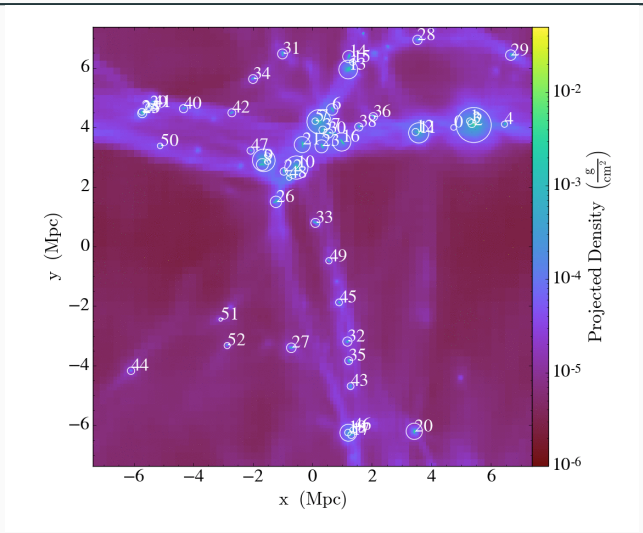
The slightest difference **may** have significant consequences

FreeSurfer (PLOS ONE, 2012)



Software environment problem
(permeability + versions differ)

Astrophysics (ResCuE-HPC, 2018)

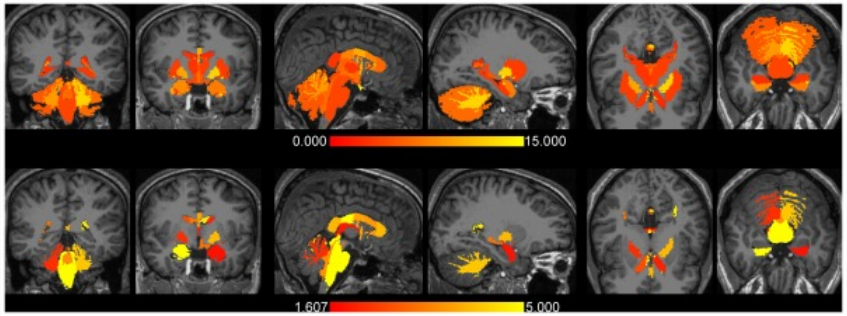


Numerical chaos problem
Compiler optimization influence

Complexity ⇝ reproducibility issues

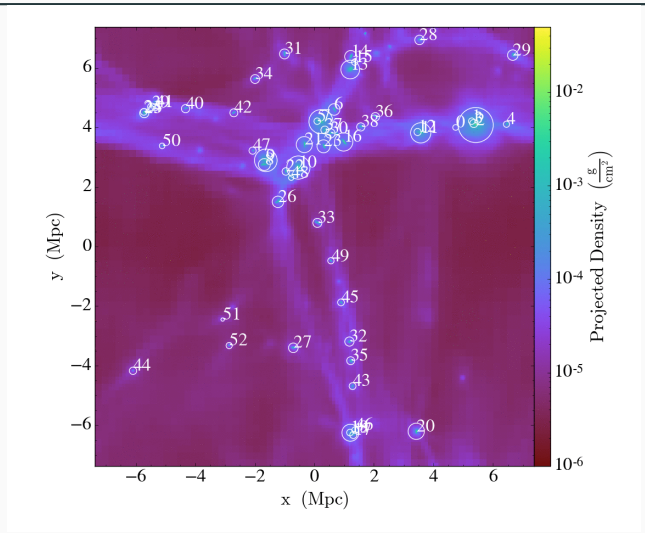
The slightest difference **may** have significant consequences

FreeSurfer (PLOS ONE, 2012)



Software environment problem
(permeability + versions differ)

Astrophysics (ResCuE-HPC, 2018)



Numerical chaos problem
Compiler optimization influence

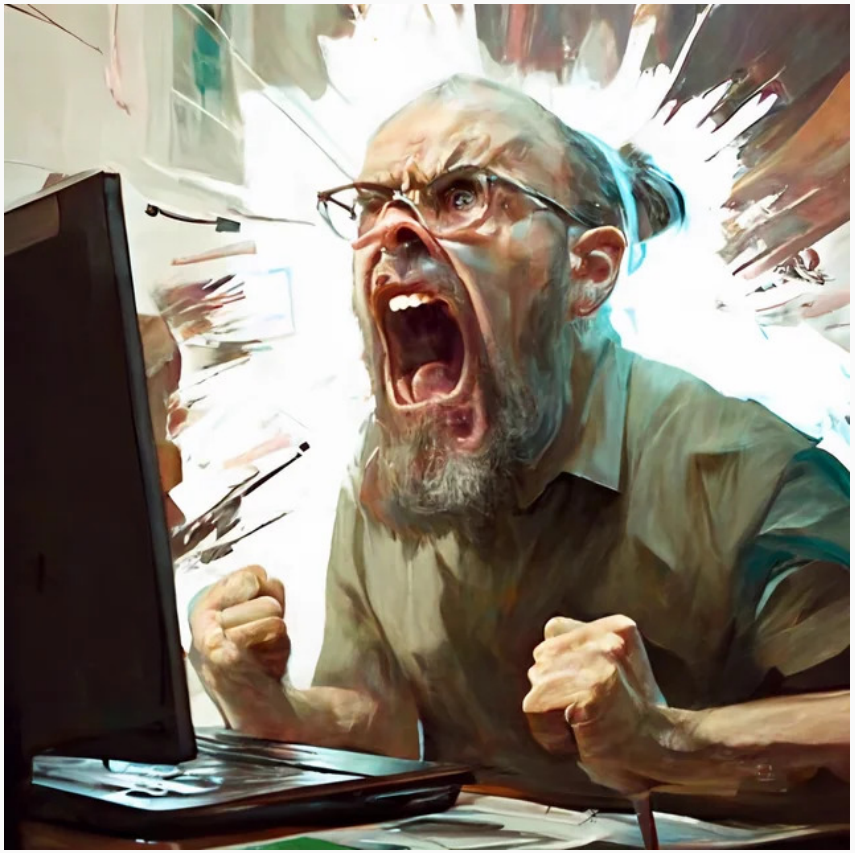
1. Controling your environment (isolation with containers)
2. Controling your dependencies (versions, reproducible recipe)

It is not a good sign if your code is sensitive. You'll need:

1. Perfect control (debug)
2. Variation generation (test)

Additional HPC issues/complaints

- Software is complex and difficult to maintain
- Users (scientists) want
  - rapid development
  - rapidly evolving environment
  - old/legacy code
  - efficient code
  - process more data in a shorter time
  - use AI/neural networks 😉
  - and reproducibility!!!
- Hardware keeps evolving



I know it's High Performance Computing, but maybe we should slow down

Changing Publishing Practices

Artifact evaluation and ACM badges

Major conferences

- ACM SIGMOD 2015-2019, Most Reproducible Paper Award...
- NeurIPS, ICLR: open reviews, reproducibility challenge

Joelle Pineau @ NeurIPS'18

- Supercomputing: Artifact Description (AD) mandatory, Artifact Evaluation (AE) optional, Double blind vs. RR

Mentalitie are evolving people care, make stuff available, errors are found and fixed

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
  - PC members nominate senior PhD students/engineers for the AE committee
  - The committee contacts the authors of accepted papers, who can submit experiments for review, and may get a Reproducible / Sharable label

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
- Realis @ ComPAS'2013 (O. Richard)
  - Additional article/submission. Authors review the work of others

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
- Realis @ ComPAS'2013 (O. Richard)
- ACM TRUST'14 @ PLDI (G. Fursin, J. Vitek)
  - AE remains optional and for accepted articles only
  - B&w *Artifact Evaluated* badge
  - Similar attempts in most major conferences (PPoPP, CGO, PACT,...)

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
- Realis @ ComPAS'2013 (O. Richard)
- ACM TRUST'14 @ PLDI (G. Fursin, J. Vitek)
- 2016: ACM organized a special taskforce (former AE chairs) to develop common methodology for artifact sharing and evaluation across all SIGS!
  1. Define terminology (Repeat, Replicate, Reproduce, Reusable)
  2. Prepare new sets of 5 badges (covering various SIGs)

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
- Realis @ ComPAS'2013 (O. Richard)
- ACM TRUST'14 @ PLDI (G. Fursin, J. Vitek)
- 2016: ACM organized a special taskforce (former AE chairs) to develop common methodology for artifact sharing and evaluation across all SIGS!
- 2020: New version of ACM badges
  - *Interchange the definitions of "Results Replicated" and "Results Reproduced" to adopt the NISO standard* 😉

Artifact Evaluation: an evolving practice

Goals

- validate experimental results from published articles
- restore trust
- promote artifact sharing (benchmarks, data sets, tools, models)
- enable fair comparison of results and techniques
- build upon others' research

History

- SIGMOD 2012: Experimental Reproducibility (P. Bonnet, J. Freire)
- Realis @ ComPAS'2013 (O. Richard)
- ACM TRUST'14 @ PLDI (G. Fursin, J. Vitek)
- 2016: ACM organized a special taskforce (former AE chairs) to develop common methodology for artifact sharing and evaluation across all SIGS!
- 2020: New version of ACM badges
- 2021: ACM EIG on Reproducibility and Replicability

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

Appendix A
Artifact Description Appendix: [Paper Title]

A. Abstract

If a paper has no computational results and submits this appendix, the authors only need to complete this abstract subsection and mention that the paper has no computational results. This text is sufficient: ***"This paper contains no computational results."*** *Other subsections can be removed.*

B. Description

1) Check-list (artifact meta information): Fill in whatever is applicable with some informal keywords and remove the rest

- **Algorithm:**
- **Program:**
- **Compilation:**
- **Transformations:**
- **Binary:**
- **Data set:**
- **Run-time environment:**
- **Hardware:**
- **Run-time state:**
- **Execution:**
- **Output:**
- **Experiment workflow:**
- **Experiment customization:**
- **Publicly available?:**

2) How software can be obtained (if available): Obligatory if the paper contains computational results.

3) Hardware dependencies:

4) Software dependencies:

5) Datasets:

C. Installation

Obligatory if the paper contains computational results.

D. Experiment workflow

Obligatory if the paper contains computational results.

E. Evaluation and expected result

Obligatory if the paper contains computational results.

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

- A submission cannot be disqualified based on information provided or not provided in this appendix, nor if no appendix is provided.
- The inclusion and quality of an appendix may be considered in evaluating a submission, particularly in ranking two submissions of similar quality.
- 9 papers with Artifact Descriptions in the proceedings

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

2017 AD is required to be considered for best paper

- Article reviewers only check whether there is an AD or not
- Boutique environments prevent replication
- ⇝AE appendix (<4 pages) with information on verification and validation of experiments and provide extra assurance that the results are correct

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

2017 AD is required to be considered for best paper

2019 AD becomes mandatory

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts
2016 AD appendix (<2 pages) is possible
2017 AD is required to be considered for best paper
2019 AD becomes mandatory
2021 Chairs recommend providing containers/spack/...

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

2017 AD is required to be considered for best paper

2019 AD becomes mandatory

2021 Chairs recommend providing containers/spack/...

2022 3 kind of badges (artifact available/compilable, reproducible result) are delivered

- Double-blind is a huge pain (data, code, machine access)
- Evaluators may rely on ChameleonCloud
- Reproduction = announced factor 10? similar ranking of algorithms?
  - Some results require M CPU.hours and TB of data
  - Verification = can the code be rerun, not obtain the same result
- 40% of applying articles got the badges!
- Crazy evaluation workload!

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

2017 AD is required to be considered for best paper

2019 AD becomes mandatory

2021 Chairs recommend providing containers/spack/...

2022 3 kind of badges (artifact available/compilable, reproducible result) are delivered

2024 A badge requires the reproduction of the *main result*

- Which result? As decided by the 5 reviewers, not by the AD
- ~~Double-blind~~ (authors may provide access to their machine, discussions are allowed, no more AE chair bottleneck)
  - As G. Fursin experienced and recommended in 2016
- Non-anonymous reproducibility report explaining what has been achieved
- Publication of (positive) reproducibility reports as an appendix to the original article
- Work budget: no more than 8 hours!
  - ×2 more evaluators, >2 evaluators/article, ≈ 2 articles/evaluator

Evolution at SC (AFAIK/recall/understood)

2015 SC reproducibility initiative starts

2016 AD appendix (<2 pages) is possible

2017 AD is required to be considered for best paper

2019 AD becomes mandatory

2021 Chairs recommend providing containers/spack/...

2022 3 kind of badges (artifact available/compilable, reproducible result) are delivered

2024 A badge requires the reproduction of the *main result*

Conclusion

- *Reproduced result* much harder to get
- More information available and evaluator recognition
- Still on voluntary basis

Thanks!

Thanks a lot to:

- Grigori Fursin
- Michael A. Heroux
- Guillaume Pallez