HPC and AI benchmark *workflow and objectives*

Benchmarking from vendor point of view

Ludovic ENAULT
Head of Applications and Performance
Eviden BDS HPC, AI and Quantum business support
ludovic.enault@eviden.com
2023-09-26

an atos business

Abstract

- In this talk, HPC and AI benchmark activity will be presented from a vendor context and point of view:
  - What are the different steps in benchmark activities ?

- An opening will be made on the main question: to what end ?
  - What is the objective ?

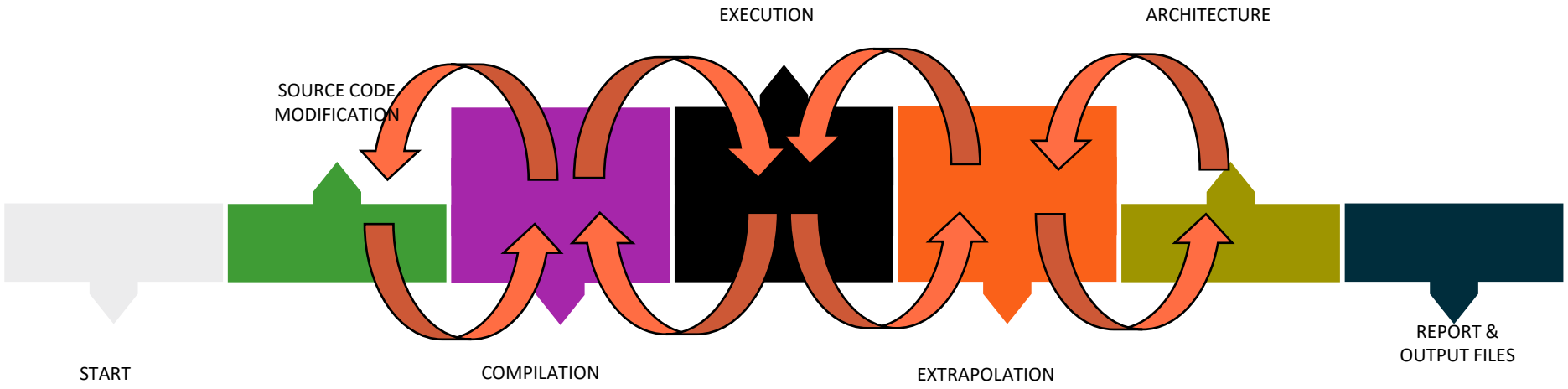
EVIDEN

Content overview

EVIDEN

01 HPC and AI benchmark Worklow

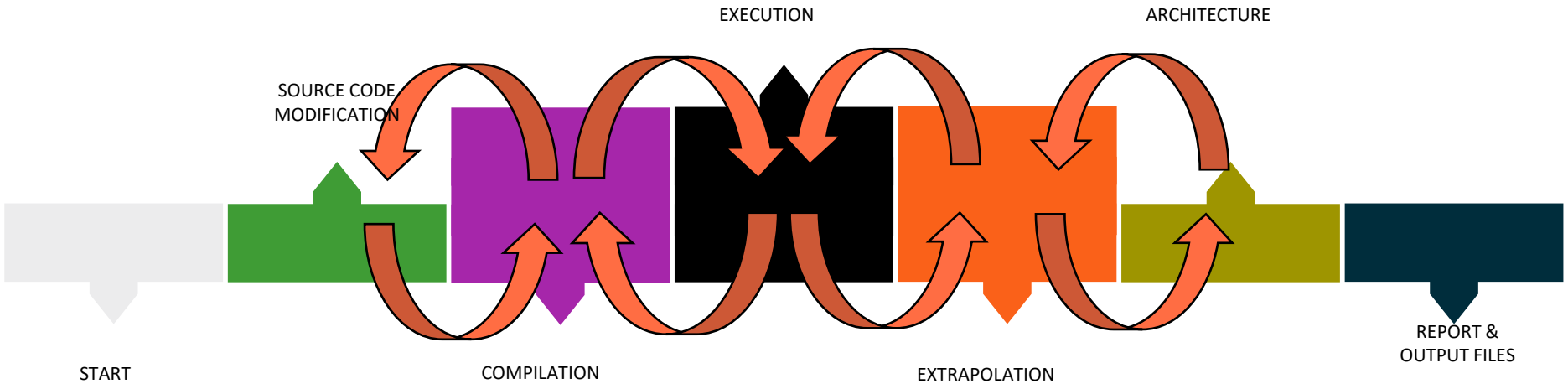
HPC and AI Benchmark workflow



- Overall benchmark activity is typically 6 to 12 weeks
- Many different steps:
  - Run “as is” to get a baseline
  - Basic profiles
  - Targeted profiles

HPC and AI Benchmark workflow

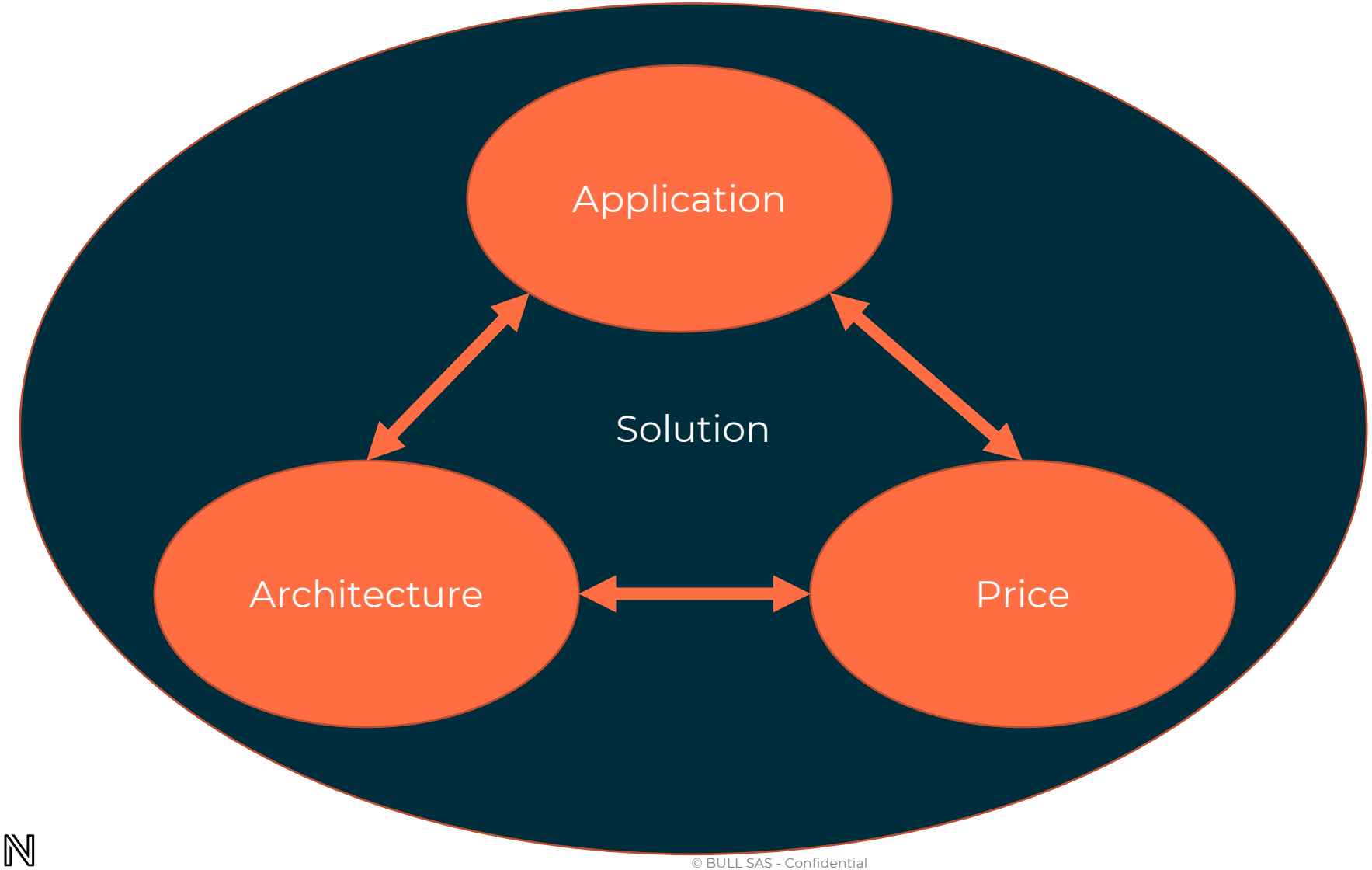
Benchmark activity requires a lot of preparation!



- System (build large “N” cluster, “N+1” sample nodes)
- Technology (Work with providers AMD, Graphcore, Intel, NVIDIA, …)
- Methodology (MPI, directives, container …)

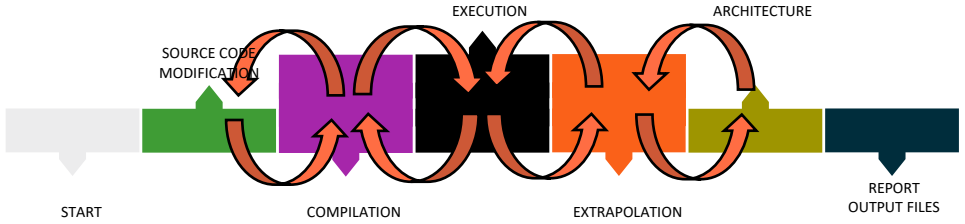
HPC and AI Benchmark workflow

Benchmark is only part of a bigger activity



Benchmark activities

Best Practicies



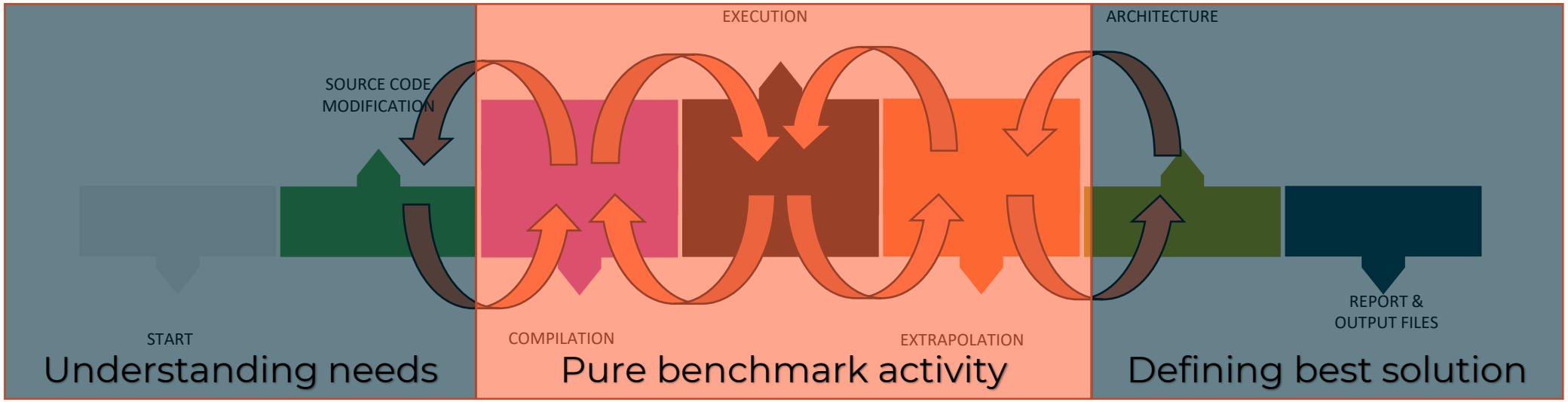
Best Practices

- Ready-to-use applications:
  - Source code available and/or Licenses pre-negotiated with ISV
- Reasonable in terms of system size / runtime / memory (while representative of an actual workload):
  - Runtime: ~10-60 minutes (at a representative scale)
  - Scale for CPU: 1000s of cores, 64-(low) 100s of nodes
  - Scale for GPU: 16-64s of GPUs, 10s of nodes
- Reduced test case available (to run on ~single node)
- Small I/O component (Difficult to keep coherency between various benchmark environments)
- Clear criteria (elapsed time, loop time, energy, …)

What is to be avoided

- Artificial long workload
  - Multiple small iterations that won't scale
- Grid specific :
  - Application/test case that fits a very specific decomposition
- "too-far scale":
  - Projections from 100nodes to 1000 or 10k nodes requires extended work on applications
  - It will more likely include heavy source code modifications
  - Estimation work would be very long (1-2 years)
  - Typical work for post-sales activity
- Applications and architecture requirement mis-match:
  - E.g.: non-GPU application and GPU only system

HPC and AI Benchmark workflow



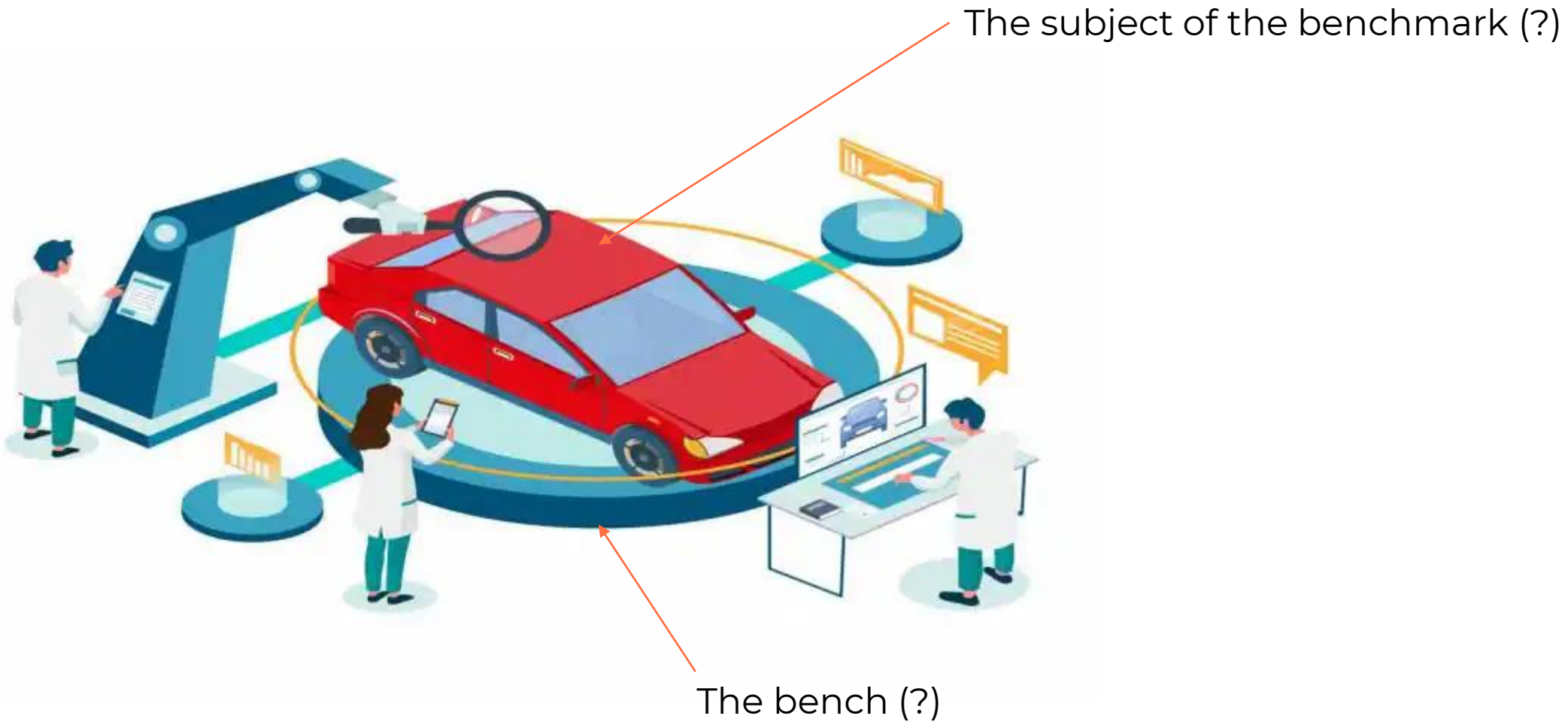
- Benchmark is directed by requirements

EVIDEN

02 HPC and AI benchmark Objective

HPC and AI Benchmark purpose

Give the best number!



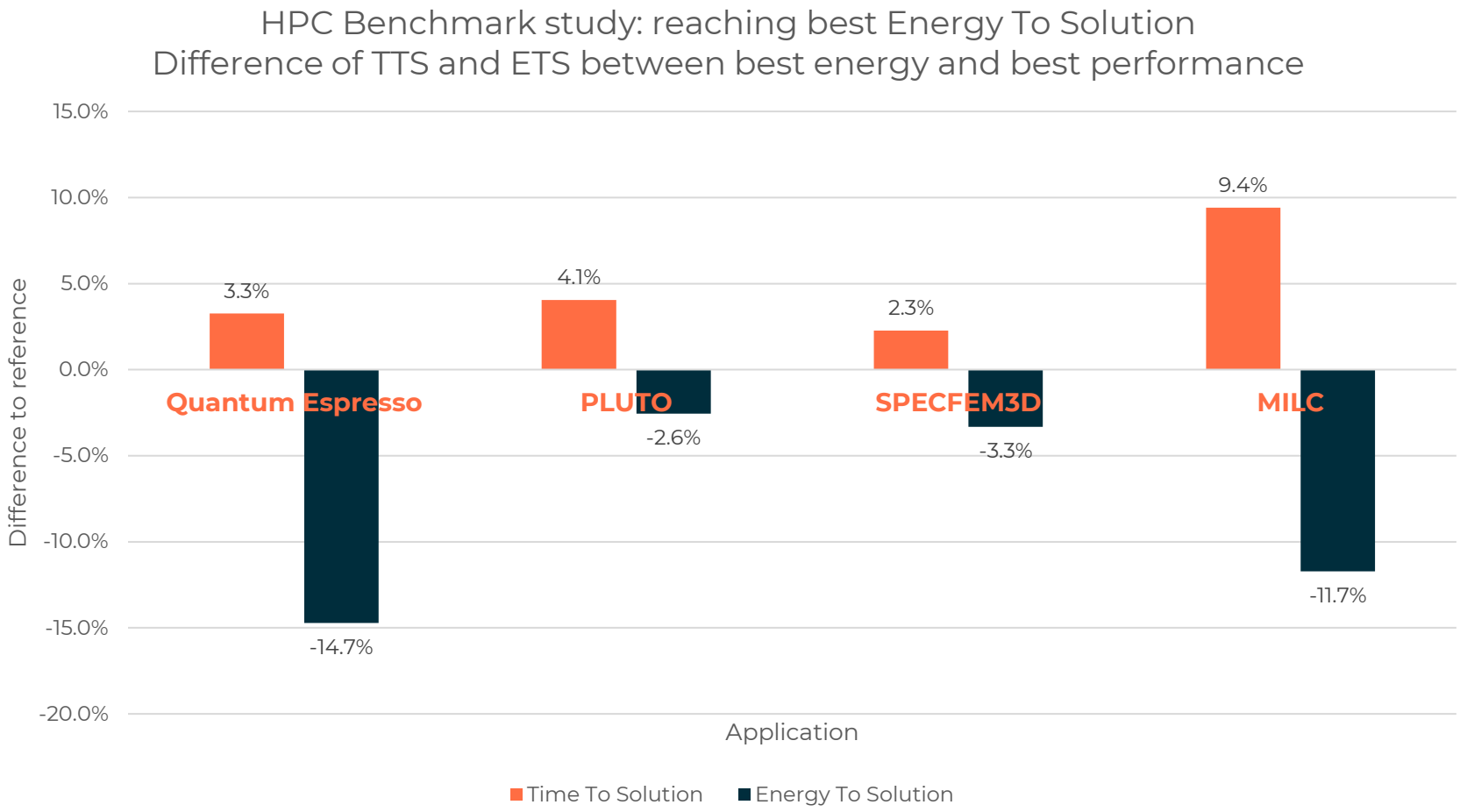
HPC and AI Benchmark purpose

Give the best number!

- What is measured
  - Pure performance
  - Performance/€
  - Performance/watt
  - Performance/m²
  - Energy
  - Price
  - Carbon footprint
  - …

Example of benchmark study

Best Energy To Solution



HPC and AI benchmark

Vendor perspective takeaways

- Main objective is, still, to fit a contract
  - The harder to estimate the performance, the higher vendor risk margin will be
  - Clear **benchmark scoring** ease estimations

- Benchmark should represent **what you are expecting** not what you have

- Vendors are open to discuss **future technologies and timelines**

EVIDEN

Questions

EVIDEN

Thank you!