

Scientific Benchmarking of Parallel Computing Systems

Paper Reading Group

Torsten Hoefler

Roberto Belli

Presents: Maksym Planeta

21.12.2015

Table of Contents

Introduction

State of the practice

The rules

- Use speedup with Care

- Do not cherry-pick

- Summarize data with Care

- Report variability of measurements

- Report distribution of measurements

- Compare data with Care

- Choose percentiles with Care

- Design interpretable measurements

- Use performance modeling

- Graph the results

Conclusion

Table of Contents

Introduction

State of the practice

The rules

- Use speedup with Care

- Do not cherry-pick

- Summarize data with Care

- Report variability of measurements

- Report distribution of measurements

- Compare data with Care

- Choose percentiles with Care

- Design interpretable measurements

- Use performance modeling

- Graph the results

Conclusion

Reproducibility

- ▶ machines are unique
- ▶ machines age quick
- ▶ relevant configuration is volatile

Interpretability

- ▶ Weaker than *reproducibility*
- ▶ Describe an experiment in an understandable way
- ▶ Allow to draw own conclusions and generalize results

Frequently wrong answered questions

- ▶ How many iterations do I have to run per measurement?
- ▶ How many measurements should I run?
- ▶ Once I have all data, how do I summarize it into a single number?
- ▶ How do I measure time in a parallel system?

Performance report

High-Performance Linpack (HPL)

*run on 64 nodes ($N=314k$) of the Piz Daint system
during normal operation achieved 77.38 Tflops/s.*

Performance report

High-Performance Linpack (HPL)

run on 64 nodes ($N=314k$) of the Piz Daint system during normal operation achieved 77.38 Tflops/s.

Theoretical peak is 94.5 Tflops/s ... the benchmark achieves 81.8% of peak performance

Performance report

High-Performance Linpack (HPL)

run on 64 nodes ($N=314k$) of the Piz Daint system during normal operation achieved 77.38 Tflops/s.

Theoretical peak is 94.5 Tflops/s ... the benchmark achieves 81.8% of peak performance

Problems

1. What was the influence of OS noise?
2. How typical this run is?
3. How to compare with other systems?

It's worth a thousand words

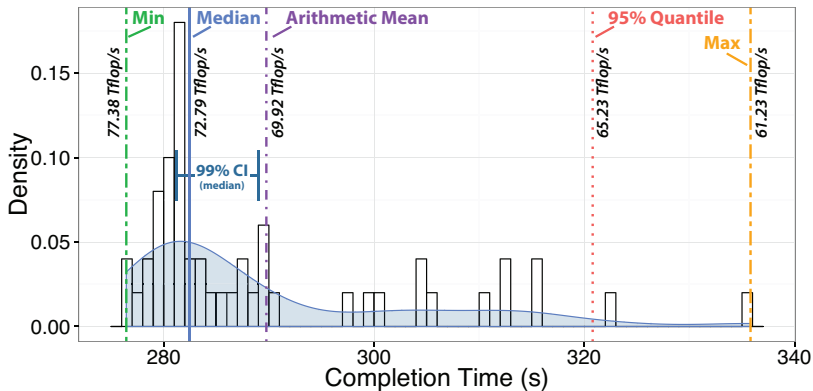


Figure 1: Distribution of completion times for 50 HPL runs.

Table of Contents

Introduction

State of the practice

The rules

- Use speedup with Care

- Do not cherry-pick

- Summarize data with Care

- Report variability of measurements

- Report distribution of measurements

- Compare data with Care

- Choose percentiles with Care

- Design interpretable measurements

- Use performance modeling

- Graph the results

Conclusion

The survey

- ▶ Pick papers from SC, PPOPP, HPDC
- ▶ Evaluate result reports from different aspects
- ▶ Categorize aspects as *covered*, *not applicable*, *missed*

Experiment report

Experimental design

1. Hardware

- 1.1 Processor Model / Accelerator (79/95)
- 1.2 RAM Size / Type / Bus Infos (26/95)
- 1.3 NIC Model / Network Infos (60/95)

2. Software

- 2.1 Compiler Version / Flags (35/95)
- 2.2 Kernel / Libraries Version (20/95)
- 2.3 Filesystem / Storage (12/95)

3. Configuration

- 3.1 Software and Input (48/95)
- 3.2 Measurement Setup (50/95)
- 3.3 Code Available Online (7/95)

Data Analysis

1. Results

Experiment report

Experimental design

1. Hardware
2. Software
3. Configuration

Data Analysis

1. Results
 - 1.1 Mean (51/95)
 - 1.2 Best / Worst Performance (13/95)
 - 1.3 Rank Based Statistics (9/95)
 - 1.4 Measure of Variation (17/95)

Outcome

- ▶ Benchmarking is important
- ▶ Study 120 papers from three conferences (25 were not applicable)
- ▶ Benchmarking usually done wrong
- ▶ Advice researchers how to do better job

If supercomputing benchmarking and performance analysis is to be taken seriously, the community needs to agree on a common set of standards for measuring, reporting, and interpreting performance results.

Table of Contents

Introduction

State of the practice

The rules

- Use speedup with Care

- Do not cherry-pick

- Summarize data with Care

- Report variability of measurements

- Report distribution of measurements

- Compare data with Care

- Choose percentiles with Care

- Design interpretable measurements

- Use performance modeling

- Graph the results

Conclusion

Use speedup with Care

When publishing parallel speedup, report if the base case is a single parallel process or best serial execution, as well as the absolute execution performance of the base case.

because speedup may be ambiguous

- ▶ Is it against best possible serial implementation?
- ▶ Or is it just parallel implementation on single processor?

because speedup may be misleading

- ▶ Higher on slow processors
- ▶ Lower on fast processors

because speedup may be misleading

- ▶ Higher on slow processors
- ▶ Lower on fast processors

Thus,

- ▶ Speedup on one computer can't be compared with speedup on another computer.
- ▶ Better avoid speedup

Do not cherry-pick

Specify the reason for only reporting subsets of standard benchmarks or applications or not using all system resources.

Do not cherry-pick

Specify the reason for only reporting subsets of standard benchmarks or applications or not using all system resources.

- ▶ Use the whole node to utilize all available resources

Do not cherry-pick

Specify the reason for only reporting subsets of standard benchmarks or applications or not using all system resources.

- ▶ Use the whole node to utilize all available resources
- ▶ Use the whole benchmark/application not only kernels

Summarize data with Care

*Use the arithmetic mean only for summarizing costs.
Use the harmonic mean for summarizing rates.*

Avoid summarizing ratios; summarize the costs or rates that the ratios base on instead. Only if these are not available use the geometric mean for summarizing ratios.

Mean

1. if all measurements are weighted equally use the *arithmetic* mean (absolute values):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

2. if the denominator has the primary semantic meaning use *harmonic* mean (rates):

$$\bar{x}^{(h)} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

3. ratios may be summarized by using geometric mean:

$$\bar{x}^{(g)} = \sqrt[n]{\prod_{i=1}^n x_i}$$

do not use geometric mean

*the geometric mean has no simple interpretation and
should thus be used with greatest care*

do not use geometric mean

the geometric mean has no simple interpretation and should thus be used with greatest care

It can be interpreted as a log-normalized average

and tell what you use

51 papers use summarizing. . .

and tell what you use

51 papers use summarizing... four of these specify the exact averaging method...

and tell what you use

51 papers use summarizing. . . four of these specify the exact averaging method. . . one paper correctly specifies the use of the harmonic mean. . .

and tell what you use

51 papers use summarizing. . . four of these specify the exact averaging method. . . one paper correctly specifies the use of the harmonic mean. . . Two papers report that they use geometric mean

and tell what you use

51 papers use summarizing. . . four of these specify the exact averaging method. . . one paper correctly specifies the use of the harmonic mean. . . Two papers report that they use geometric mean, both without a good reason.

Report variability of measurements

*Report if the measurement values are deterministic.
For nondeterministic data, report confidence intervals of
the measurement.*

Dangerous variations

Measurements may be very unpredictable on HPC systems.

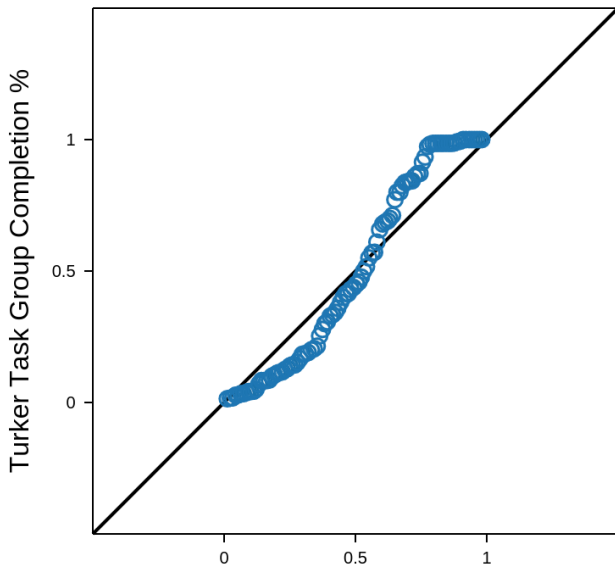
In fact, this problem is so severe that several large procurements specified upper bounds on performance variations as part of the vendor's deliverables.

Report distribution of measurements

Do not assume normality of collected data (e.g., based on the number of samples) without diagnostic checking.

Q-Q plot

Q-Q Plot of Mechanical Turk Participation Rates



Parametric measurements

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Usual central measure	Mean	Any
Data set relationships	Independent	Any ¹
Type of data	Interval or Ratio	Ordinal, Nominal, Interval, Ratio
Conclusion	More powerful	Conservative

¹Paper says opposite

Compare data with Care

Compare nondeterministic data in a statistically sound way, e. g., using non-overlapping confidence intervals or ANOVA.

None of the 95 analyzed papers compared medians in a statistically sound way.

Mean vs. Median

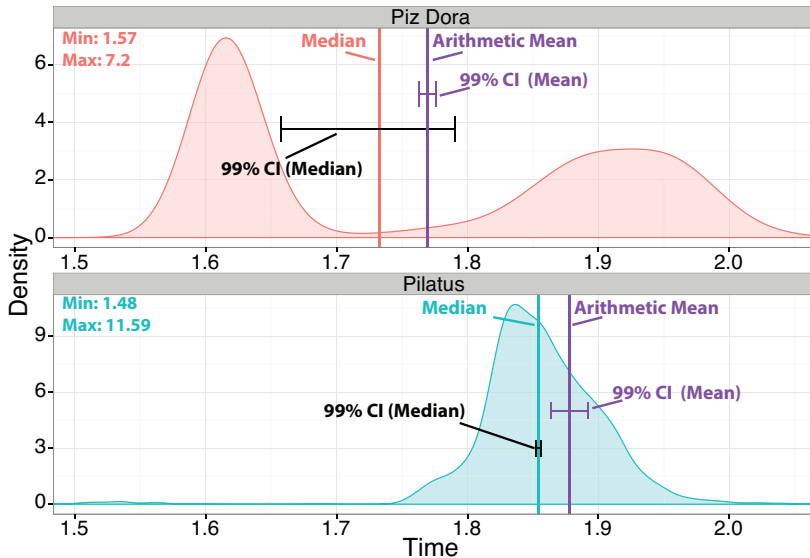


Figure 3: Significance of latency results on two systems.

Choose percentiles with Care

Carefully investigate if measures of central tendency such as mean or median are useful to report. Some problems, such as worst-case latency, may require other percentiles.

Piz Dora vs Pilatus

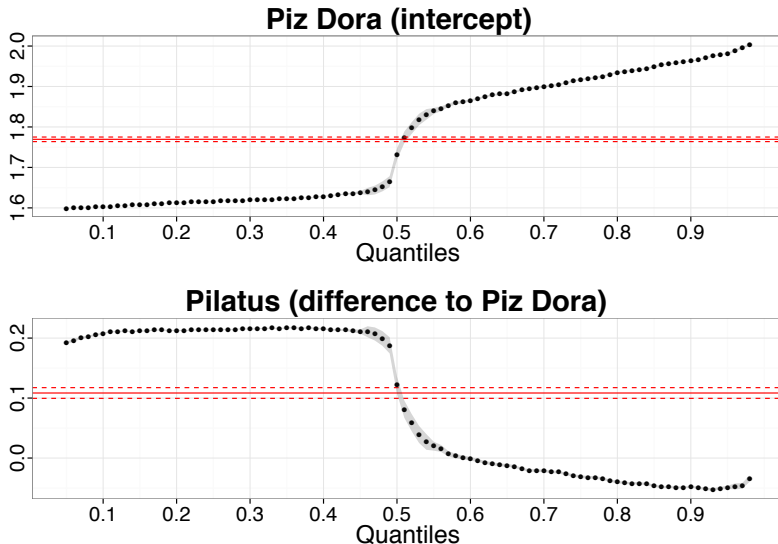


Figure 4: Quantile regression comparison of the latencies comparing Pilatus (base case or intercept) with Piz Dora.

Design interpretable measurements

Document all varying factors and their levels as well as the complete experimental setup (e. g., software, hardware, techniques) to facilitate reproducibility and provide interpretability.

Fix environments

1. Fix environment parameters

If controlling a certain parameter is not possible then we suggest randomization following standard textbook procedures.

2. Document setup

For parallel time measurements, report all measurement, (optional) synchronization, and summarization techniques.

Particular parameters may be very important

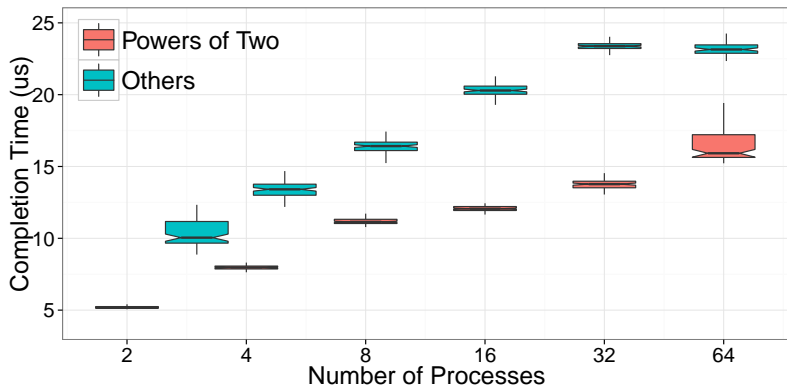
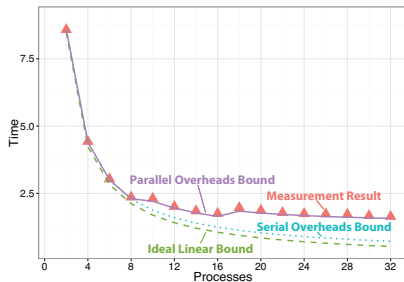


Figure 5: 1,000 MPI_Reduce runs for different process counts.

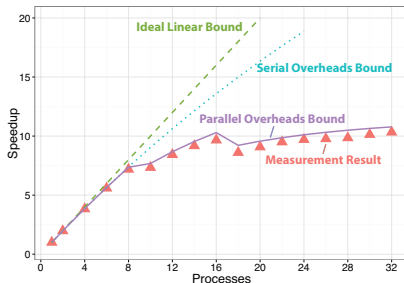
Use performance modeling

If possible, show upper performance bounds to facilitate interpretability of the measured results.

Interpretable speedup graph



(a) Time



(b) Speedup

Parallel overheads bound (based on Amdahl's law)

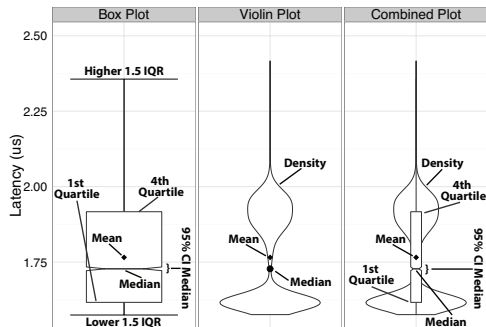
$$t = \begin{cases} 10\text{ns} & , \text{ if } p \leq 8 \\ 0.1\text{ms} \cdot \log_2 p & , \text{ if } 8 < p \leq 16 \\ 0.17\text{ms} \cdot \log_2 p & , \text{ if } 16 < p \end{cases}$$

Graph the results

Plot as much information as needed to interpret the experimental results. Only connect measurements by lines if they indicate trends and the interpolation is valid.

Use appropriate tool

- ▶ Box plots
- ▶ Histograms
- ▶ Violin plots
- ▶ Plot summary statistics
- ▶ Plot CIs
- ▶ Combinations of all



(c) Box and Violin Plots

Table of Contents

Introduction

State of the practice

The rules

- Use speedup with Care

- Do not cherry-pick

- Summarize data with Care

- Report variability of measurements

- Report distribution of measurements

- Compare data with Care

- Choose percentiles with Care

- Design interpretable measurements

- Use performance modeling

- Graph the results

Conclusion

Conclusion

- ▶ Important problem
- ▶ Good introduction
- ▶ Some of the claims have no obvious conclusion