Benchmarking Optimization Software a (Hi)Story

Hans D Mittelmann

School of Mathematical and Statistical Sciences
Arizona State University

EURO 2019
Dublin - Ireland
25 June 2019
UPDATED and EXTENDED
for talk at PolyU, March 2026

Outline

Outline

Our Service and the Rationale for Benchmarking

our "community service, part I"

- about 1996 Decision Tree started (with Peter Spellucci)
- soon after Benchmarks added
- first no commercial software, later selected codes
- extensive, very frequently updated
- lead to more transparency and competition
- both open source and commercial developers use benchmarks for advertising

Our Service and the Rationale for Benchmarking

our "community service, part II"

- after benchmarks, NEOS solvers were added
- NEOS (network-enabled optimization solver) provides large number of interactively usable optimization programs
- about 1/3 run on our computers, NEOS only gateway
- needs to be demonstrated to give impression
- both service components benefit (our) research and teaching
- department reduced computer access late 2025, all solvers moved to NEOS server

Our Service and the Rationale for Benchmarking

The Rationale for Benchmarking

- Optimization is ubiquitous
- Most number-crunching computing is done in optimization
- While mathematically most optimization is not hard, writing efficient and robust programs is
- Users of optimization are well advised to try not one but several programs on their problems
- Even some powerful commercial software is available for use: NEOS (everyone), source/binaries (certain groups)

Outline

What will be shown next

- Initially we had chosen all benchmark problems ourselves
- Later various libraries were created:
  MIPLIB2010/17, CBLIB14, QPLIB17
- To allow tracking of development over time we archived our benchmark talks starting in 2002. From them the history will be documented
- In view of the very latest developments mostly MILP results are presented, in particular for the "BIG THREE"
  CPLEX, Gurobi, XPRESS
- Note that historic MILP speedup is 10^{12} (one trillion)

What happened in the early history?

- Multicore computing becomes the standard
- After publishing CPLEX vs. XPRESS in a benchmark in 2007, XPRESS(Dash) asks not to be included
- In late 2008 at INFORMS Washington/DC Bixby/Gurobi presents first results after 18 months, during 9 of which code development by Gu and Rothberg
- Later Gurobi makes code available to academics; this forces CPLEX to make it available as well; we include Gurobi starting 2010
- FICO buys XPRESS. In 2010 they want to be included again

What is the shifted geometric mean?

- There are huge problems in using the performance profiles for several codes in one graph
- One would need to do $N - 1$ graphs for N codes
- Commercial code developers use the shifted geometric mean
- If c_i is the compute time for instance i then one computes
- $$\left(\prod_{i=1}^{N} [c_i + \text{shift}]\right)^{\frac{1}{N}} - \text{shift}$$
- For the shift typically 10 [secds] is used to avoid skewing from relatively very small c_i
- This provides a balanced averaging

Outline

What happened in the intermediate history?

- MIPLIB2010 was released
  - 361 instances, benchmark set 87, still unsolved 70
- We introduce the shifted geometric mean
- Gurobi surpasses CPLEX, XPRESS falls behind
- Standard benchmark set becomes too easy
- A new benchmark in 2013: SOCP and MISOCP (not shown, from CBLIB)
- A new code appears out of nowhere: MIPCL

Outline

In how many benchmarks are the BIG THREE?

- Pre INFORMS 2018
  - CPLEX is in 15 of 22 of our benchmarks
  - Gurobi and XPRESS are in 13 of our benchmarks (not TSP, not QCQP)
- Post INFORMS 2018
  - CPLEX, Gurobi, XPRESS are in NONE of our benchmarks
- What happened?
- This is finally the Story
  - Gurobi advertised aggressively
  - CPLEX (IBM) and XPRESS (FICO) reacted

This is what happened at INFORMS2018

The Story part I

- Over many years Gurobi had used our benchmark results for advertising making bargraphs from the tables
- At INFORMS 2018 the library MIPLIB2017 was released. We had just used it in our benchmark. It has 240 instances and only the full set is a benchmark set
- Instance selection of MIPLIB2017 uses a sophisticated computer program
- Gurobi was represented on the MIPLIB2017 committee
- At INFORMS2018 Gurobi claimed that we had used certain 99 MIPLIB2017 instances in our benchmark showing they are 2.69 times faster than CPLEX and 5.51 times faster than XPRESS

This is what happened at INFORMS2018

The Story part II

- On the last day of the conference in our session Gurobi apologized to IBM, FICO, ourselves and the community
- Tobias Achterberg and Zonghao Gu draft a paper analyzing what had happened
- After INFORMS2018 both IBM and FICO request from me to remove their numbers from all benchmarks
- We decide to also omit the Gurobi numbers
- See the following slides documenting these developments

Gurobi Optimizer 8.1: The Fastest Solver in the World

2.69X

Faster than CPLEX

5.51X

Faster than Xpress

"Benchmarks on the 99 models in the new 2017 MIPLIB demonstrate the purest objective comparison of speed."

Independent performance tests performed by Professor Hans Mittelmann using all new models from the recently released MIPLIB 2017 benchmark set show that Gurobi Optimizer 8.1.0 is 2.69X faster than IBM® CPLEX 12.8.0 and 5.51X faster than FICO® Xpress 8.5.1.

- ✓ The new 2017 MIPLIB is a standard test set used to compare the performance of Mixed-Integer Programming (MIP) solvers.
- ✓ These results look at performance on all 99 new models in the set.
- ✓ Considering only the newest models in the set gives the fairest, most objective speed comparison, since none of the vendors have had a chance to tune to these models.
- ✓ These numbers show geometric mean runtime ratios, calculated using the standard PAR-10 performance testing methodology.
- ✓ These results confirm Gurobi Optimizer's position as the world's fastest math programming solver.

Outline

What did we learn?

- Optimization Software is a cutthroat business
- IBM claims that Gurobi had their license for years while refusing to grant them a license for Gurobi
- Gurobi has similar accusations against the others
- Sometimes even very smart people overstep the mark
- Now users have to benchmark themselves again
- Our benchmarks are less exciting but to make up a bit for the loss we list ballpark geomeans for best commercial codes

From INFORMS 2023

COMBINATORIAL OPTIMIZATION

- *Concorde-TSP with different LP solvers (3-3-2023)*

LINEAR PROGRAMMING

The Simplex and Barrier benchmarks are replaced by benchmarks that show how well solvers find a primal-dual feasible point (as barrier methods in general do) or an optimal basic solution (as simplex methods in general do). Codes do not have to use a textbook version of either method.

- *LPfeas Benchmark (find a PD feasible point) (10-4-2023)*
- *LPopt Benchmark (find optimal basic solution) (10-5-2023)*
- *Large Network-LP Benchmark (commercial vs free) (10-4-2023)*

MIXED INTEGER LINEAR PROGRAMMING

- *MILP Benchmark - MIPLIB2017 (10-4-2023)*
- *MILP cases that are slightly pathological (10-5-2023)*
- *Infeasibility Detection for MILP Problems (10-4-2023)*

SEMIDEFINITE/SQL PROGRAMMING

- *SQL problems from the 7th DIMACS Challenge (8-8-2002)*
- *Several SDP codes on sparse and other SDP problems (9-28-2023)*
- *Infeasible SDP Benchmark (8-24-2023)*
- *Large SOCP Benchmark (9-27-2023)*

- *MISOCP Benchmark (9-27-2023)*

NONLINEAR PROGRAMMING

- *AMPL-NLP Benchmark (1-18-2022)*

MIXED INTEGER QPS AND QCPS

- *Non-commercial convex QP Benchmark (9-16-2021)*
- *Binary Non-Convex QPLIB Benchmark (7-12-2023)*
- *Non-Convex QUBO-QPLIB Benchmark (9-20-2023)*
- *Discrete Non-Convex QPLIB Benchmark (non-binary) (7-14-2023)*
- *Continuous Non-Convex QPLIB Benchmark (7-18-2023)*
- *Convex Continuous QPLIB Benchmark (9-28-2023)*
- *Convex Discrete QPLIB Benchmark (9-27-2023)*

MIXED INTEGER NONLINEAR PROGRAMMING

- *MINLP Benchmark (7-10-2023)*

PROBLEMS WITH EQUILIBRIUM CONSTRAINTS

- *MPEC Benchmark (4-12-2022)*

LPfeas Benchmark (find PD feasible point)

Top of benchmark table shown

- Total of 65 instances (16 hidden)
- Sizes up to 30m/30m/35m rows/cols/nonzeros
- Intel i7-11700K, 3.6GHz, 64GB, 15,000 secs wall clock
- Own instance selection

	Gurob	COPT	MDOPT	MOSEK	HiGHS	NITRO	PDLP
65 probs	1.11	1	1.80	2.60	19.2	22.9	17.0
solved	65	65	62	64	52	43	49

LPopt Benchmark (find optimal basic solution)

- Same own instance selection as in LPfeas benchmark

  COPT-7.0.0, MindOpt-1.0.0, HiGHS-1.6.0, Gurobi-10.0, OptVerse-0.2.13

  MOSEK-10.1.9, SOPLEX-6.0.0

- Intel i7-11700K, 3.6GHz, 64GB, 15,000 secs wall clock

```
65 probs 26.1  1.36   1  1.89  3.95  5.61  17.4 87.2
solved     40    65  65    63    57    52    51   32
=============================================================
probs      CLP Gurob COPT MDOPT OPTV MOSEK HiGHS SPLX
=============================================================
```

The MIPLIB2017 Benchmark Instances

- Total of 240 instances
- Sizes up to 1.5m/1m/43m rows/cols/nonzeros
- Intel i7-11700K, 3.6GHz, 64GB, 7,200 secs wall clock

	CBC	Gurob	COPT	SCIP	SCIPC	HiGHS	SMOO	XSMOO
unscal	1328	81.5	126	888	727	720	612	510
scaled	16.3	1	1.54	10.9	8.92	8.83	7.51	6.26
solved	107	227	212	137	152	159	163	172

(X)SMOO: (X)Smoothie (FiberSCIP+HiGHS, Cplex/Soplex)

Outline

The Benchmarks at a Turning Point

We are in the age of AI

- Potential use of over-tuning/machine learning
- benchmarks prone to AI except those with unknown datasets (LP, MIPLIB)
- Optverse (the AI code) is unavailable and cannot be compared to other codes; its numbers will be removed after the conference
- MIPLIB2024 to be published soon; will it be AI-proof?
- utilization of GPUs (accelerated computing), so far in LP, QP, SDP
- more solvers can handle nonconvexity (globally)

LPfeas Benchmark (find a primal-dual feasible point)

	Intel CPU		NVIDIA H100 GPU				
65 pr	COPT	MOSEK	HiGHS	KNTRO	PDLP	XOPT	OPTV
unscd	25.3	83.3	556	431	421	146	44.1
scald	1.19	3.93	26.2	20.3	20.0	6.87	2.08
solvd	65	59	49	49	50	59	64

65 pr	CUPDL	CUOPT	CUPDLX	COPTG	HPRLP
unscd	47.1	22.4	41.1	21.2	64.3
scald	2.22	1.06	1.94	1	3.03
solvd	56	62	61	64	59

Performance on known and unknown datasets

- This comparison attempts to sense any tuning effort.

```
LPFEAS-Benchmark on 49 public or all 65 instances
=================================================
   probs   COPT MOSEK HiGHS XOPT OPTV
=================================================
PUBLIC     1.25 3.63  25.7  4.87 1.74
 ALL       1.19 3.93  26.2  6.87 2.08
=================================================
```

- LPOPT-Benchmark on 49 public or all 65 instances

```
=============================================================
probs  CLP  COPT OPTV MOSEK HiGHS SPLX XOPT
=============================================================
PUBLIC 27.3  1  1.45  5.95  17.0   116 4.15
ALL    27.4  1  1.76  7.49  17.1  94.1 6.63
=============================================================
```

The MIPLIB2017 Benchmark at INFORMS2025

- Total of 240 instances
- Sizes up to 1.5m/1m/43m rows/cols/nonzeros
- AMD Ryzen 9 5900X (12 cores, 128GB), 7,200 secs wall clock
- modification through random row and column reordering
- NOTE! OPTV received on 10/15, COPT on 9/21

	COPT	SCIP	SCIPC	HiGHS	XOPT	OPTV	XSMOO
unscl	133	1175	935	880	677	[117]	551
scald	1.14	10.1	8.02	7.55	5.81	[1]	4.73
solvd	218	128	145	157	160	[218]	172

MIPLIB redone after INFORMS

- Since the previous reordering of the MIPLIB dataset was done a while ago, the random seed was changed and the benchmark rerun

- COPT-8.0.3, SCIP-10.0, HiGHS-1.13.0, OptVerse-2.0.1

	COPT	SCIP	SCIPC	HiGHS	OPTV
unscal	101	1003	853	743	173
scaled	1	9.93	8.45	7.36	1.72
solved	219	136	150	162	210

Outline

go from H100 to B200 GPU

the LPFEAS instances

- cuOpt-26.02, cuPDLPx-0.2.5, COPT-8.0.3, HPR-LP-C-0.1.1

	CUOPT	CUPDX	COPTG	HPRLP
unscaled	18.0	26.3	15.1	39.3
scaled	1.19	1.74	1	2.60
solved	61	58	64	54

go from H100 to B200 GPU

prob	CUOPT	CUPX	COPTG	HPRLP
heat-source-easy	507	464	m	t
mcf_2500_100_500	f	2681	395	t
mcf_5000_100_400	f	t	748	t
mcf_5000_50_500	f	t	376	t
mediterranean-shipping	f	t	m	t
prod_100_300_02	1160	1124	m	t
production-inventory	343	114	826	940
qap-tho-150	f	m	m	2486
qap-wil-100	1132	496	m	477
supply-chain	f	4611	694	t
synthetic-design-match	118	115	m	189
tsp-gaia-10m	476	456	m	532

How big are the problems?

prob	constraints	variables	nonzeros
heat-source-easy	15625000	31628008	125000000
mcf_2500_100_500	1512600	126250100	253750100
mcf_5000_100_400	2525100	202500100	407500100
mcf_5000_50_500	2775050	126250050	253750050
mediterranean-sh	7490593	208479461	628927462
prod_100_300_02	4650850	18270600	500049700
production-inv	1610550	6110200	66809900
qap-tho-150	6705300	249783750	1005795000
qap-wil-100	1980200	49015000	198020000
supply-chain	2210100	201000100	403000100
synthetic-des	5500135	10000000	690000000
tsp-gaia-10m	1701668	60601996	475701996

Latest benchmarks added

a pure QP benchmark on GPU

```
18 Feb 2026  ================================
             Convex Continuous QPLIB Benchmark
             ================================
```

- The benchmark has QP and QCQP instances
- These QP codes can only handle the subset of QPs

21 probs	cuOpt			HPR-QP		
mean	1	3.48	8.29	3.05	8.89	10.1
solved	19	16	13	19	15	15
prob#	e-4	e-6	e-8	e-4	e-6	e-8

on NVIDIA B200

Latest benchmarks added

a MIP feasibility benchmark on CPU and GPU

- The goal is to determine how fast solvers find good feasible solutions in a given time as measured by the PRIMARY INTEGRAL. These codes were run through GAMS-53.1 on the 233 feasible MIPLIP2017 instances
- CBC-2.10.11, cuOpt-26.02 (CPU/GPU), HiGHS-1.12.0, SCIP(spx)-10.0.1; VMCS: virtual mean commercial solver
- For details see https://www.gams.com/blog/2026/03/expanding-the-focus-introducing-the-mipfeas-benchmark/

solver	cbc	cu-gb10	cu-h100	highs	scip	VMCS
GeoM.	0.1121	0.0887	0.0651	0.0567	0.0755	0.0132
GeoM.Sc.	8.4603	6.6948	4.9135	4.2774	5.6979	1.0000
Feas. found	192	223	225	212	204	227
Proven Opt.	77	60	57	97	77	180

THANK YOU

Questions?

Slides of talk at
https://plato.asu.edu/talks/hongkong2026.pdf