Findings of the WMT 2022 Shared Task on Efficient Translation

Kenneth Heafield Biao Zhang Graeme Nail Jelmer van der Linde
Nikolay Bogoychev

February 8, 2023

University of Edinburgh
{Kenneth.Heafield,b.zhang,graeme.nail,jelmer.vanderlinde,n.bogoych}@ed.ac.uk

Table of contents

1. Task description

2. Participants

3. Results

4. Conclusion

Task description

Fifth edition of the WMT efficiency shared task

Training setup:

- WMT 21 English-German Data condition
- 4x Ensemble of transformer-big teachers
- Distilled training data provided: 314M sentence pairs
- Participants can distill their own data
- Evaluation is performed on the pareto frontier: translation speed / translation quality tradeoff, as measured by wall clock and comet score on WMT22.

Testing data

Corpus	Sentences
WMT 08–19	32,477
WMT 20 under 150 tokens	1,416
WMT 20 sentence split	2,048
WMT 21 sentence split	1,096
WMT 21 inc. additional tests	14,938
WMT 22	2,037
Khresmoi Summary Test v2	1,000
IWSLT 2019	2,278
SimpleGen	2,664
WinoMT	3,888
TED 2020 v1	293,562
Tilde RAPID 2019	663,922
Total	1,021,326
Deduplicated	1,000,000

Table 1: 1M lines, 19,926,744 English words

Submissions and hardware

Participants submit dockerized systems

CPU setting

- Intel Xeon Gold 6354, from Oracle Cloud BM.Optimized3.36
- Throughput setting where all 36 Cores are used, input provided in bulk.
- Latency setting where one CPU core is used, input provided line by line, unbuffered.

Submissions and hardware

Participants submit dockerized systems

CPU setting

- Intel Xeon Gold 6354, from Oracle Cloud BM.Optimized3.36
- Throughput setting where all 36 Cores are used, input provided in bulk.
- Latency setting where one CPU core is used, input provided line by line, unbuffered.

GPU setting

- NVidia A100, from Oracle Cloud BM.GPU4.8
- Throughput and Latency setting

Participants

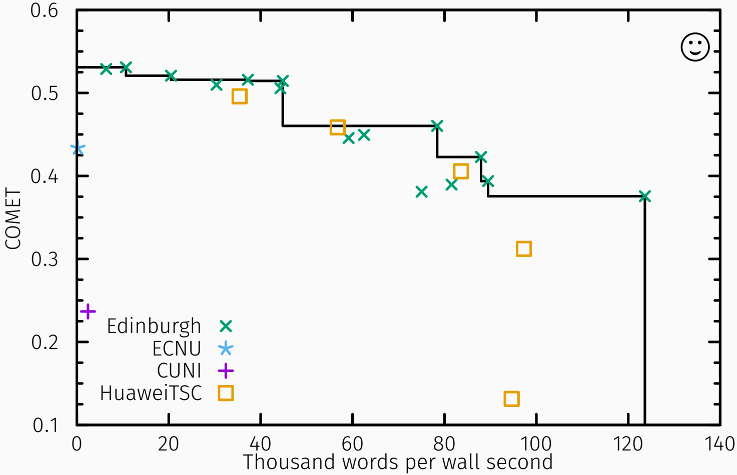
Contributions

	Throughput		Latency	
	CPU-ALL	GPU	CPU-1	GPU
CUNI (non-AG)	1	1	1	1
ECNU	1	1	1	1
Edinburgh	15	11	15	11
HuaweiTSC	5		5	
RoyalFlush (semi non-AG)				6

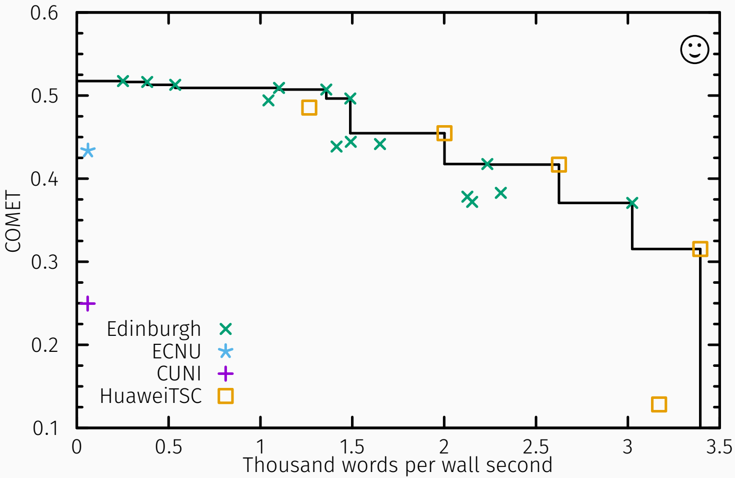
Table 2: 76 systems submitted in total by 5 distinct participants for the different hardware and batching conditions. CPU-ALL refers to the 36-core hardware setting.

Results

CPU Throughput

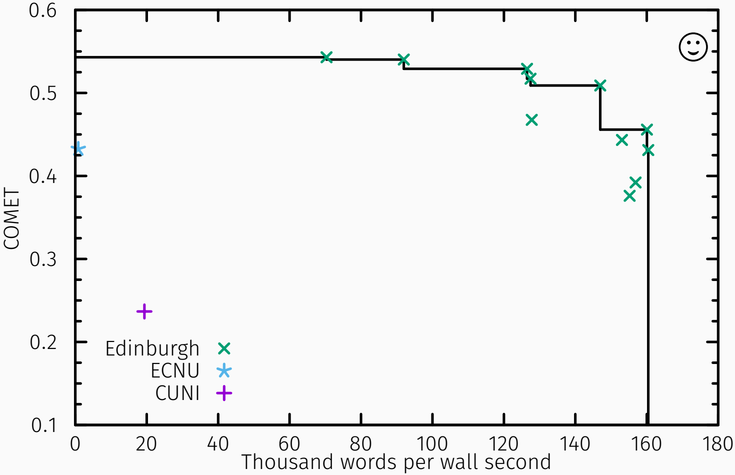


CPU Latency

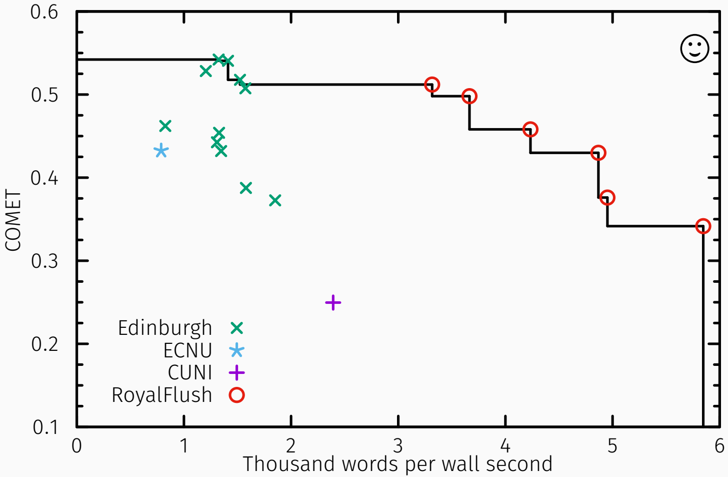


0.6
0.5
0.4
0.3
0.2
0.1
COMET
0
0.5
1
1.5
2
2.5
3
3.5
Thousand words per wall second
☺
Edinburgh ×
ECNU *
CUNI +
HuaweiTSC □

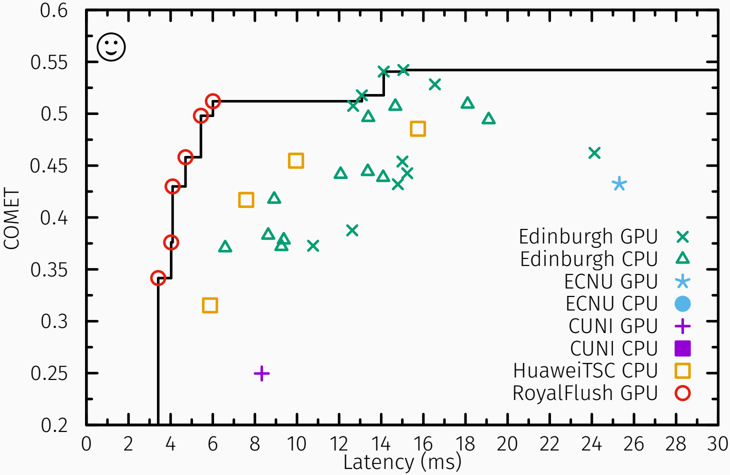
GPU Throughput



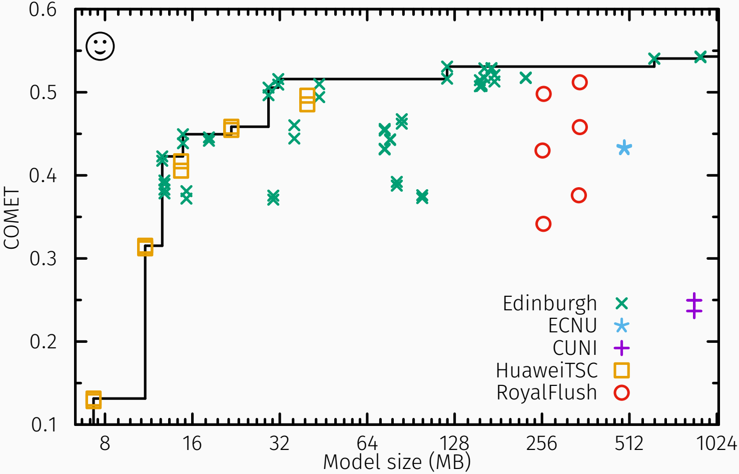
GPU Latency


0.6
0.5
0.4
0.3
0.2
0.1
COMET
0
1
2
3
4
5
6
Thousand words per wall second
☺
Edinburgh ×
ECNU ✱
CUNI +
RoyalFlush ○

GPU CPU Latency

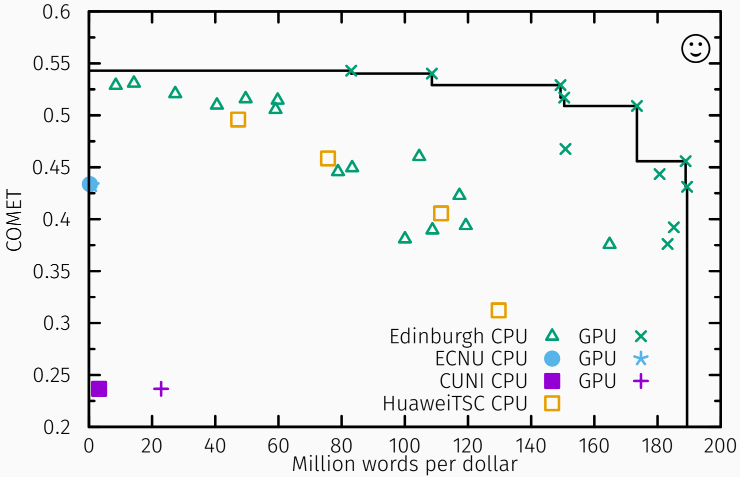


Model sizes



☺
0.6
0.5
0.4
0.3
0.2
0.1
COMET
8
16
32
64
128
256
512
1024
Model size (MB)
Edinburgh ×
ECNU *
CUNI +
HuaweiTSC □
RoyalFlush ○

Financial stats



Conclusion

Conclusion

- Semi non-AG makes GPU competitive on latency in terms of speed, but not in terms of $$$
- Fully non-AG systems have poor quality
- GPU throughput is the cheapest way to translate large quantities of text
- That is $0.002/million characters. By comparison, Google Translate's cost is $20/million characters

Conclusion

- Semi non-AG makes GPU competitive on latency in terms of speed, but not in terms of $$$
- Fully non-AG systems have poor quality
- GPU throughput is the cheapest way to translate large quantities of text
- That is $0.002/million characters. By comparison, Google Translate's cost is $20/million characters

Thank you for your time!

Special thanks to the participants in the task!