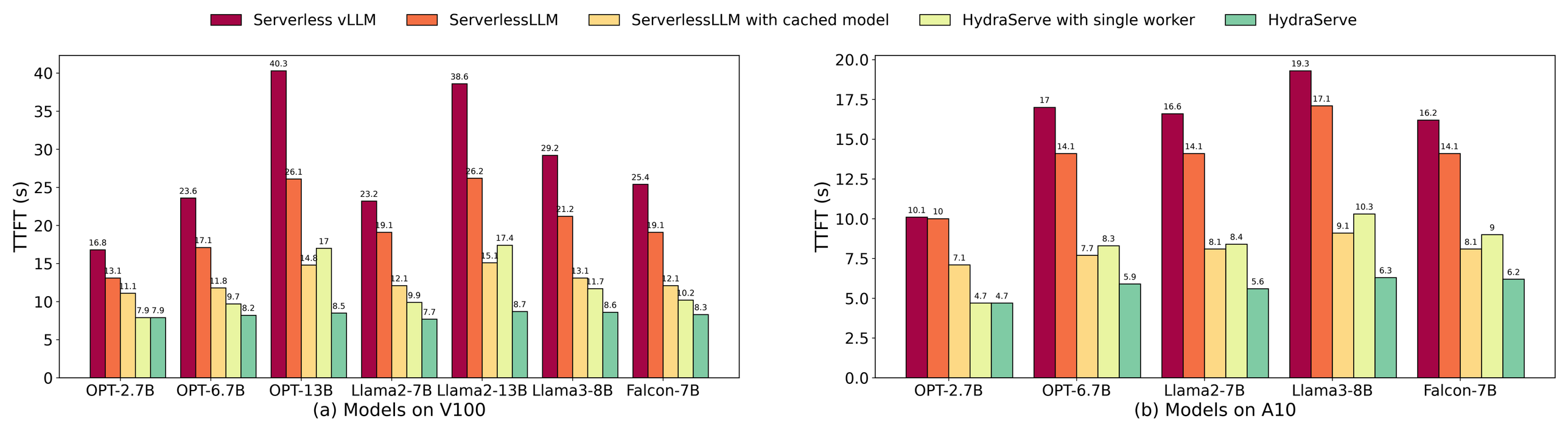
Serverless vLLM
ServerlessLLM
ServerlessLLM with cached model
HydraServe with single worker
HydraServe
TTFT (s)
0
5
10
15
20
25
30
35
40
16.8
13.1
11.1
7.9
7.9
23.6
17.1
11.8
9.7
8.2
40.3
26.1
14.8
17
8.5
23.2
19.1
12.1
9.9
7.7
38.6
26.2
15.1
17.4
8.7
29.2
21.2
13.1
11.7
8.6
25.4
19.1
12.1
10.2
8.3
OPT-2.7B
OPT-6.7B
OPT-13B
Llama2-7B
Llama2-13B
Llama3-8B
Falcon-7B
(a) Models on V100
TTFT (s)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
10.1
10
7.1
4.7
4.7
17
14.1
7.7
8.3
5.9
16.6
14.1
8.1
8.4
5.6
19.3
17.1
9.1
10.3
6.3
16.2
14.1
8.1
9
6.2
OPT-2.7B
OPT-6.7B
Llama2-7B
Llama3-8B
Falcon-7B
(b) Models on A10