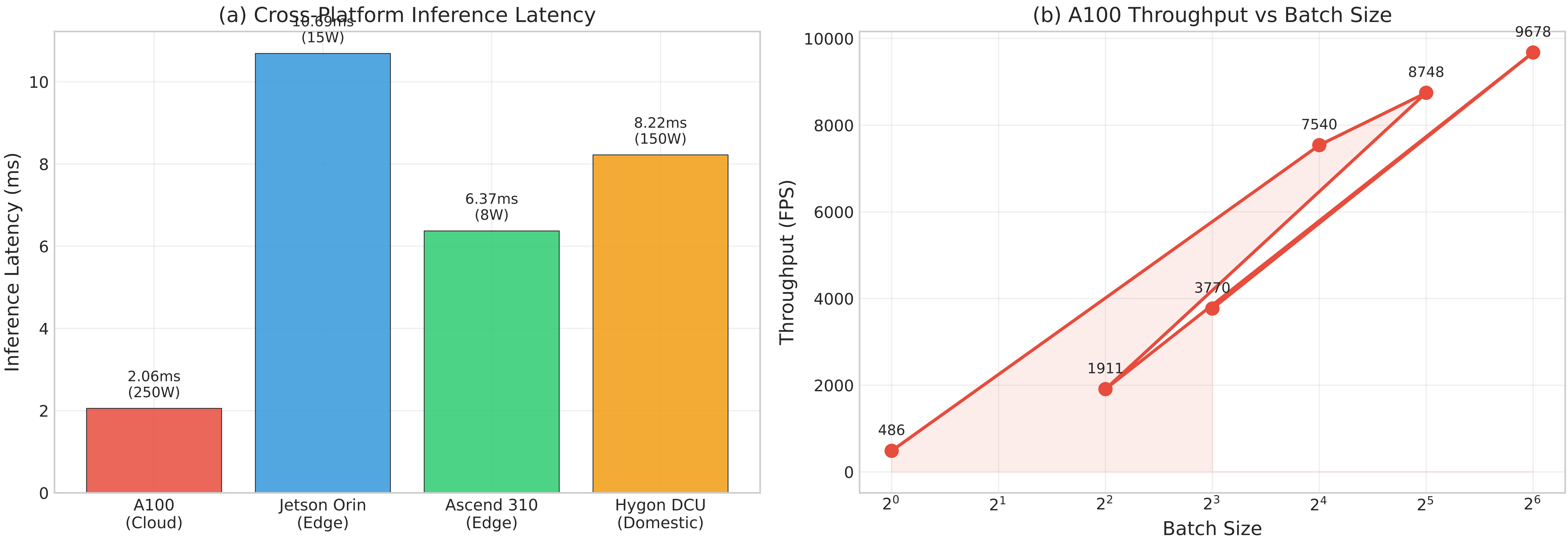
(a) Cross-Platform Inference Latency
Inference Latency (ms)
0
2
4
6
8
10
2.06ms
(250W)
10.69ms
(15W)
6.37ms
(8W)
8.22ms
(150W)
A100
(Cloud)
Jetson Orin
(Edge)
Ascend 310
(Edge)
Hygon DCU
(Domestic)
(b) A100 Throughput vs Batch Size
Throughput (FPS)
0
2000
4000
6000
8000
10000
486
1911
3770
7540
8748
9678
Batch Size
2^0
2^1
2^2
2^3
2^4
2^5
2^6