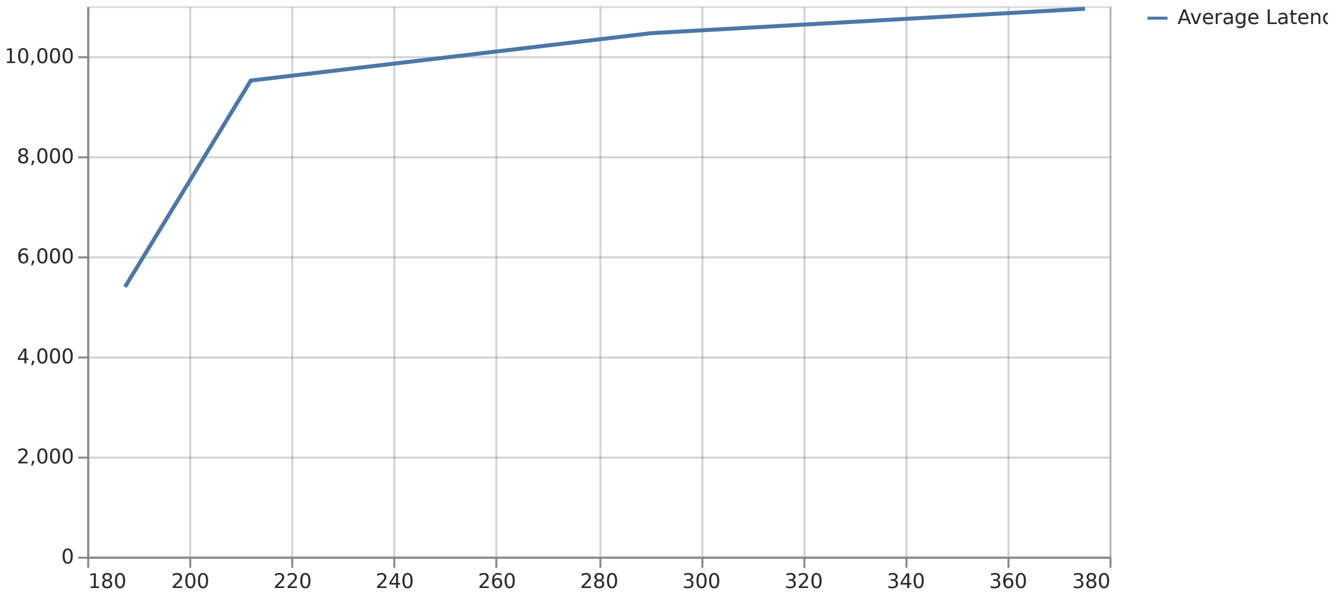
Nirbhay Sharma (B19CSE114)

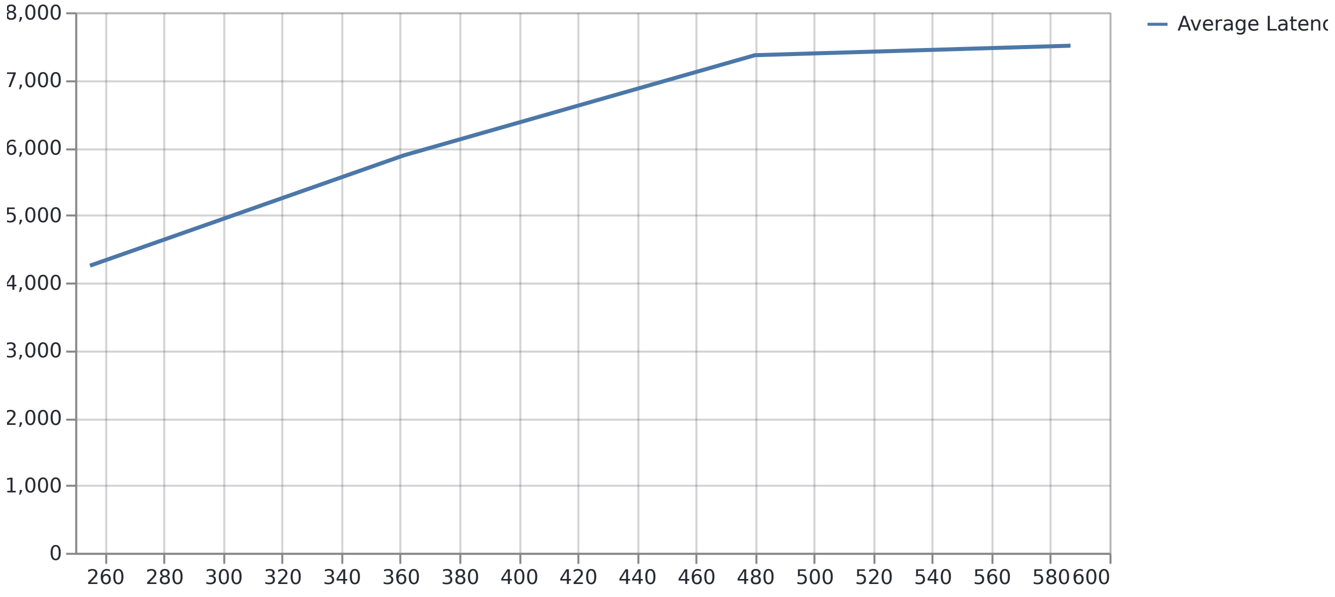
DL-Ops - lab assignment - 4

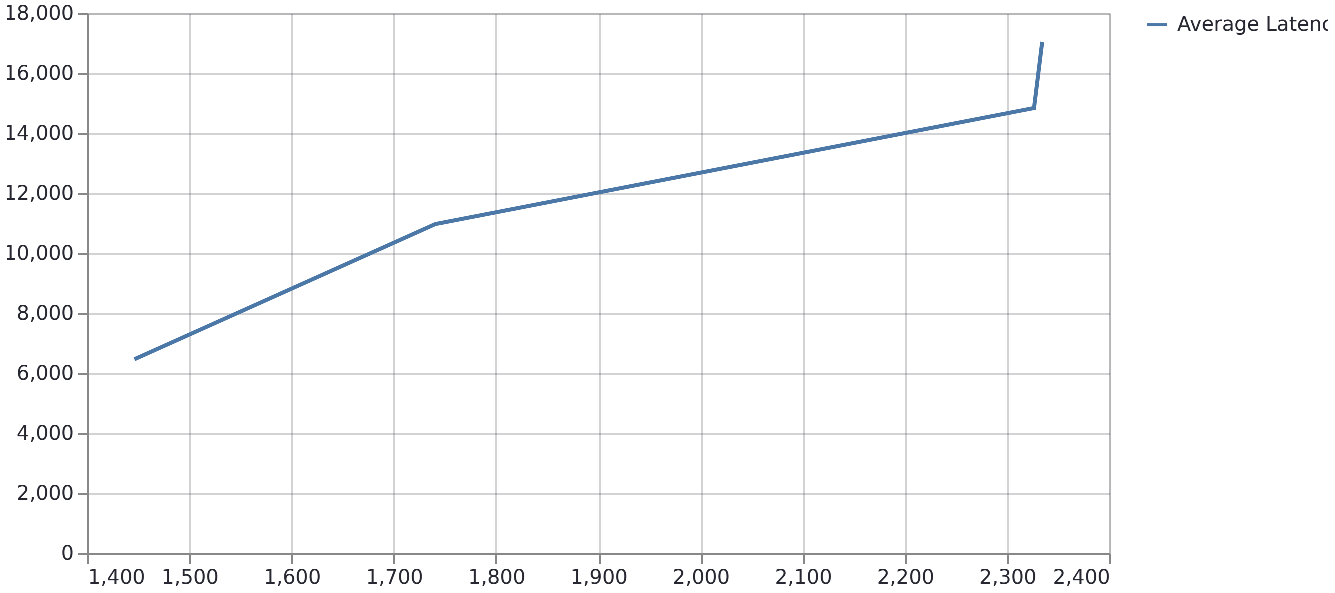
Testing on various values of batchsize and concurrency endpoint

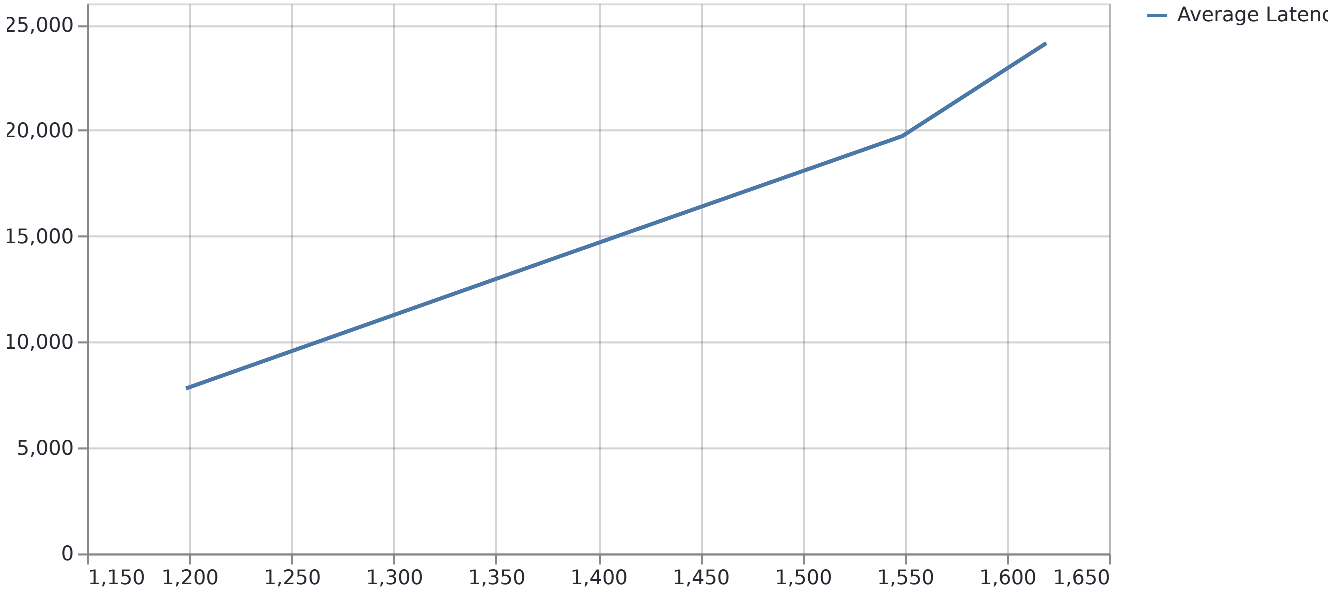
Model	Batchsize	Concurrency range	Throughput	Latency
torch	8	4		
onnx	8	4		
tensorrt 32	8	4		
tensorrt 16	8	4		
tensorrt 8	8	4		

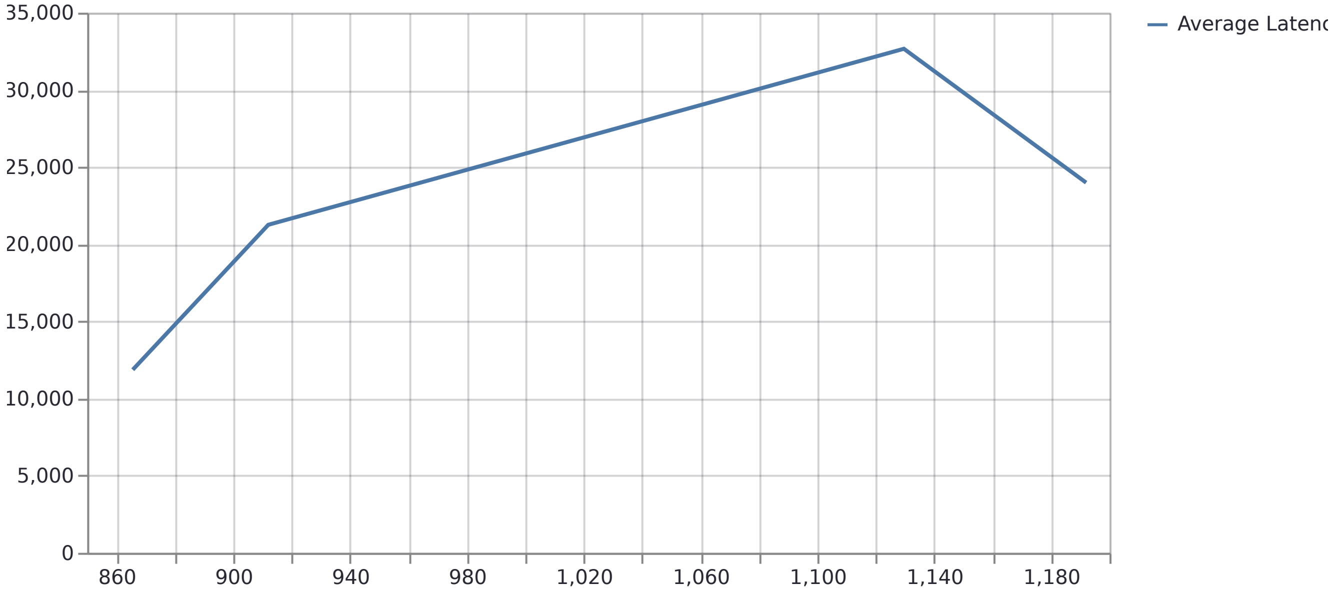
Latency vs Throughput curve for torch, onnx, fp32, fp16, fp8 respectively







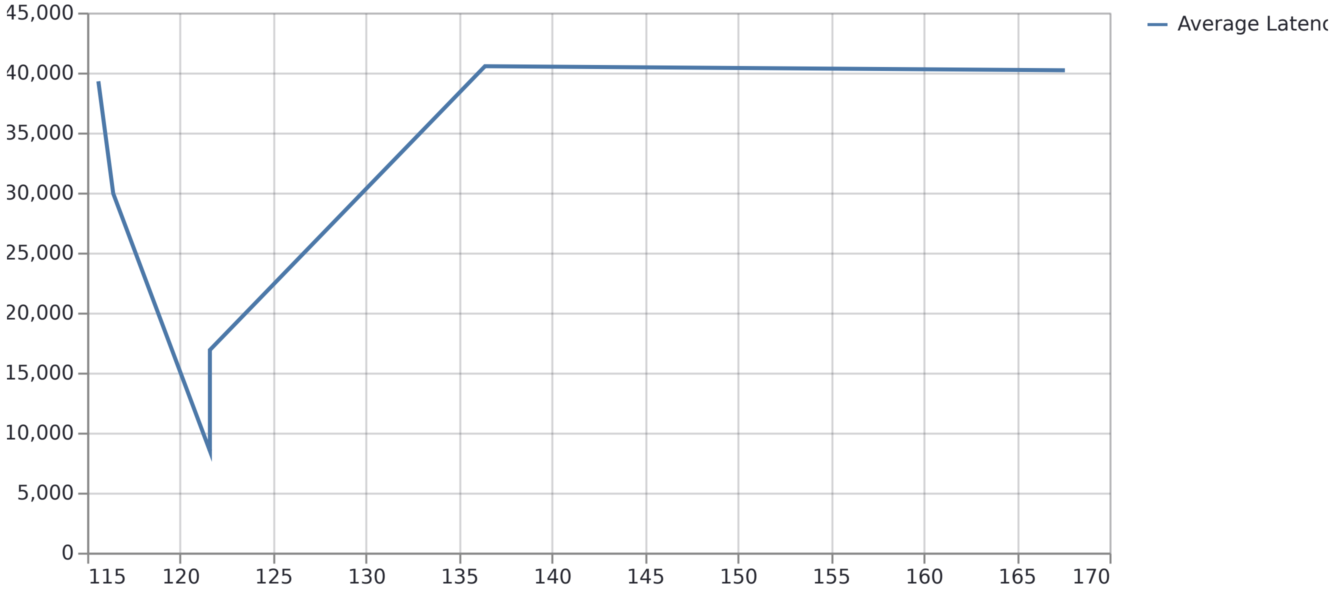


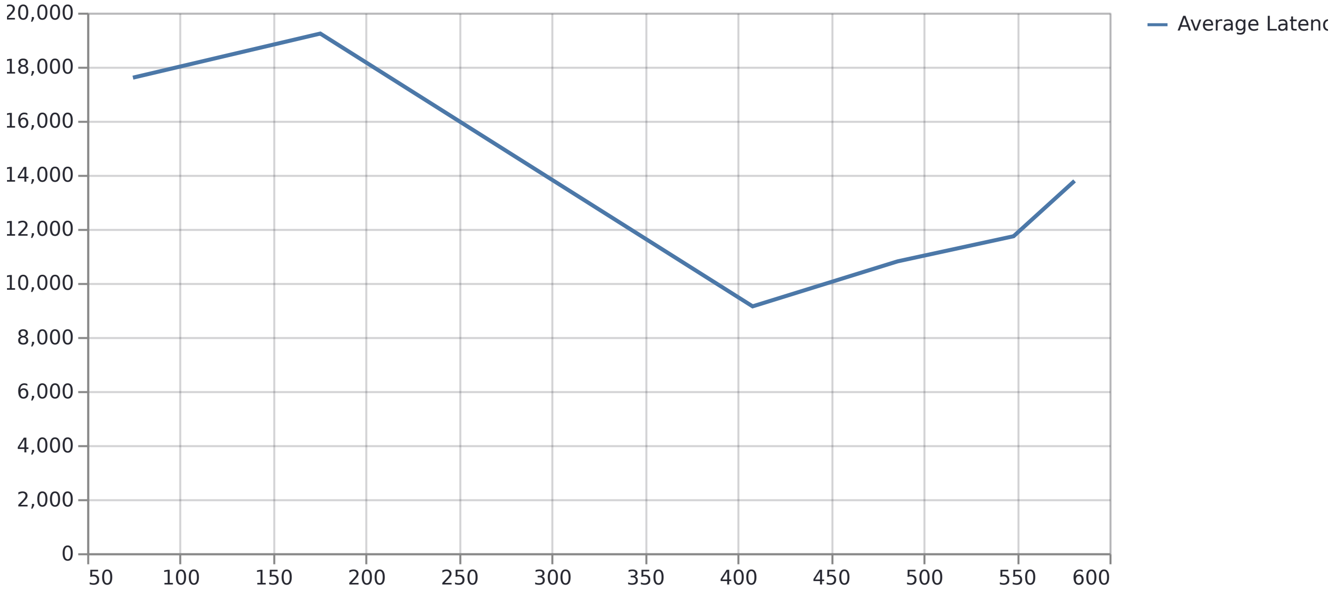


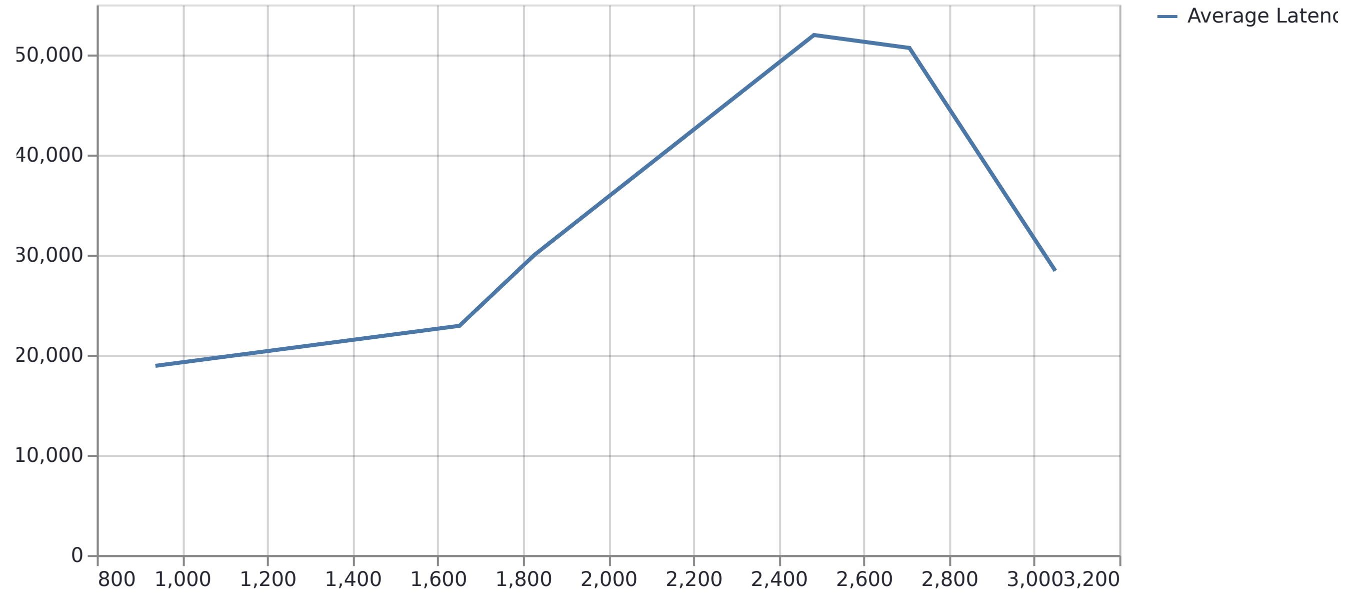
Model	Batchsize	Concurrency range	Throughput	Latency
torch	16	6		

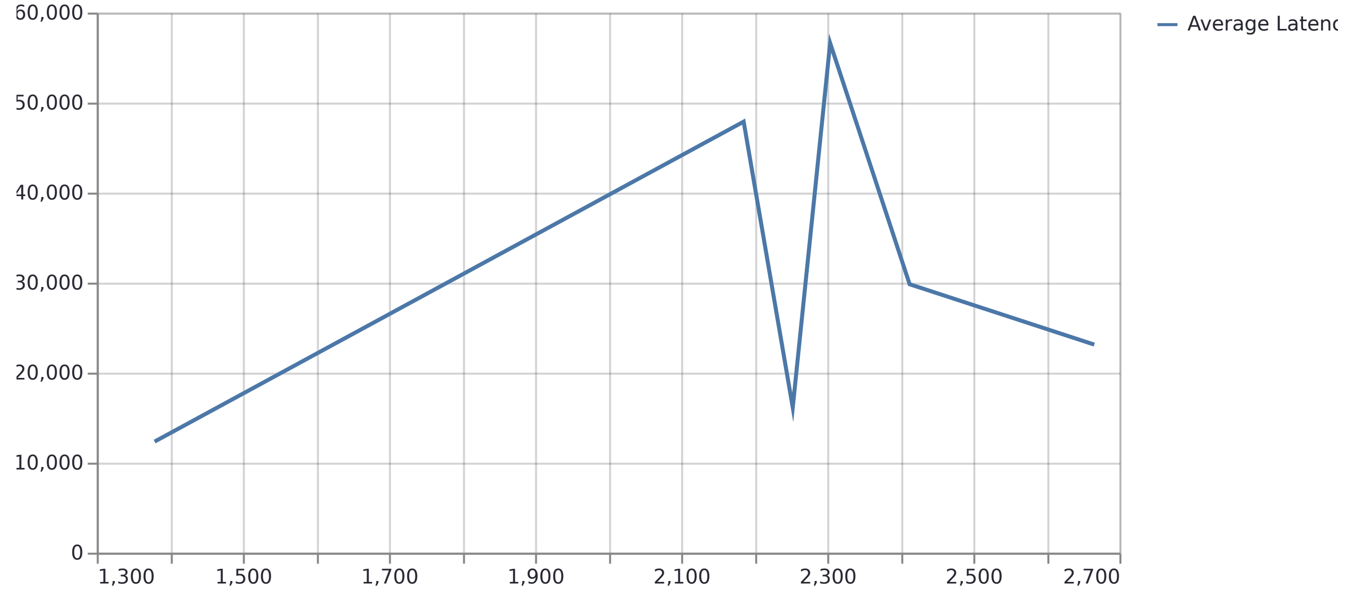
Model	Batchsize	Concurrency range	Throughput	Latency
onnx	16	6		
tensorrt 32	16	6		
tensorrt 16	16	6		
tensorrt 8	16	6		

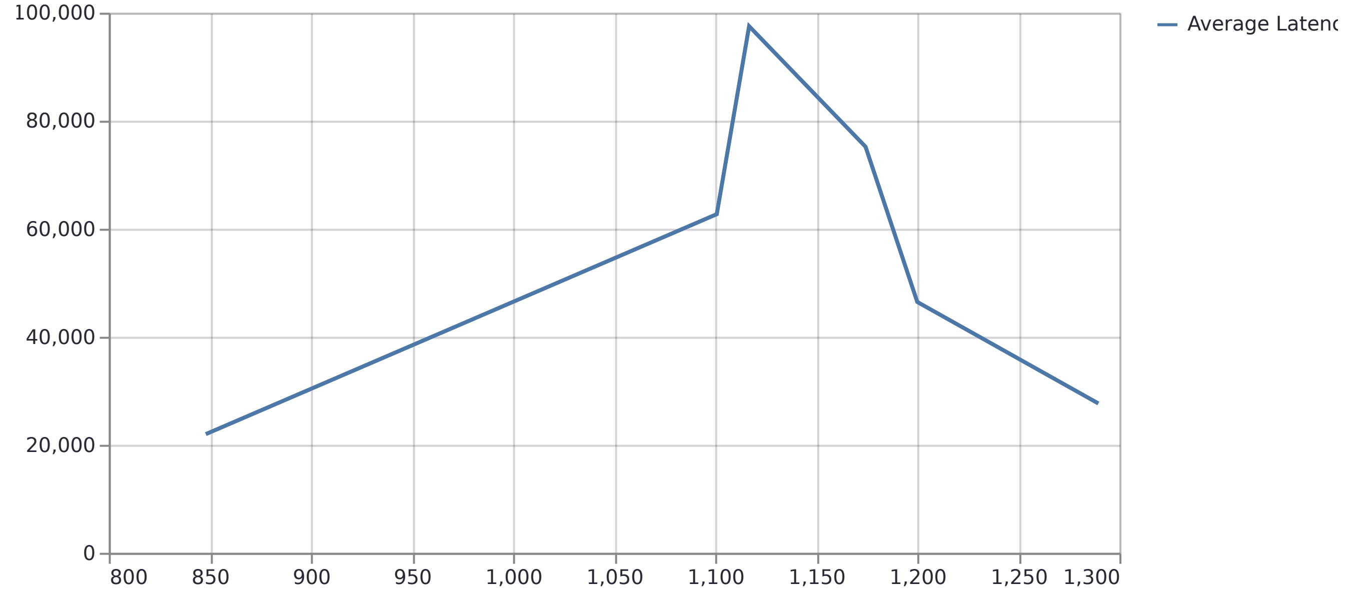
Latency vs Throughput curve for torch, onnx, fp32, fp16, fp8 respectively







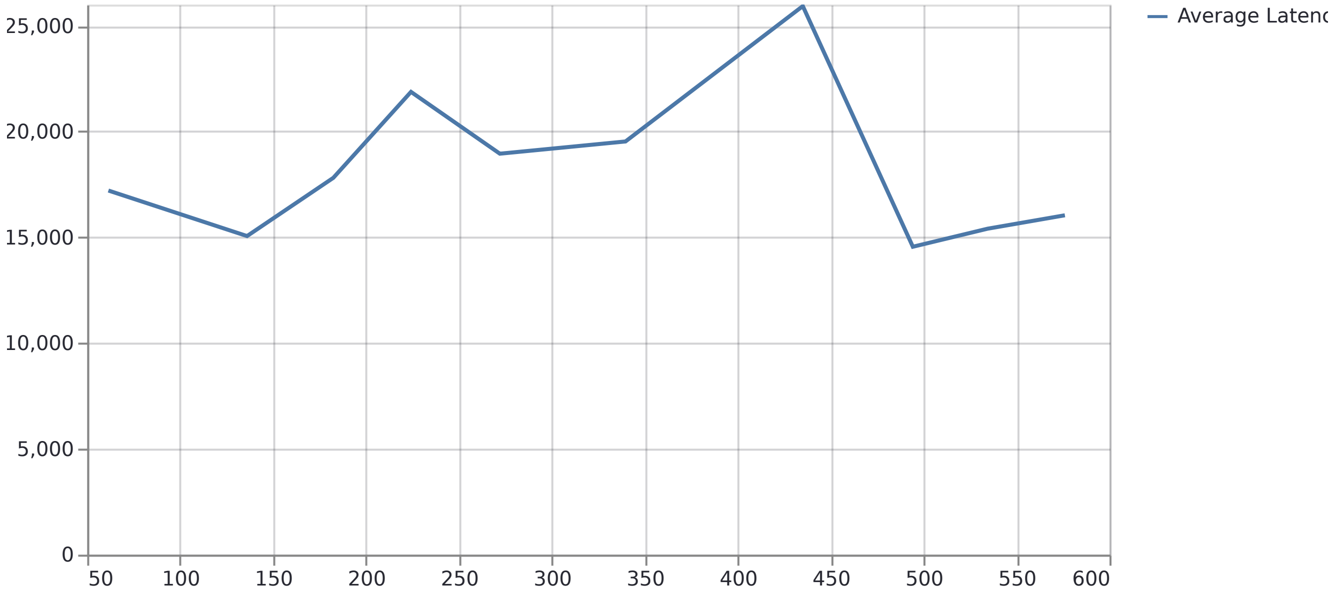


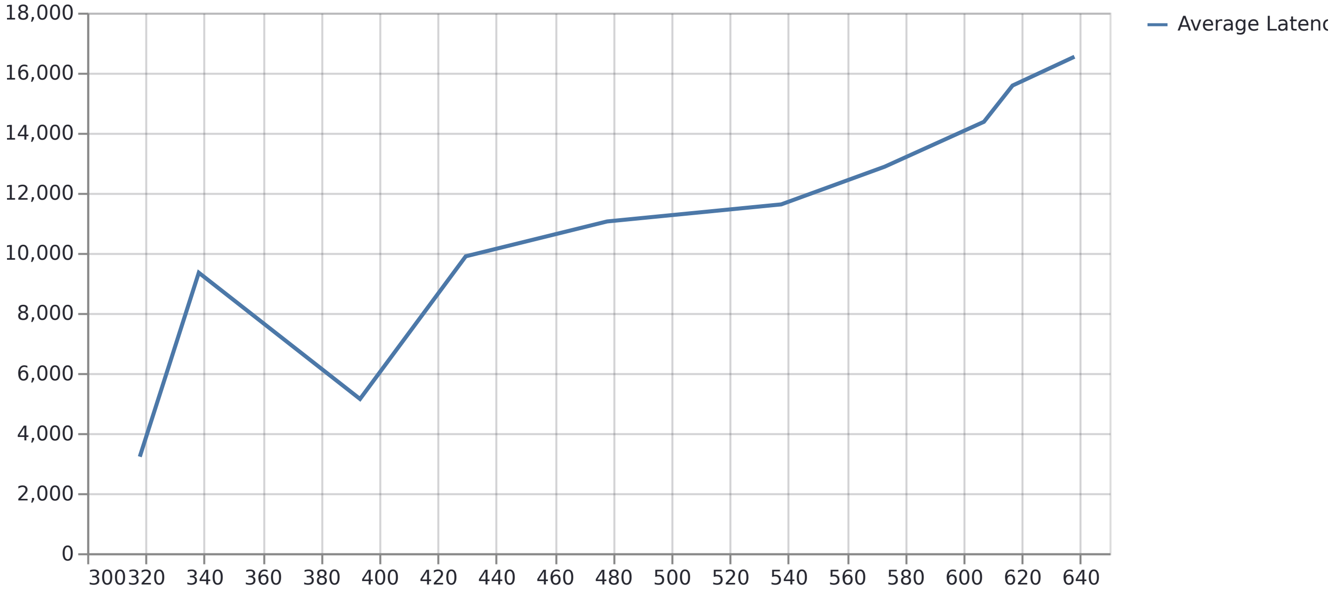


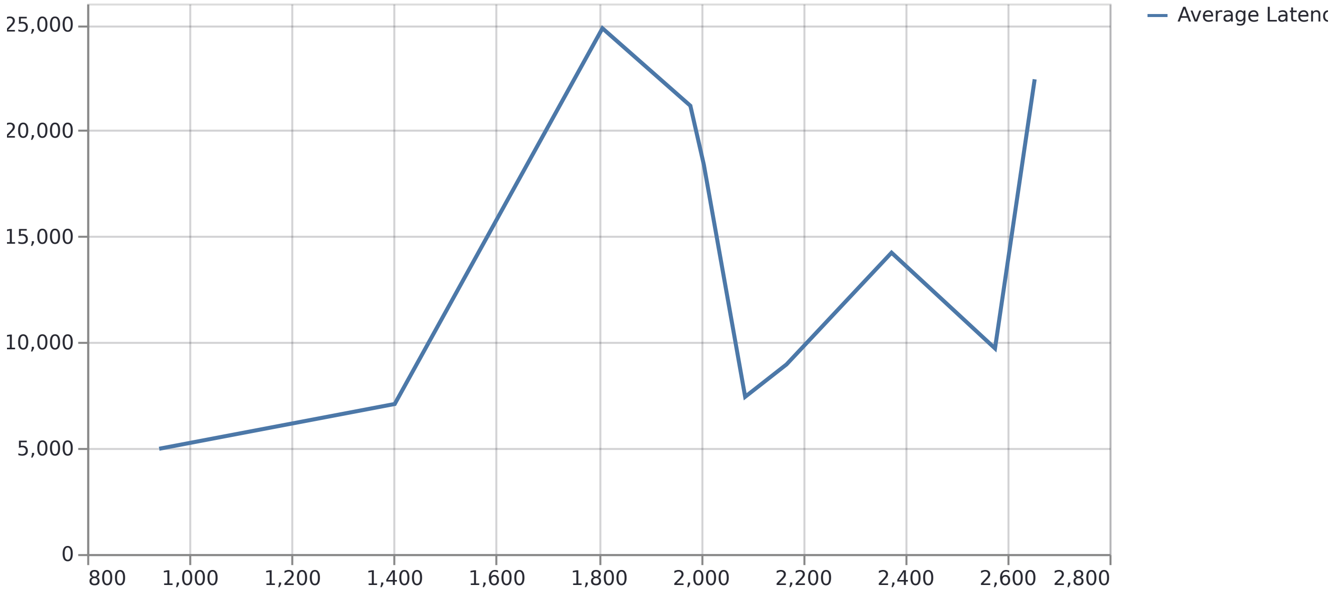
Model	Batchsize	Concurrency range	Throughput	Latency
torch	4	10		

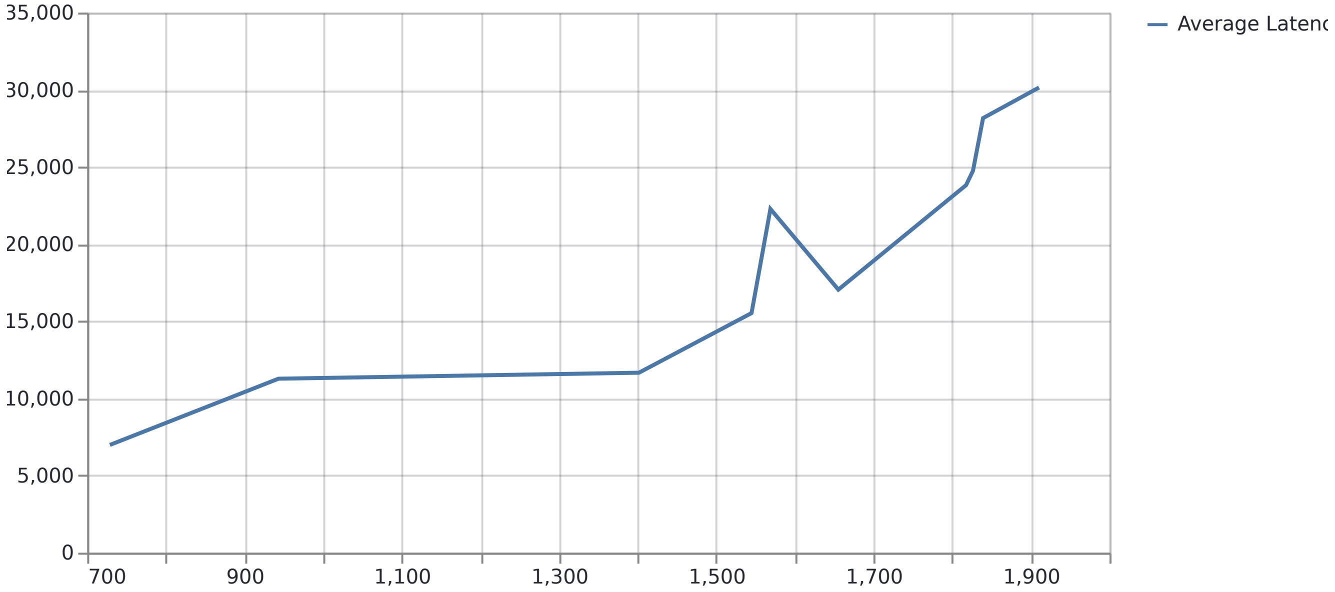
Model	Batchsize	Concurrency range	Throughput	Latency
onnx	4	10		
tensorrt 32	4	10		
tensorrt 16	4	10		
tensorrt 8	4	10		

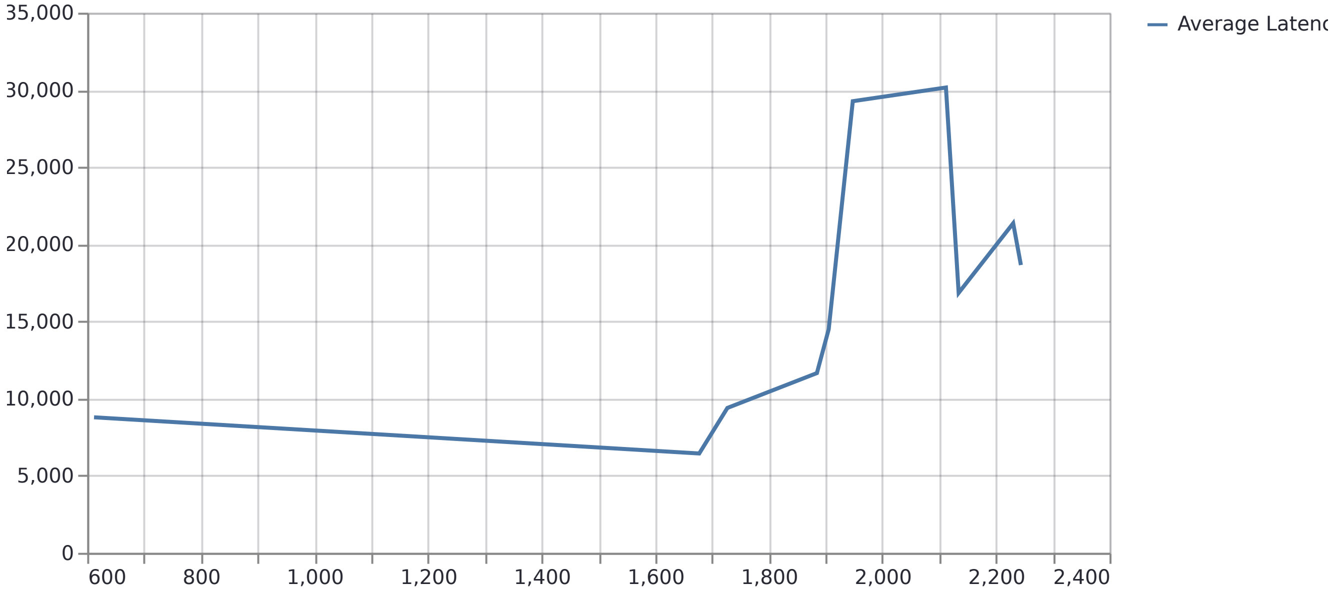
Latency vs Throughput curve for torch, onnx, fp32, fp16, fp8 respectively






Average Latenc
25,000
20,000
15,000
10,000
5,000
0
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
2,800


Average Latenc
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
700
900
1,100
1,300
1,500
1,700
1,900


Average Latenc
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400