



$$\text{system_latency} = T 1 / t$$

$$\text{system_throughput} = b / \text{system_latency}$$

$$\text{n_percentile_time} = \max(\text{nth\% of inferenceTimes})$$

example : `np.percentile(inferenceTimes, n)`

$$\text{max_time} = \max(\text{inferenceTimes})$$

$$\text{min_time} = \min(\text{inferenceTimes})$$