DeltaServe(4bit)+I/O+Prefetch, $deltaserve$, $\lambda = 3.0$, NVMe

Time Spent On
- Loading
- Inference
- Queueing

Time (s) vs Request ID