vLLM, $\lambda = 0.5$, NVMe

Time Spent On
- Loading
- Inference
- Queueing

Time (s) vs Request ID