|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 目标 | 模型 | QPS | GPU类型 | 个数 |
| 部署NLU模型 | BERT词向量,BILSTM+CRF,LAC,相似度匹配。 | 20 | T4（16G) | 1-2个 |
| 部署ASR模型 |  |  |  |  |