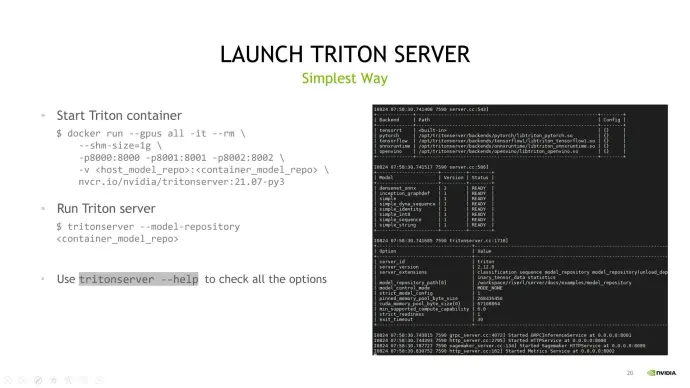
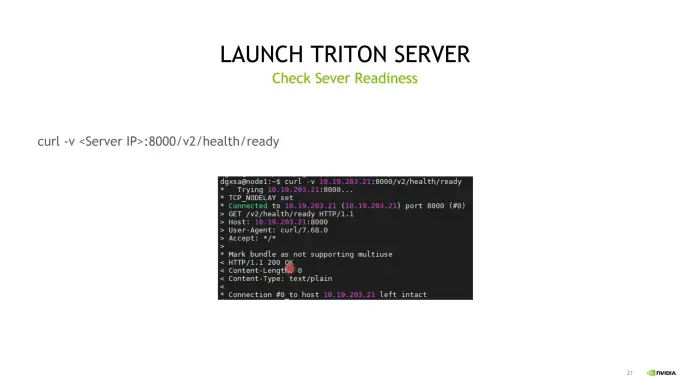
****一、最简单的方式****



* 开启Triton容器
* 运行Triton服务

****二、检查服务是否准备就绪****



****三、其他启动选项****

--log-verbose <integer>: 开启verbose日志信息。

--strict-model-config <boolean>: 是否需要配置模型。

--strict-readiness <boolean>: ready状态显示状况。

--exit-on-error <boolean>: 模型加载部分失败，是否也启动。

--http-port <integer>: 指定http服务端口，默认是8000。

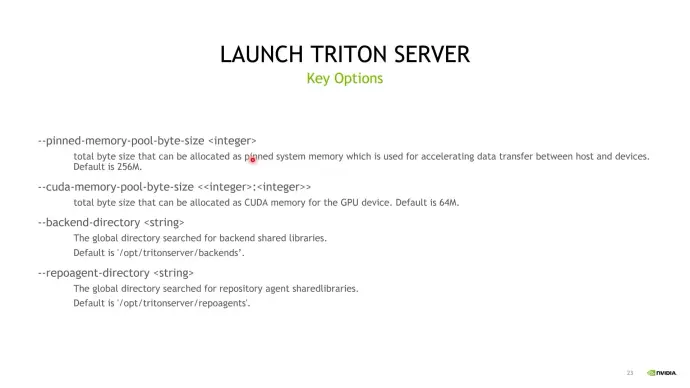
--grpc-port <integer>: 指定GRPC服务端口，默认是8001。

--metrics-port <integer>: metrics报告端口，默认8002。

****--model-control-mode <string>****: 模型管理模式，默认是none，把模型库中所有模型都load进来，并且无法动态卸载或者更新。explicit，server启动时不加载模型，可以通过api进行加载或者卸载模型；poll，动态更新模型，增加新的版本或者修改配置，服务都会动态去加载模型。

--repository-poll-secs <integer>: 模型控制模式为poll时，自动检查模型库变动的时间。

--load-model <string>: 模型控制模式为explicit时指定启动时加载的模型。



--pinned-memory-pool-byte-size <integer>：可以被Triton服务使用的锁页内存大小，关于锁页内存可以参考：https://cloud.tencent.com/developer/article/2000487。

--cuda-memory-pool-byte-size <<integer>:<integer>>：可以被Triton使用的cuda memory 大小。

--backend-directory <string>: backend搜索路径，可在使用custom backend的时候指定自己的库。

--repoagent-directory <string>：预处理模型库的库，譬如在load模型的时候进行加密。

****发送请求到Triton服务****

**同步方式推理**：



**异步方式推理：**



**通过share memory去推理：**

