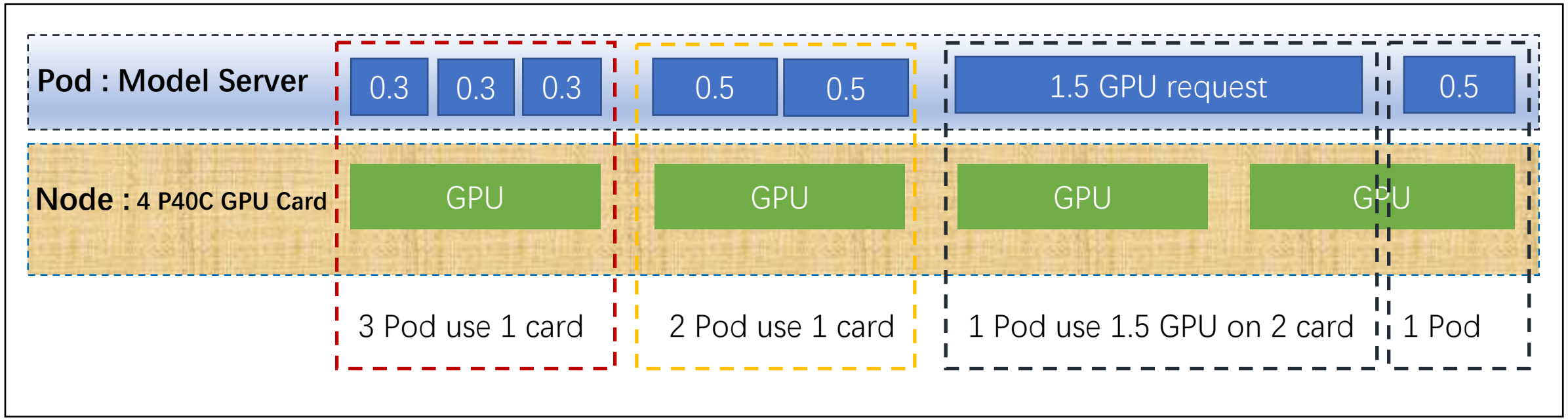
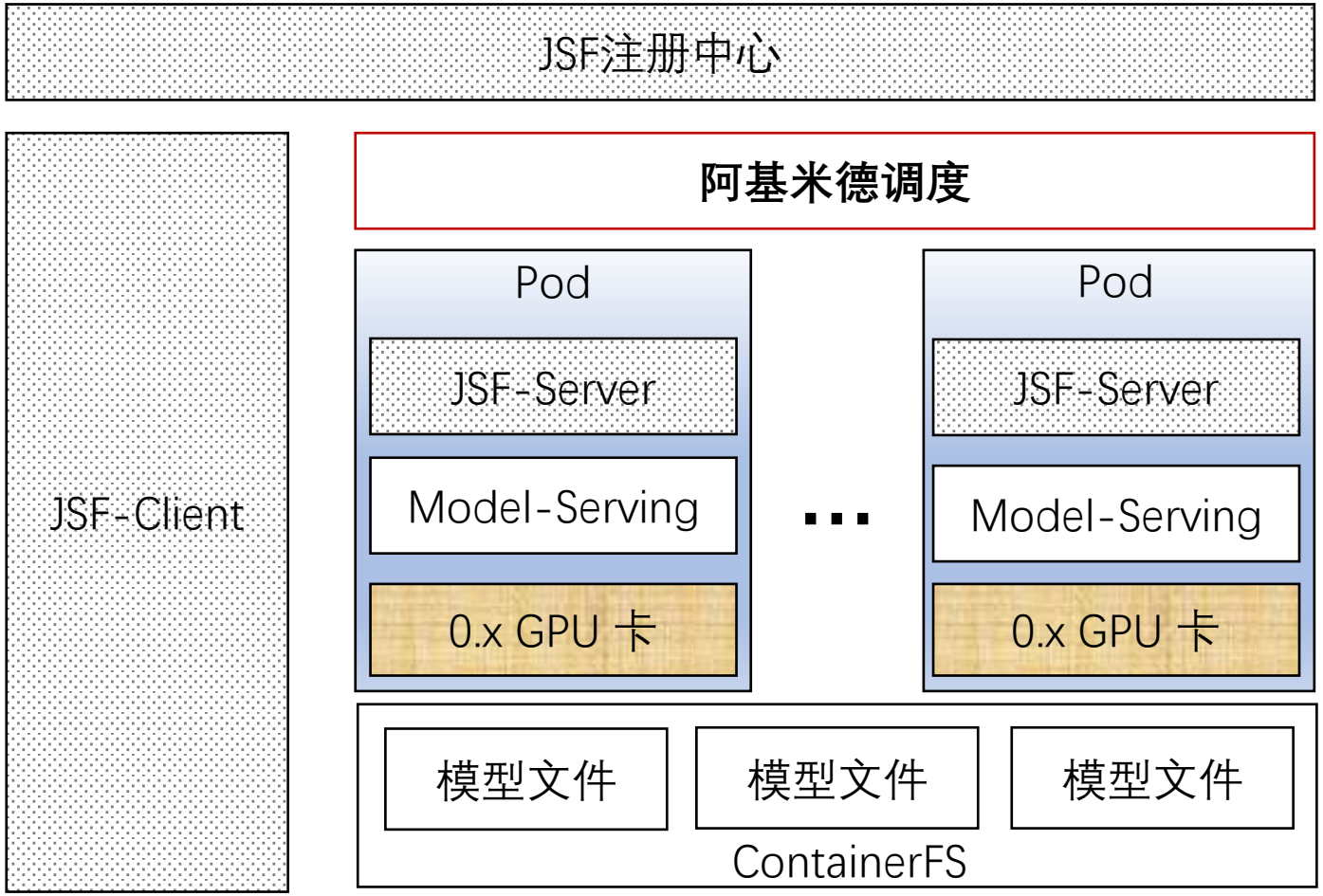
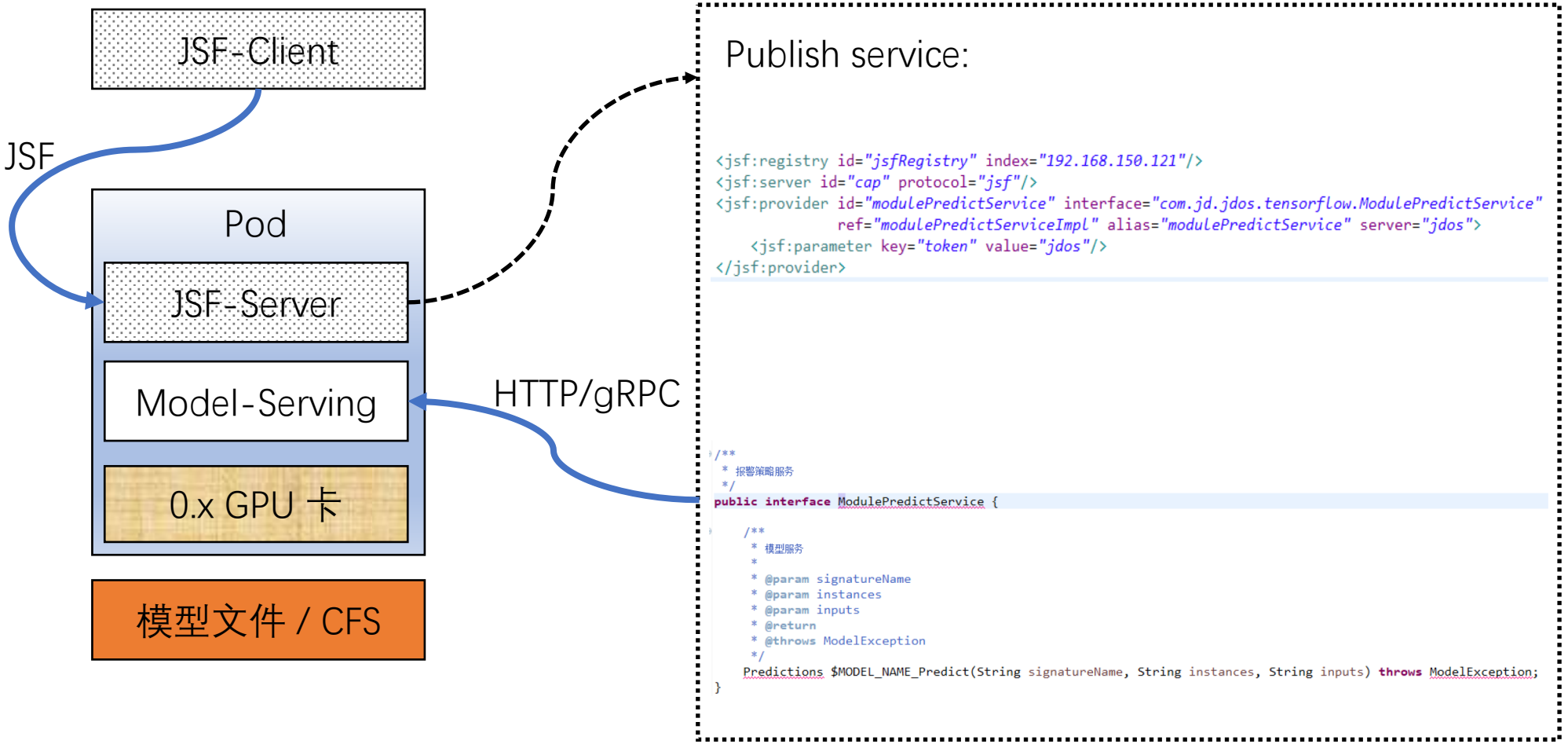
单卡多容器技术-控制GPU显存使用



TIG AI Serving Architecture

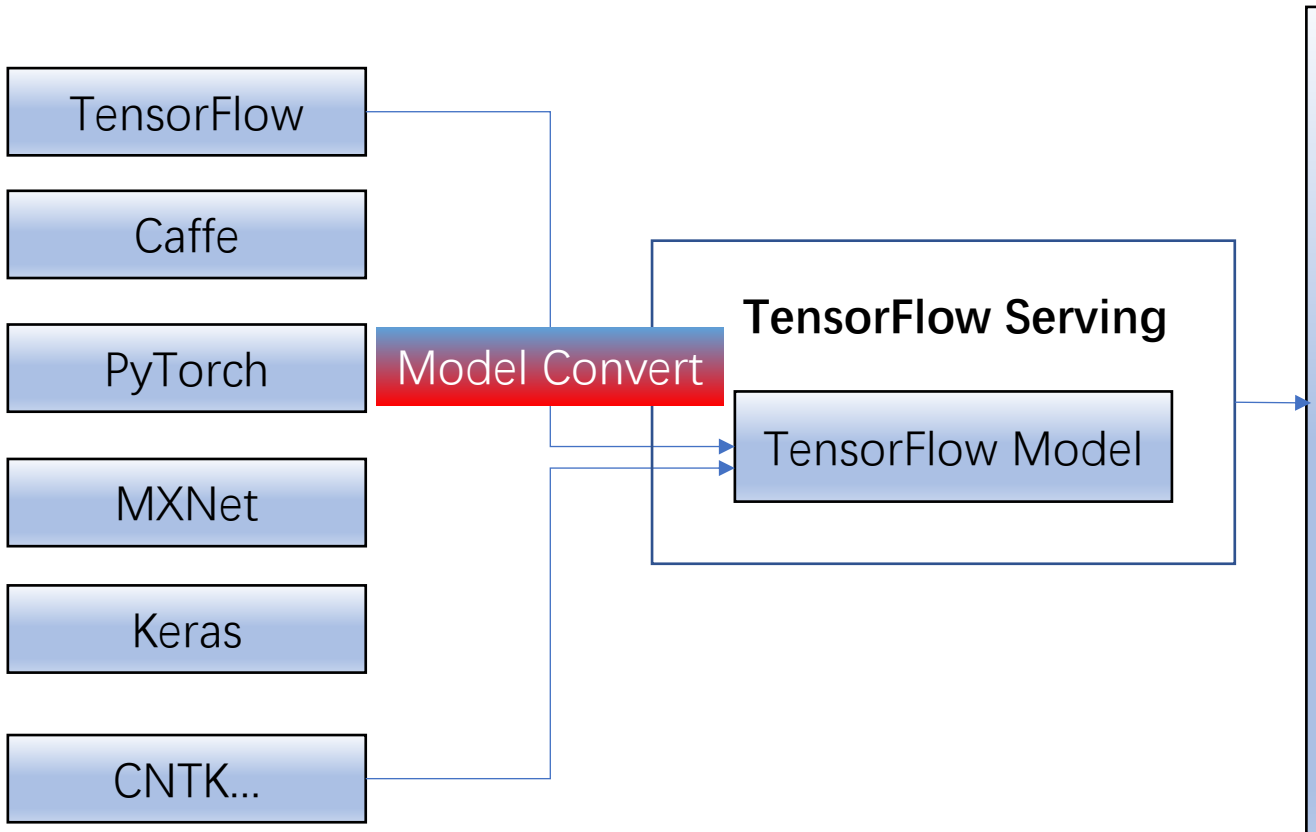


How JSF Helps Model Serving



JSF Service 可以模板生产Server端代码 模型开发者也可以在JSF Server内开发前置处理逻辑代码

How Model Serving work



Unified Predict Api

/v1/models/$MODEL_NAME:predict

Standard Request Format:

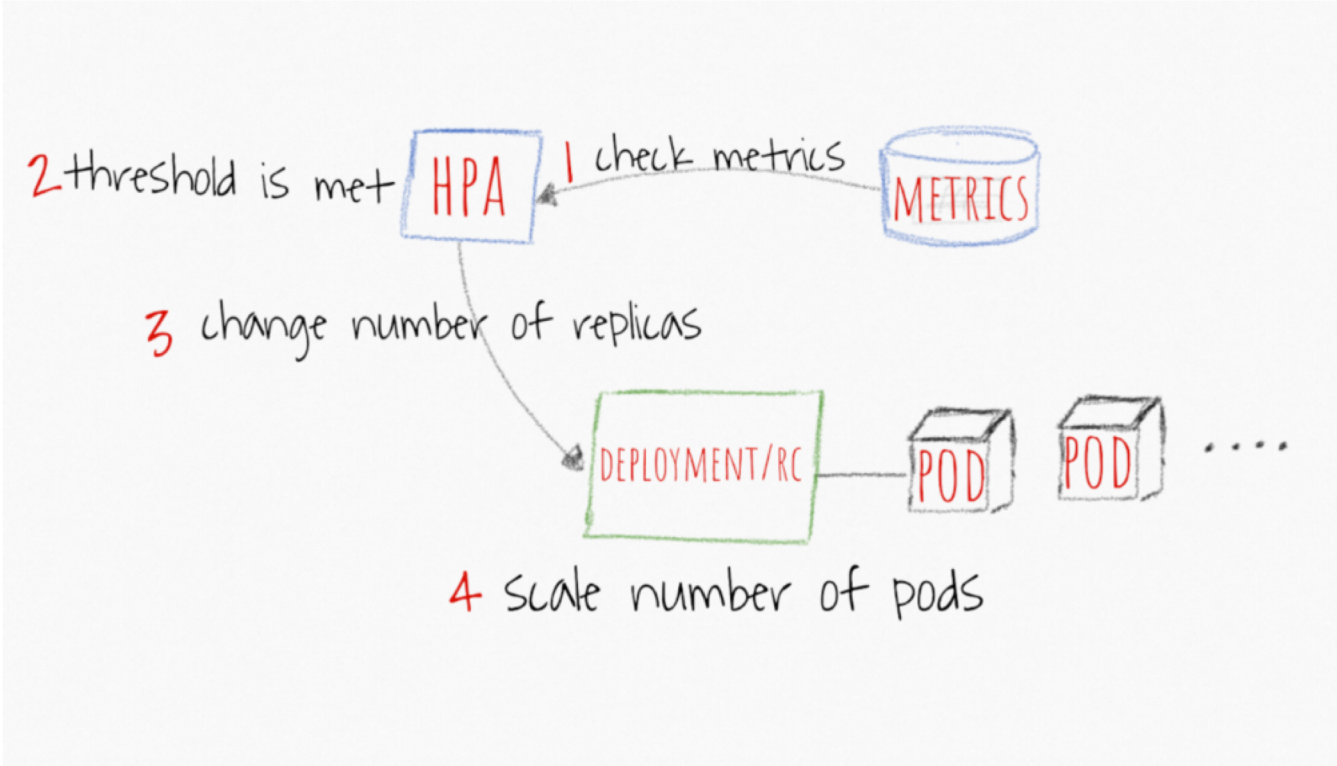
```
{
  // (Optional) Serving signature to use.
  // If unspecifed default serving signature is used.
  "signature_name": <string>,

  // Input Tensors in row ("instances") or columnar ("inputs") format.
  // A request can have either of them but NOT both.
  "instances": <value>|<(nested)list>|<list-of-objects>
  "inputs": <value>|<(nested)list>|<object>
}
```

Standard Response Format

```
{
  "predictions": <value>|<(nested)list>|<list-of-objects>
}
```

https://developer.nvidia.com/deep-learning-frameworks

How Archimedes increases GPU average usage



阿基米德调度器根据JSF反馈tp情况和Pod聚合负载决定伸缩Pod【每个Pod share 0.x GPU卡】