Bringing the app to production …

1. **aspects** to make the application provided ready for production:

**Scaling**

Tornado runs a single thread application, and this could limit the CPU cores/thread usage of the hosting server.

Possible workarounds are:

- using application load balancers and multiple tornado instances (as a process, containers/pod),

- deploying multiple instances of Tornado exploiting multiple threads,

- deploying multiple containers,

- deploying Tornado on Kubernetes (e.g. using pod replicas)

**Monitoring**

Single process monitoring could be difficult and expensive in case the instances number is high.

Load balancer monitor could be a viable solution.

Monitoring a Kubernetes cluster could be the easiest way to go.

**High availability**

Deploying more instances on different locations could help – not sure if with Azure this is possible across Italy (it looks like that we will have [one region – in lombardia](https://news.microsoft.com/europe/2023/06/05/microsoft-announces-its-first-cloud-region-in-italy-accelerating-innovation-and-economic-opportunity/))

2. **actions** to make the application provided ready for production:

**Scaling**

Running Tornado in a ***Kubernetes cluster***, using an Application Load Balancer (ALB) in front of the Tornado deployments, e.g. NGINX, HAProxy, Traefik or Istio. These are the so called **Ingress controllers** (or External load balancer)and they operate at Layer 7. For example Azure has its own Application Gateway. This kind of load balancer exposes the application to external users or services outside the cluster and they require an Ingress controller to be installed (as additional step to the K8s configuration).

Kubernetes makes also use of Internal Load Balancer, which operates at layer 4, to route traffic within the cluster (Azure’s one is azure-load-balancer-internal) [refs.[1](https://spacelift.io/blog/kubernetes-load-balancer), [2](https://spacelift.io/blog/kubernetes-ingress)].

**Monitoring**

Deploying a monitoring system, e.g. Grafana (open-source) or Dynatrace (commercial).

Better would it be on a Kubernetes cluster – but also feasible on a single VM instance, or on a single process.

**High availability**

Go together the scaling approach, caring about the Vms/Nodes locations – interesting for an international deployments, less interesting for a national scale one (since there are not many Azure datacenters in Italy, unless we want to proceed with an hybrid cloud/on-prem approach).

3. **additional developments** to make the application provided ready for production:

The application provides substantially a single endpoint that covers the 5 required actions, namely getting the whole list of student, or a single student by using its id, create a new user, modify or delete and existing one.

The single endpoint makes use of the 4 method involved (GET, POST, PUT, DELETE).

It can be extended to multiple endpoints following the schema:

- GET “/list/” : provide the full students list

- GET “/getStudent/\_id\_student”: provide a student by its id

- POST “/addStudent/”: add a new student using the body of the request as an input,

- PUT “/modStudent/”: modify a student using the body of the request as an input,

- DELETE “/delStudent/\_id\_student”: delete a student using its id.

Improving tests:

- using the tornado.testing framework more tests can be developed.

- Jmeter can be used to stress the application to check if there is a breaking point, e.g., creating new users (e.g. the MongoDB collection max.16mb size, or max Tornado open connections – maybe tens of thousands?), it looks like Jmeter is highly suitable for running parallel http requests.

- Postman semplifies the operation of creating, and modifying users since it sets Header of the requests automatically – further tests can be created.