IONICA

High Availability

Do you need high availability?

If you depend on your applications or web sites to run your business, they need to be online, fully operational, secure, and fast all the time – 24 hours a day, 7 days a week. That's high availability.

Is uptime the same as high availability?

Uptime is the amount of time an application or service is online and available. That's important, but you also need your applications and services to be fully functional, responsive (fast enough), and secure as well – otherwise, their utility to your customers, employees, and other stakeholders is limited . . . or not present at all. High availability means that all essential services and data are fully redundant (backed up and on more than one server or cloud instance), so that if a hardware component or server goes down, there are additional resources available to do the same jobs.

Measuring availability

Measuring service availability and uptime is essential to ensure services are up all the time, and to catch issues before they actually disrupt delivery of services.

At IONICA, our comprehensive monitoring and logging systems check hundreds of data points on every application environment that we care for. These systems check key web pages to ensure they are online, error-free, responsive, and loading the content our clients expect them to display. They check database health, load, and responsiveness, filesystem cluster synchronisation status, CPU utilisation, swap / memory, disk I/O, availability of all other services in a given environment (like Redis, Solr), and much more.

IONICA security monitoring services aggregate data from host and network-based intrusion detection systems, web application firewalls, and system logs.

Of course, monitoring is only of so much use without alerting. If our monitoring systems detect an issue, IONICA engineers are notified instantly, so that we can take action before an issue becomes a serious problem that can hinder service availability, security, or performance.

How IONICA designs for high availability

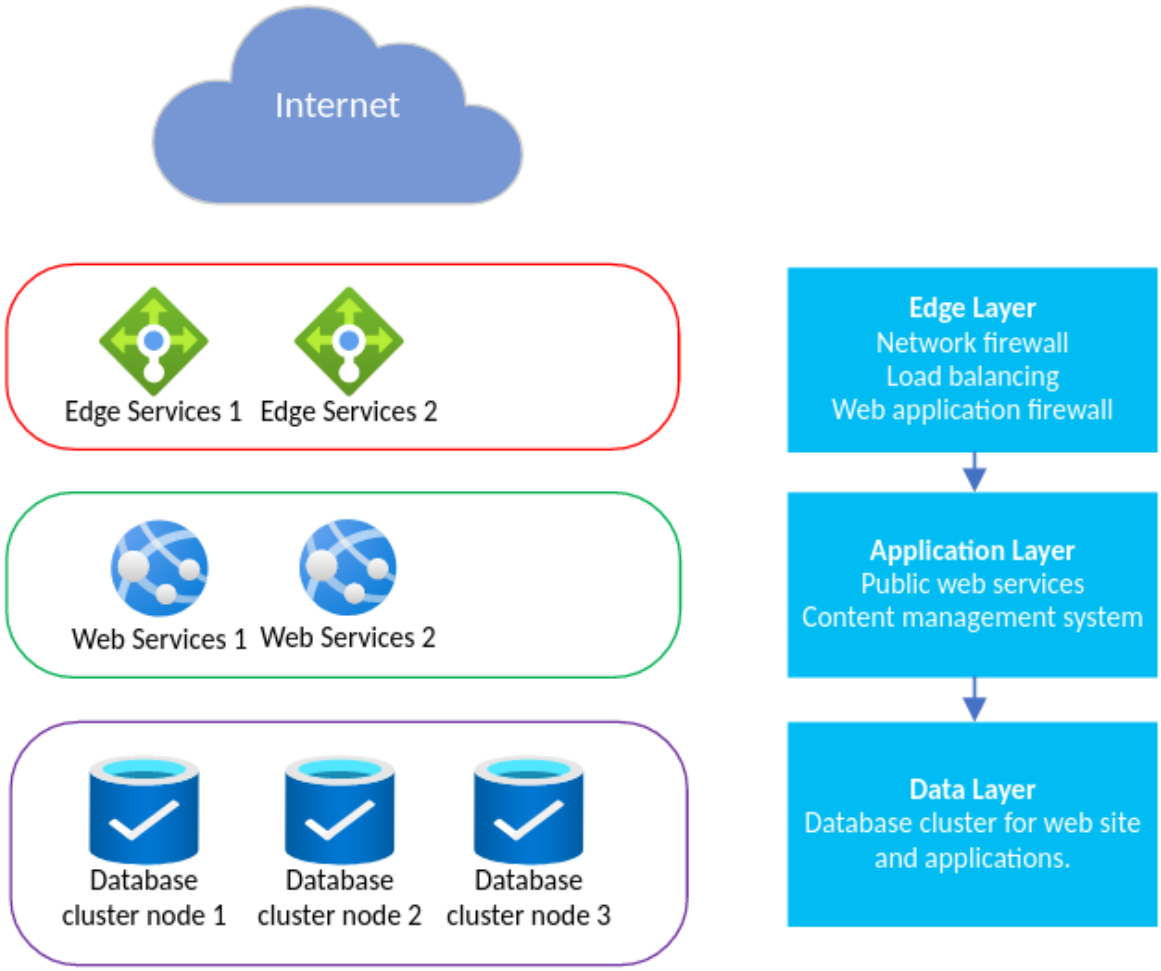
First, we talk to you and learn your business needs – what is important to your business operations, why it is important, what security compliance standards your applications and data storage must meet (if any) and couple that with information on traffic patterns and volume. Then we design a comprehensive solution to meet your goals and budget.

An example of a relatively simple high availability environment for an application with moderate traffic:

- Two load balancers, each with their own network and web application security firewalls.
- Two static asset cache nodes.
- Two to five application servers (or web servers). Files are synchronised among application servers instantly, with filesystem clustering.
- A three–node database cluster. High availability and high availability + high performance configurations are available.

For clients who truly need to be prepared for any contingency, we design and maintain disaster recovery environments in geographically distant locations. That way if there is a disaster at the primary data centre used to host your applications – a fire, earthquake, flood, or other catastrophic incident -- the disaster recovery environment is activated, ensuring business continuity and continuous delivery of services.

Example diagram:



Putting it all together – high availability at scale

High scale use cases require more capacity and infrastructure *(more cloud instances, containers, virtual machines, and more storage)*, though high scale environments have several other key differences from smaller ones in architecture design as well. For example, high scale use cases may involve geographically dispersed resources to serve large numbers of client connections in multiple locales or jurisdictions. Database sharding and/or federation may be required, clustering of search and caching services may also prove beneficial. Network capacity and bandwidth are often overlooked in ensuring environments are ready for scale – network capacity and throughput must be sufficient to accommodate peak traffic levels. Finally, a content distribution network (CDN) may also provide benefits in reducing load, increasing performance, and reducing latency.

As they often require large numbers of cloud instances, containers, applications, and virtual machines, high scale environments also often require sophisticated provisioning, automation, and change management through systems like Ansible, Puppet, Jenkins, or Azure DevOps tools.