CCD: Laboratorio Web

Diagnóstico infraestructura

Descripción del problema

En nuestro sitio principal, experimentamos una tendencia de aprox. 250 usuarios activos diariamente. En nuestros días de mayor actividad, podemos llegar al orden de los 1000 usuarios activos. En ocasiones, se han generado cuellos de botella en los que algunas personas no puedan acceder a los servicios, o que estos operen deficientemente.

Un ejemplo: el 25 de octubre, tras una invitación a en twitter a visitar el sitio del CCD, el tráfico aumentado causó un bloqueo en el CPU, por lo que que muchas personas no pudieran ver el sitio.

Además, algunas de nuestras aplicaciones más recientes tienen un mayor consumo de recursos, ya que:

- tienen un mayor grado de interacción con la base de datos por parte de los usuarios
- utilizan notificaciones en tiempo real (via socket.io)
- dichos datos en tiempo real no pueden ser cacheados
- realizan consultas más complejas a la base de datos

Como ejemplo: al probar nuevas funcionalidades de nuestra reciente Zona Hipermedial, cuando hemos intentado correr pruebas con arriba de 50 usuarios simultáneos, hemos experimentado saturación del 100% en el CPU. Tras monitoreo y diagnóstico, la principal carga ocurre en en el software de balanceo de carga (nginx) y en la aplicación back-end.

Propuesta de solución

Tras investigar posibles mecanismos para mitigar estos cuellos de botella, y hacer viables las cargas de trabajo, decidimos que la mejor solución es implementar un cluster de servidores que facilite y/o automatice el escalado de los servicios dependiendo las fluctuaciones en la carga.
Por esta razón, nos encontramos transicionando a una arquitectura en la nube, usando un cluster de máquinas virtuales, concretamente, con la plataforma Kubernetes.
Concretamente, hemos trabajado en DigitalOcean.

Esto nos permite mantener múltiples contenedores ofreciendo servicio en paralelo, asignar los recursos de un modo flexible, restaurar o actualizar aplicaciones sin downtimes, respaldar instancias cuando hay servicios que crashean, balancear la carga entre los contenedores, permitiendo a las aplicaciones escalar o desescalar su consumo de recursos, garantizando que haya alta disponibilidad en el servicio.

Las plataformas que van a recibir más tráfico en este 2020, ya van a ser desplegadas en Kubernettes.

- Compás Creativo
- Zona Hipermedial

Esto es necesario para el Laboratorio Web por varias razones:

- Evitar caídas en servidores, garantizar alta disponibilidad (Actualmente, tenemos "downtime" en promedio 3 ocasiones mensuales)
- Relevantado automático de nodos no saludables / con fallos
- Balanceo de carga que pueda redireccionar usuarios a nodos saludables
- Disponibilidad de recursos (CPU,RAM, disco duro, etc) para las aplicaciones
- Simplificar los flujos de trabajo
- Evitar trabajos de emergencia

En momentos de mayor tráfico, podemos manualmente incrementar la cantidad de nodos

- El cobro es por el tiempo (segundos) en que los servidores están activos, no por mes entero
- Para festivales o eventos grandes, se puede escalar según sea necesario, y luego reducir.
- En esta infraestructura, podemos albergar
    - Productos en línea (en producción)
    - Versiones de desarrollo
    - Versiones de montaje
    - Servidor de mail

Infraestructura Actual:

A. **Situación actual/previa:** Servidores tradicionales

1. Servidor de producción - Servidor provisto por DGTIC
    - Sistema Operativo CENTOS
    - 4gb ram, 1cpu
2. Servidor de imágenes / medios - Servidor provisto por DGTIC
    - Sistema Operativo CENTOS
    - 4gb ram, 1cpu
        - (Limitado a un único subdominio. complicando su uso como servidor de desarrollo/dev)

subdominios.)

Servidor de Producción: provisto por DGTIC

- Sistema Operativo CENTOS
- 4gb ram, 1cpu

- Un certificado SSL para el dominio principal
- apache reverse proxy
    - este servidor usa un servicio poco mantenido y descontinuado: debe sustituirse por un proyecto con más soporte, p.ej., nginx-reverse-proxy
- letsencrypt
    - nos posibilita la creación de certificados automatizada para los subdominios *.centroculturadigital.mx adicionales al principal

- Todas las aplicaciones se corren en contenedores, utilizando Docker y docker-compose.
- 40+ apps antiguas (legado)
    - Algunas son páginas estáticas (HTML, js, css)
    - Algunas usan un stack
        - mongo
        - node.js
        - express
    - Servidor de medios (vision)

Todas las nuevas apps desarrolladas desde 2019 utilizan:

- nginx-proxy
- letsencrypt
- mongo
- node.js
- keystone.js

    - para construcción de back-ends
- sapper / svelte para el frontend
- cypress para creación de pruebas E2E

Estudio de costos de Kubernettes en Digital Ocean:

Costos:

En DigitalOcean (proveedor más barato)

- Costo base **$1900 MXN mensuales** ($95 USD mensuales)
- Fondo para eventualidades **$800 MXN mensuales** ($40 USD mensuales) (*permite subir la cantidad de nodos para momentos en los que se anticipa alto tráfico. Ver abajo*)
- El costo mínimo
    - para un nodo balanceador es de $200 MXN mensuales ($10 usd mensuales)
    - para un nodo “de trabajo” (worker) es de $400 MXN mensuales ($20 usd mensuales)
        - Capacidad de 1 nodo worker: 4GB ram, 2 cpus. 50GB de disco.
    - para un “Space”, unidad de espacio en CDN de 250gb es de $100 ($5 USD mensuales)
- La configuración mínima recomendada es:
    - 1 nodo balanceador: $200 MXN mensuales ($10 usd mensuales)
    - 4 nodos worker: $1600 MXN mensuales ($80 usd mensuales)
    - 1 “Space”, unidad de espacio en CDN ($5 usd mensuales)