Escalabilidad

Capacidad de un sistema p/ **adaptarse** a diferentes ambientes **modificando los recursos** del sistema.

Objetivo

Crecimiento respecto de:

- **Tamaño.** +Usuarios/recursos.
- **Distribución geográfica.** Dispersión.
- **Objetivos administrativos del sistema.**

Características de plataformas

- **Plataformas p/ alta concurrencia.**
  - Patrones conocidos.
  - Escalamiento automático.
  - Fuerte vínculo con infra/producto.
- **Arquitecturas ad-hoc y personalizadas.**
  - Necesidad de configuración y soporte.
  - Escalamiento manual / automatizado x humanos.
  - Posibilidad de migraciones.

Patrones de carga

- Predictable Burst.
- Unpredictable Burst.
- Periodic Processing (avg usage + períodos de inactividad).
- Start Small, Grow Fast.

Limitantes

- Arquitectura y algoritmos.
- Datos.
- Red (latencia, ancho de banda).
- Restricciones de negocio / locales.
- **Presupuesto.**

Técnicas

- Escalamiento **vertical**. +Recursos.
- Escalamiento **horizontal**.
  - Redundancia,
  - Balanceadores de carga,
  - Proximidad geográfica.
- **Fragmentación de datos.**
- **Optimizar algoritmos.** Performance, mensajería.
- **Asincronismo.** Limitado por negocio.
- Componentización -> separar servicios.

Elasticidad

Capacidad de un sistema p/ **modificar dinámicamente los recursos adaptándose a patrones** de carga.

Componentes

- **Application Load Balancer.** Ver a qué servicio/instancia le mandamos tráfico.
- **Autoscaler.**
  - **Scale In:** decrementar instancias.
  - **Scale Out:** incrementar instancias.
  - En función de **métricas**.
- **Monitoring Automático.** Métricas sobre CPU, memoria, I/O, networking, etc. p/ c/ servicio/instancia.
- Ejemplos:
  - **AWS:** Amazon ELB, Amazon Autoscaling, Amazon CloudWatch.
  - **K8s:** K8s Service, Horizontal Pod Autoscale, K8s Metrics Server.