Auto Scaling

Auto Scaling je metoda správy IT prostředků v rámci cloudu, která automaticky upravuje kapacitu běžících služeb (výpočetní výkon, paměť, počet instancí) na základě aktuální poptávky. Cílem je zajistit vysokou dostupnost aplikace při zachování minimálních nákladů.

Vertikální vs. Horizontální škálování

V encyklopedii je důležité rozlišovat dva základní přístupy k růstu výkonu:

1. Vertikální (Scaling Up)

Znamená navýšení výkonu stávajícího stroje (přidání CPU, RAM).

- **Omezení:** Každý fyzický server má svůj strop. Často vyžaduje restart systému.

2. Horizontální (Scaling Out)

Znamená přidání dalších identických strojů (instancí) do sítě.

- **Výhoda:** Téměř neomezený růst. Probíhá za plného chodu aplikace. Toto je standard pro moderní webové služby.

Jak Auto Scaling funguje?

Proces automatického škálování se obvykle opírá o tři základní komponenty:

1. Metriky (Monitoring)

Systém neustále sleduje vytížení zdrojů. Mezi nejčastější metriky patří:

- Procentuální vytížení **CPU**.
- Využití operační paměti (**RAM**).
- Počet síťových požadavků za sekundu (**Requests per Second**).

2. Pravidla a politiky (Policies)

Definují hranice (prahové hodnoty), kdy má systém zasáhnout:

- **Scale-out (Rozšíření):** „Pokud průměrné vytížení CPU překročí 70 % po dobu 5 minut, přidej 2 nové servery.“
- **Scale-in (Zmenšení):** „Pokud vytížení klesne pod 30 %, odeber 1 server (pro úsporu nákladů).“

3. Skupina automatického škálování (ASG)

Kolekce instancí, se kterými systém pracuje jako s jedním celkem. Správce nastavuje:

- **Minimum:** Nejmenší počet serverů, které musí běžet vždy.
- **Maximum:** Horní hranice, přes kterou systém nepůjde (kontrola nákladů).
- **Desired Capacity:** Ideální aktuální stav.

Výhody Auto Scalingu

- **Optimalizace nákladů:** Platíte pouze za výkon, který skutečně využíváte. V noci, kdy je provoz nízký, servery „vypnete“.
- **Spolehlivost:** Pokud jeden server selže (crash), Auto Scaling jej rozpozná jako nezdravý, ukončí ho a automaticky spustí nový.
- **Uživatelská zkušenost:** Aplikace se nezpomaluje ani v momentech nečekaných špiček (např. marketingová kampaň nebo Black Friday).

Typy škálovacích politik

- **Target Tracking:** Systém se snaží udržet metriku na konkrétní hodnotě (např. „udržuj CPU na 50 %“).
- **Step Scaling:** Reaguje skokově podle závažnosti (např. při 70 % přidej jeden stroj, při 90 % přidej tři).
- **Scheduled Scaling:** Plánované škálování podle času (např. „každé pondělí v 8:00 ráno zdvojnásob počet serverů“).
- **Predictive Scaling:** Využívá strojové učení k předpovědi zátěže na základě historických dat.

Propojení s Load Balancerem

Auto Scaling úzce spolupracuje s **Load Balancerem** (rozptylovačem zátěže). Když Auto Scaling přidá nový server, Load Balancer jej automaticky zaregistruje a začne na něj posílat část uživatelského provozu.

Související pojmy: Cloud Computing, Load Balancing, AWS, Azure, Microservices, High Availability.