EC2 FUNDEMENTALS

ON DEMAND INSTANCES
- → SHORT WORKLOAD, PREDICTABLE PRICING, PAY BY SECOND
- → HAS THE ↑ $$ BUT ☒ UPFRONT COST
- → NO LONG TERM COMMITMENT
- → GOOD WHEN YOU HAVE NO IDEA HOW THE APPLICATION WILL BEHAVE

RESERVED INSTANCE (1 OR 3 YEARS)
- → GOOD FOR DATABASE + LONG WORKLOADS
- → NEED TO RESERVE SPECIFIC INSTANCE ATTRIBUTES (EX: OS, TYPE)
- → PAYMENT OPT. (PARTIAL/ALL/NO UPFRONT)
- → RESERVED INSTANCE'S SPOT - REGIONAL/ZONAL
- → GOOD FOR STEADY STATE USAGE APPS LIKE DATABASE
- → BUY AND SELL IN THE RESERVED INSTANCE MARKETPLACE

CONVERTIBLE RESERVED INSTANCE
- → CAN CHANGE EC2 TYPE, FAMILY, OS, SCOPE, TENANCY
- → ↓ DISCOUNT

EC2 SAVING PLANS
- → GET A DISCOUNT BASED ON LONG TERM USAGE
- → COMMIT TO A CERTAIN TYPE OF USAGE ($10/hr for 1/3 years)
- → USAGE BEYOND EC2 SAVINGS PLAN IS BILLED AT ON-DEMAND PRICE
- → LOCKED TO A SPECIFIC INSTANCE FAMILY AND AWS REGION
- → HOWEVER FLEXIBLE OVER: INSTANCE SIZE, OS, TENANCY (HOST, DEFAULT)

SPOT INSTANCES
- → MOST COST EFFICENT INSTANCE IN AWS
- → USEFUL FOR WORKLOADS THAT ARE RESILIENT TO FAILURE:
  - → BATCH JOBS, DATA ANALYSIS, IMAGE PROCESSING
- → ☒ SUITABLE FOR CRITICAL JOBS OR DBs

DEDICATED HOSTS (TENANCY: PHYSICAL SERVER)
- → USEFUL FOR COMPLIANCE REQUIREMENTS AND/OR USING EXISTING SERVER BOUND SOFTWARE LICENSES (PRE SOCKET, PRE CORE, VM LICENSES)
- → MOST $$$ OPTION
- → TWO KINDS OF PURCHASING OPTIONS
  - → ON DEMAND
  - → RESERVED (PARTIAL/ALL/NO UPFRONT $$$)

DEDICATED INSTANCES (TENANCY: INSTANCE RUNNING ON HARDWARE)

→ NO CONTROL OVER INSTANCE PLACEMENT (AFTER START & STOP)
→ MAY SHARE HARDWARE W/ OTHER INSTANCES IN SAME ACCOUNT

EC2 CAPACITY RESERVATIONS

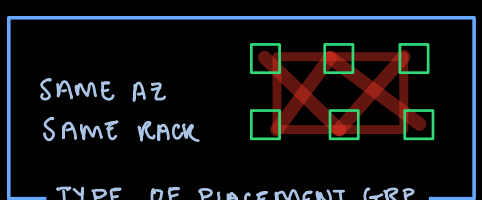
→ GOOD FOR SHORT-TERM, UNINTERRUPTED WORKLOADS THAT NEEDS TO BE IN A SPECIFIC AZ.
→ YOU'RE CHARGED ON-DEMAND RATE WHETHER YOU RUN INSTANCES OR NOT.
→ TYPICALLY COMBINED WITH REGIONAL RESERVED INSTANCES AND SAVINGS PLANS TO BENEFIT FROM BILLING DISCOUNTS

WHAT ARE SPOT FLEETS?

→ SET OF SPOT INSTANCES + OPTIONAL ON-DEMAND INSTANCES
→ SPOT FLEET WILL TRY TO MEET TARGET CAPACITY W/ PRICE CONSTRAINTS
  → DEFINE POSSIBLE LAUNCH POOLS INSTANCE TYPE, OS, AZ
  → CAN HAVE MULTIPLE LAUNCH POOLS SO THAT THE FLEET CAN CHOOSE
→ STRATEGIES TO ALLOCATE SPOT INSTANCES
  → LOWEST PRICE, DIVERSIFIED, CAPACITY OPTIMIZED, PRICE CAPACITY OPTIMIZED

→ SPOT FLEETS ALLOWS US TO AUTOMATICALLY REQUEST SPOT INSTANCES WITH THE LOWEST PRICE

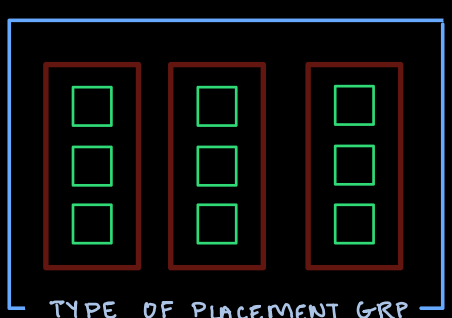
WHAT ARE CLUSTERS?

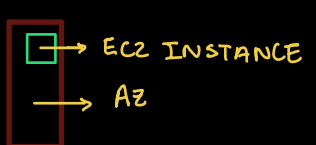


□ → EC2 INSTANCE

PROS: GREAT NETWORK
CONS: NOT FAULT TOLERANT
USE CASE:
→ BIG DATA JOB THAT NEEDS TO BE DONE FAST
→ APP THAT NEEDS EXTREMELY ↓ LATENCY AND ↑↑ NETWORK THROUGHPUT

WHAT ARE SPREADS?

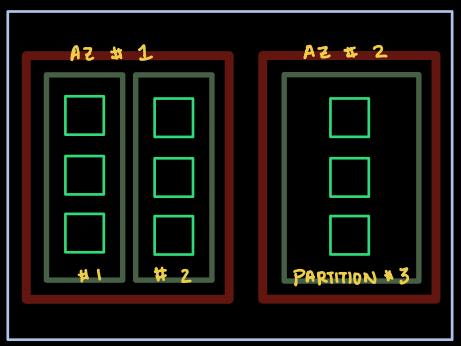




PROS: FAULT TOLERANT
  • EC2 INSTANCES ON DIFFERENT HARDWARE
CONS: LIMITED TO 7 INSTANCES/AZ/PLACEMENT
  → LIMITS ON HOW BIG YOUR PLACEMENT GROUP CAN BE.

USE CASE:
→ APP THAT NEEDS TO MAXIMIZE HIGH AVAIL.
→ CRITICAL APPS WHERE EACH INSTANCE MUST BE ISOLATED FROM FAILURE FROM EACH

WHAT ARE PARTITIONS?



→ THINK OF PARTITIONS LIKE PHYSICAL RACKS
→ TO PREVENT RACK FAILURE
→ UP TO 7 PARTITIONS/AZ & UP TO 100 EC2s.
→ EC2 INSTANCES GET ACCESS TO THE PARTITION INFORMATION AS METADATA

USE CASES:
→ APP THAT IS PARTITION AWARE TO DISTRIBUTE THE DATA AND SERVERS ACROSS PARTITIONS.
Ex: HDFS, HBASE, CASSANDRA, KAFKA

WHAT IS A VIRTUAL PRIVATE CLOUD?

→ VPCs BELONG TO A REGION
↳ LOGICALLY ISOLATED PORTION OF THE AWS CLOUD THAT ONE CAN USE TO DEPLOY RESOURCES LIKE EC2.
↳ SERVICES LIKE S3 RESIDE OUT OF VPC.

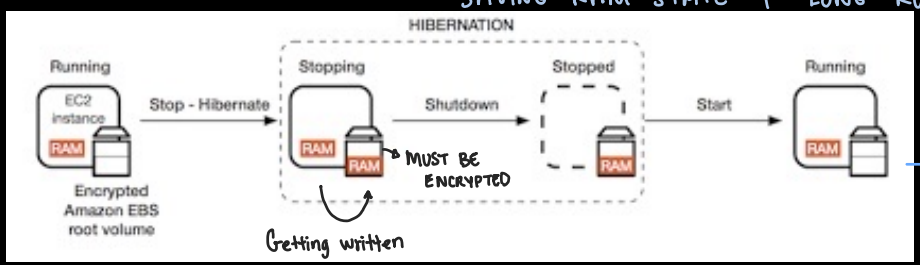
WHAT IS A SUBNET?

→ A SUBNET IS A RANGE OF IP ADDRESSES IN YOUR VPC
→ A SUBNET MUST RESIDE IN A SINGLE AZ
↳ HOWEVER, YOU CAN DEPLOY MULTIPLE SUBNETS PER AZ.

WHAT ARE ELASTIC NETWORK INTERFACES?

→ LOGICAL COMPONENT IN A VPC THAT REPRESENTS A VIRTUAL NETWORK CARD.
→ YOU CAN CREATE ENI INDEPENDENTLY AND ATTACH THEM ON THE FLY (MOVE THEM) ON EC2 INSTANCES FOR FAILOVER.
→ BOUND TO A SPECIFIC AZ LIKE SUBNETS.
→ ENIs ALLOW US TO DO A VERY QUICK AND EASY NETWORK FAILOVER BETWEEN TWO OR MORE EC2 INSTANCES BY MOVING ENIs.

EC2 HIBERNATING STATE: USEFUL FOR SERVICES THAT TAKE TIME TO INITIALIZE, SAVING RAM STATE & LONG RUNNING PROCESSING



→ BOOT UP FAST
→ MAX RAM: 150 GB
→ NO MORE THAN 60 DAYS.

WHAT ARE EBS VOLUME ?

→ ELASTIC BLOCK STORE (EBS) VOLUME IS A NETWORK DRIVE THAT YOU ATTACH TO A EC2 INSTANCE.

→ ALLOWS EC2s TO PERSIST DATA EVEN AFTER TERMINATION

→ ALSO BOUND TO AZ LIKE ENIs AND SUBNETs