









# EC2, Launch Types, Spot Instances e Placement Groups



# EC2: Visão Geral

- EC2 permite executar máquinas virtuais na AWS com diferentes configurações
- As instâncias variam por tipo (t2, m5, c6g...) e são otimizadas para diferentes casos
- É possível escolher tipos baseados em CPU, memória, armazenamento ou rede

## EC2: Prefixos dos Tipos de Instância

PREFIXO	OTIMIZAÇÃO	DESCRIÇÃO RESUMIDA	PALAVRA-CHAVE
<b>t</b>	Uso geral (burst)	Instância econômica com créditos de CPU	 Econômico
<b>m</b>	Uso geral balanceado	Bom equilíbrio entre CPU, memória e rede	 Balanceado
<b>c</b>	Computação intensiva	Alta performance de CPU	 Processamento
<b>r</b>	Memória intensiva	Ideal para bancos em memória	 RAM
<b>i</b>	Armazenamento local	Alta IOPS com SSD local (NVMe)	 Storage
<b>g / p</b>	GPU	Machine Learning, render, jogos	 GPU
<b>inf / trn</b>	IA (Inferência/Treinamento)	Instâncias otimizadas para ML	 AI
<b>x / z</b>	Altíssima memória	Banco em memória denso (ex: SAP)	 SAP/HANA

## Comparativo: Tipos de Instância EC2

TIPO	OTIMIZADO PARA	CASOS DE USO	RELEVÂNCIA NA PROVA
t3 / t4g	Uso geral com burst	Dev/teste, workloads leves e intermitentes	<ul style="list-style-type: none"><li>• Possuem crédito de CPU (pegadinha comum)</li><li>• Modelo econômico para workloads variáveis</li></ul>
m5 / m6g	Uso geral balanceado	Apps corporativas, microsserviços	<ul style="list-style-type: none"><li>• Equilíbrio entre CPU, memória e rede</li><li>• Ideal quando não há foco específico em I/O ou computação</li></ul>
c5 / c6g	Computação intensiva	CI/CD, HPC, processamento paralelo	<ul style="list-style-type: none"><li>• Alta performance por núcleo</li><li>• ⚠ Prova pode confundir com tipos de GPU</li></ul>
r5 / r6g	Memória intensiva	Banco de dados em memória, cache, Redis	<ul style="list-style-type: none"><li>• Mais RAM por vCPU</li><li>• Confunde com "storage optimized" na prova</li></ul>
i3 / i4i	Armazenamento em disco NVMe	NoSQL, data warehousing local, OLAP	<ul style="list-style-type: none"><li>• ⚠ Prova pode confundir com EBS otimizados</li><li>• Alta IOPS com armazenamento local SSD</li></ul>
g4 / p3	GPU (machine learning / inferência)	IA, jogos, renderização, computação gráfica	<ul style="list-style-type: none"><li>• ⚠ Saber quando usar GPU vs computação intensiva (CPU)</li><li>• G4 = inferência   P3 = treinamento</li></ul>
inf1 / trn1	Inferência e treinamento de IA	Modelos deep learning otimizados	<ul style="list-style-type: none"><li>• ⚠ Diferenciar Inf1 (inferência) de Trn1 (treinamento)</li><li>• Exclusivo para cargas específicas de IA</li></ul>



# EC2 Launch Types (Formas de Execução)

- On-Demand: paga por hora ou segundo; ideal para uso flexível ou imprevisível
- Reserved Instances: compromisso de 1 ou 3 anos com desconto; ideal para workloads constantes
- Spot Instances: aproveita capacidade ociosa da AWS com até 90% de desconto
- Savings Plans: opção de desconto com mais flexibilidade que Reserved Instances

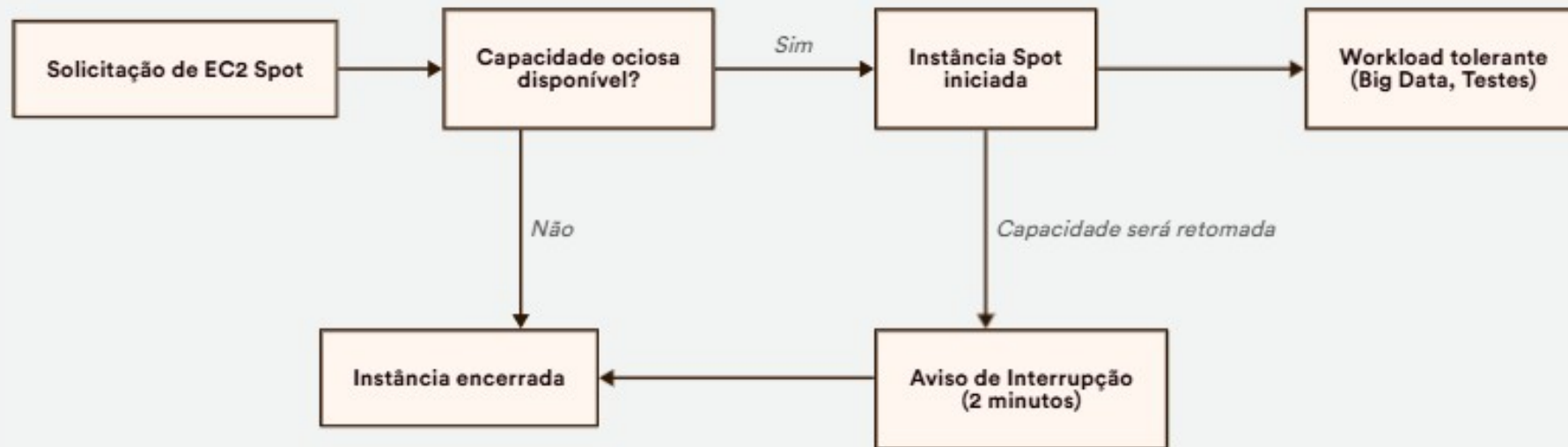
## EC2 Launch Types – Formas de Execução

TIPO	DESCRIÇÃO	VANTAGENS	QUANDO USAR	RELEVÂNCIA NA PROVA
<b>On-Demand</b>	Pagamento por hora ou segundo	Flexibilidade total, sem compromisso	Ambientes imprevisíveis, testes, uso esporádico	<ul style="list-style-type: none"><li>⚠️ Mais caro a longo prazo</li><li>✓ Simples e flexível</li></ul>
<b>Reserved Instances</b>	Compromisso de 1 ou 3 anos com instância específica	Desconto até 75% comparado ao On-Demand	Workloads constantes e previsíveis	<ul style="list-style-type: none"><li>⚠️ Menos flexível (instância e região fixas)</li><li>✓ Ótimo custo-benefício para uso contínuo</li></ul>
<b>Spot Instances</b>	Capacidade ociosa com até 90% de desconto	Custo extremamente baixo	Jobs tolerantes a interrupção, processamento em lote	<ul style="list-style-type: none"><li>⚠️ Pode ser interrompida a qualquer momento</li><li>✓ Excelente para IA, render, big data</li></ul>
<b>Savings Plans</b>	Compromisso por tempo com mais flexibilidade	Desconto similar a RIs, mas com liberdade de família/instância	Quando se quer desconto sem amarrar instância exata	<ul style="list-style-type: none"><li>✓ Flexível entre instâncias, zonas e serviços</li><li>⚠️ Exige compromisso de uso (ex: 1 ano)</li></ul>



# EC2 Spot Instances

- Utilizam capacidade ociosa da AWS com grandes descontos
- Podem ser encerradas pela AWS com aviso prévio de 2 minutos
- Boas para workloads tolerantes a interrupção, como big data e testes





# Placement Groups

- Cluster: instâncias próximas fisicamente para baixa latência e alto throughput
- Spread: distribui instâncias em AZs para maior tolerância a falhas
- Partition: divide instâncias em grupos lógicos com isolamento de falha

TIPO	DESCRIÇÃO	VANTAGEM	CASOS DE USO	ALERTA NA PROVA
<b>Cluster</b>	Instâncias fisicamente próximas	Baixa latência e alto throughput de rede	HPC, Big Data, jobs distribuídos de alta performance	<ul style="list-style-type: none"><li>• ⚠ Pode ter falha em cascata</li><li>• ✔ Performance máxima</li></ul>
<b>Spread</b>	Distribui instâncias entre diferentes racks ou AZs	Alta tolerância a falhas	Ambientes críticos com poucas instâncias (ex: bancos)	<ul style="list-style-type: none"><li>• ⚠ Limite de 7 instâncias por AZ</li><li>• ✔ Máximo isolamento</li></ul>
<b>Partition</b>	Divide instâncias em grupos lógicos (partições)	Isolamento de falhas por partição	Grandes clusters, Hadoop, Cassandra, HDFS	<ul style="list-style-type: none"><li>• ✔ Ideal para 10+ instâncias</li><li>• ⚠ A prova pode confundir com Spread</li></ul>



## Estados da Instância EC2 – Comportamento e Custos

ESTADO	A INSTÂNCIA EXISTE?	GERA COBRANÇA?	PODE SER REINICIADA?	IP PÚBLICO MANTIDO?	VOLUME EBS MANTIDO?	OBSERVAÇÕES PARA A PROVA
running	✓ Sim	💰 EC2 + EBS	✓ Sim	✓ Sim	✓ Sim	Estado ativo com cobrança total
stopped	✓ Sim	💰 Apenas EBS	✓ Sim	✗ Não	✓ Sim	IP público é liberado; ideal para pausas planejadas
hibernated	✓ Sim	💰 EBS + snapshot da RAM	✓ Sim (mantém RAM)	✓ Sim	✓ Sim	Requer volume root criptografado e tipo compatível
terminated	✗ Não	✗ Não	✗ Não	✗ Não	✗ Não (a menos que configurado para manter)	⚠ Estado final, irreversível; apaga IP, instância e EBS (se não marcado para preservar)

# Boas Práticas e Dicas de Prova

- On-Demand é a escolha padrão; Spot é ideal para custo e interrupção tolerável
- Reserved é ótimo para workloads previsíveis; Savings Plans oferecem flexibilidade
- Cluster Placement maximiza performance de rede; Spread foca em resiliência
- Spot pode ser encerrada sem aviso — use com auto scaling ou tarefas tolerantes
- A prova costuma testar cenários de escolha ideal entre os tipos