



Auto Scaling

Center of Electrical Engineering and Informatics
Federal University of Campina Grande

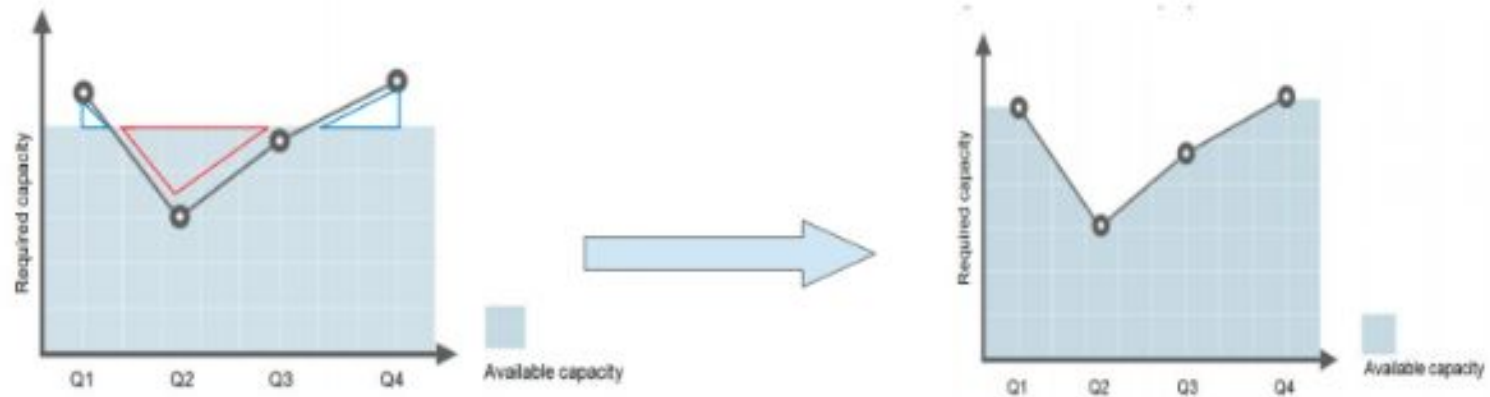


- Visão Geral
- Criação
- Gerenciamento

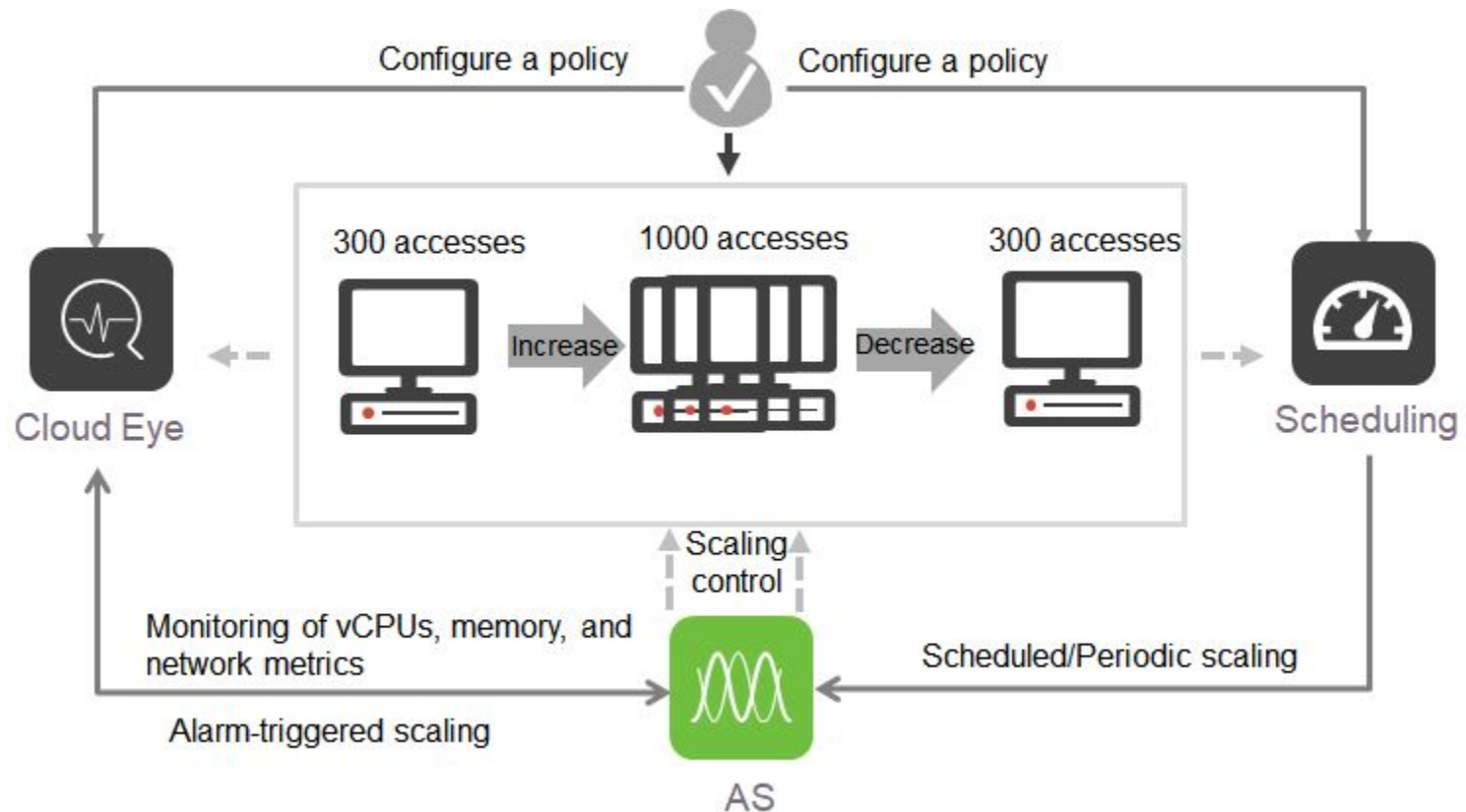
“Ajuste automático de recursos computacionais baseados em requisitos de serviços e políticas pré-configuradas para assegurar estabilidade e custos otimizados”

- Ajuste dinâmico baseados em agendamentos, períodos ou alarmes
- Checagem automática de status nos grupos de AS com substituição imediata de instâncias defeituosas
- Gerenciamento através de gráficos
- O serviço de AS não é cobrado, apenas as instâncias do grupo são pagas

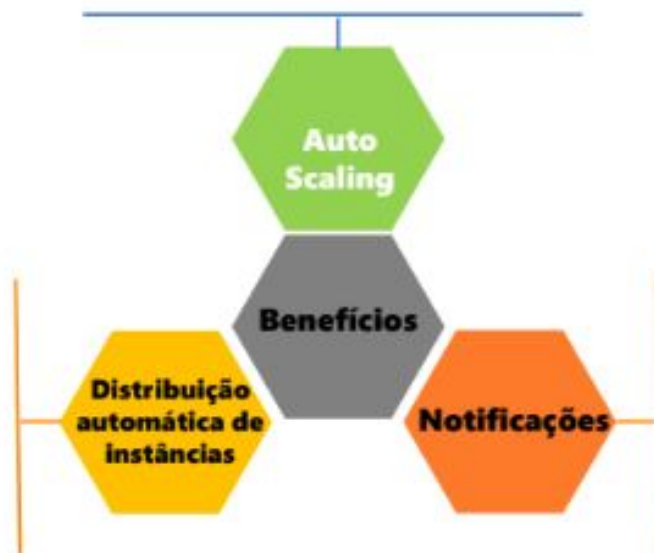
- Melhor utilização da capacidade disponível



Visão Geral - Arquitetura

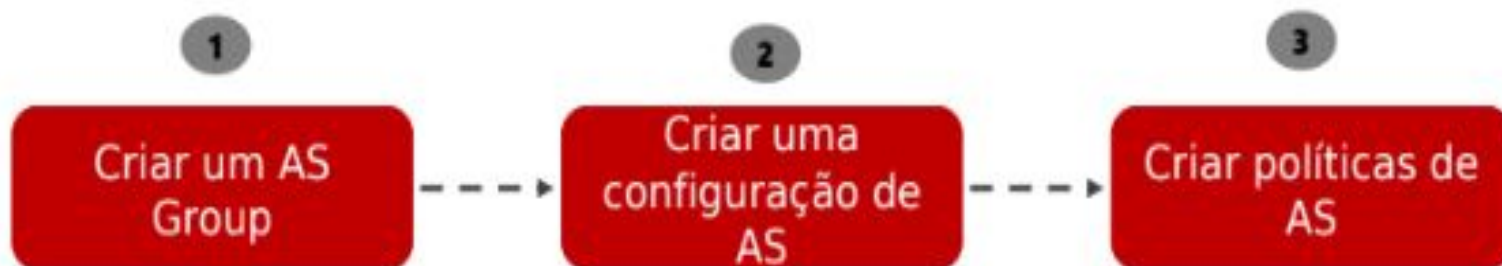


Funcionalidades



- Aplicações/Serviços WEB
 - O **AS** escala servidores lógicos de serviços web comuns como sites empresariais, educação, mídia...
 - Requisições dos clientes são distribuídas entre instâncias diferentes através do balanceamento de carga
 - Serviço escala para cima ou para baixo de acordo com o número de requisições

- Deploy de clusters de alta performance
 - O serviço escala aumentando ou diminuindo recursos para servidores de backend de aplicações baseadas em tempo real com grande volumes de dado
 - Os servidores incluem nós de computação de análises de Big Data e serviços para mineração de dados em clusters



Configuração de AS - Especificação

1 Especificar detalhes

2 Configuração AS

3 Política AS

4 (Optional) Configure Notificação

5 Confirmar especificações

6 Concluir

* Nome do grupo AS

as-group-gbj4

* Máx. Instâncias

1

* Instâncias esperadas

0

* Min. Instâncias

0

Duração de refrigeração (s) ?

900

* AZ ?

* VPC ?

vpc-2326(192.168.0.0/16)

Exibir VPC

* Sub-rede

* Grupo de segurança ?

default (Inbound: - | Outbound: -)

Gerenciar grupo de segurança Saiba mais sobre como configurar um grupo de segurança

Entrada: - | Saída: -

* Balanceamento de carga

☒ Não utilizar ☐ Usar ELB

* Método de verificação de integridade ?

Verificação de integridade de ECS

* Intervalo de verificação de integridade

5 minutos

* Política de remoção de instâncias

Instância mais antiga criada com base ...

* Liberar EIP em Remoção da instância

☒ ?

- Grupos de AS
 - Consiste em uma coleção de instâncias ECS e Políticas de AS que possuem atributos similares e se aplicam ao mesmo cenário de aplicação
 - Habilitar/Desabilitar políticas de AS e realizar ações de escalabilidade
 - Criar um AS Group
 - Adicionar/Substituir configurações
 - Adicionar balanceamento de carga
 - Habilitar/Desabilitar

- Configurações AS
 - Consiste em um template que lista especificações para instâncias que serão adicionadas ao grupo
 - Operações
 - Criar uma nova configuração
 - Utilizar uma instância ECS existente para configurar AS
 - Copiar configuração de outro AS
 - Deletar uma determinada configuração

- Criar uma nova configuração AS

Criar configuração AS ②

1 Especificar detalhes

2 Confirmar especificações

3 Concluir

Informações básicas

* Nome da configuração

* Modelo de configuração

Criar um novo modelo de especificações

Utilizar as especificações de um ECS existente

Especificações

* Tipo de ECS

General-purpose

Computação I

Computação II

Memória otimizada

Primeira geração

* vCPU

1 vCPUs

2 vCPUs

4 vCPUs

8 vCPUs

16 vCPUs

32 vCPUs

* Memória

4 GB

Especificaciones Seleccionadas: s1.medium | 1 vCPUs | 4 GB

Imagem

- Criar uma nova configuração AS

Imagem

Imagem pública Imagem privada Imagem compartilhada

* Imagem

mongo-img(40GB)



Disco

Disco do sistema

I/O Comum



40



GB Não criptografado

[+ Adicionar um disco de dados](#) Você pode adicionar mais 23 discos.

EIP

EIP 


Não obrigatório

Atribuir automaticamente

Um ECS sem um EIP não pode acessar a Internet. No entanto, ele ainda pode ser usado como um serviço do ECS implantado em um cluster ou em uma rede privada.

- Criar uma nova configuração AS

Logon

* Par de chaves KeyPair-a5ad [Exibir par de chaves](#) 

☐ Confirmo que tenho o arquivo de chave privada KeyPair-a5ad.pem e que, sem esse arquivo, não conseguirei fazer login em meu ECS.

Logon e introdução de arquivos

Logon e introdução de arquivos

Não configurar Configurar agora

Criar agora

- Alterando as configurações
 - Clique no botão modificar nas informações do grupos de AS

Alterar Configuração do AS do Grupo de AS

★ Nome do grupo AS

as-group-gbj4

★ Máx. Instâncias

1

★ Instâncias esperadas

0

★ Mín. Instâncias

0

Duração de refrigeração (s)

900

★ Método de verificação de integridade

Verificação de integridade de ...

★ Intervalo de verificação de integridade

5 minutos

★ Política de remoção de instâncias

Instância mais antiga criada c...

Modo de notificação

☐ Via e-mail ?

OK


Cancelar

- Movendo instância para um grupo de AS
 - Clique no botão “Adicionar” na aba de instância da página de grupos AS


Adicionar instância

Você pode adicionar até 10 instâncias em um lote.
Você pode selecionar apenas instâncias pertencentes ao mesmo VPC que o grupo de AS, cujo AZ está incluído nos AZs do grupo de AS e não foi adicionado a outro grupo de AS.

Instâncias disponíveis

<input type="checkbox"/>	Nome	ID	AZ
 Não há dados disponíveis			

Instâncias selecionadas

Nome	ID	AZ	Operaç...
 Nenhuma instância foi selecionada.			

- Uma ação de dimensionamento é necessária para expandir recursos quando a demanda de serviço aumenta. Existem as seguintes maneiras de expansão de recursos:
 - Expansão Dinâmica
 - Expansão Planejada
 - Expansão Manual

Ações de Dimensionamento



- Visualização
 - Na página de detalhes do grupo AS, clique na aba de monitoramento
 - Depois, clique no diagrama ou tabela para visualizar o registro de ações de dimensionamento

Grupo AS > as-group-gbj4

Informações básicas

Modificar

Grupo AS(ID): 4cd991bc-6f77-474e-8141-8bd7e265243a

AZ: AZ1

Estado: Habilitado [Desabilitar](#)

VPC: [vpc-2326](#)

Nome da configuração: [as-config-jhht](#) [Modificar](#)

Sub-rede: subnet-2326

ID da configuração: 30784508-0891-4415-86af-94e1d1bc3a68

Grupo de segurança: default (Inbound:TCP/1337, 22 | Outbound: -)

Instâncias esperadas: 0

Ouvinte: --

Duração de refrigeração (s): 900

Método de verificação de integridade: Verificação de integridade de ECS

Política de remoção de instâncias: Instância mais antiga criada com base na configuração do AS mais antiga

Intervalo de verificação de integridade: 5 minutos

Modo de notificação: --

Criado: 16/09/2018 14:58:47 GMT-03:00

Liberar EIP em Remoção da instância: Sim

Número de instâncias:
Atual: 0
0
1
Mín. Máx.

- Gancho do ciclo de vida
 - Quando um gancho de ciclo de vida é adicionado a um grupo AS e ocorre uma ação de dimensionamento, a instância que está sendo adicionada/removida é suspensa para realizar as configurações customizadas
 - A instância entrará em um período de espera
 - É possível realizar atividades como por exemplo instalar ou configurar softwares na instância recém-criada

- Gerenciamento de políticas de AS
 - Uma política de AS especifica a condição para ativar uma ação de dimensionamento

Política de AS

Adicionar política ×

*

 Nome da política

as-policy-v9sp|

*

 Tipo de política

Alarme

Agendado

Periódico

*

 Alarme

Criar regra de alarme ▾

*

 Nome do alarme

as-alarm-oprw

*

 Condição de disparador

Uso de CPU ▾

Máx. ▾

> ▾

%

Para verificar se o Uso de memória, a Taxa de saída do grupo ou a Taxa de entrada do grupo são compatíveis com diferentes sistemas operacionais, consulte o [Guia do usuário do Elastic Cloud Server](#).

*

 Intervalo de monitoramento

5 minutos ▾

*

 Ocorrências repetidas

?

Ação de regulação da capacidade

Adicio... ▾

1

instâncias ▾

Duração de refrigeração (s) ?

900

OK

Cancelar

- Gerenciamento de políticas de AS
- Alarme
 - Automaticamente redimensiona as instâncias de um grupo de acordo com uma determinada métrica
 - Exemplo:
 - **“Ao atingir 80% de cpu, incluir mais uma instância”**

- Gerenciamento de políticas de AS - Alarme

Adicionar política

* Nome da política

as-policy-j54t

* Tipo de política

Alarme

Agendado

Periódico

* Alarme

Criar regra de alarme

* Nome do alarme

as-alarm-vmjv

* Condição de disparador

Uso de CPU

Máx.

>

80

%

Para verificar se o Uso de memória, a Taxa de saída do grupo ou a Taxa de entrada do grupo são compatíveis com diferentes sistemas operacionais, consulte o [Guia do usuário do Elastic Cloud Server](#).

* Intervalo de monitoramento

5 minutos

* Ocorrências repetidas

2

?

Ação de regulação da capacidade

Adicio...

1

instâncias

Duração de refrigeração (s) ?

900

OK

Cancelar

- Gerenciamento de políticas de AS -
- Agendado
 - Automaticamente redimensiona as instâncias de um grupo de acordo com uma data especificada
 - Exemplo:
 - **“No dia 22/12/2018 aumente o número de instâncias”**

Ações de Dimensionamento


- Gerenciamento de políticas de AS
- Agendado

Ouvinte: ×

Adicionar política

★ Nome da política


★ Tipo de política Alarme **Agendado** Periódico

★ Disparado às × 

A data especificada deve ser posterior à hora padrão e ao tempo de inicialização do sistema.

Adicionar

Ação de regulação da capacidade Adicio... instâncias

Duração de refrigeração (s) 

OK Cancelar

- Gerenciamento de políticas de AS
- Periódico
 - Automaticamente redimensiona as instâncias de um grupo de acordo com um intervalo especificado
 - Exemplo:
 - **“A cada dois dias aumente o número de instâncias”**


- Gerenciamento de políticas de AS - Periódico

Ouvinte: ×

Adicionar política ×

* Nome da política

* Tipo de política

* Selecionar horário × 

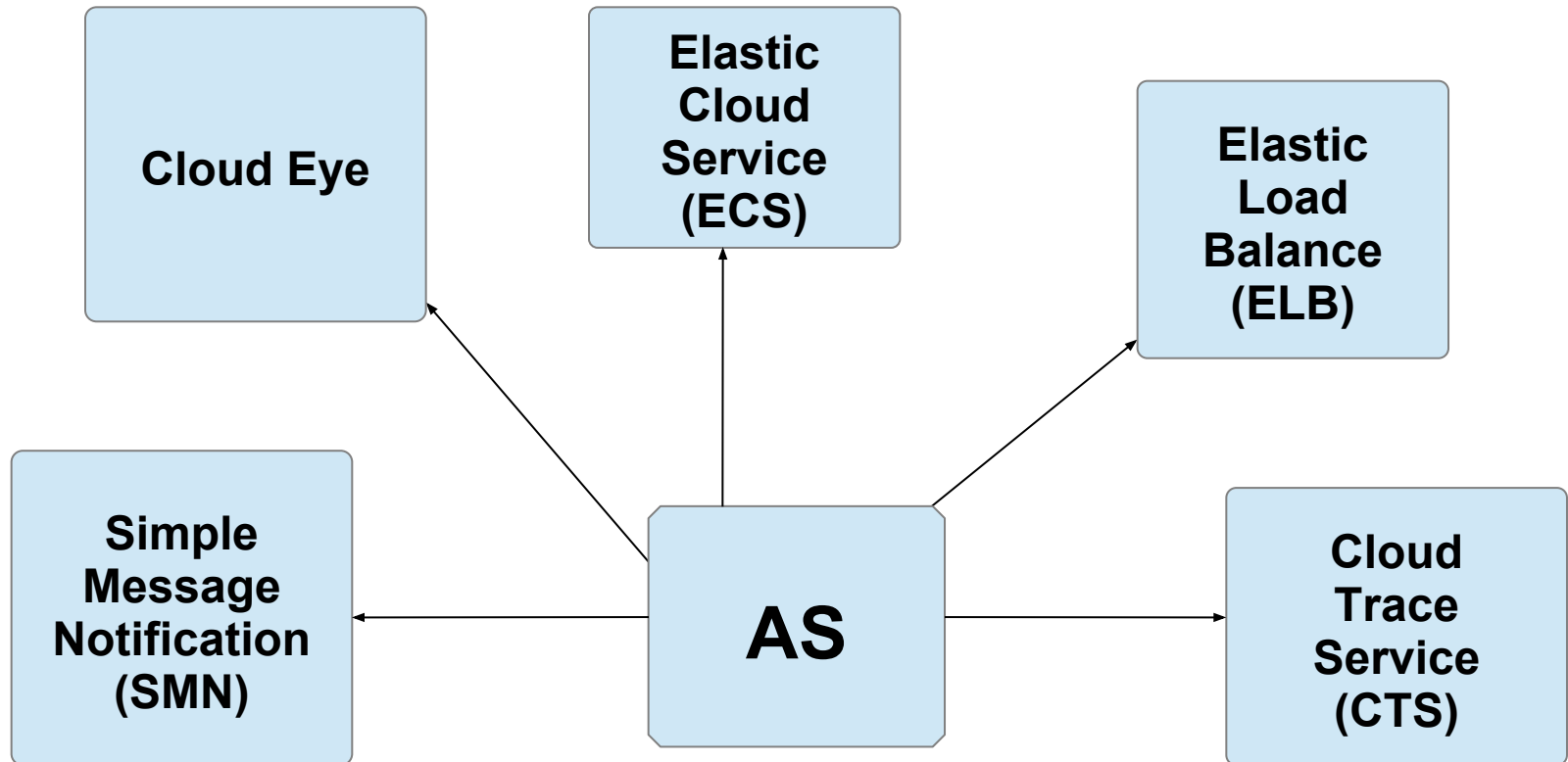
* Intervalo ▼

* Ativado às

Ação de regulação da capacidade ▼ ▼

Duração de refrigeração (s) 

- Para executar em um grupo de AS, as aplicações precisam suportar escalonamento horizontal
- Limites:
 - Quantidade de grupos AS: 10
 - Quantidade de configurações de AS: 100
 - Quantidade de políticas de AS: 10
 - Quantidade de instâncias no grupo: 300



- Acessar o projeto my-project na instância ECS
- Instalar o stress-ng
 - `sudo apt install stress-ng`
- Instalar o PM2
 - `sudo npm install pm2 -g`

- Alterar o projeto para simular carga de CPU no ECS
 - Criar o controlador `api/controllers/EcsController.js`

```
const { spawn } = require('child_process');
var LOAD = 0;
var process;
function updateLoad(){
  if(LOAD > 100 || LOAD < 0) LOAD = 20;
  if(process){
    process.stdin.pause();
    process.kill();
  }
  process = spawn('stress-ng', ['-c', '0', '-l', LOAD.toString()], {detached: true});
}
module.exports = {
  increase: function(req, res){
    LOAD+=20;
    updateLoad();
    res.status(200).send();
  },
  decrease: function(){
    LOAD-=20;
    updateLoad();
    res.status(200).send();
  }
}
```

- Alterar o projeto para simular carga de CPU no ECS
 - Permitir acesso a API REST sem autenticação modificando o arquivo `config/policies.js`

```
'*': true
```

- Adicionar as rotas da API REST no arquivo `config/routes.js`

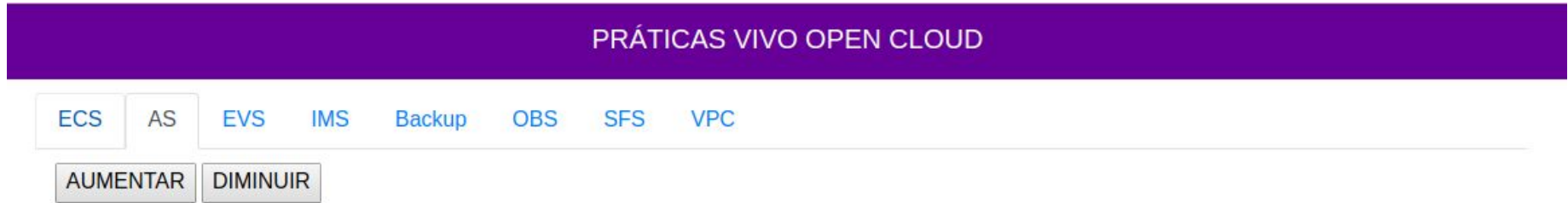
```
'GET /api/ecs/increase': { controller: 'ecs', action: 'increase' }  
'GET /api/ecs/decrease': { controller: 'ecs', action: 'decrease' }
```

- Incluir os botões para manipular a carga de CPU na página `views/pages/homepage.ejs`

```
<button onclick="$.ajax({url: '/api/ecs/increase'})">AUMENTAR</button>  
<button onclick="$.ajax({url: '/api/ecs/decrease'})">DIMINUIR</button>
```

- Configurar o PM2 para inicializar os projetos cadastrados no boot do ECS
 - pm2 startup
 - Copiar a saída fornecida pelo comando e executar
- Iniciar e salvar o projeto my-project no PM2
 - cd /home/linux/workspace/my-project
 - pm2 start app.js
 - pm2 save
- Reiniciar o ECS
 - sudo reboot

- Verificar se a aba ECS apresenta a UI para manipular a carga de CPU



- Criar uma imagem com estado atual do ECS
- Criar uma configuração de AS utilizando essa imagem
 - Definir no mínimo 1 instância e no máximo 2
 - Definir o alarme para aumentar 1 instância caso a CPU ultrapasse 60%
 - Definir o alarme para diminuir 1 instância caso a CPU fique abaixo 40%
- Acessar a aba ECS e simular carga de 80% CPU
- Verificar no console de gerenciamento se a nova instância foi adicionada no grupo



Contact

Angelo Perkusich, D.Sc.

Professor, CEO

angelo.perkusich@embedded.ufcg.edu.br

+55 83 8811.9545

Hyggo Almeida, D.Sc.

Professor, CTO

hyggo.almeida@embedded.ufcg.edu.br

+55 83 8875.1894

Rohit Gheyi

Professor, Program Manager

rohit.gheyi@embedded.ufcg.edu.br

+55 83 8811 3339

