



# **Hot-Plug Heaven: Dynamic AI Workloads in a Mini IaaS Cloud**

# The Problem: Inefficient Resource Utilization

*Requested != Used*



## Overestimates

Requesting 8 or 16 cores but actively using only a few



## Waiting Time

Increased wait times for other users



## Underutilization

Reduced overall efficiency

# Why Mini-Cloud (IaaS)?

*Mini-cloud brings cloud elasticity to on-premise HPC environments*

01

## Elastic Usage

Mimics cloud's use-as-needed principle, so CPUs are only active when required.

03

## Local Simulation

Built using VirtualBox to simulate a private IaaS setup.

## Real-Time Scaling

Dynamically adds or removes CPUs based on workload intensity.

02

## AI-Friendly Design

Perfect for variable AI workloads needing on-demand compute power.



# Approach's Components

1

## Usage Monitoring

Track live CPU utilization using `psutil`(Python Lib.)

2

## Dynamic Control

Hot-plug or unplug CPUs automatically via `VBoxManage`

3

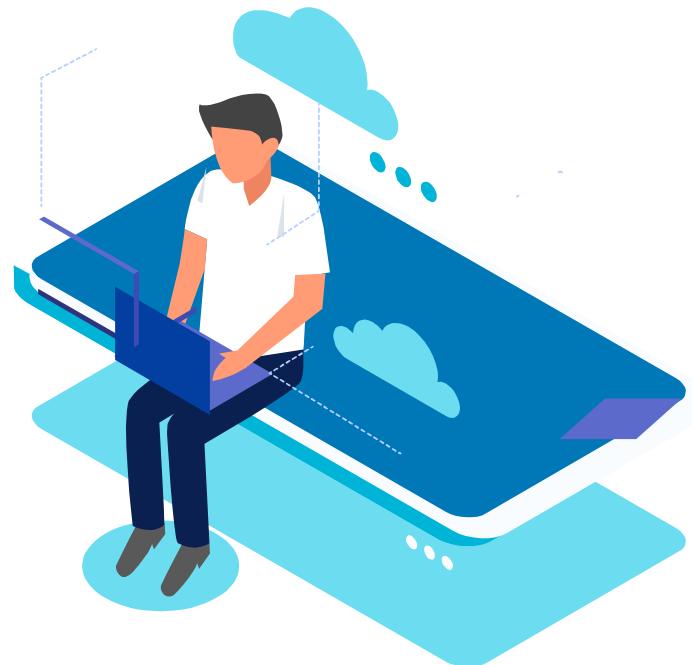
## Smart Decision Logic

Scale up or down using threshold-based rules.

4

## Performance Tracking

Log every scaling event for later optimization.



# Progress So Far



## VM Setup Complete

Ubuntu VMs with Guest Additions successfully configured



## Hot-Plug Verified

CPUs can be dynamically added and removed without reboot



## Live Monitoring Working

Real-time CPU usage collection running smoothly



## Automation Achieved

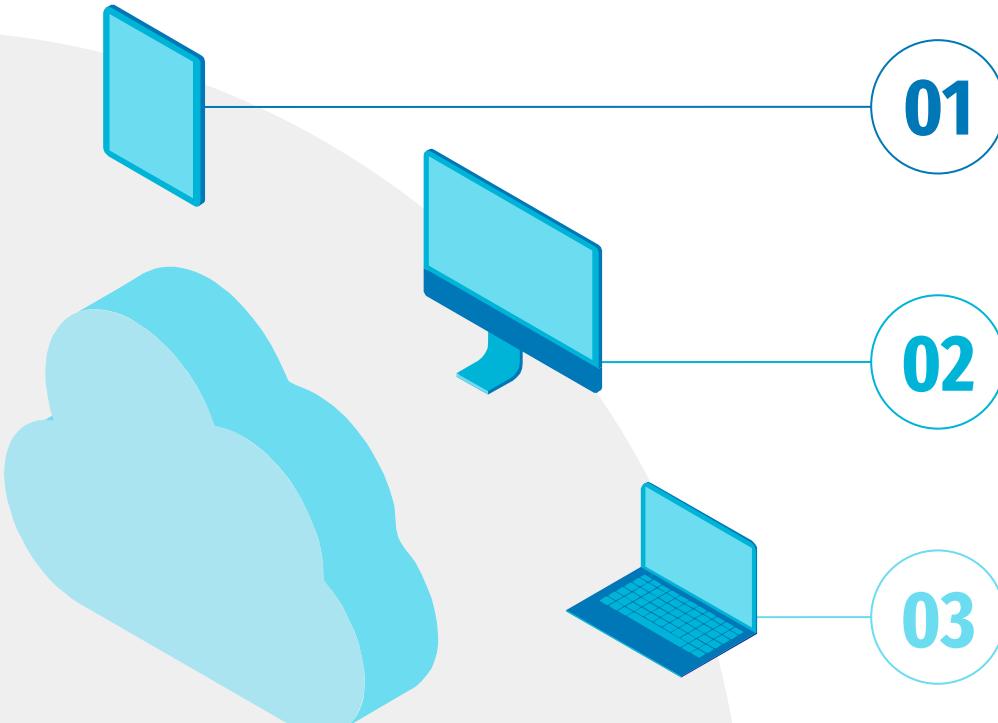
Scaling script executes independently and consistently





**What is happening?  
Let's Visualise it.....**

# Expected Outcome



## Better Efficiency

Optimized CPU usage across all running jobs

## Adaptive Support

Automatically adjusts resources per task demand

## Cloud Behavior

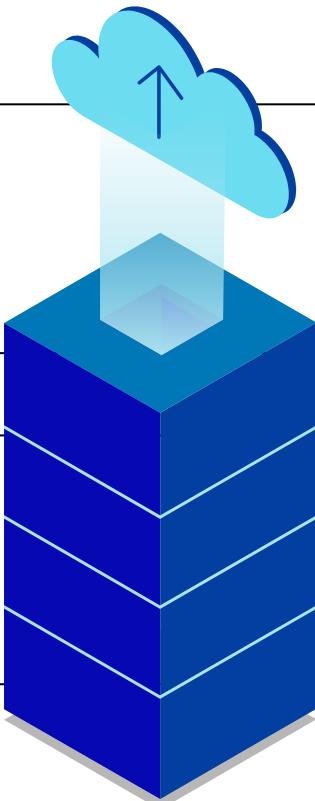
Achieve elasticity within local infrastructure

# References

Xiao, Z., Song, W., & Chen, Q. (2012). Dynamic resource allocation using virtual machines for cloud computing environment. *IEEE transactions on parallel and distributed systems*, 24(6), 1107-1117.

Liu, H., Jin, H., Liao, X., Deng, W., He, B., & Xu, C. Z. (2014). Hotplug or ballooning: A comparative study on dynamic memory management techniques for virtual machines. *IEEE Transactions on parallel and distributed systems*, 26(5), 1350-1363.

Qiu, H., Mao, W., Wang, C., Franke, H., Youssef, A., Kalbarczyk, Z. T., ... & Iyer, R. K. (2023). {AWARE}: Automate workload autoscaling with reinforcement learning in production cloud systems. In 2023 USENIX Annual Technical Conference (USENIX ATC 23) (pp. 387-402).



**Presented By:** Urjit Mehta | AU2444007

**Email:** [urjit.m@ahduni.edu.in](mailto:urjit.m@ahduni.edu.in)

**LinkedIn:** [urjitmehta](https://www.linkedin.com/in/urjitmehta/)