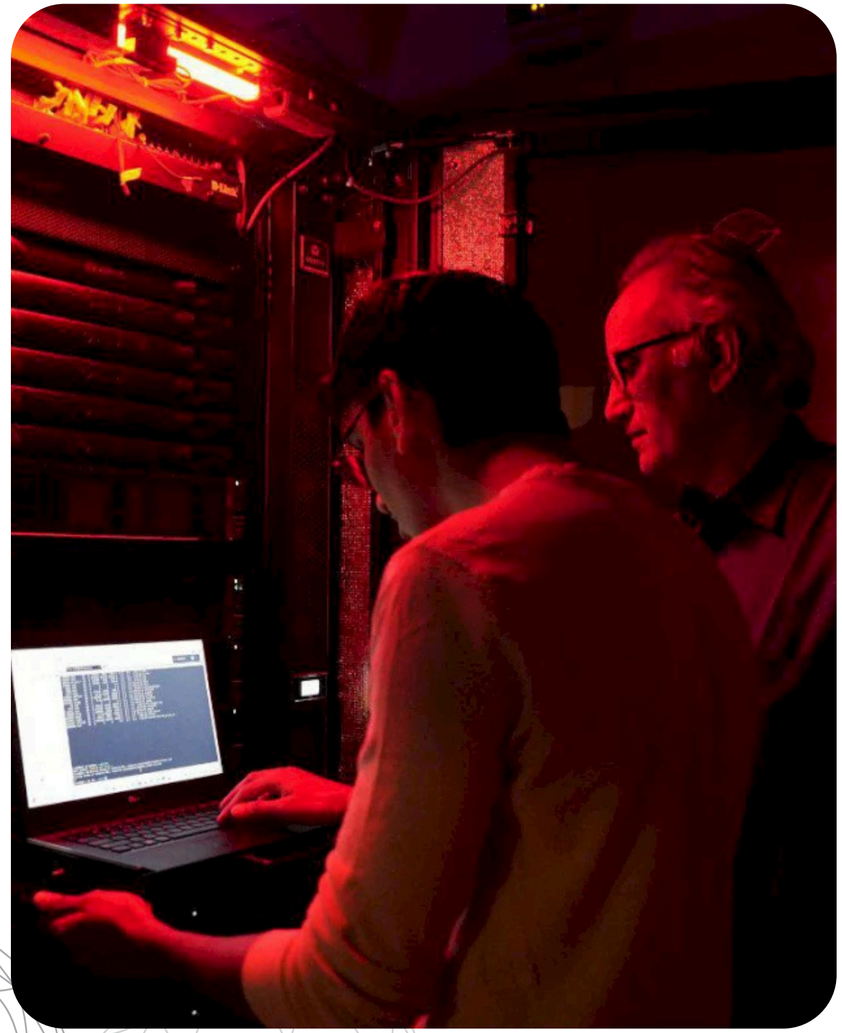
Hot-Plug Heaven: Dynamic AI Workloads in a Mini IaaS Cloud

Need-as-you-go

Presented by Urjit Mehta



Are We Fully Utilizing it?

Overestimation

- Requesting 16* cores but using only a few
- **Booked the whole theatre, watched alone**

Waiting Time

- Increased wait times for other users
- **When one feasts, others fast**

Underutilization

- Reduced overall efficiency
- **Looks busy, works lazy**

Stepwell HPC Plugged Cloud

Elastic Usage

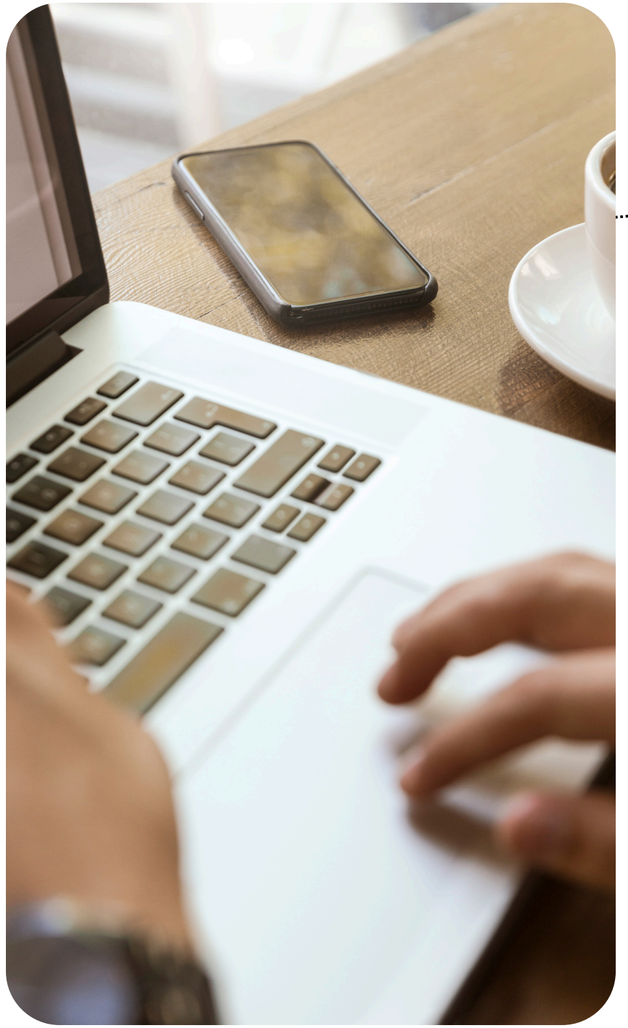
Mimics cloud’s use-as-needed principle, so CPUs are only plugged in when required to utilised

Real-Time Scaling

Dynamically adds or removes CPUs based on workload intensity

Local Simulation :(

Built using VirtualBox to simulate a private IaaS setup



See It in Action

A behind-the-scenes look at how everything comes together, one step at a time.

Thanks for plugging in!

urjit.m@ahduni.edu.in

urjit-mehta.web.app

linkedin.com/in/urjitmehta