Cloud Cost Optimization

Project Guide: Dr. Debabrata Sarddar

Date 7 July, 2025

A Practical System for Infrastructure Cost Analysis and Hybrid Cloud Strategy

From Research to Implementation:
Tool for the Hybrid Equilibrium

The Infrastructure Cost Crisis

Organizations are spending 2–5x more than necessary on infrastructure while achieving suboptimal performance for predictable workloads.

The Challenge We Face

Cloud-first strategies are breaking budges

20–30% of all cloud spending is wasted

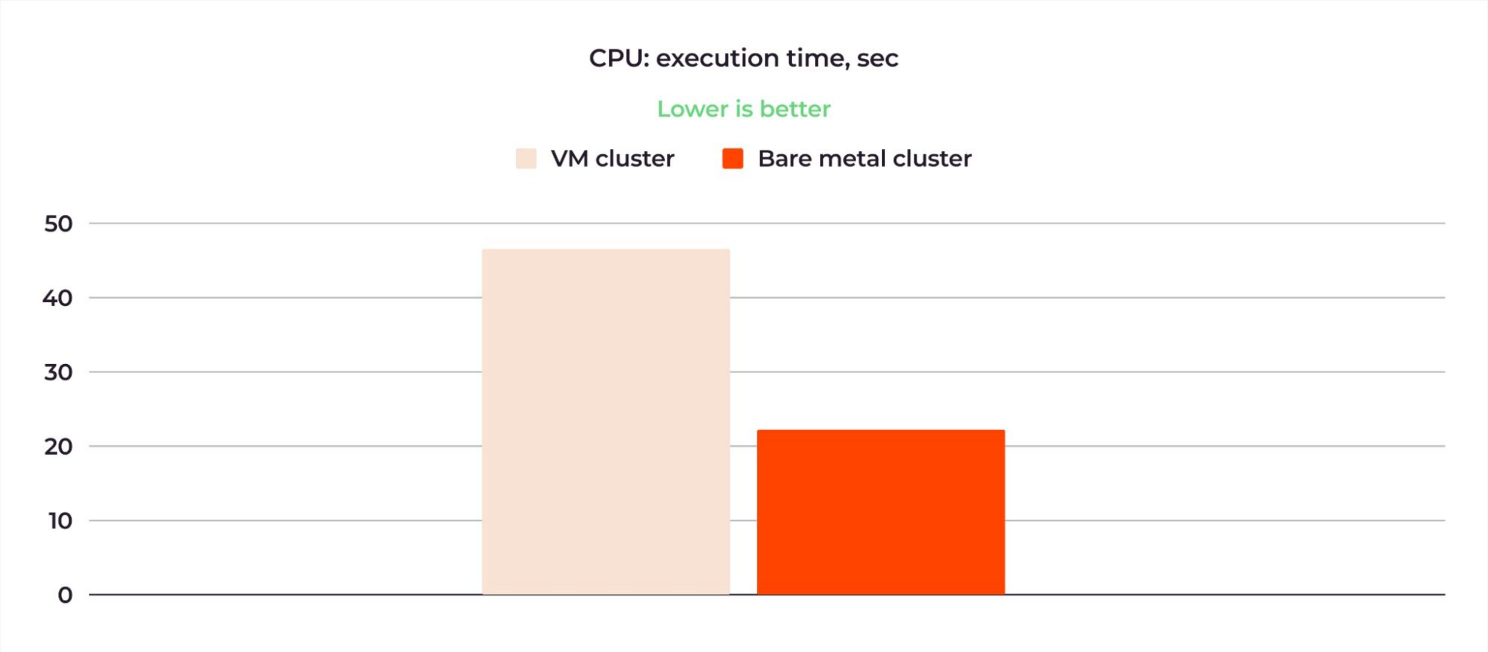
Source: IDC Future Enterprise Planning Guides, Control Cloud Costs and Expand Transparency with FinOps, IDC #US50654223, May 19, 2023

Test Results

K8s cluster type	CPU		RAM latency (write/read)	Storage		Network	
	Speed (execution time)	Utilization		TPC (transactions per second)	Latency	Bandwidth	Latency
VM	47.07 sec	86.81%	174.53 / 173.75 ms	4,636	55.21 ms	6.52 MB/sec	145 us
Bare metal	21.46 sec	43.75%	62.02 / 47.33 ms	12,029	21.28 ms	31 MB/sec	24.5 us

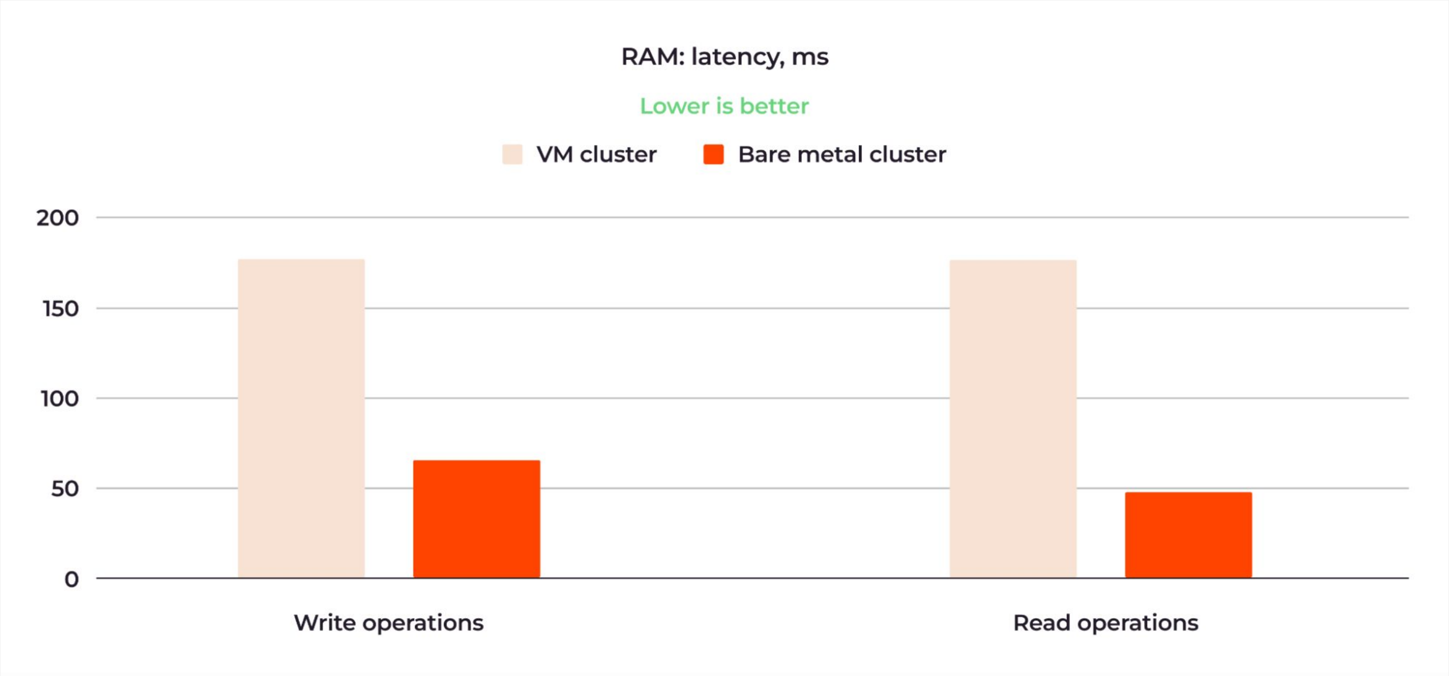
Source: Does Kubernetes Really Perform Better on Bare Metal vs. VMs? - https://thenewstack.io/does-kubernetes-really-perform-better-on-bare-metal-vs-vms

CPU speed comparison results



Source: Does Kubernetes Really Perform Better on Bare Metal vs. VMs? - https://thenewstack.io/does-kubernetes-really-perform-better-on-bare-metal-vs-vms

RAM Latency results



Source: Does Kubernetes Really Perform Better on Bare Metal vs. VMs? - https://thenewstack.io/does-kubernetes-really-perform-better-on-bare-metal-vs-vms

Instance Configuration

Instance type	Configuration	Storage
Bare metal	(x3) Intel Xeon E-2388 8C/16T 3.2 GHz / 64 GB	2x 480 GB SSD (RAID 1)
Virtual machine	(x3) 16 vCPU / 64 GiB	480 GB High IOPS SSD

Source: Does Kubernetes Really Perform Better on Bare Metal vs. VMs? - https://thenewstack.io/does-kubernetes-really-perform-better-on-bare-metal-vs-vms

What's Wrong

01

Blind Cost adoption without cost analysis

02

Lack of usage monitoring

03

Lack of optimization

04

No comparison with available resources

05

Missing tools for hybrid decision-making

What We Need

01

Cost-aware infrastructure decisions

02

Real-time system monitoring

03

Data-driven optimization tools

04

Scheduling infrastructure

05

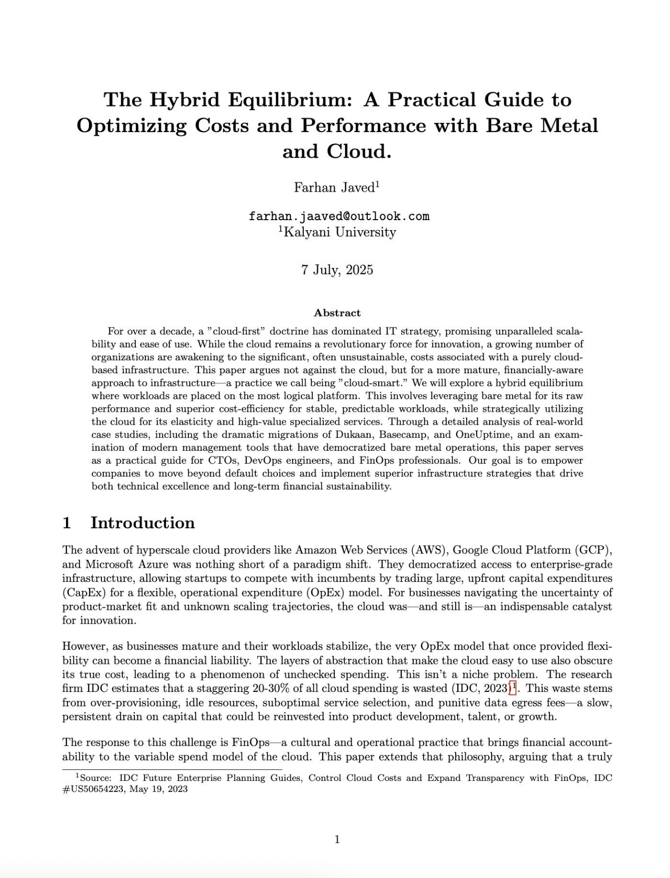
Hybrid cloud strategy guidance

I combined academic research with practical tool development to address infrastructure cost and optimization holistically.

Research Component

- Analyzed real-world case studies
- Documented cost savings strategies
- Developed hybrid decision framework
- Provided practical implementation guidance

Key Finding: Strategic Hybrid approach can reduce infrastructure costs by 40-60% while maintaining performance



The Hybrid Equilibrium: A Practical Guide to Optimizing Costs and Performance with Bare Metal and Cloud.

Farhan Javed[1]

farhan.jaaved@outlook.com
[1]Kalyani University

7 July, 2025

Abstract

For over a decade, a "cloud-first" doctrine has dominated IT strategy, promising unparalleled scalability and ease of use. While the cloud remains a revolutionary force for innovation, a growing number of organizations are awakening to the significant, often unsustainable, costs associated with a purely cloud-based infrastructure. This paper argues not against the cloud, but for a more mature, financially-aware approach to infrastructure—a practice we call being "cloud-smart." We will explore a hybrid equilibrium where workloads are placed on the most logical platform. This involves leveraging bare metal for its raw performance and superior cost-efficiency for stable, predictable workloads, while strategically utilizing the cloud for its elasticity and high-value specialized services. Through a detailed analysis of real-world case studies, including the dramatic migrations of Dukaan, Basecamp, and OneUptime, and an examination of modern management tools that have democratized bare metal operations, this paper serves as a practical guide for CTOs, DevOps engineers, and FinOps professionals. Our goal is to empower companies to move beyond default choices and implement superior infrastructure strategies that drive both technical excellence and long-term financial sustainability.

1 Introduction

The advent of hyperscale cloud providers like Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure was nothing short of a paradigm shift. They democratized access to enterprise-grade infrastructure, allowing startups to compete with incumbents by trading large, upfront capital expenditures (CapEx) for a flexible, operational expenditure (OpEx) model. For businesses navigating the uncertainty of product-market fit and unknown scaling trajectories, the cloud was—and still is—an indispensable catalyst for innovation.

However, as businesses mature and their workloads stabilize, the very OpEx model that once provided flexibility can become a financial liability. The layers of abstraction that make the cloud easy to use also obscure its true cost, leading to a phenomenon of unchecked spending. This isn't a niche problem. The research firm IDC estimates that a staggering 20-30% of all cloud spending is wasted (IDC, 2023)[1]. This waste stems from over-provisioning, idle resources, suboptimal service selection, and punitive data egress fees—a slow, persistent drain on capital that could be reinvested into product development, talent, or growth.

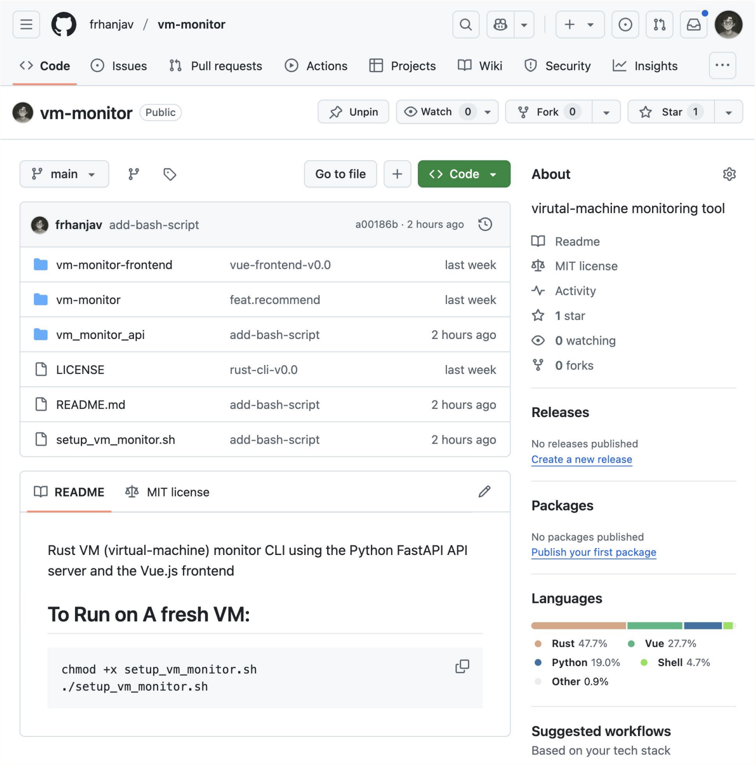
The response to this challenge is FinOps—a cultural and operational practice that brings financial accountability to the variable spend model of the cloud. This paper extends that philosophy, arguing that a truly

[1]Source: IDC Future Enterprise Planning Guides, Control Cloud Costs and Expand Transparency with FinOps, IDC #US50654223, May 19, 2023

Implementation Component

- Real-time system monitoring
- Cost recommendation engine
- Virtual Machine analysis

Key Value: Helps developers monitor systems in real time, compare infrastructure costs, and choose the best cloud provider based on data.



System Architecture

01

CLI Agent (Rust)

Collets CPU, memory, disk, and network metrics with high performance.

02

API Server (FastAPI)

Secure data aggregation and serving.

03

Dashboard (Vue.js)

Interactive visualization of system performance data

Running the Backend API

```
(venv) ➜ vm_monitor_api git:(main) uvicorn app.main:app --reload --host 0.0.0.0 --port 8000
INFO:     Will watch for changes in these directories: ['/home/zua/Projects/vm-monitor/vm_monitor_api']
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
INFO:     Started reloader process [4497] using WatchFiles
INFO:     Started server process [4499]
INFO:     Waiting for application startup.
VM Monitor API starting up...
INFO:     Application startup complete.
New agent registration: 02f88de8-aca2-47e3-ae34-95d7a6adb356
Agent 'Test-VM-THETA' (02f88de8-aca2-47e3-ae34-95d7a6adb356) registered with API key prefix: cUthTmEK...
INFO:     127.0.0.1:37338 - "POST /v1/agent/register HTTP/1.1" 201 Created
Agent 02f88de8-aca2-47e3-ae34-95d7a6adb356 authenticated successfully.
Received metrics batch (count: 3) for agent 02f88de8-aca2-47e3-ae34-95d7a6adb356.
INFO:     127.0.0.1:39798 - "POST /v1/agent/metrics HTTP/1.1" 202 Accepted
Agent 02f88de8-aca2-47e3-ae34-95d7a6adb356 authenticated successfully.
Received metrics batch (count: 3) for agent 02f88de8-aca2-47e3-ae34-95d7a6adb356.
INFO:     127.0.0.1:53860 - "POST /v1/agent/metrics HTTP/1.1" 202 Accepted
```

Running the CLI

```
➜  vm-monitor git:(main) ./target/debug/vm-monitor start --interval=5
[2025-07-07T06:15:59Z INFO  vm_monitor] Starting VM Monitor Agent for instance ID: 02f88de8-aca2-47e3-ae34-95d7a6adb356
[2025-07-07T06:15:59Z INFO  vm_monitor] Monitoring interval: 5s, Batch size: 3
VM Monitor agent started. Interval: 5s, Batch Size: 3. Press Ctrl+C to stop.
[2025-07-07T06:16:04Z INFO  vm_monitor] Collected metrics. Buffer size: 1
[2025-07-07T06:16:09Z INFO  vm_monitor] Collected metrics. Buffer size: 2
[2025-07-07T06:16:14Z INFO  vm_monitor] Collected metrics. Buffer size: 3
[2025-07-07T06:16:14Z INFO  vm_monitor] Batch limit reached (3 items). Sending metrics...
[2025-07-07T06:16:14Z INFO  vm_monitor] Successfully sent batch of 3 metrics.
[2025-07-07T06:16:19Z INFO  vm_monitor] Collected metrics. Buffer size: 1
[2025-07-07T06:16:25Z INFO  vm_monitor] Collected metrics. Buffer size: 2
[2025-07-07T06:16:30Z INFO  vm_monitor] Collected metrics. Buffer size: 3
[2025-07-07T06:16:30Z INFO  vm_monitor] Batch limit reached (3 items). Sending metrics...
[2025-07-07T06:16:30Z INFO  vm_monitor] Successfully sent batch of 3 metrics.
[2025-07-07T06:16:35Z INFO  vm_monitor] Collected metrics. Buffer size: 1
[2025-07-07T06:16:40Z INFO  vm_monitor] Collected metrics. Buffer size: 2
```

Running the Frontend Dashboard

```
➜  vm-monitor-frontend git:(main) npm run dev

> vm-monitor-frontend@0.0.0 dev
> vite


  VITE v7.0.0  ready in 534 ms

  ➜  Local:   http://localhost:5173/
  ➜  Network: use --host to expose
  ➜  press h + enter to show help
```

VM Monitor Dashboard

API Status: Connected

Select Agent

Test-VM-THETA (02f88de8...)

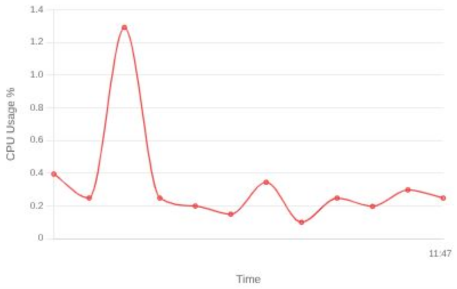
Refresh Agent List

Refresh Metrics Data

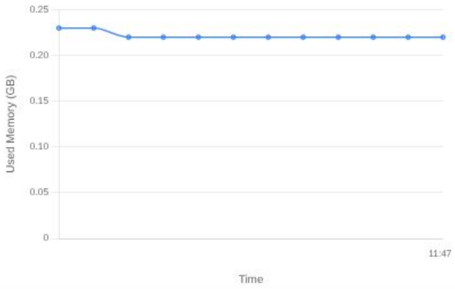
Metrics for: Test-VM-THETA

Last refreshed: 11:47:05 AM

CPU Usage (%)



Memory Usage (Used GB)



System Info Snapshot (Latest)

Hostname:	penguin	OS:	Debian GNU/Linux 12
Kernel:	6.6.76-08174-g2f3b34fb3650	Uptime:	0d 0h 39m 8s
Cloud:	N/A	Total Memory:	14.17 GB
Available Memory:	13.95 GB	CPU Cores:	4

```
➜ vm-monitor git:(main) ./target/debug/vm-monitor recommend
Collecting system usage data for 60 seconds. Please wait...

--- Usage Analysis Complete ---
Average CPU Usage: 0.25%
Average Memory Used: 0.21 GB
Physical CPU Cores on this machine: 4
-----------------------------

Loading VM instance dataset...
Finding recommendations...
Based on average usage, recommending for ~1.00 vCPUs and 2.00 GB Memory...
Top VM Recommendations (lower score is better):
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Provider      | Instance Name   | Region        | vCPUs | Memory (GB) | Est. Hourly Cost ($)  | Efficiency Score |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Hetzner       | cx21            | fsn1          | 2     | 4           | 0.0120                | 0.003200         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| GCP           | e2-medium       | us-central1   | 2     | 4           | 0.0269                | 0.007162         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| GCP           | e2-medium       | europe-west1  | 2     | 4           | 0.0295                | 0.007878         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| AWS           | t4g.medium      | us-east-1     | 2     | 4           | 0.0336                | 0.008960         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| AWS           | t3a.medium      | us-west-2     | 2     | 4           | 0.0376                | 0.010027         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Azure         | Standard_B2s    | eastus        | 2     | 4           | 0.0416                | 0.011093         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Azure         | Standard_F2s_v2 | eastus        | 2     | 4           | 0.0850                | 0.022667         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Hetzner       | g2-gpubox-small | fsn1          | 2     | 8           | 0.1500                | 0.040000         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Equinix Metal | c3.small.x86    | sv            | 2     | 8           | 0.5000                | 0.133333         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
| Equinix Metal | s3.xlarge.x86   | ny            | 8     | 64          | 1.7500                | 0.466667         |
+---------------+-----------------+---------------+-------+-------------+-----------------------+------------------+
```

Thank you!

Name: Farhan Javed