Marrying HPC and Cloud For Long Term Happiness

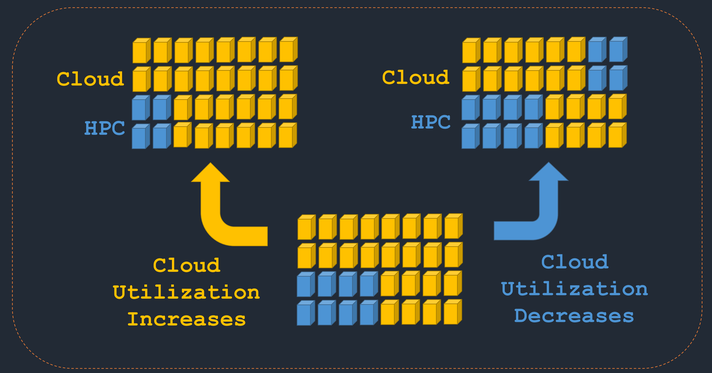
Massachusetts Open Cloud

Apoorve Mohan[†], Ravi S. Gudimetla[†], Ata Turk[‡], Sourabh Bollapragada[†], Rajul Kumar[†], Jason Hennesey[‡], Evan Weinberg[‡], Dmitri Markigiorgos[‡], Christopher N. Hill*, Gene Cooperman[†], Peter Desnoyers[†], Richard Brower[‡], Orran Krieger[‡]
{[†]Northeastern University, [‡]Boston University, *Massachusetts Institute Of Technology} Boston, MA

Motivation And Goals

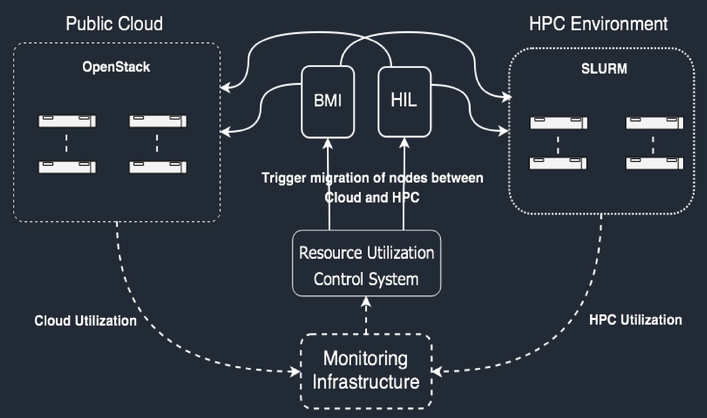
- **Public Clouds** often have **low utilization**:
  - On-demand usage model, no tolerance for latency.
  - Accomodate any potential customers.
  - To provide sufficient performance isolation.
- **HPC clusters** have **steady high utilization**:
  - Batch queues always ready to supply new jobs.
  - Tolerance for job start/completion latencies.
  - Almost infinite workload.
- **GOAL: Improve Overall Data Center Utilization**
  - Co-locate Cloud and HPC clusters, share resources intelligently to drive utilization.
  - Sharing physical resources (servers) between clusters is non-trivial.
- **Issues addressed here**:
  - Switch servers between different frameworks.
  - Rapidly re-provision servers between different clusters.
  - Minimize downtime while re-provisioning.

- We **developed the Resource Utilization Control System (RUCS)**, a framework that provides a way to share infrastructure hosting a cloud with the help of hardware isolation and rapid provisioning tools.



Resource Utilization Control System (RUCS)

- **Resource Utilization Control System**:
  - Monitors utilization of both the **Cloud** and **HPC**.
  - Based on utilization thresholds, migrates nodes between Cloud and HPC using **Hardware Isolation Layer (HIL)**.
  - Rapidly Provision / De-provision nodes for both Cloud and HPC clusters using **Bare Metal Imaging Service (BMIS)**.
- **Hardware Isolation Layer**, shimming layer for datacenters:
  - Hardware (server) reservation.
  - Network isolation.
  - Rapid switching of servers between networks.
- **Bare Metal Imaging Service**:
  - Rapid bare-metal provisioning via network mounting images that contain OS and applications.
  - Image management capabilities for Bare-Metal akin to those available in virtualized systems (e.g. (De)Provisioning, Cloning/Snapshotting).
  - BMIS catalyzes bare-metal cloud development.



Results And Experimental Evaluation

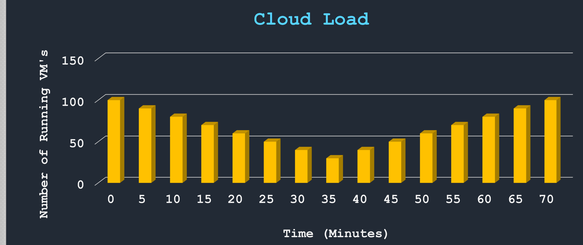


Figure 1(a)

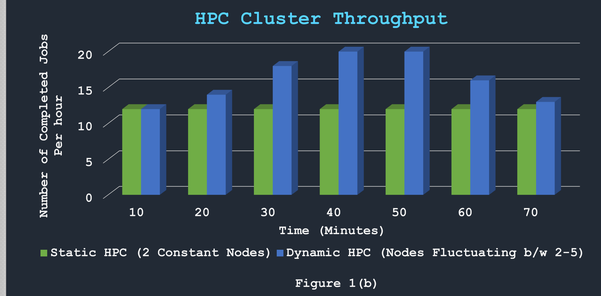


Figure 1(b)

- **Figure 1(a)** shows a simulated time-varying workload for the Cloud, following a cosine curve pattern.
- **Figure 1(b)** compares throughput of a static HPC cluster vs a dynamically resizing HPC cluster based on cloud workload.
  - As the cloud utilization varies, the number of nodes in the HPC cluster varied from 2 to 5.
  - HPC utilization always >90%.

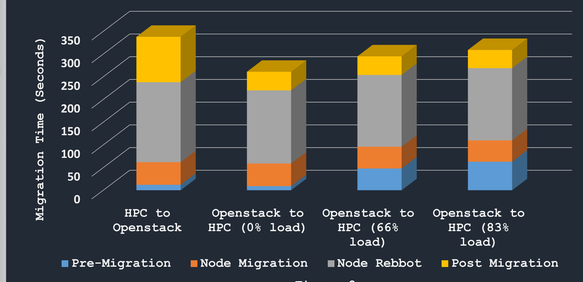


Figure 2

- **Figure 2** presents server migration times under different load scenarios.
  - HPC (>90% load) to Cloud transition: ~330s.
  - Cloud (~83% load) to HPC transition: ~310s

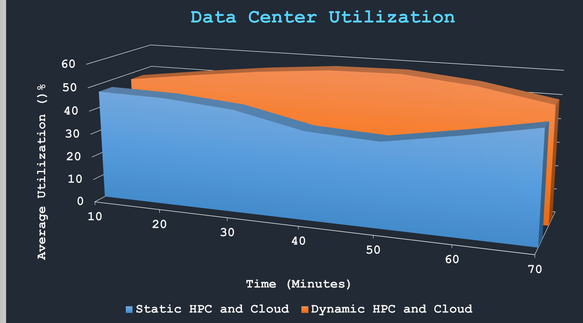


Figure 3

- **Figure 3** presents the utilization comparison between static clusters vs dynamically scaling HPC and Cloud for 11 Nodes.
  - Average utilization can be improved as much as 20% even with conservative policies