FIU
FLORIDA INTERNATIONAL UNIVERSITY

IBM

LA Grid
Latin American Grid

National Science Foundation

CI-PIRE

LA Grid Summit 2011
November 3rd & 4th 2011
Florida Atlantic University

Partnership for International Research and Education
A Global Living Laboratory for Cyberinfrastructure Application Enablement

Efficiency Assessment of Parallel Workloads on Virtualized Resources

Javier Delgado, PhD candidate, Florida International University
FIU Advisors: Malek Adjouadi and S. Masoud Sadjadi, College of Engineering and Computing
PIRE Industrial Partner Advisor: Liana L. Fong, IBM

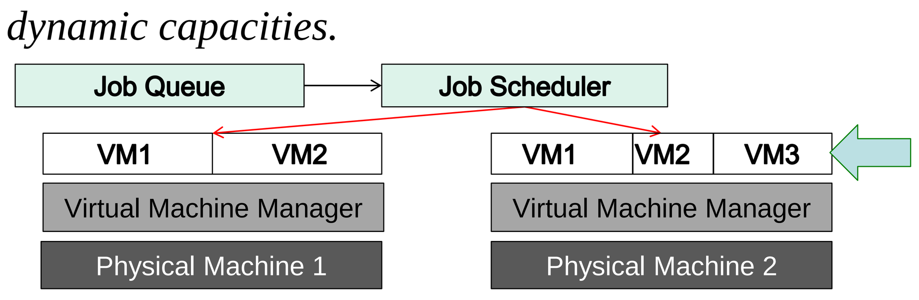
I. Research Overview and Outcome

Abstract

- Virtualization technology simplifies deployment of applications; users just create *virtual appliances*
- Therefore, the most common way of **deploying applications** in compute clouds is **through virtual machines (VMs)**
- To optimize profits, resource providers need to maximize utilization of their resources without degrading users' perceived performance
- To maximize resource utilization, providers generally *pack* **multiple virtual appliances** into **each physical machine**
- Packing VMs results in **heterogeneous CPU allotments** for different VMs, which is a problem when the VMs are workers of **tightly-coupled parallel applications**
- Early analysis through simulation showed that resources can be underutilized by nearly 50%, even when all physical machines are used, when running tightly coupled applications in this kind of environment
- **In this work, we measure the impact of the *ripple effect* and *elasticity constraint effect* that occur when running parallel jobs in virtualized environments**
- **We address these effects by applying some heuristics to a scheduling algorithm, increasing utilization to around 65%**

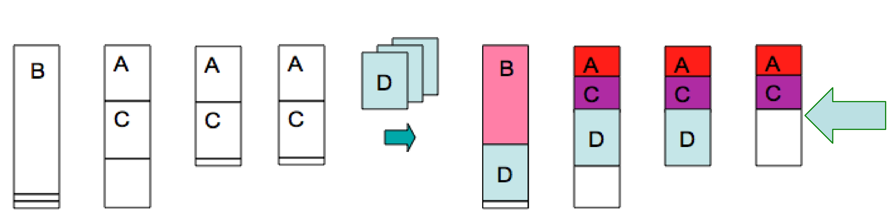
Problem

First Problem: Scheduling jobs on nodes with *heterogeneous, dynamic capacities.*



Think of horizontal length of VM as its CPU capacity

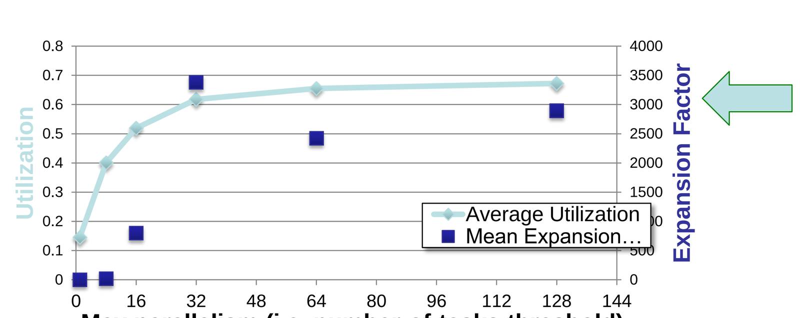
Picking the best node for a job is not trivial when they are VMs.



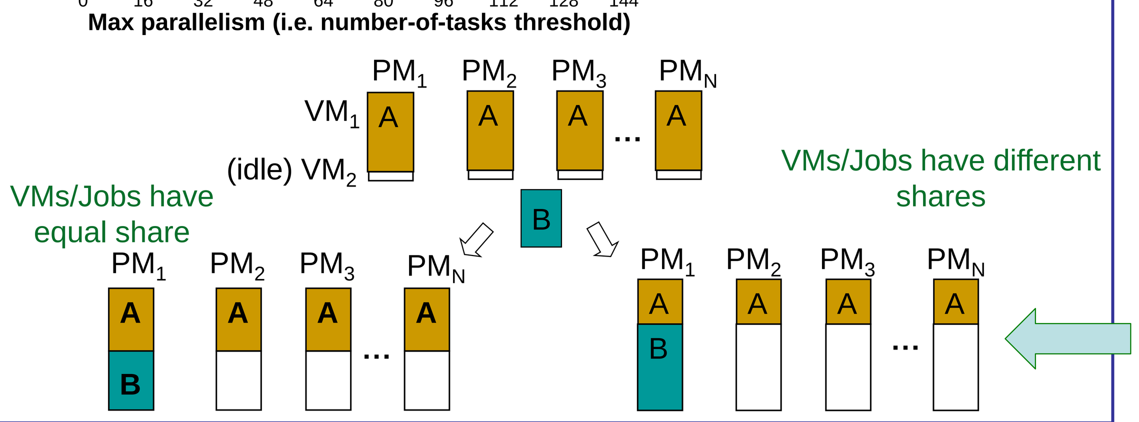
Which nodes should be selected for 3-task job D?

The dynamic capacity issue was first addressed in [1].

Second Problem: Even if **all machines** are **used**, the **cluster** may **still** be **underutilized** due to task synchronization.



Average job's expansion factor increases, but cluster utilization does not exceed 70%



The **ripple effect** results in one or more nodes being left underutilized, since Job A in PM1 needs to synchronize with Job A on other PMs.
Elasticity constraint effect is a similar phenomenon, but happens as a result of a capacity limit on a subset of the executers

Proposed Solution

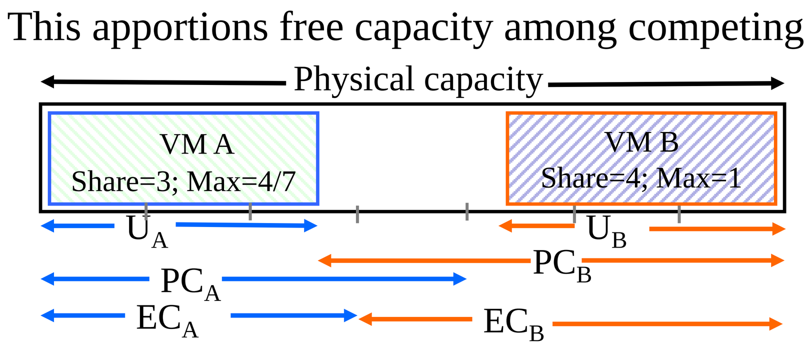
Scheduling Algorithm. We use First come first serve (FCFS) with backfill, for its compromise of fairness and scheduling performance.

The **first problem**, VM selection, was addressed last year (see [1]) by defining 2 metrics for assigning jobs to VMs:

- **Equilibrium capacity (EC)** - lower-limit on a VM's capacity
- **Potential capacity (PC)** - upper limit on a VM's capacity

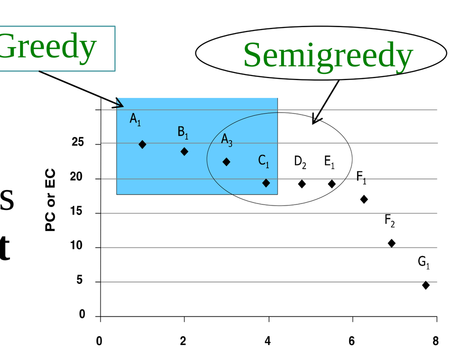
These values are based on the number of VMs competing for CPU, their CPU utilization, and on each VM's *sharing parameters*:

- `min` - Minimum CPU capacity the VM is guaranteed
- `max` – Maximum CPU capacity that a VM can obtain
- `share` - This apportions free capacity among competing VMs



For the **second problem**, we apply additional heuristics to address the ripple and elasticity constraint effects

- **Semigreedy** – Leave higher-capacity VMs for smaller jobs
- **Free capacity based** – Use free capacity as a secondary metric, to **reduce ripple effect** when assigning a serial job to nodes with parallel jobs (see ripple effect example)



Experiments

Overview. The ripple and elasticity constraint effects were quantified and the efficacy of the new scheduling heuristics evaluated via simulation using real-world workload traces.

Platform. A cluster with 64 physical machines with 4 VMs each was simulated. **Heterogeneous** and **homogeneous** cluster platforms were used in the evaluation.

- Heterogeneous - different shares => `Share =` 1 | 1 | 2 | 4
- Homogeneous - same share => `Share =` 1 | 1 | 1 | 1

Workloads. 10,000 jobs from the Grid'5000 and Cornel Theory Center (CTC) were scheduled on the cluster.

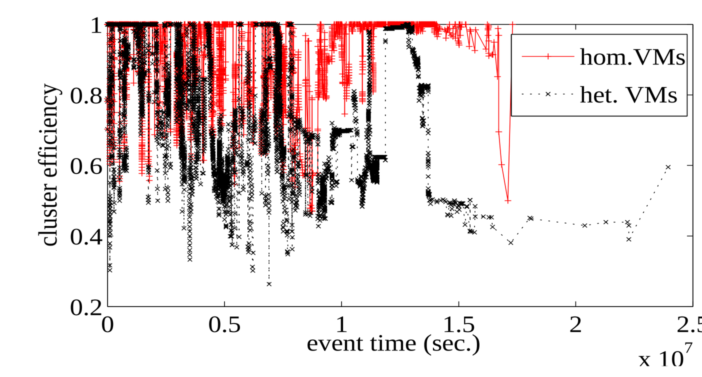
Scheduling Algorithms

- **PC-g**: Select VMs with the **highest PC** ranking
- **PC-sg**: Select VMs according to their PC ranking using the **semi-greedy** algorithm
- **PCFC-sg**: Extend **PC-sg with** the **free capacity heuristic** when scheduling serial jobs: First select the VM with the highest PC ranking from the VMs whose PC is not bigger than the free capacity (FC) of the underlying physical machine. If no such VM is found, select the one with the highest PC ranking.
- **EC**: Select VMs with the **highest EC** ranking.
- **PMVM**: Prioritize PMs with **the least number of active VMs**; use VM's `share` as the second priority (i.e. "tie-breaker").

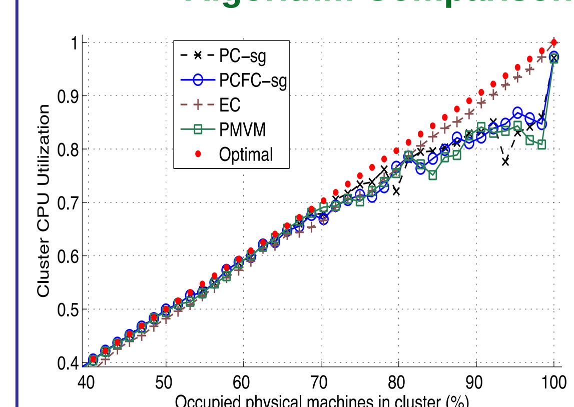
Evaluation

Grid'5000

Cluster efficiency, Heterogeneous vs. homogeneous VMs

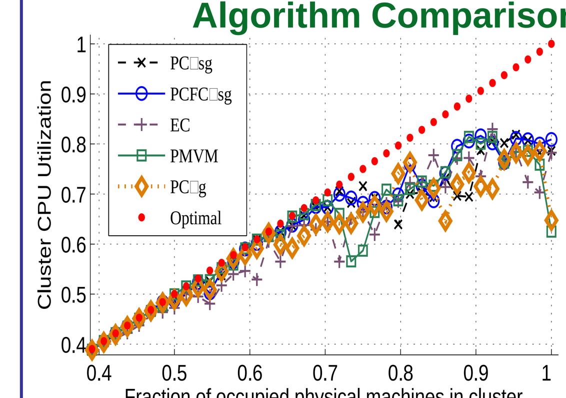


Algorithm Comparison, Homogeneous VMs



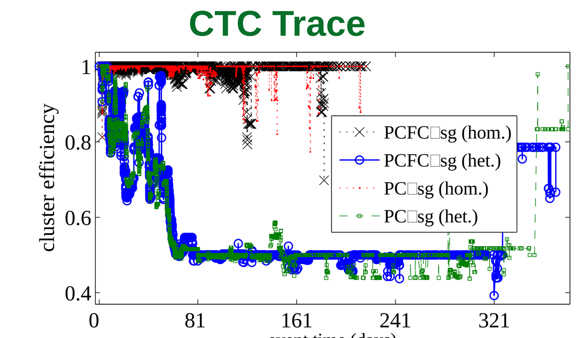
Algorithm	Utilization	Efficiency	XF (median)	Makespan (days)
PC-g	0.719	0.88	7.9	204.77
PCFC-sg	0.729	0.89	5.98	201.78
EC	0.706	0.95	11.31	208.28

Algorithm Comparison, Heterogeneous VMs



Algorithm	Util.	Eff.	XF (median)	Makespan (days)
PC-g	0.490	0.53	31	299.8
PCFC-sg	0.529	0.61	8.0	277.9
EC	0.559	0.64	8.31	263.0

CTC Trace



Conclusions

- Simulated different job placement metrics and heuristics with FCFS scheduler on real-world traces
- Discoveries
  - Serial jobs: Much better performance with the PC metric when there are few job arrivals
  - Tightly-coupled parallel jobs: Ripple effect caused significant waste of CPU cycles, especially when there is a high ratio of parallel jobs
  - Some heuristics to address the ripple effect helped, but need further improvements, especially if VMs are heterogeneous

Future Work

- Optimal algorithms for scheduling tightly coupled jobs and for bags of tasks
- Empirical experiments to validate simulation assumptions

References and Publications

[1] *New Metrics for Scheduling Jobs on a Cluster of Virtual Machines*, IPDPS SMTPS 2011
[2] *Efficiency Assessment of Parallel Workloads on Virtualized Resources*, UCC 2011

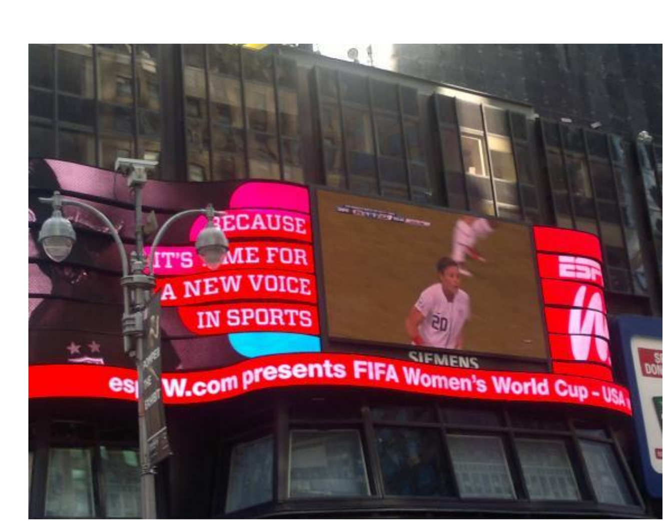
II. Industry and local Experience

First "peak" of Breakneck ridge– and a nice reminder that I am spending the summer at my home country

Another nice view from one of the many IBM Tuesday Night Hikes, which have been going on for several decades

Personal Impact

- Experienced full-time IBM internship thanks to PIRE
- Last year, I got a feel for the early stages of a research project.
- This year, we went through another milestone (i.e. planning and writing a paper)
- Publishing milestone was accelerated by eliminating physical barrier
- Also started a possible collaboration with other experts within IBM for a related topic we found interesting
- Overall, this experience left me better prepared for a professional research career in either academia or industry



In Times Square watching the US women's team "almost" win the World Cup

A nice view from Croton Point Park, where we had a department picnic

III. Acknowledgement

The material presented in this poster is based upon the work supported by the National Science Foundation under Grant Numbers OISE-0730065, CNS- BPC-AE-1042341, CNS-MRI-R2-0959985, and HRD-CREST-0833093. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.