Virtual Machine Placement in Cloud Data Centers

Common Pitfalls and Possible Research Directions

Abdelkhalik Mosa

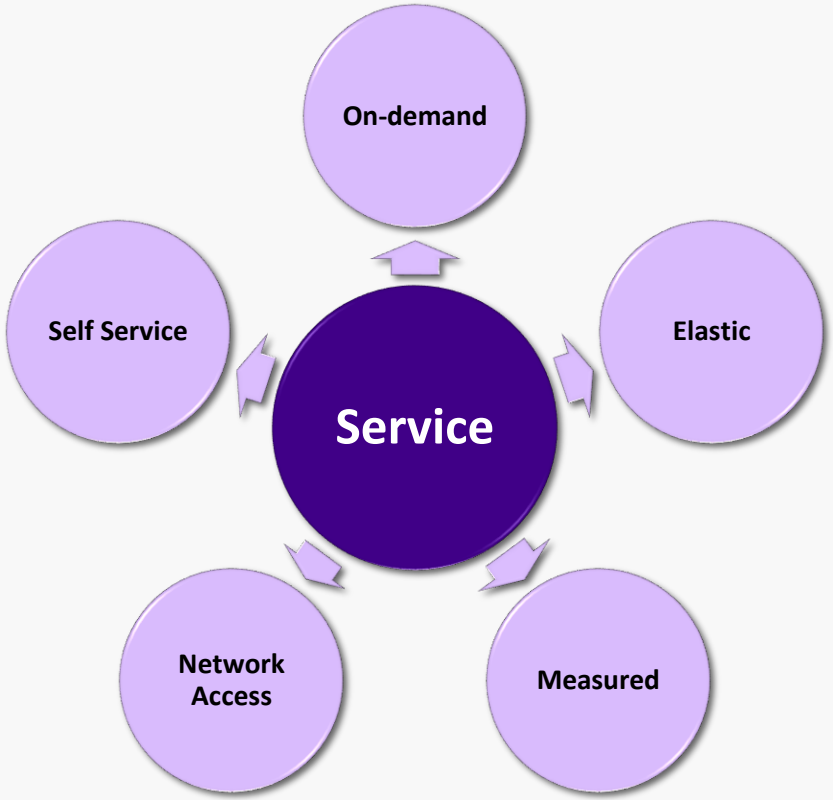
Sessional Tutor A at SCMDT, MMU
PhD candidate, School of Computer Science
The University of Manchester
abdelkhalik.mosa@manchester.ac.uk

Overview

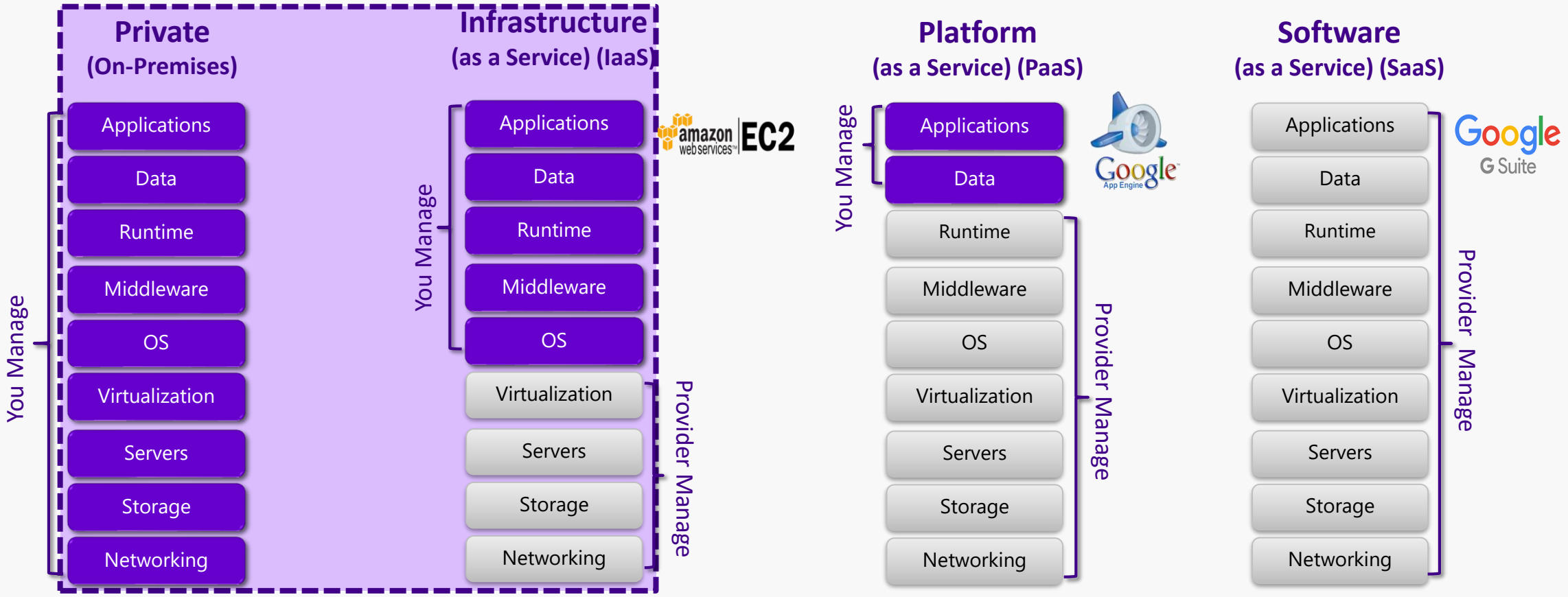
- Introducing Cloud Computing
- Cloud Enabling Technologies
- Virtualization and Cloud Management
- Energy and Server Utilization

Cloud Computing

Computing resources are provided a service over a network

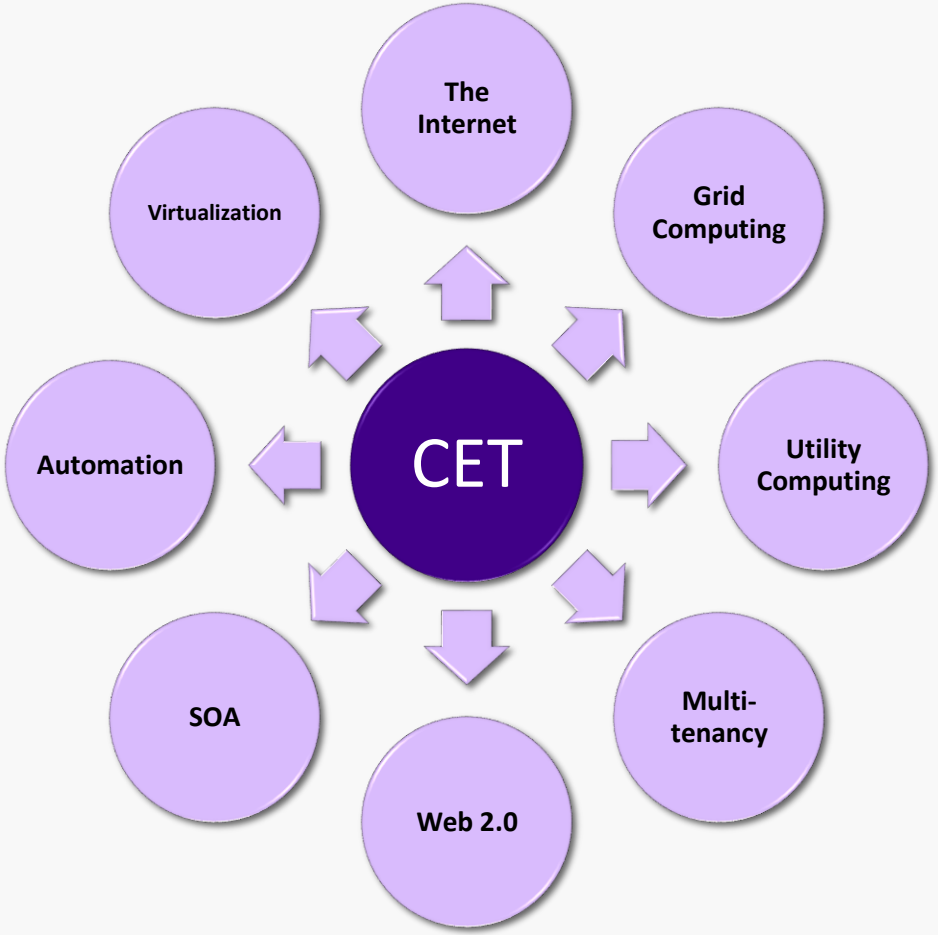


Cloud Service Models (XaaS)



Cloud Enabling Technologies (CET)

Cloud is not a revolution, it's a normal evolution!

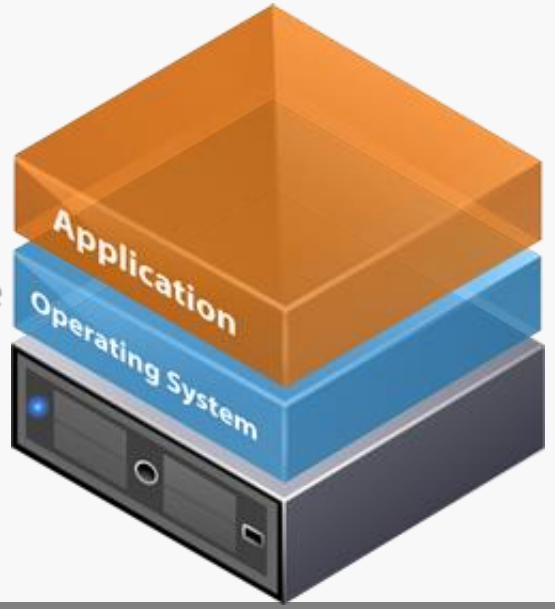


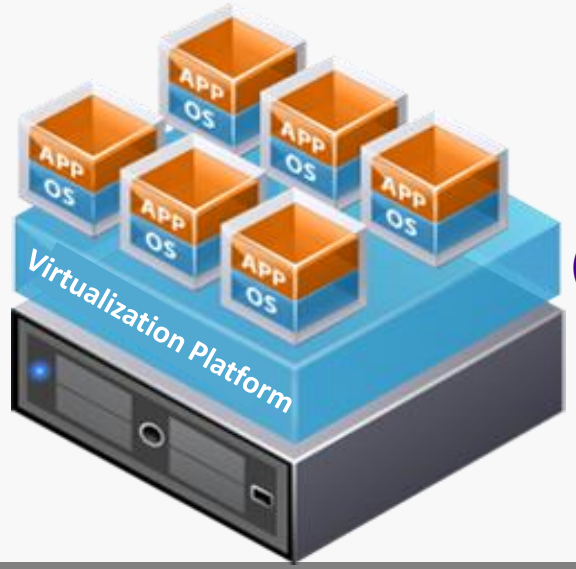
Virtualization

"Creating a virtual version of something, including a virtual computer hardware platform, OS, storage, or computer network resources."

Type 1 (bare metal) full virtualization

Traditional Architecture (One PM)



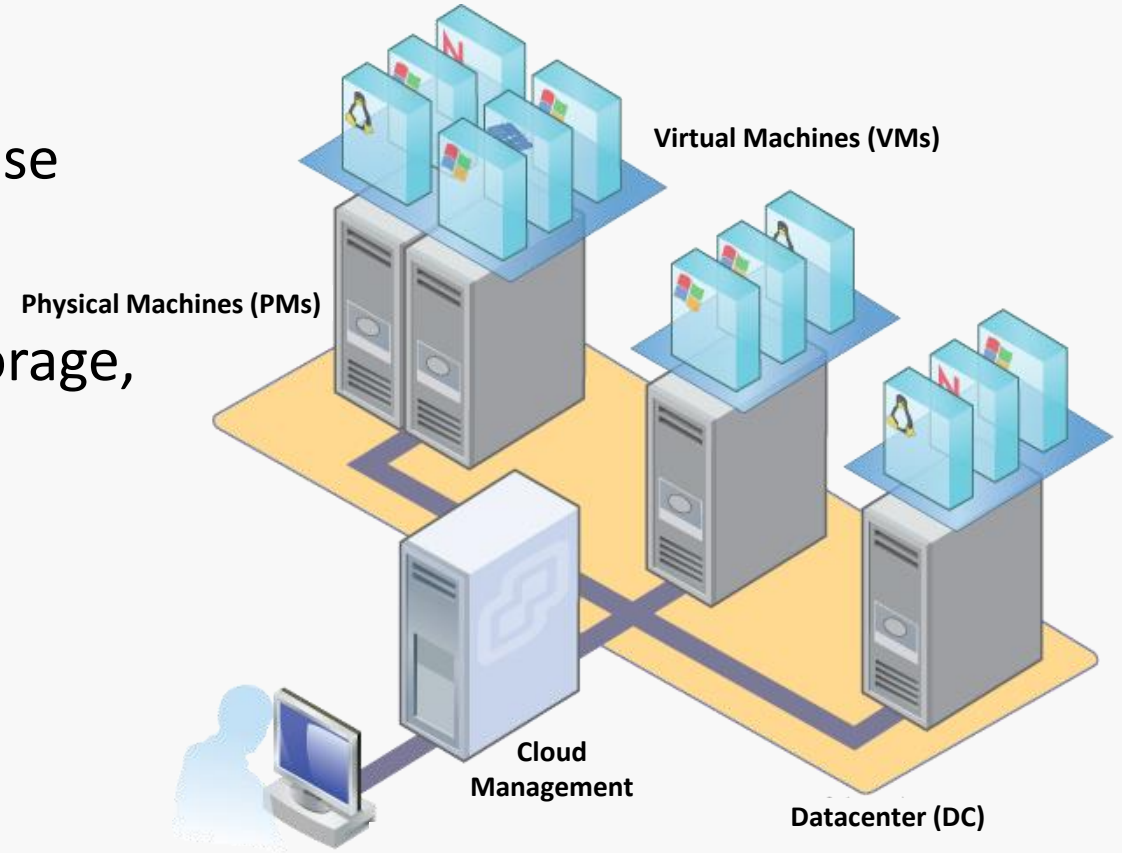


Virtual Architecture (One PM, Multiple VMs)

Credit: http://www.vmware.com

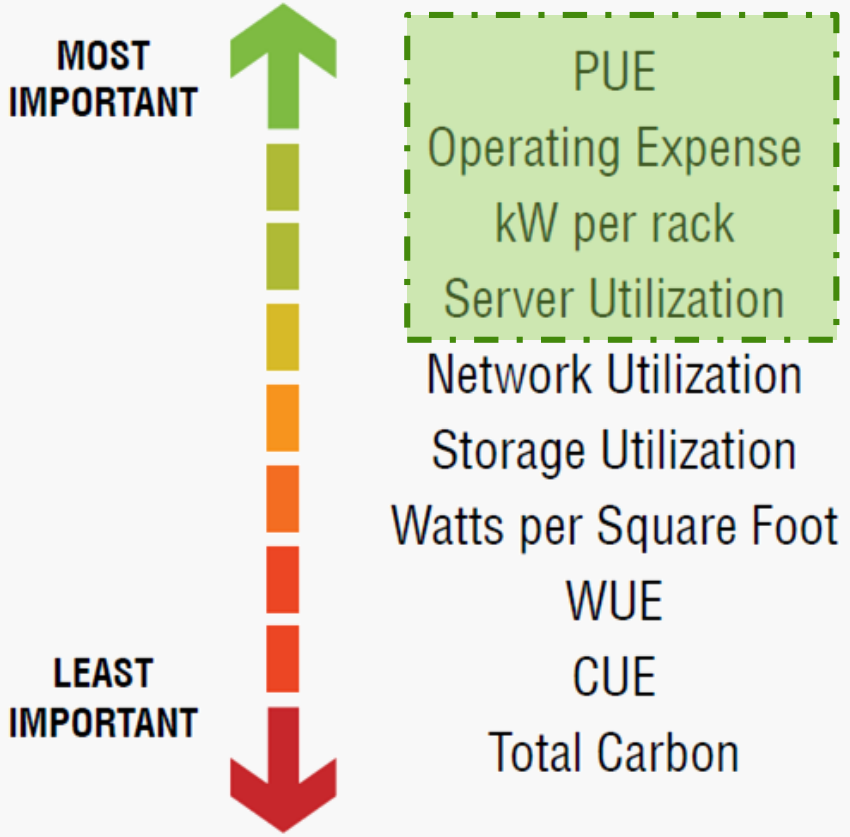
Virtualization and Cloud Management

- A cloud data center is a facility used to house servers (PMs) and associated components.
- Cloud management controls PMs, VMs, storage, and networking throughout a data center.
- One of the cloud management tasks is to manage the mapping of VMs to the PMs (Virtual Machine Placement (VMP)).



Credit: http://pubs.vmware.com/vsphere-4-esxi-embedded-vcenter/topic/com.vmware.vsphere.gsembedded.doc_40/getting_started/images/home_page_setup_02.png

Important Metrics of Data Centers



DC metrics in terms of importance by 1,000 DC operators (Feb–Apr 2015). ***The Uptime Institute survey.***

Energy and Server Utilization

More than fifth of IT devices do nothing rather than warming the planet. (NRDC 2015)

A server consumes about 70% of the power even if it is entirely idle.

Server	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
HP ProLiant G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
HP ProLiant G5	93.7	97	101	105	110	116	121	125	129	133	135

Energy and Server Utilization

A typical data center wastes large amounts of energy powering equipment doing little or no work. **The average server operates at only 12-18% of capacity!**

Main cause of energy waste is the underutilization of the servers. NRDC (2013)

Major cloud providers only accounts for about 5%; accomplishing efficiency within the big providers does not solve the problem.

Action is needed to accelerate adoption of energy efficiency best-practices.

Achieving just half of technologically feasible savings could cut electric use by 40% and save U.S. businesses $3.8 billion annually.

https://www.nrdc.org/sites/default/files/data-center-efficiency-assessment-IP.pdf

Virtual Machine Placement (VMP)

- Introduction to VMP
- VMP Problems
- VMP Strategies

Virtual Machine Placement

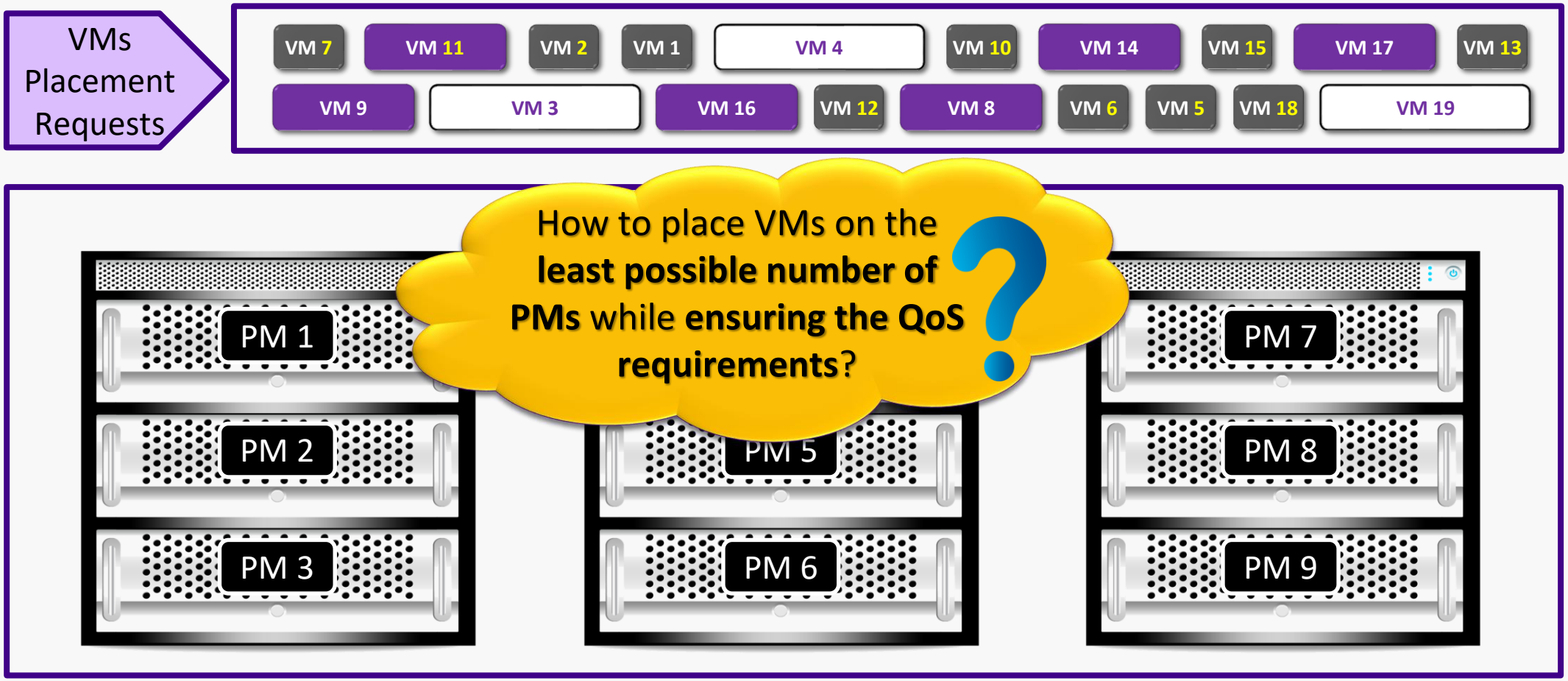
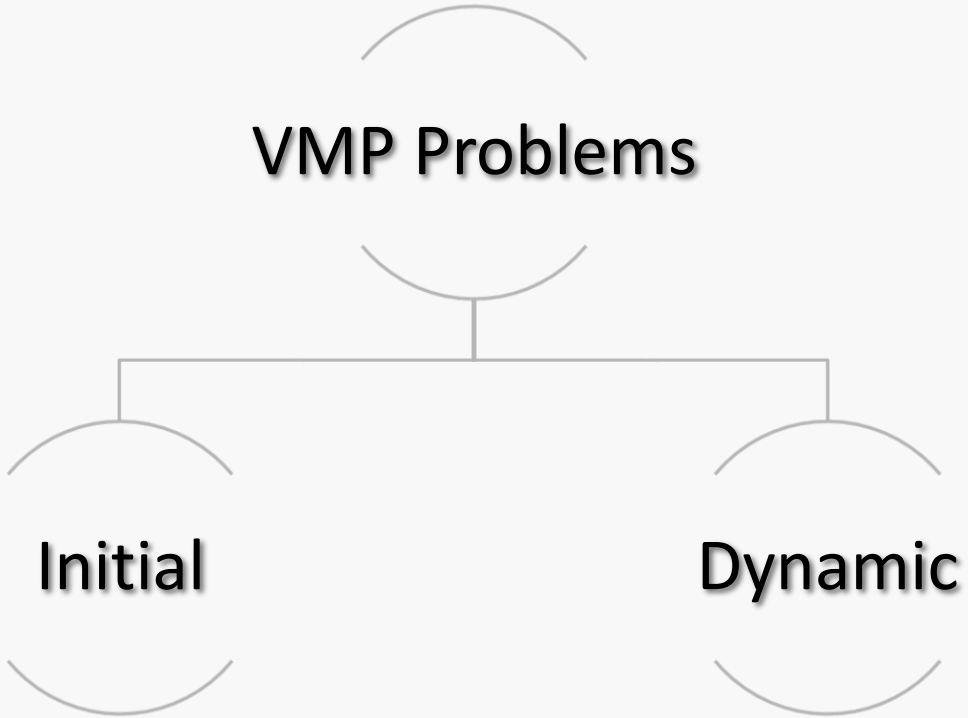


Photo Credit: https://forrst-live.s3.amazonaws.com/posts/snaps/59837/original.jpg

Virtual Machine Placement Problems



Initial Virtual Machine Placement

An initial (static) VMP accepts or rejects VMP requests and then assigns each accepted VM to an appropriate PM based on the requested VM size.

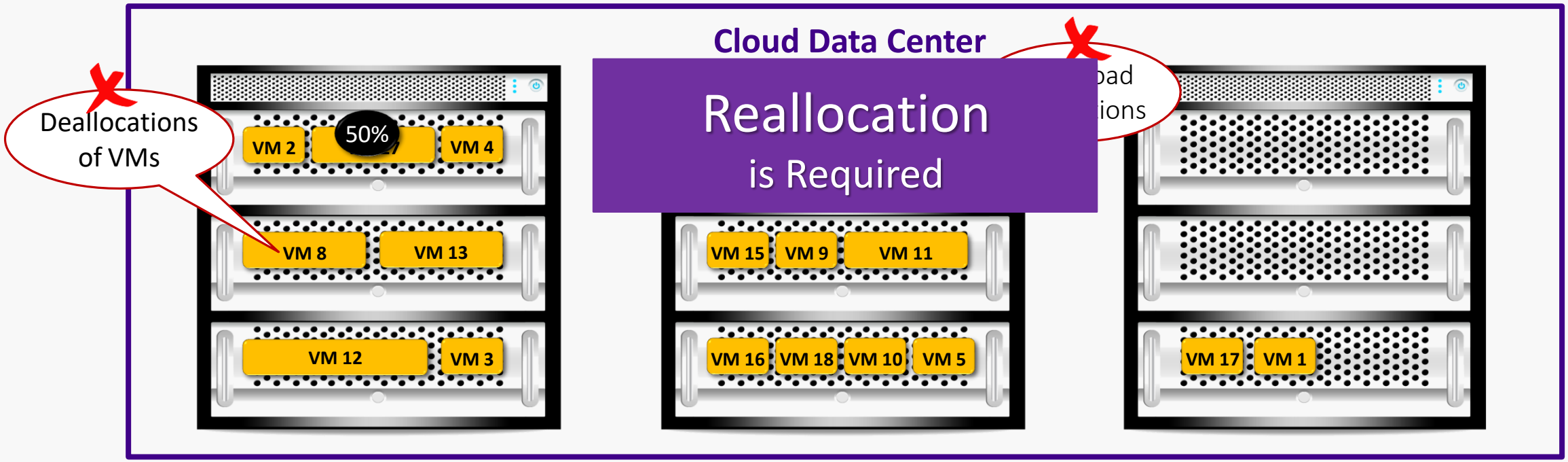
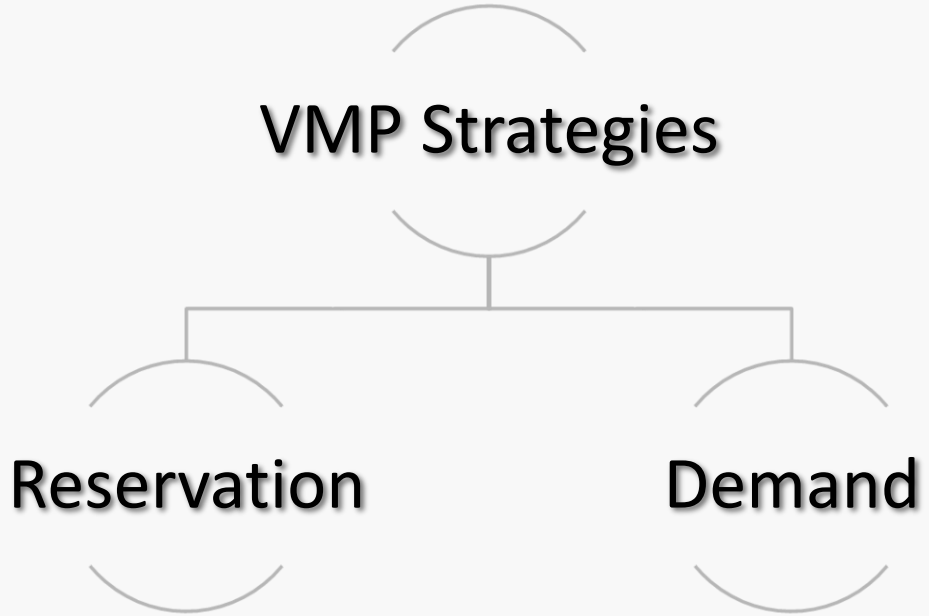


Photo Credit: https://forrst-live.s3.amazonaws.com/posts/snaps/59837/original.jpg

Dynamic Virtual Machine Placement

Dynamic VM placement uses live migrations to reallocate VMs.

On which basis are we going to reallocate VMs?



Dynamic Reservation-based VMP

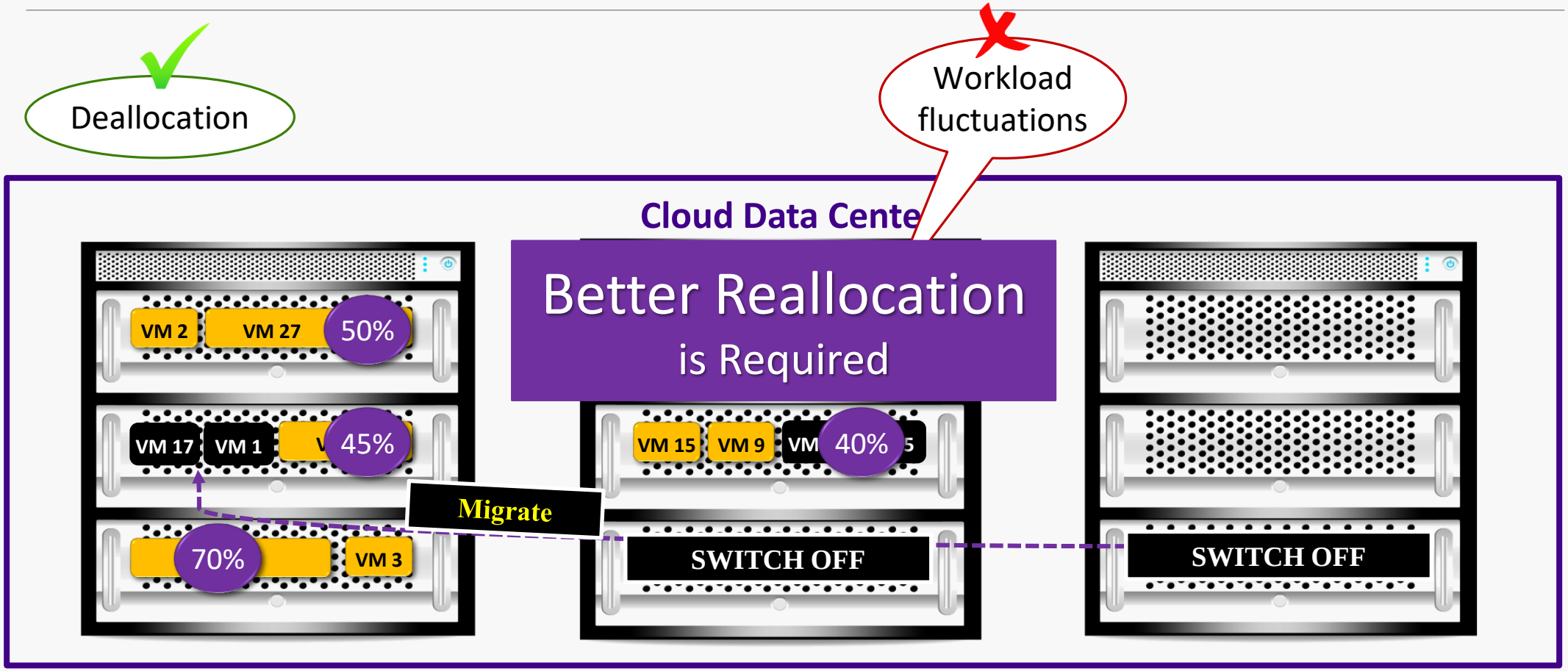


Photo Credit: https://forrst-live.s3.amazonaws.com/posts/snaps/59837/original.jpg

Dynamic Demand-based VMP

More efficient utilization

Energy Savings & minimizing Co2

SLA violations (SLAVs)
- VMs' Migration
- PMs' Overutilization

Demand-based reallocation

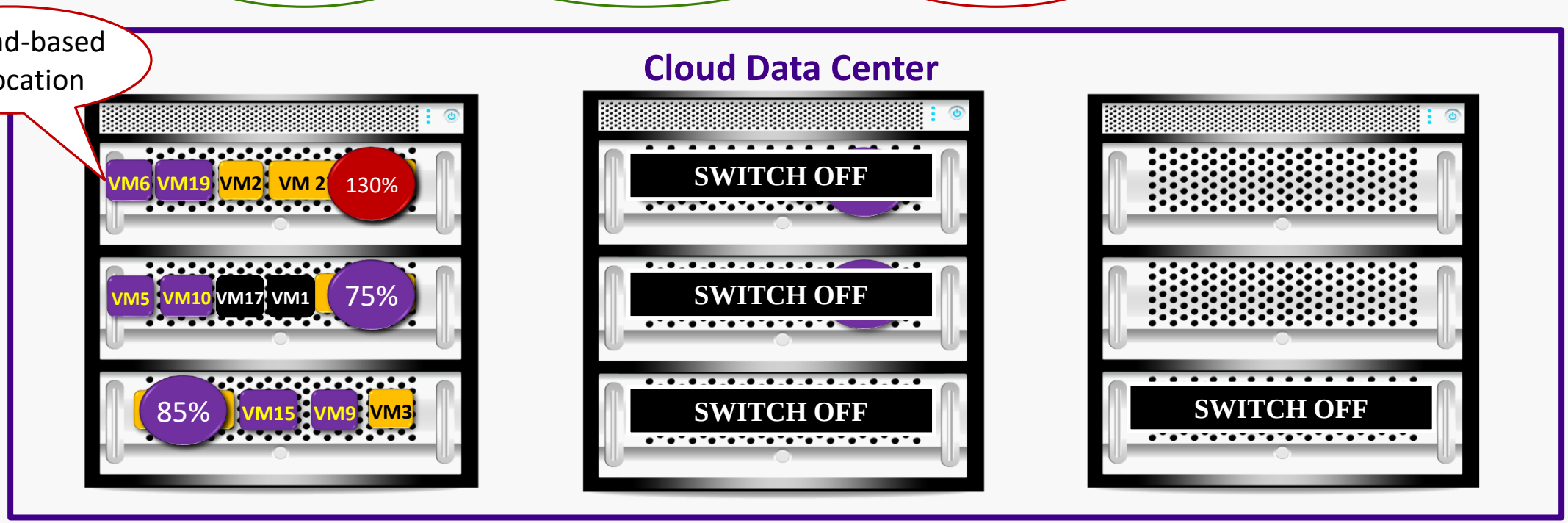


Photo Credit: https://forrst-live.s3.amazonaws.com/posts/snaps/59837/original.jpg

Utilization vs SLAVs

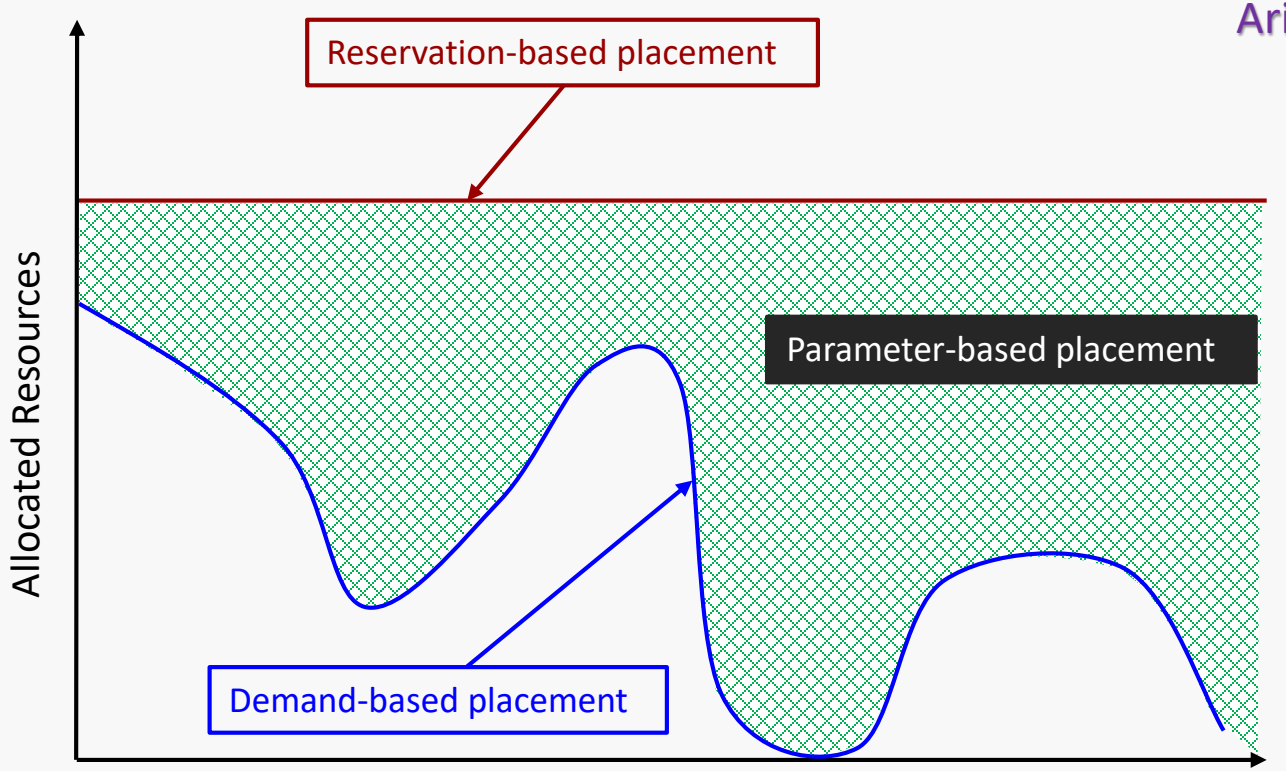
Parameter-based VMP Strategy

- Introduction to Parameter-based VMP
- Static Parameter
- Dynamic Parameter

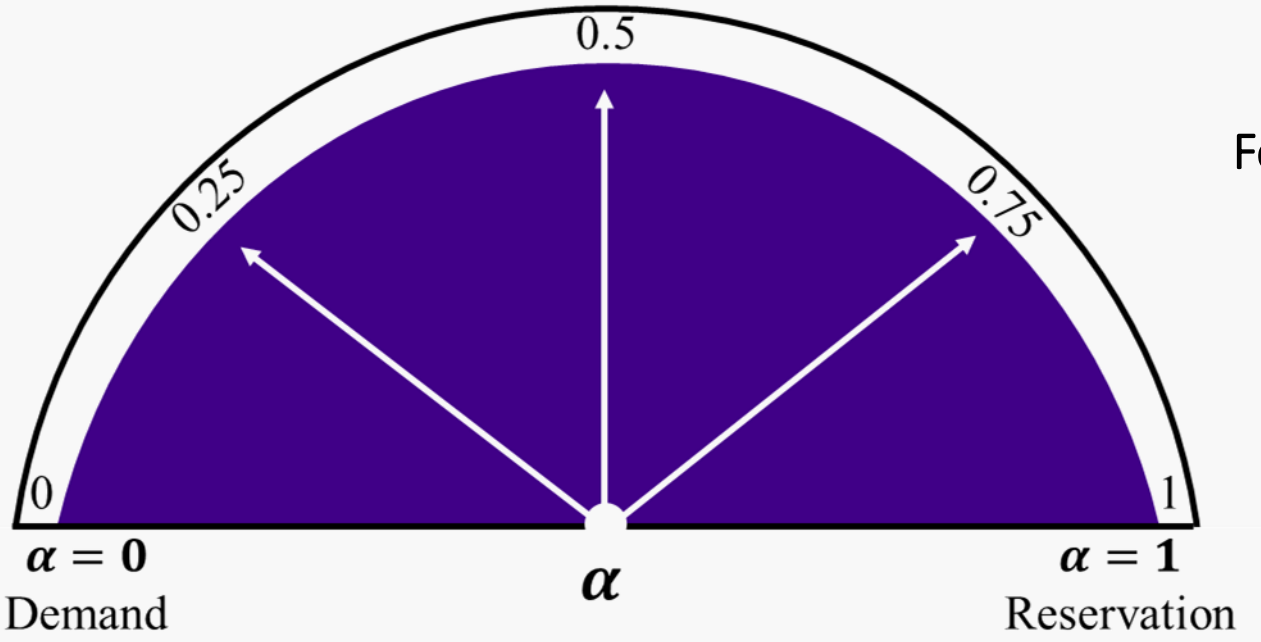
Parameter-based Virtual Machine Placement

"The golden mean or golden middle way is the desirable middle between two extremes, one of excess and the other of deficiency"

Aristotle's philosophy



Parameter-based Virtual Machine Placement



$$a_{jt}^r = \alpha \cdot (c_j^r - d_{jt}^r) + d_{jt}^r$$

Reservation

Demand

For more details, please see these papers:

1. **Mosa, Abdelkhalik, and Rizos Sakellariou. "Virtual machine consolidation for cloud data centers using parameter-based adaptive allocation." Proceedings of the Fifth European Conference on the Engineering of Computer-Based Systems. ACM, 2017. https://dl.acm.org/citation.cfm?id=3123807**

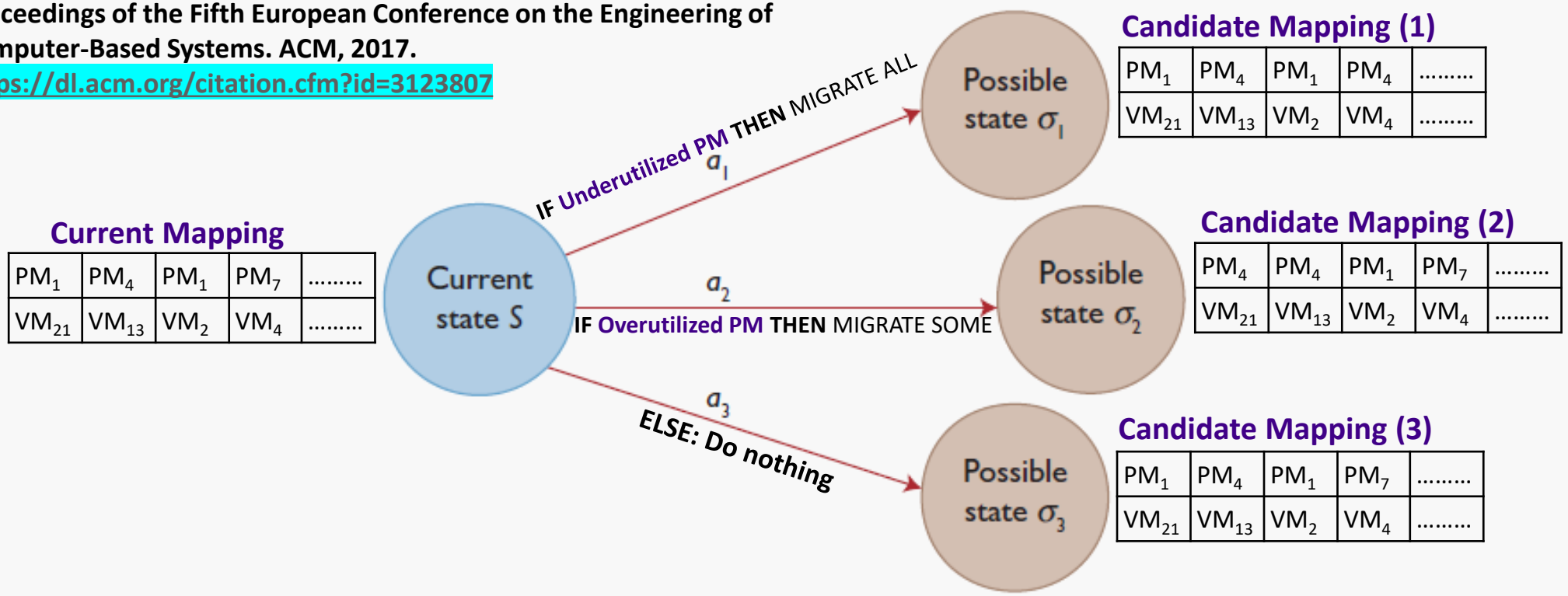
2. **Mosa, Abdelkhalik, and Rizos Sakellariou. "Dynamic Tuning for Parameter-Based Virtual Machine Placement." 2018 17th International Symposium on Parallel and Distributed Computing (ISPDC). IEEE, 2018. https://ieeexplore.ieee.org/abstract/document/8452018**

Virtual Machine Placement Policies

- Action (Rule-based) Policies
- Utility Function Policies

VMP using Action-based Policy

Mosa, Abdelkhalik, and Rizos Sakellariou. "Virtual machine consolidation for cloud data centers using parameter-based adaptive allocation." Proceedings of the Fifth European Conference on the Engineering of Computer-Based Systems. ACM, 2017.
https://dl.acm.org/citation.cfm?id=3123807

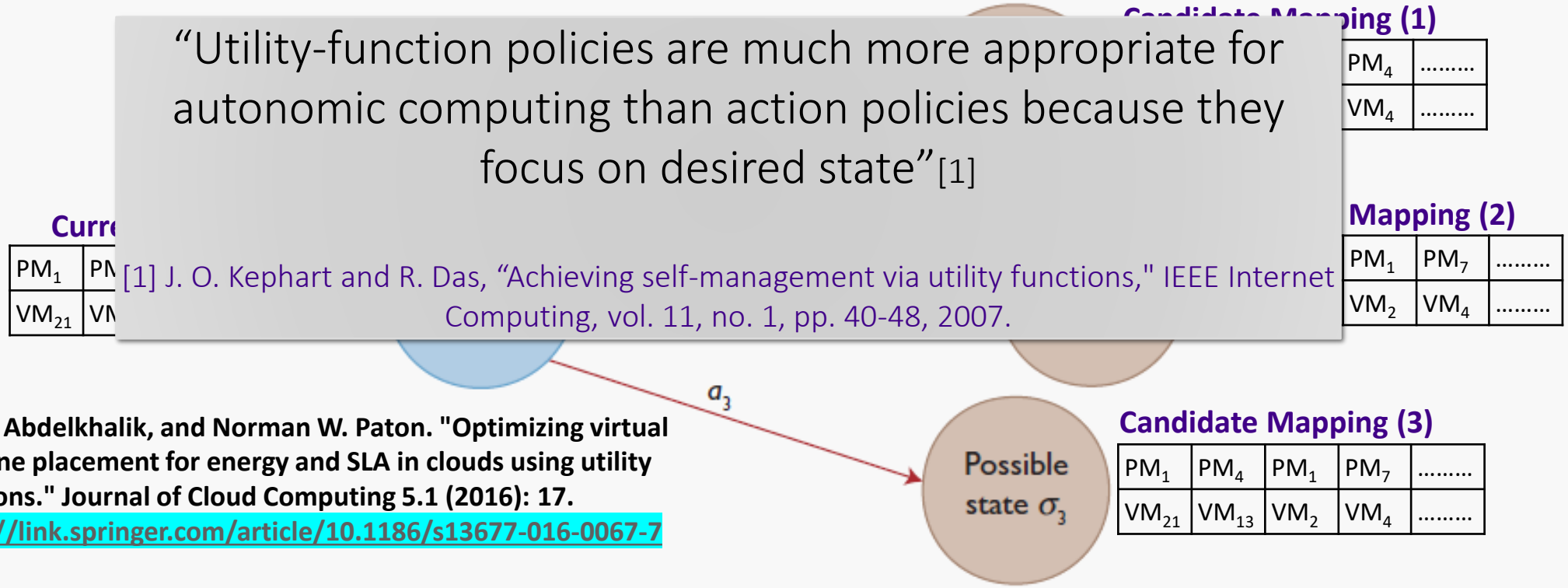


VMP using Utility Functions

- Compute the desired state by selecting the state with the highest utility.
- Requires: Objective (Utility) function + Optimization algorithm.

"Utility-function policies are much more appropriate for autonomic computing than action policies because they focus on desired state"[1]

[1] J. O. Kephart and R. Das, "Achieving self-management via utility functions," IEEE Internet Computing, vol. 11, no. 1, pp. 40-48, 2007.

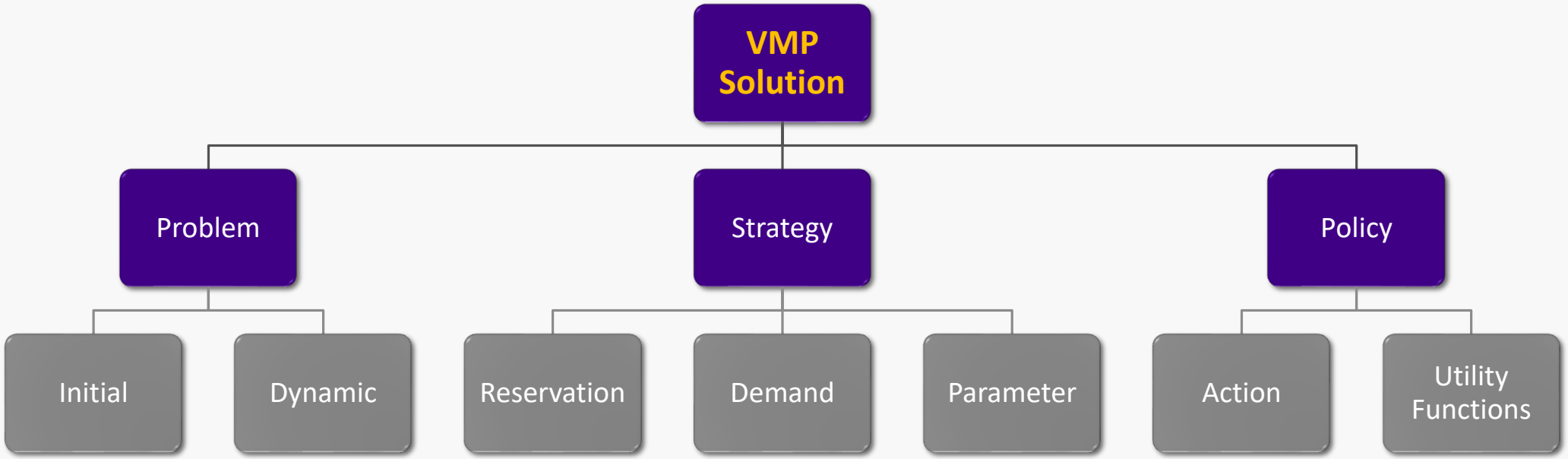


Mosa, Abdelkhalik, and Norman W. Paton. "Optimizing virtual machine placement for energy and SLA in clouds using utility functions." Journal of Cloud Computing 5.1 (2016): 17.
https://link.springer.com/article/10.1186/s13677-016-0067-7

Conclusion

- The Big Picture
- Common Pitfalls
- Possible Research Directions

The Big Picture



Common Pitfalls

- Unclear variant of the VMP (which deployment and service model; objectives).
- Working only on initial VMP.
- Ignoring the on-line vs the off-line nature of the initial and the dynamic VMP.
- Reinventing the wheel.
  - It may count, but does not add a real value.

Possible Research Directions

- Dynamic parameter estimation of the parameter-based VMP strategy.
- Considering multiple resources (CPU, memory, network and storage).
- Considering the hierarchical structure instead of just the flat one.
- Considering migrations among multiple data centers.
- Real-world performance evaluation of existing solutions.
- Developing an evaluation and verification framework
- Considering application-level orchestration (using containers)

Resources

- Mosa, Abdelkhalik, and Norman W. Paton. "Optimizing virtual machine placement for energy and SLA in clouds using utility functions." Journal of Cloud Computing 5.1 (2016): 17.
- Beloglazov, Anton, and Rajkumar Buyya. "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers." Concurrency and Computation: Practice and Experience 24.13 (2012): 1397-1420.
- Mosa, Abdelkhalik, and Rizos Sakellariou. "Virtual machine consolidation for cloud data centers using parameter-based adaptive allocation." Proceedings of the Fifth European Conference on the Engineering of Computer-Based Systems. ACM, 2017.
- Mosa, Abdelkhalik, and Rizos Sakellariou. "Dynamic Tuning for Parameter-Based Virtual Machine Placement." 2018 17th International Symposium on Parallel and Distributed Computing (ISPDC). IEEE, 2018.


THANKS