ITU KALEIDOSCOPE
SANTA FE 2018
Machine learning for a 5G future

Machine Learning Opportunities in Cloud Computing Datacenter Management for 5G Services

Benjamín Barán

National University of the East,

Ciudad del Este, Paraguay

bbaran@pol.una.py

26-28 November
Santa Fe, Argentina

- Introduction and Motivation
- Two-Phase Optimization Scheme for VMP Problems
- Machine Learning Opportunities
- Conclusions and Future Directions

- 5G networks and services will increase exponentially in data traffic, storage and processing.
- Smartphones as gateways to remotely access resources through cloud computing.
- Several challenges should be addressed to further advance cloud computing in order to serve as a basis to integrate 5G components and protocols.
- For cloud computing datacenters, main research challenges could be addressed by designing management solutions based on Machine Learning (ML) techniques.

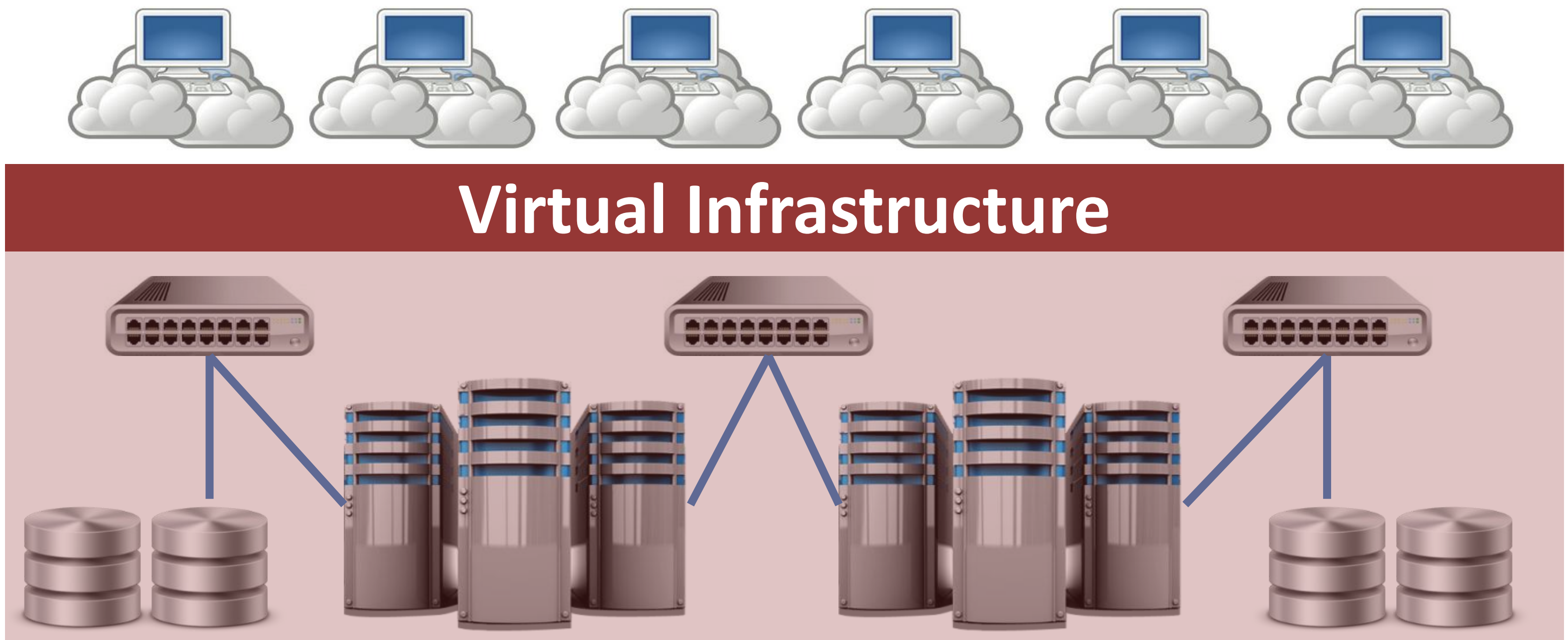
- Large-scale infrastructures providing computational services.
- Energy consumption, carbon emissions, among others.
- Quality of Service
  - High availability
  - Security
  - …
- Resource Management
  - Resource Allocation
  - Resource Adaptation
  - …

Google's Datacenter [www.google.com/about/datacenters]

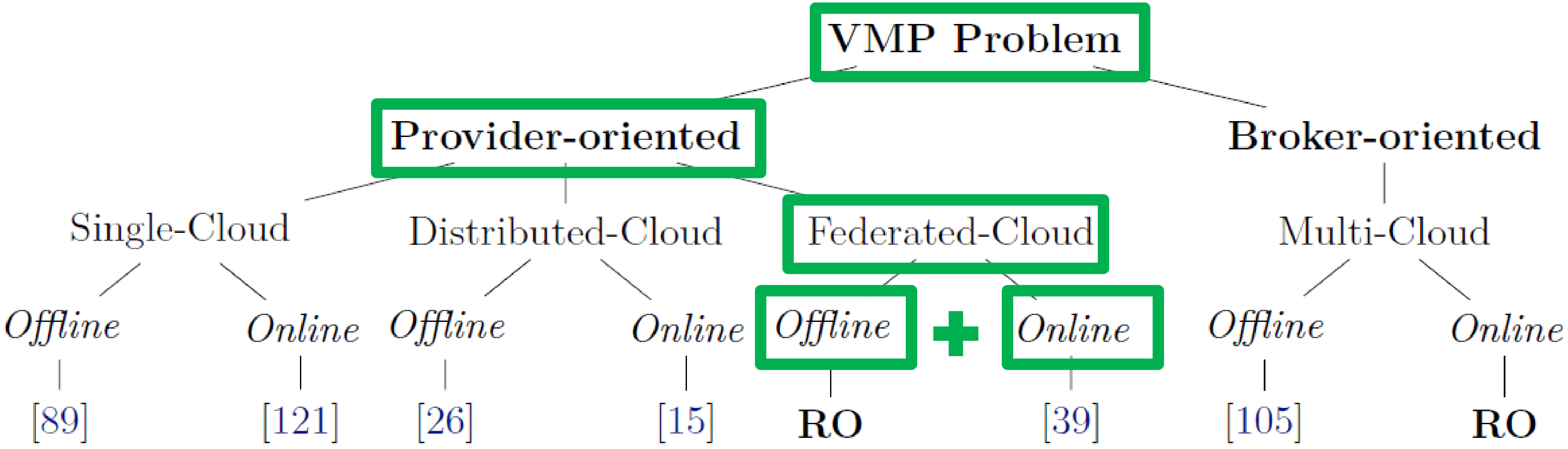
Virtual Machine Placement (VMP)

Which virtual machines should be located at each physical machine?

Virtual Machines (VMs)



Physical Machines (PMs)



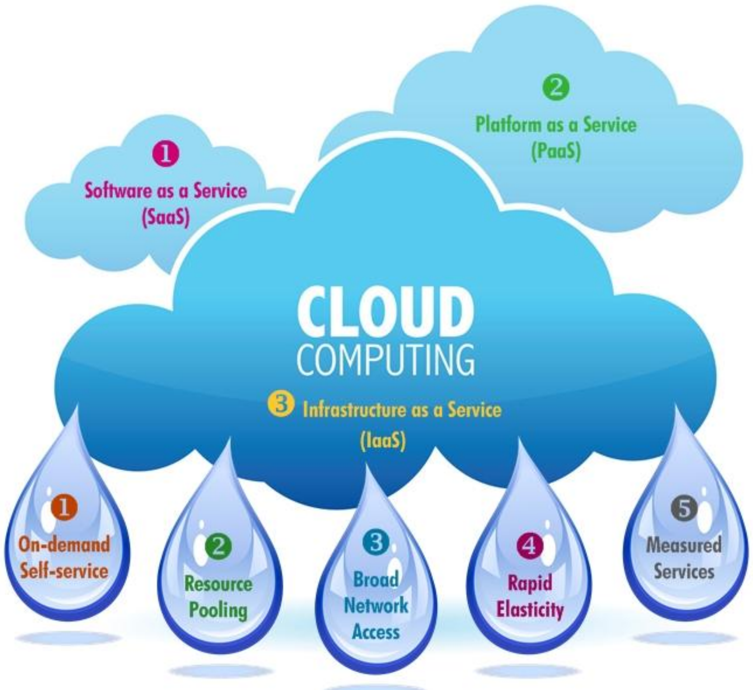
- VMP for Cloud Computing under uncertainty:

  Provider-oriented VMP in Federated-Cloud Deployments.

- VMP for Cloud Computing under uncertainty:

MAM Optimization, Heuristics + Meta-Heuristics and Multiple Objectives.

Technique	Approach	Objective Functions				
		$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$
Deterministic Algorithms	MOP	[53]	[114]	[30]	[15]	[84]
	MAM	[6, 131]	[6, 129]	**RO**	**RO**	[129, 131]
	PMO	**RO**	**RO**	**RO**	**RO**	**RO**
Heuristics	MOP	[39]	[33]	[125]	[55]	[84]
	MAM	[37, 121]	[37, 61]	[29, 61]	[121, 146]	[20, 21]
	PMO	**RO**	**RO**	**RO**	**RO**	**RO**
Meta-Heuristics	MOP	[135]	**RO**	[105]	[142]	**RO**
	MAM	[26, 150]	[26, 126]	[1, 150]	[126]	[1, 150]
	PMO	[51, 89]	[89, 100]	[89, 100]	**RO**	[51]
Approximation Algorithms	MOP	[148]	**RO**	**RO**	**RO**	**RO**
	MAM	[43]	**RO**	**RO**	**RO**	**RO**
	PMO	**RO**	**RO**	**RO**	**RO**	**RO**



Deployment Models

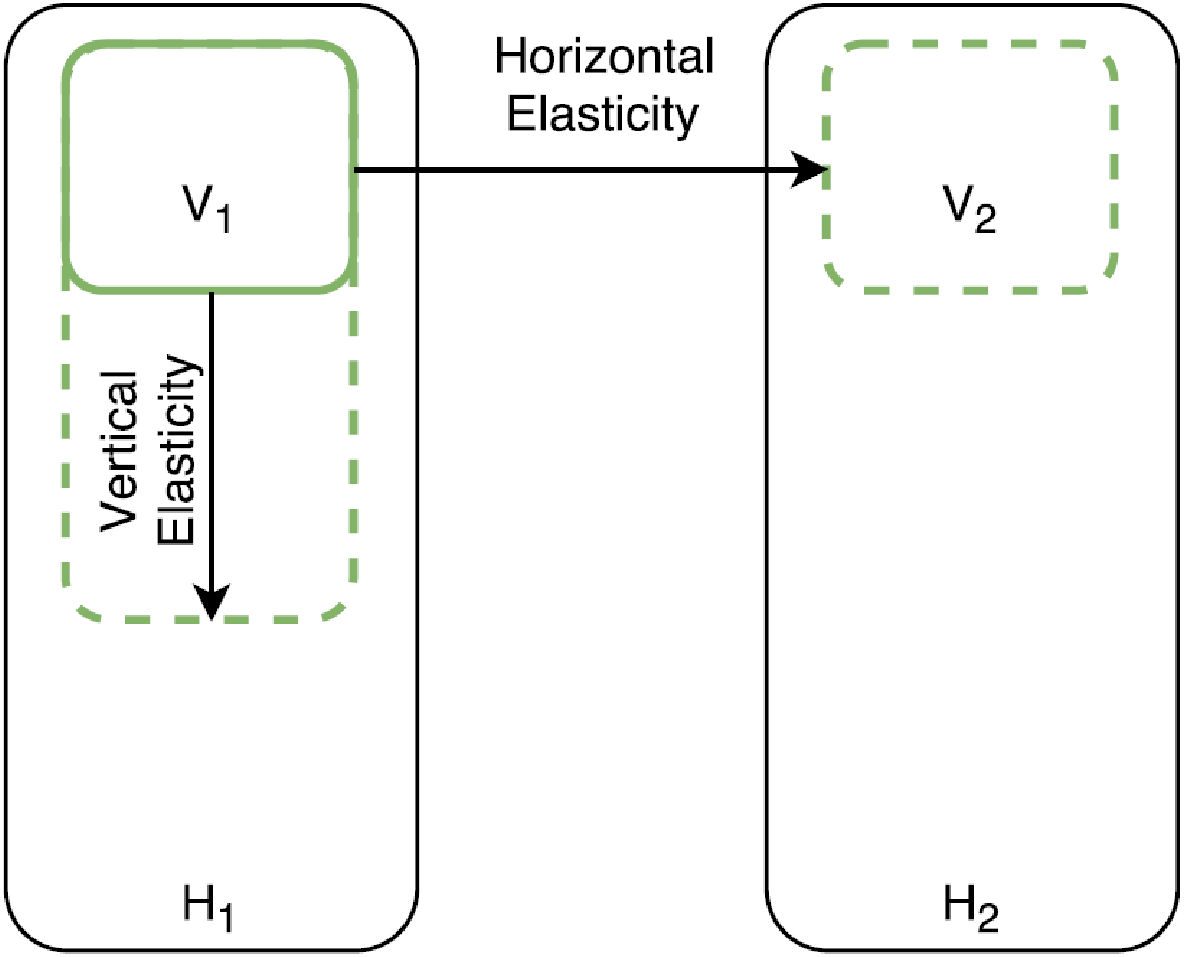
1. Private cloud
2. Community cloud
3. Public cloud
4. Hybrid cloud

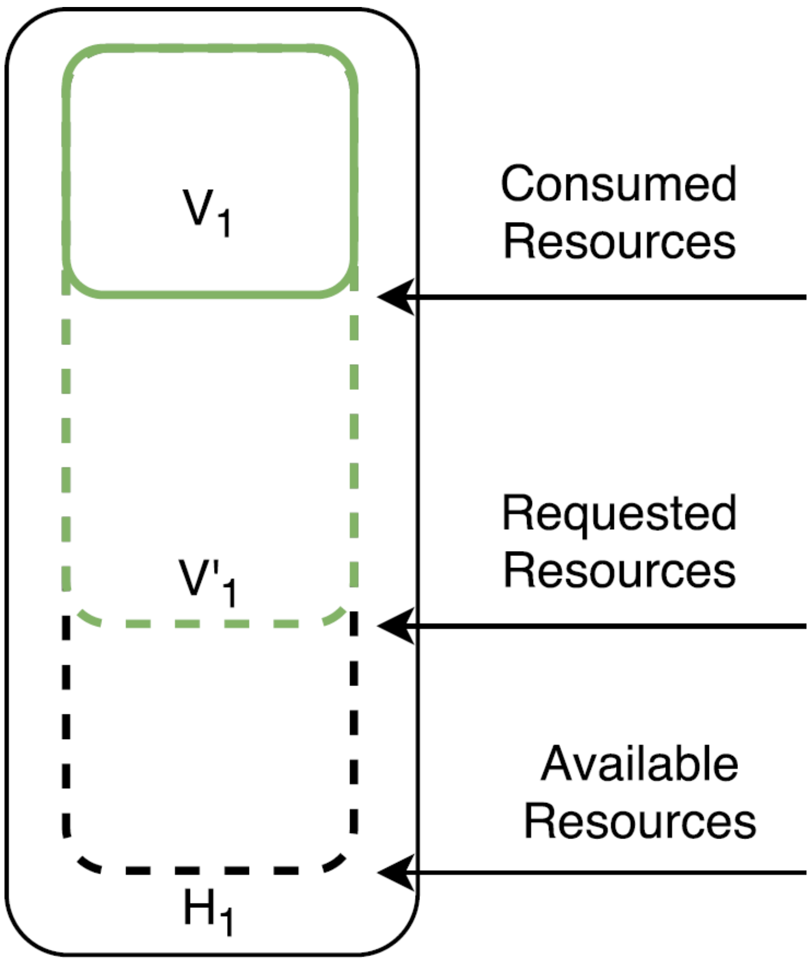
Cloud computing is a **model** for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of **5 essential characteristics**, **3 service models, and 4 deployment models** [107].

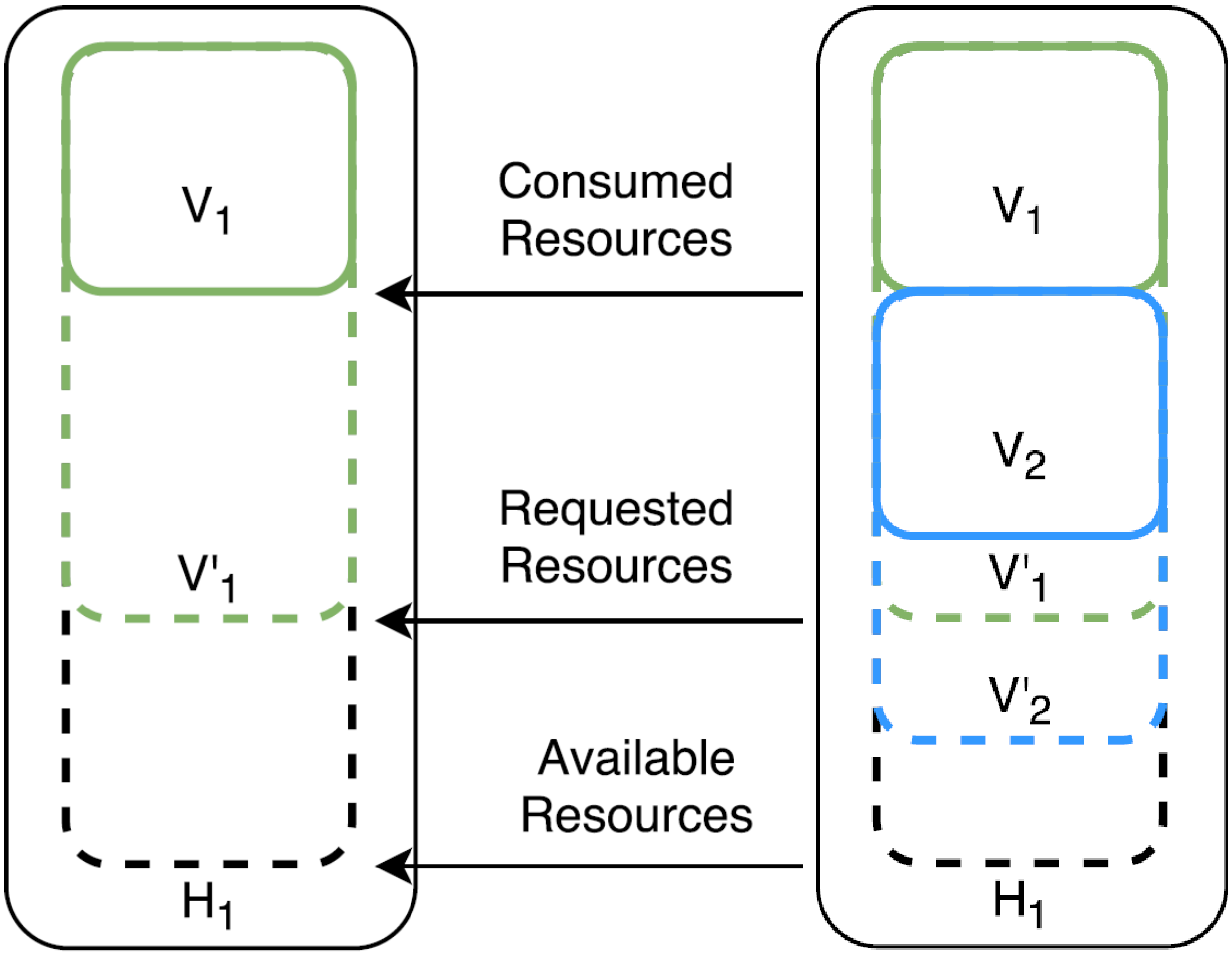
[107] Mell and Grance. “The NIST definition of Cloud Computing”, 2011.

- Placement of Cloud Services, mostly composed by more than just one VM towards a Software-Defined Datacenter.
- Consider relevant characteristics of Infrastructure as a Service model of Cloud Computing for the VMP problem such us:
  - On-Demand Self-Service (dynamic)
  - Rapid Elasticity (dynamic)
- Most relevant dynamic parameters in VMP problems are [112]:
  - Resource Capacities of VMs (vertical elasticity)
  - Number of VMs of a Cloud Service (horizontal elasticity)
  - Utilization of Resources of VMs (overbooking)

[112] Ortigoza, López-Pires and Barán. "A Taxonomy on Dynamic Environments for provider-oriented Virtual Machine Placement" in IC2E


Horizontal
Elasticity
V_1
V_2
Vertical
Elasticity
H_1
H_2


V1
V'1
H1
Consumed Resources
Requested Resources
Available Resources


V1
V'1
H1
Consumed Resources
Requested Resources
Available Resources
V1
V2
V'1
V'2
H1

- IaaS environments for provider-oriented VMP problems could be classified by one or more of the following classification criteria [112]:

Elasticity	Overbooking
0. no elasticity	0. no overbooking
1. horizontal elasticity	1. server resources overbooking
2. vertical elasticity	2. network resources overbooking
3. both horizontal and vertical elasticity	3. both server and network overbooking

- The are 16 possible VMP environments for Cloud Computing.

[112] Ortigoza, López-Pires and Barán. "A Taxonomy on Dynamic Environments for provider-oriented Virtual Machine Placement" in IC2E

- IaaS environments for provider-oriented VMP problems could be classified by one or more of the following classification criteria [112]:

Elasticity	Overbooking
0. no elasticity	0. no overbooking
1. horizontal elasticity	1. server resources overbooking
2. vertical elasticity	2. network resources overbooking
3. both horizontal and vertical elasticity	3. both server and network overbooking

[112] Ortigoza, López-Pires and Barán. "A Taxonomy on Dynamic Environments for provider-oriented Virtual Machine Placement" in IC2E

- A VMP problem formulation for the optimization of:
  1. Power Consumption
  2. Economical Revenue
  3. Quality of Service
  4. Resource Utilization



- Online Algorithms: Heuristics
- Offline Algorithms: Meta-Heuristics

(Online) A1: First-Fit	(Online) A4: First-Fit Decreasing
(Online) A2: Best-Fit	(Online) A5: Best-Fit Decreasing
(Online) A3: Worst-Fit	(Offline) A6: Memetic Algorithm

Table 5.2: Objective Function Costs of Evaluated Algorithms.

Workload \ Algorithm	A1: FF	A2: BF	A3: WF	A4: FFD	A5: BFD	A6: MA
W_1: *Poisson* $\lambda = 10$	3.2927	3.3098	3.5250	3.0205	3.1392	**2.6096**
W_2: *Poisson* $\lambda = 50$	2.4602	2.5112	2.4811	2.4602	2.4555	**2.0001**
W_3: *Poisson* $\lambda = 70$	1.7054	1.6458	1.7054	1.6458	1.6458	**1.3588**
W_4: *Uniform*	3.1875	3.1556	3.0489	3.0907	3.1556	**2.3420**
Average	2.6615	2.6556	2.6901	2.5543	2.5990	**2.0776**
Ranking	5th	4th	6th	2nd	3th	1st

(Online) A1: First-Fit	**(Online) A4: First-Fit Decreasing**
(Online) A2: Best-Fit	**(Online) A5: Best-Fit Decreasing**
(Online) A3: Worst-Fit	**(Offline) A6: Memetic Algorithm**

Table 5.2: Objective Function Costs of Evaluated Algorithms.

Workload \ Algorithm	A1: FF	A2: BF	A3: WF	A4: FFD	A5: BFD	A6: MA
W_1: *Poisson* $\lambda = 10$	3.2927	3.3098	3.5250	3.0205	3.1392	**2.6096**
W_2: *Poisson* $\lambda = 50$	2.4602	2.5112	2.4811	2.4602	2.4555	**2.0001**
W_3: *Poisson* $\lambda = 70$	1.7054	1.6458	1.7054	1.6458	1.6458	**1.3588**
W_4: *Uniform*	3.1875	3.1556	3.0489	3.0907	3.1556	**2.3420**
Average	2.6615	2.6556	2.6901	2.5543	2.5990	**2.0776**
Ranking	5th	4th	6th	2nd	3th	1st

(Online) A1: First-Fit	(Online) A4: First-Fit Decreasing
(Online) A2: Best-Fit	(Online) A5: Best-Fit Decreasing
(Online) A3: Worst-Fit	(Offline) A6: Memetic Algorithm

Table 5.2: Objective Function Costs of Evaluated Algorithms.

Workload \ Algorithm	A1: FF	A2: BF	A3: WF	A4: FFD	A5: BFD	A6: MA
W_1: *Poisson* $\lambda = 10$	3.2927	3.3098	3.5250	3.0205	3.1392	**2.6096**
W_2: *Poisson* $\lambda = 50$	2.4602	2.5112	2.4811	2.4602	2.4555	**2.0001**
W_3: *Poisson* $\lambda = 70$	1.7054	1.6458	1.7054	1.6458	1.6458	**1.3588**
W_4: *Uniform*	3.1875	3.1556	3.0489	3.0907	3.1556	**2.3420**
Average	2.6615	2.6556	2.6901	2.5543	2.5990	**2.0776**
Ranking	5th	4th	6th	2nd	3th	1st

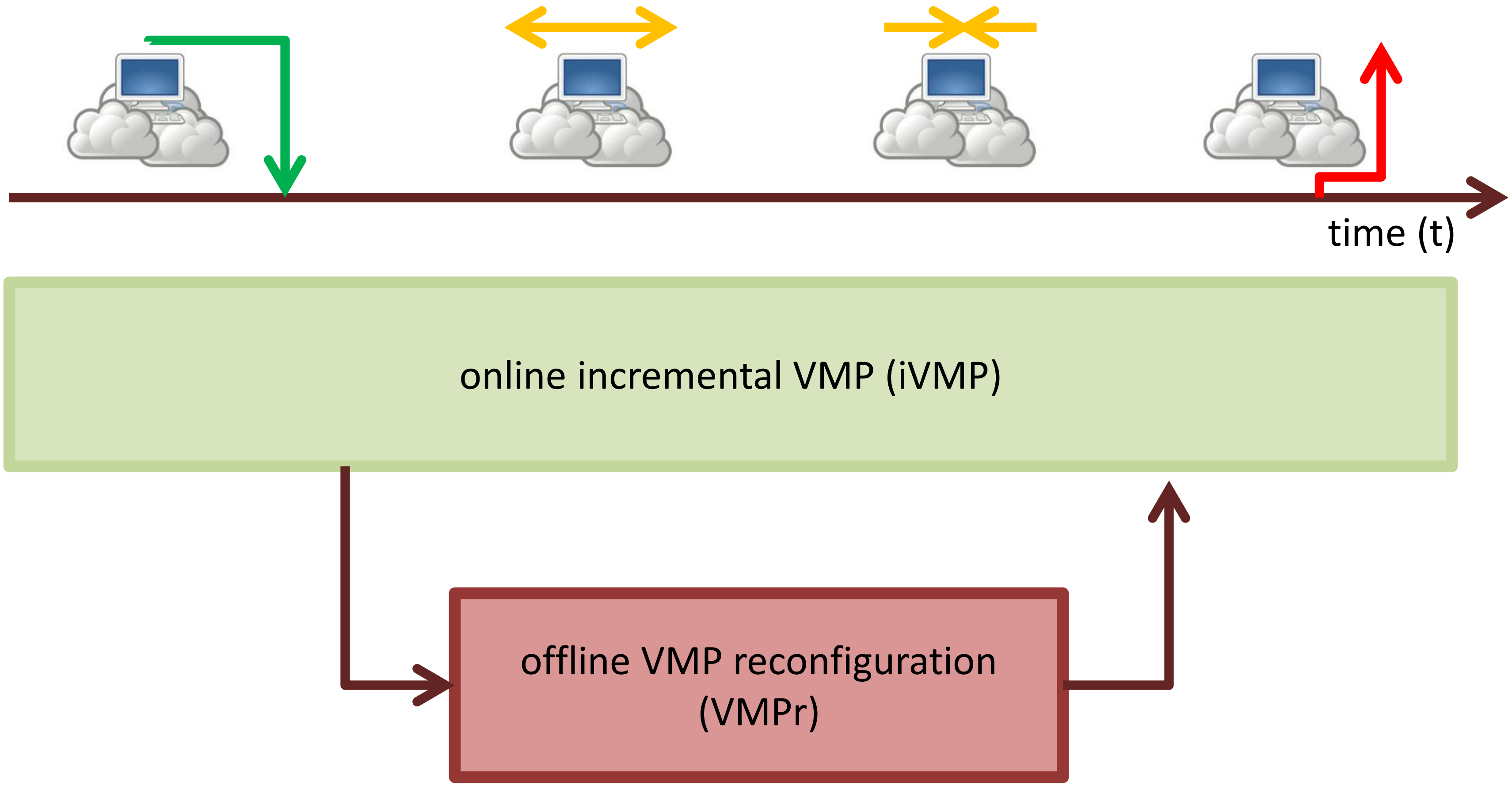
Algorithm	Advantages	Disadvantages
offline meta-heuristic (A6)	• better quality of solutions when comparing to online alternatives.	• not appropriate for highly dynamic environments of VMP problems.
online heuristics (A1 – A5)	• short execution time • unknown cloud service requests are considered.	• online decisions negatively affects the quality of solutions.

Algorithm	Advantages	Disadvantages
offline meta-heuristic (A6)	• better quality of solutions when comparing to online alternatives.	• not appropriate for highly dynamic environments of VMP problems.
online heuristics (A1 – A5)	• short execution time • unknown cloud service requests are considered.	• online decisions negatively affects the quality of solutions.

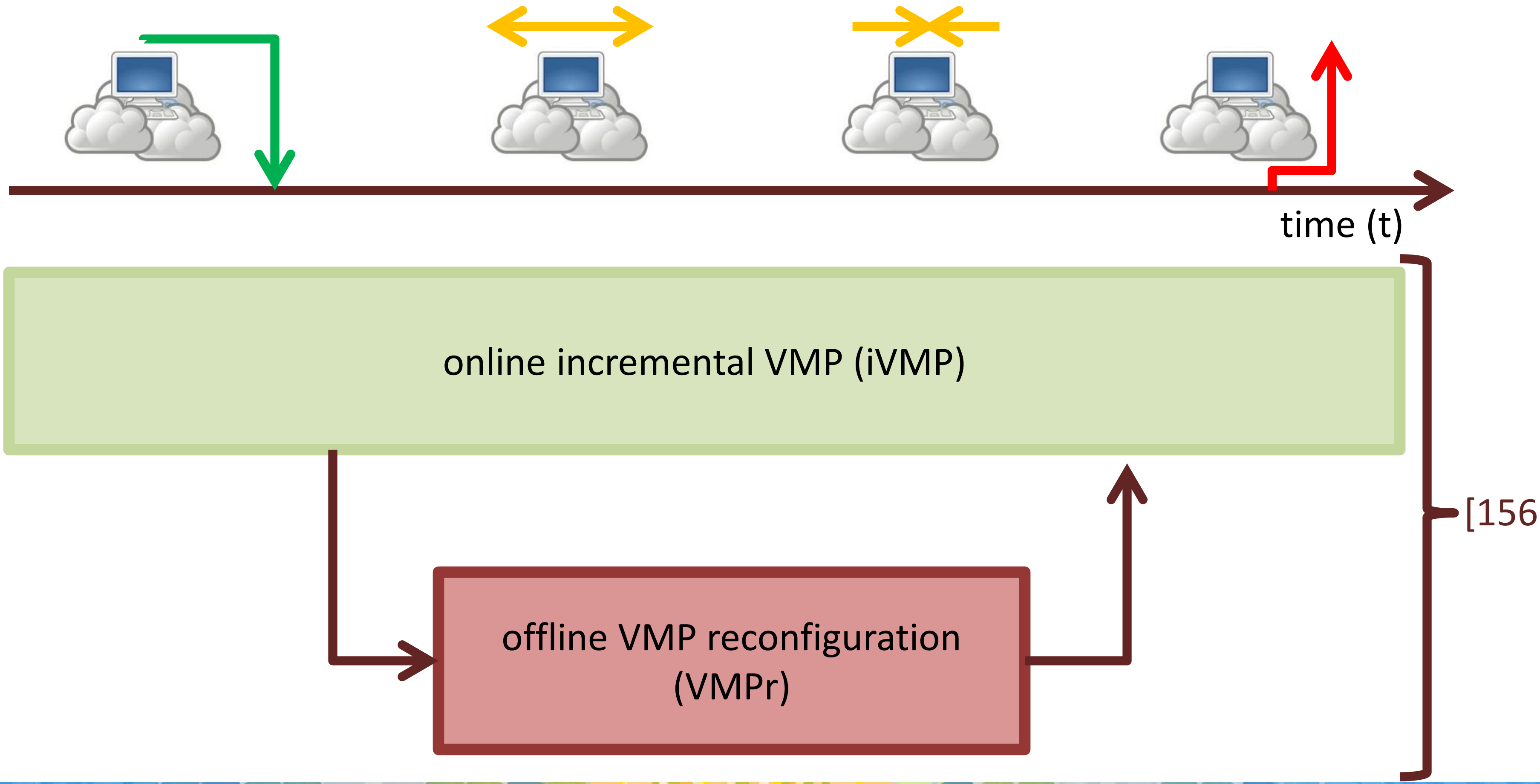
Algorithm	Advantages	Disadvantages
offline meta-heuristic (A6)	• better quality of solutions when comparing to online alternatives.	• not appropriate for highly dynamic environments of VMP problems.
online heuristics (A1 – A5)	• short execution time • unknown cloud service requests are considered.	• online decisions negatively affects the quality of solutions.

Algorithm	Advantages	Disadvantages
offline meta-heuristic (A6)	• better quality of solutions when comparing to online alternatives.	• not appropriate for highly dynamic environments of VMP problems.
+		
online heuristics (A1 – A5)	• short execution time • unknown cloud service requests are considered.	• online decisions negatively affects the quality of solutions.

IaaS environments for provider-oriented VMP problems



IaaS environments for provider-oriented VMP problems



[156] Zheng, et al. "Virtual machine placement based on multi-objective biogeography-based optimization" in Future Generation Computer Systems, 2015.

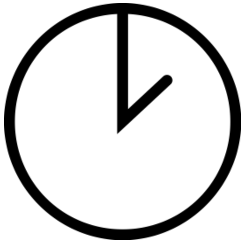
- IaaS environments for provider-oriented VMP problems could be classified by one or more of the following classification criteria [112]:

Elasticity	Overbooking
0. no elasticity	0. no overbooking
1. horizontal elasticity	1. server resources overbooking
2. vertical elasticity	2. network resources overbooking
3. both horizontal and vertical elasticity	3. both server and network overbooking

[112] Ortigoza, López-Pires and Barán. "A Taxonomy on Dynamic Environments for provider-oriented Virtual Machine Placement" in IC2E

- A VMP problem formulation for the optimization of:
  1. Power Consumption
  2. Economical Revenue
  3. Resource Utilization
  4. Reconfiguration Time





- Uncertain parameters considered:
  1. Virtual resources capacities (vertical elasticity)
  2. Number of VMs in cloud services (horizontal elasticity)
  3. Utilization of virtual CPU / RAM (server overbooking)
  4. Utilization of networking virtual (network overbooking)

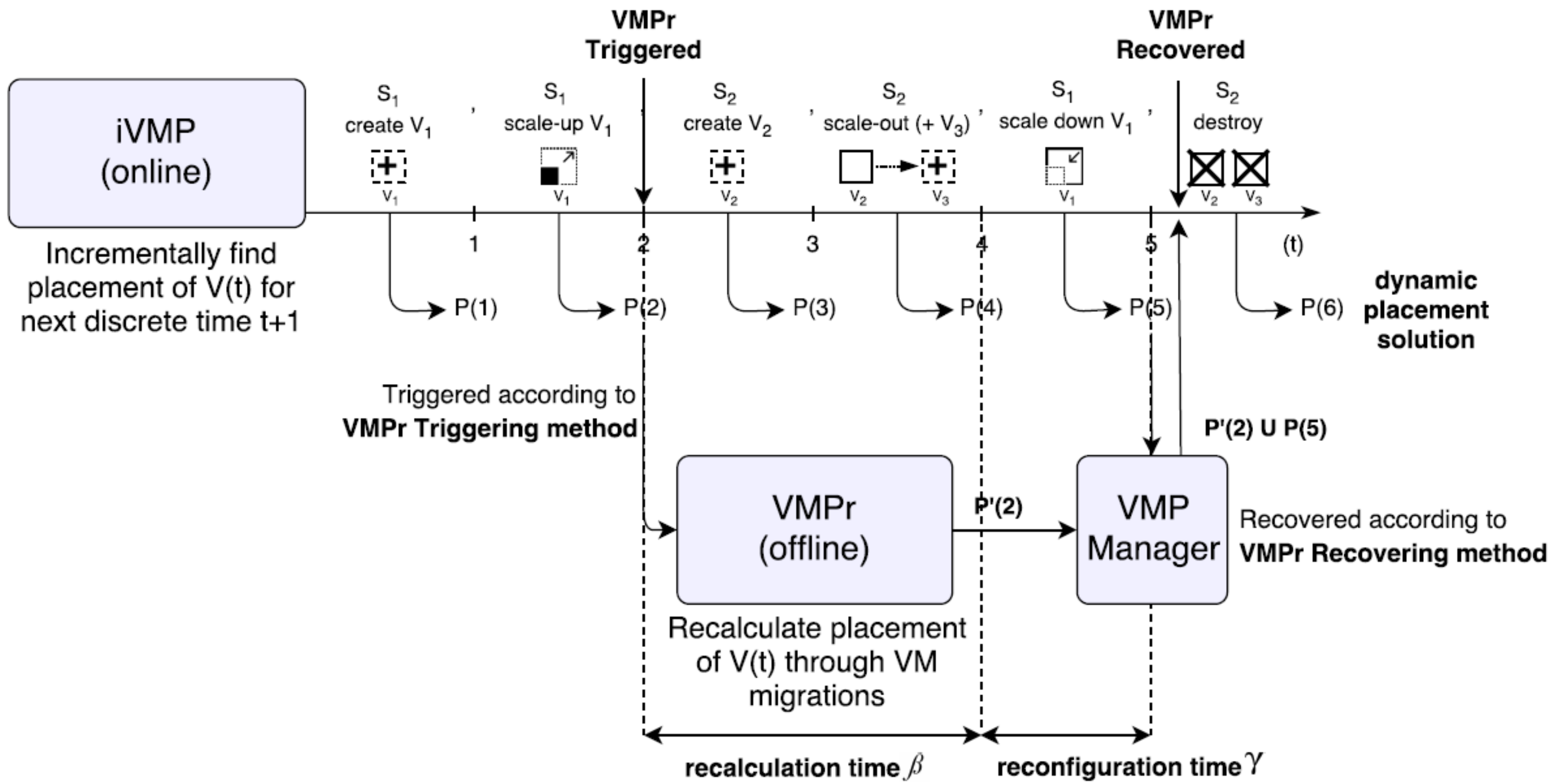


Figure 1: Two-phase optimization scheme for VMP problems considered in this work, presenting a basic example with a placement recalculation time of $\beta = 2$ (from $t = 2$ to $t = 4$) and a placement reconfiguration time of $\gamma = 1$ (from $t = 4$ to $t = 5$).

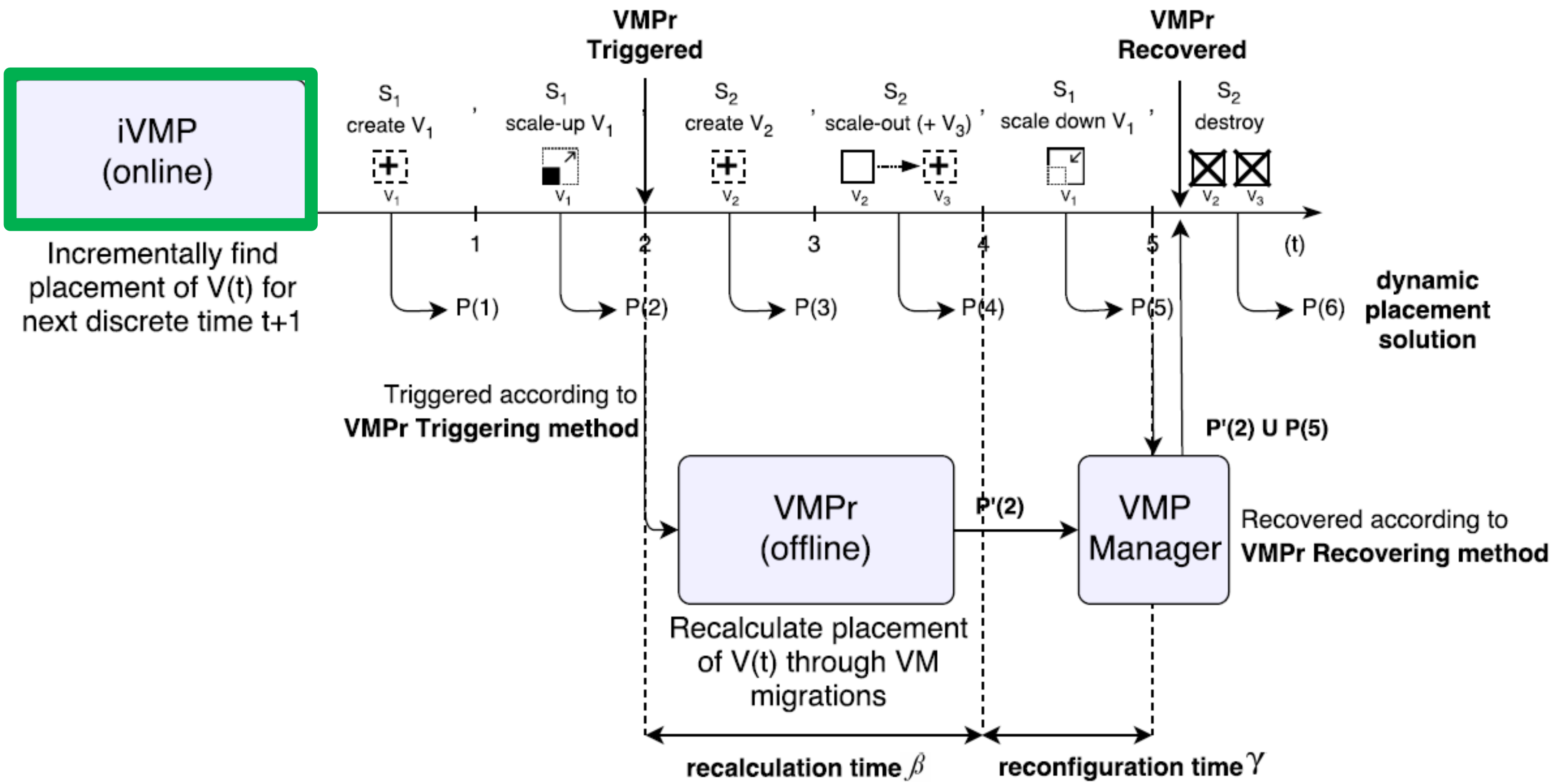


Figure 1: Two-phase optimization scheme for VMP problems considered in this work, presenting a basic example with a placement recalculation time of $\beta = 2$ (from $t = 2$ to $t = 4$) and a placement reconfiguration time of $\gamma = 1$ (from $t = 4$ to $t = 5$).

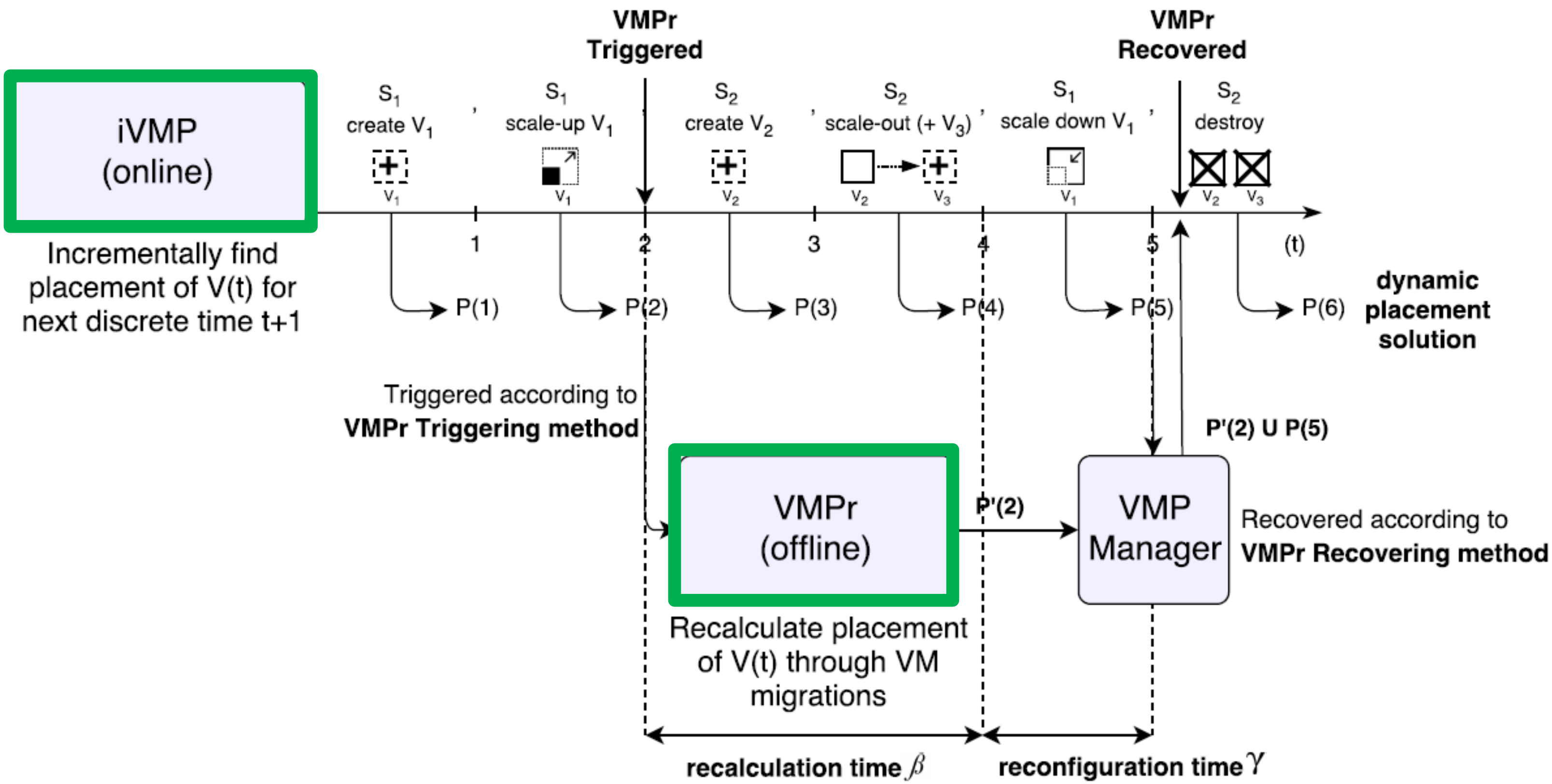


Figure 1: Two-phase optimization scheme for VMP problems considered in this work, presenting a basic example with a placement recalculation time of $\beta = 2$ (from $t = 2$ to $t = 4$) and a placement reconfiguration time of $\gamma = 1$ (from $t = 4$ to $t = 5$).

when or under which circumstances the VMPr phase should be triggered?

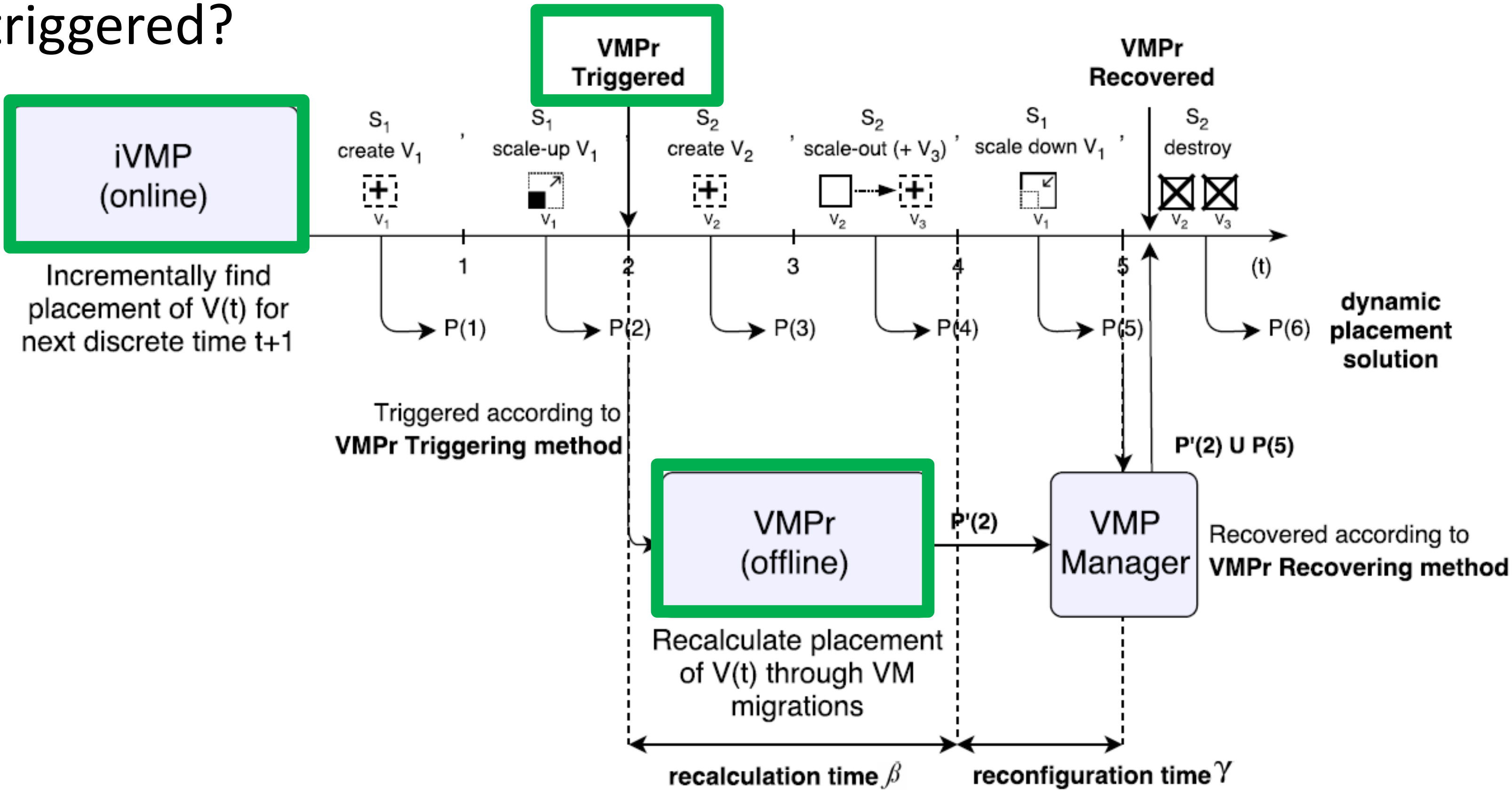


Figure 1: Two-phase optimization scheme for VMP problems considered in this work, presenting a basic example with a placement recalculation time of $\beta = 2$ (from $t = 2$ to $t = 4$) and a placement reconfiguration time of $\gamma = 1$ (from $t = 4$ to $t = 5$).

what should be done with cloud service requests arriving during recalculation time?

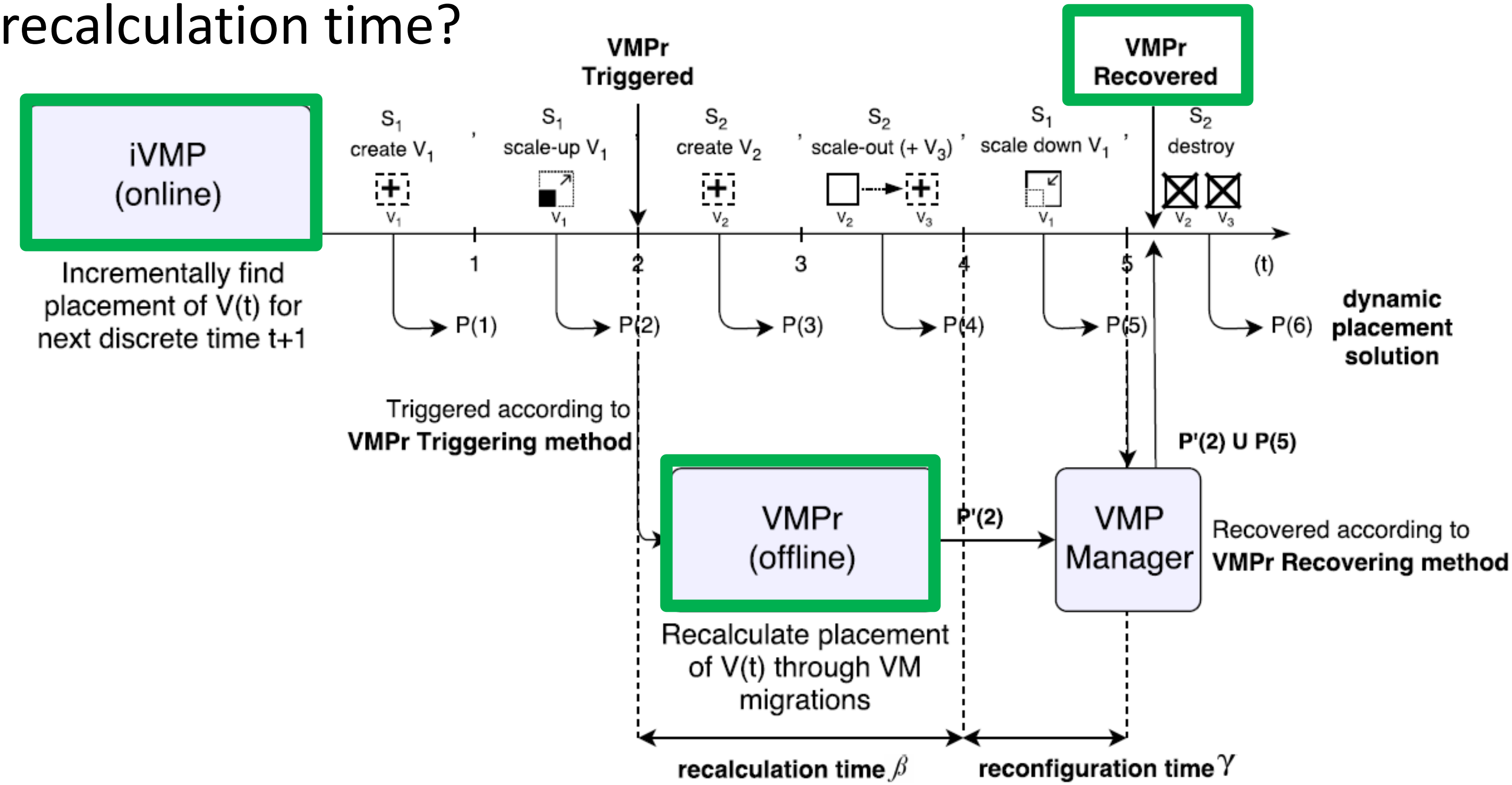


Figure 1: Two-phase optimization scheme for VMP problems considered in this work, presenting a basic example with a placement recalculation time of $\beta = 2$ (from $t = 2$ to $t = 4$) and a placement reconfiguration time of $\gamma = 1$ (from $t = 4$ to $t = 5$).

Reference	Overbooking Type	Elasticity Type	VMPr Triggering	VMPr Recovering
[20]	CPU	Not Considered	Periodically	Cancellation
[151]	Not Considered	Not Considered	Periodically	Not Considered
[46]	Not Considered	Not Considered	Periodically	Not Considered
[87]	Not Considered	Not Considered	Periodically	Not Considered
[44]	CPU and RAM	Not Considered	Periodically	Not Considered
[156]	Not Considered	Not Considered	Periodically	Not Considered
[132]	Not Considered	Not Considered	Continuously	Not Considered
[13]	CPU	Not Considered	Threshold-based	N/A
[123]	CPU, RAM and Network	Not Considered	Threshold-based	N/A
[137]	CPU	Horizontal	Threshold-based	N/A
This work	*CPU, RAM and Network*	*Vertical and Horizontal*	*Prediction-based*	*Update-based*

Reference	Overbooking Type	Elasticity Type	VMPr Triggering	VMPr Recovering
[20]	CPU	Not Considered	Periodically	Cancellation
[151]	Not Considered	Not Considered	Periodically	Not Considered
[46]	Not Considered	Not Considered	Periodically	Not Considered
[87]	Not Considered	Not Considered	Periodically	Not Considered
[44]	CPU and RAM	Not Considered	Periodically	Not Considered
[156]	Not Considered	Not Considered	Periodically	Not Considered
[132]	Not Considered	Not Considered	Continuously	Not Considered
[13]	CPU	Not Considered	Threshold-based	N/A
[123]	CPU, RAM and Network	Not Considered	Threshold-based	N/A
[137]	CPU	Horizontal	Threshold-based	N/A
This work	*CPU, RAM and Network*	*Vertical and Horizontal*	*Prediction-based*	*Update-based*

Reference	Overbooking Type	Elasticity Type	VMPr Triggering	VMPr Recovering
[20]	CPU	Not Considered	Periodically	Cancellation
[151]	Not Considered	Not Considered	Periodically	Not Considered
[46]	Not Considered	Not Considered	Periodically	Not Considered
[87]	Not Considered	Not Considered	Periodically	Not Considered
[44]	CPU and RAM	Not Considered	Periodically	Not Considered
[156]	Not Considered	Not Considered	Periodically	Not Considered
[132]	Not Considered	Not Considered	Continuously	Not Considered
[13]	CPU	Not Considered	Threshold-based	N/A
[123]	CPU, RAM and Network	Not Considered	Threshold-based	N/A
[137]	CPU	Horizontal	Threshold-based	N/A
This work	*CPU, RAM and Network*	*Vertical and Horizontal*	*Prediction-based*	*Update-based*

A contribution:

A novel prediction-based VMPr Triggering method to decide when or under what circumstances the VMPr phase should be triggered.

- Considers main weaknesses of existing VMPr Triggering methods.
- Considers Double Exponential Smoothing (DES) [63] as a statistical technique for predicting values of the objective functions.
- Predicts the next values of objective functions and triggers the VMPr phase in case objectives are predicted to consistently increase (in a minimization context).

[63] Huang et al. "Resource prediction based on double exponential smoothing in cloud computing" in CECNet 2012.

- **ML for VMPr Triggering:**
  - RQ 1: which ML techniques could be considered more appropriate for VMPr Triggering methods?
  - RQ 2: how important is to accurately predict when to trigger a VMPr phase in VMP problems?
  - RQ 3: rather than predicting future objective function values, what other parameters could be evaluated for VMPr Triggering methods?

- **ML for Network Management:**
  - RQ 4: which ML techniques could be considered more appropriate for predicting Network Routing Reconfiguration (NRR) as part of VMP problems in SDN implementations?
  - RQ 5: which ML techniques could be considered more appropriate for clustering VMs for supporting placement decisions?

ITU KALEIDOSCOPE
SANTA FE 2018

Thank you