TA24

DRS Deep Dive and Technology Preview of Distributed Power Management

Minwen Ji

Staff Engineer

VMware

This session may contain product features that are currently under development.

This session/overview of the new technology represents no commitment from VMware to deliver these features in any generally available product.

Features are subject to change and must not be included in contracts, purchase orders, or sales agreements of any kind.

Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new technologies or features discussed or presented have not been determined.

Overview

- **What DRS does**
  - DRS stands for Distributed Resource Scheduler
  - Maps cluster-wide SLA to host-level SLA
  - Recommends initial placement of VMs
  - Recommends dynamic load balancing with VMotion
- **What DPM does**
  - DPM stands for Distributed Power Management
  - Recommends powering off hosts to save energy
  - Recommends powering on hosts when workload increases
  - Recommends powering on hosts for certain events

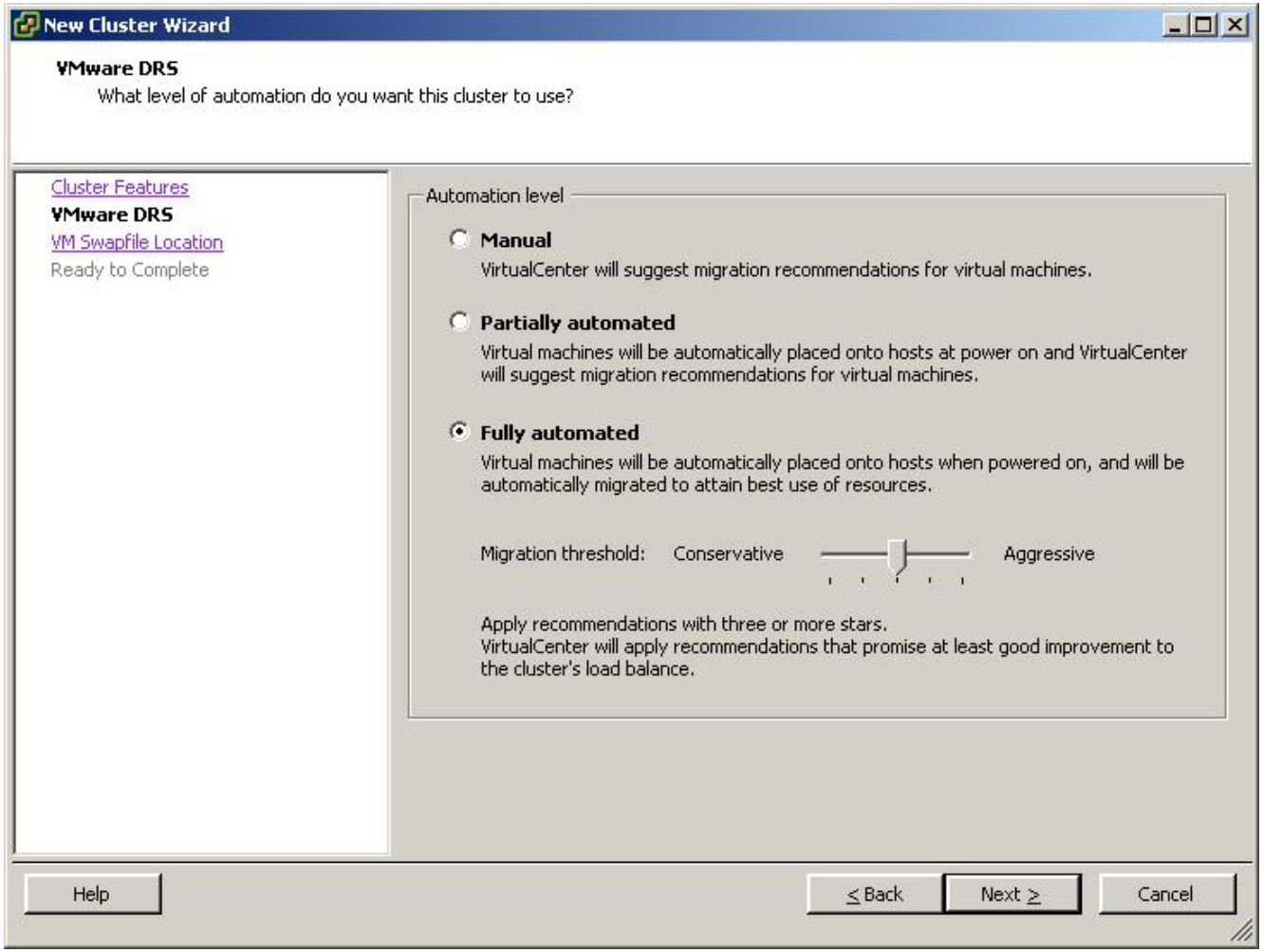
How to Influence DRS/DPM Decisions

- **Many ways to influence the decisions of DRS/DPM:**
- **Service Level Agreement**
  - e.g., reservations, shares and limits
- **Migration threshold**
- **Automation level**
- **HA (High Availability) requirements**
- **Affinity / anti-affinity rules**
- **Some advanced options**

Outline

- **Create and configure a DRS cluster**
- Compound and dependent recommendations
- Enable DPM – power recommendations
- Power on VMs – host recommendations
- Maintain failover level for High Availability (HA)
- Keep VMs together or separate them – affinity rules

Configure DRS Clusters



Migration Threshold

- **Reflects tolerance of cluster load imbalance**
  - Aggressive: less tolerance and more migrations
  - Conservative: more tolerance and fewer migrations
- **Internal load imbalance metric**
  - Standard deviation of host load/capacity
- **Internal load imbalance threshold**
  - Proportional to external threshold (stars) / sqrt (number of hosts)
    - Consistent across clusters with different numbers of hosts
    - Consistent across clusters with different numbers of VMs
    - Effective for a wide range of loads

Load Balancing Algorithm

While (load imbalance metric > threshold) {

Select the best migration in the current state;

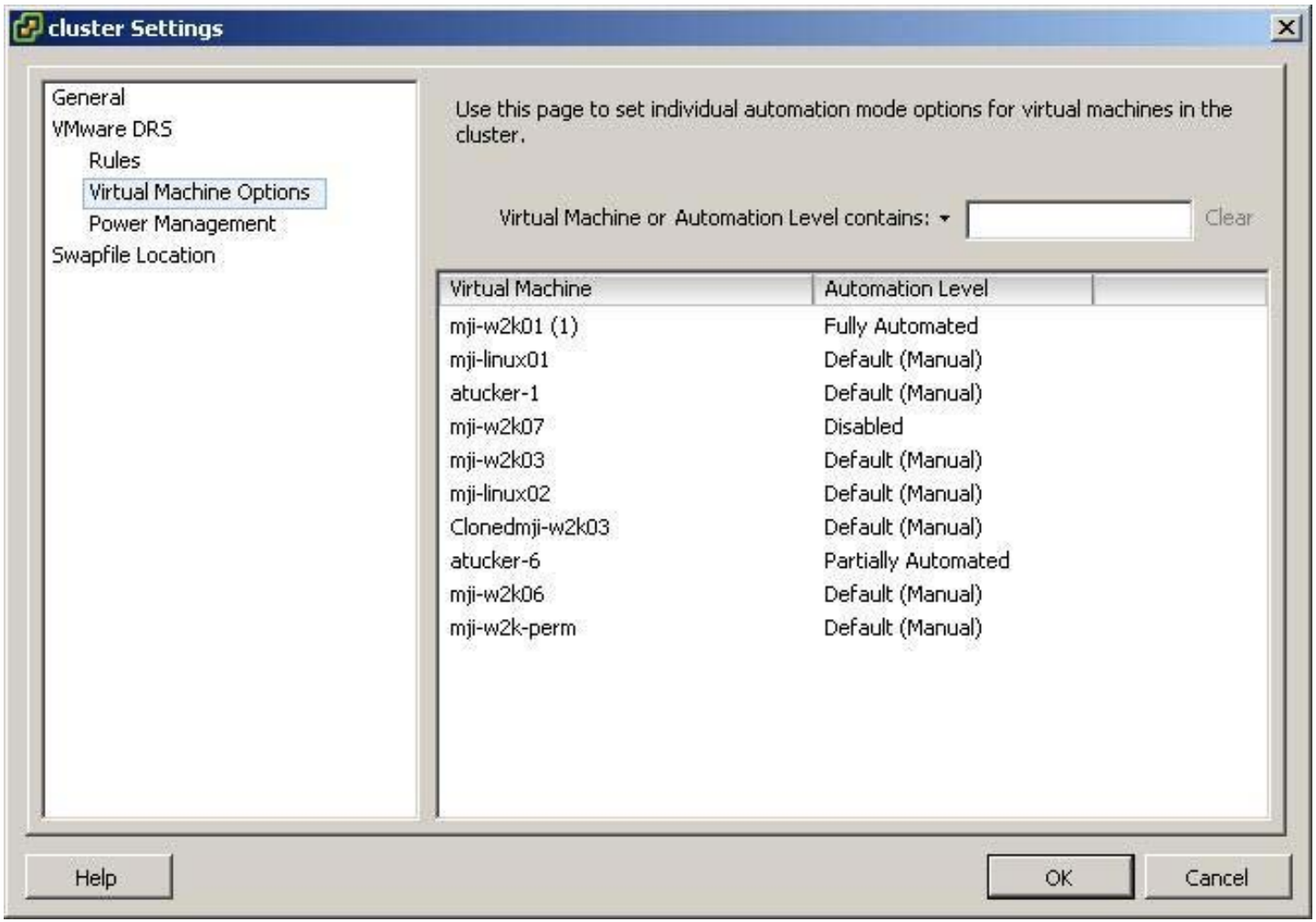
If no good migration is found, stop;

Else recommend the best migration;

Update cluster to the state after the recommended migration is executed;

}

VM Automation Level



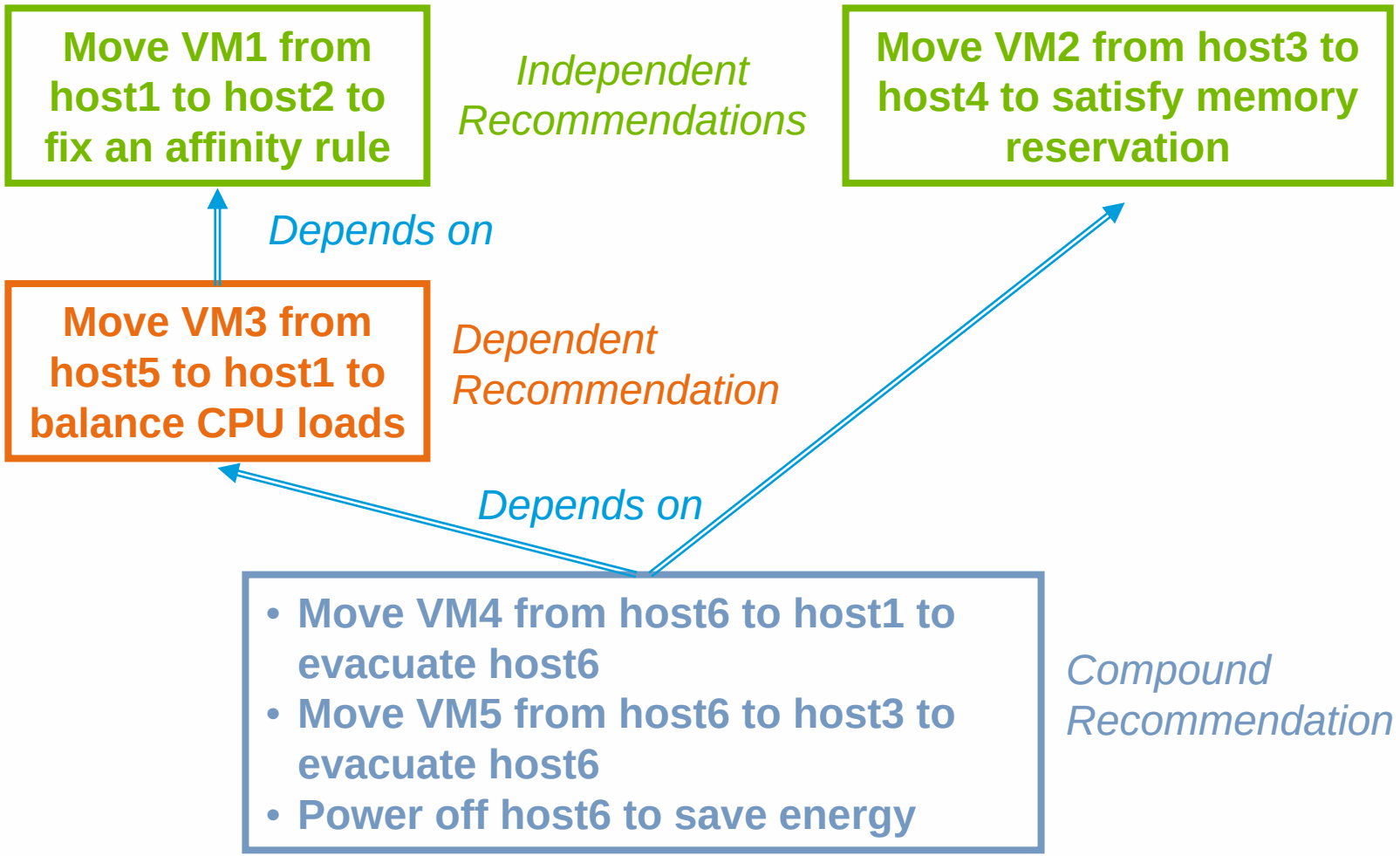
VM Automation Levels

- **Manual VMs**
  - User needs migrate privilege on manual VMs in order to migrate them.
- **Automatic VMs**
  - DRS can migrate automatic VMs without user's privilege or approval.
- **In selecting migration candidates, DRS prefers automatic VMs to manual ones**

Outline

- Create and configure a DRS cluster
- **Compound and dependent recommendations**
- Enable DPM – power recommendations
- Power on VMs – host recommendations
- Maintain failover level for High Availability (HA)
- Keep VMs together or separate them – affinity rules

Workflow of Recommendations



Attributes of Recommendations

- **Automatic vs. manual recommendations**
  - A recommendation is manual if and only if any VM or host involved in it has a manual automation level
- **Compound recommendations**
  - Have multiple actions, all of which must be executed atomically
- **Independent vs. dependent recommendations**
  - Independent ones can start execution as soon as approved
  - Dependent ones cannot start execution until the ones they depend on complete execution

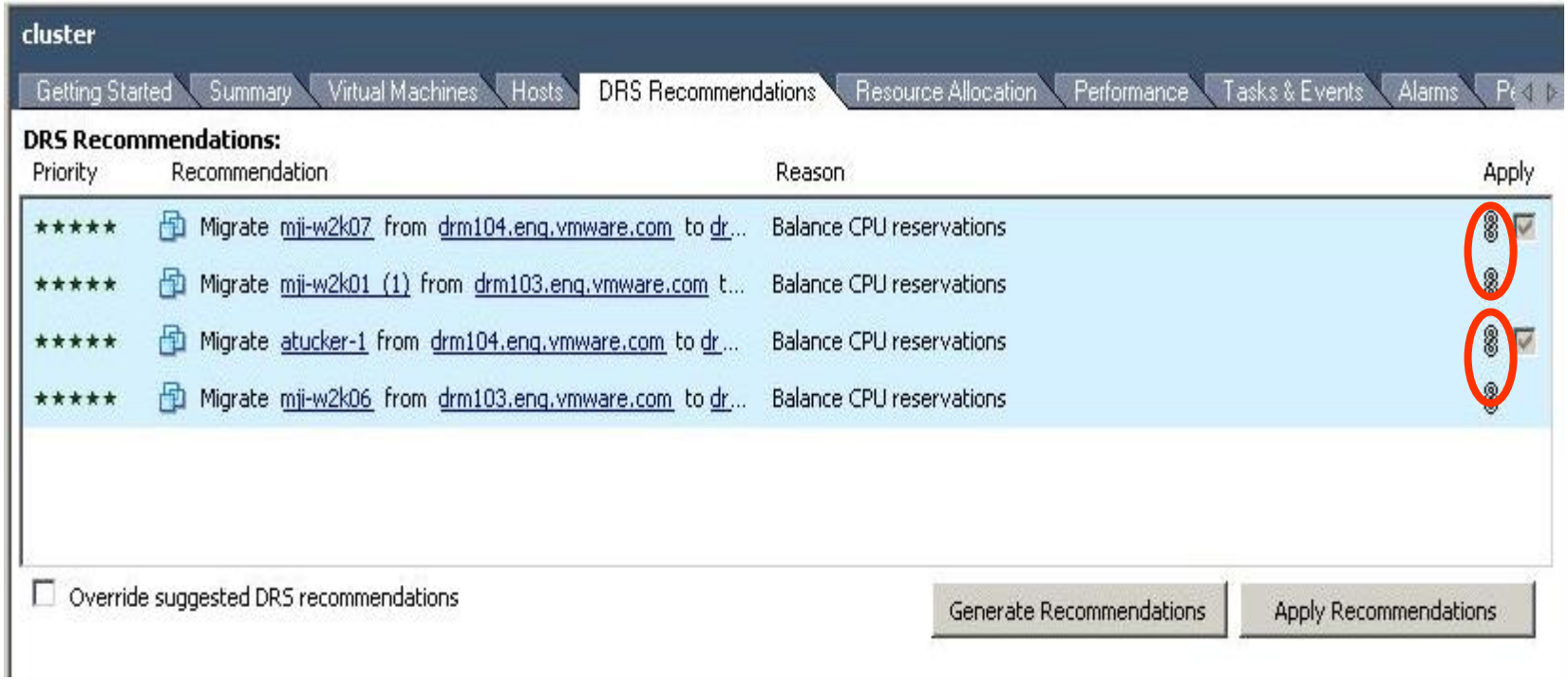
Dependent Recommendations

- **Recommendation A depends on B if and only if**
  - A is moving a VM into a host that B is moving a VM out of, or
  - A is moving a VM into a host that B is powering on.
- **If A is executed before B completes, A may fail due to**
  - Admission failure on host
  - Invalid host state
  - Other reasons …

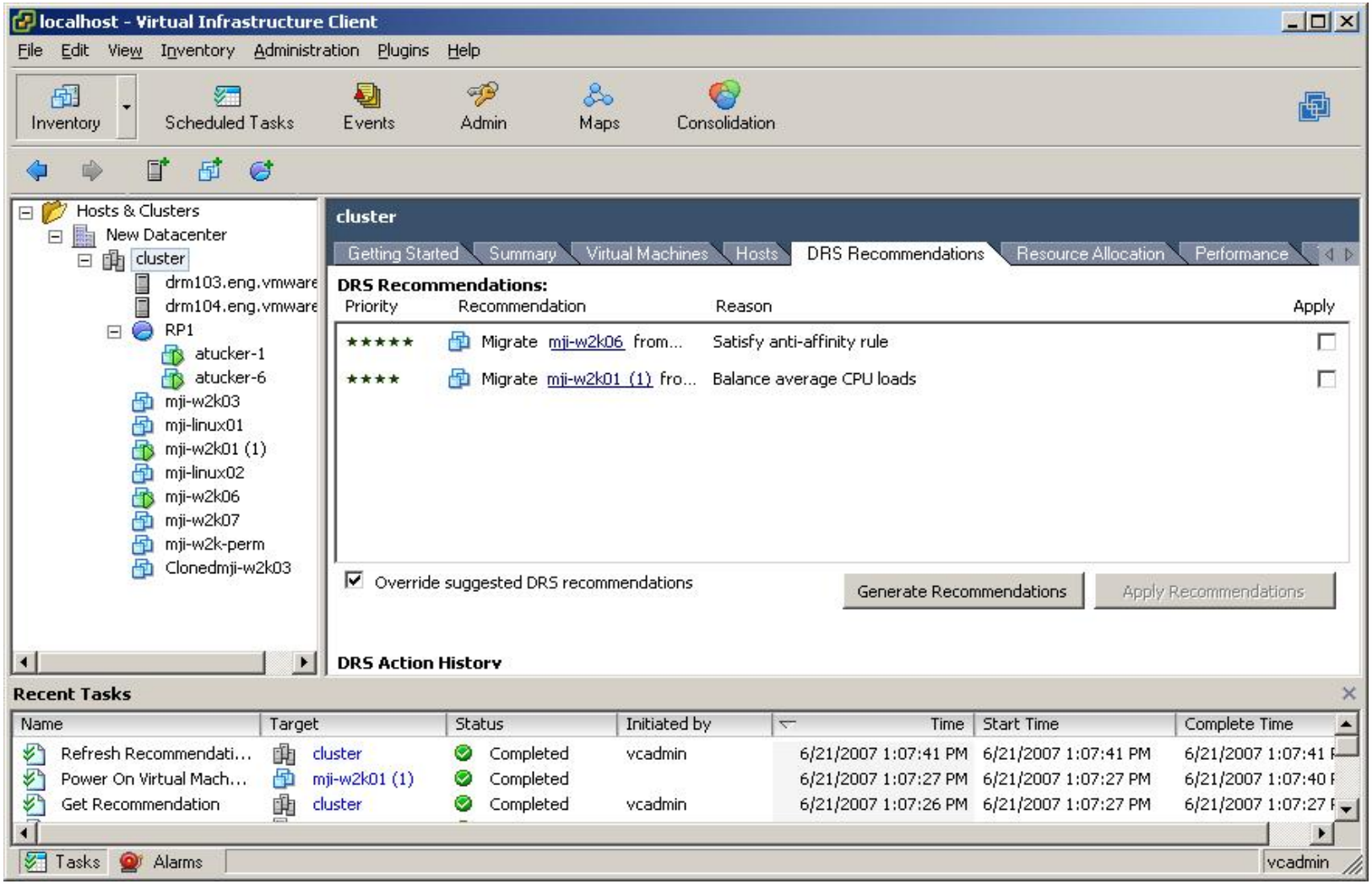
Compound Recommendations

- A compound recommendation has multiple actions A, B, C, …
- All actions share the same goal (or reason) and the same star rating
- All actions must be executed in order to achieve the goal
- Execution of subset may leave the cluster in a worse state than before the execution

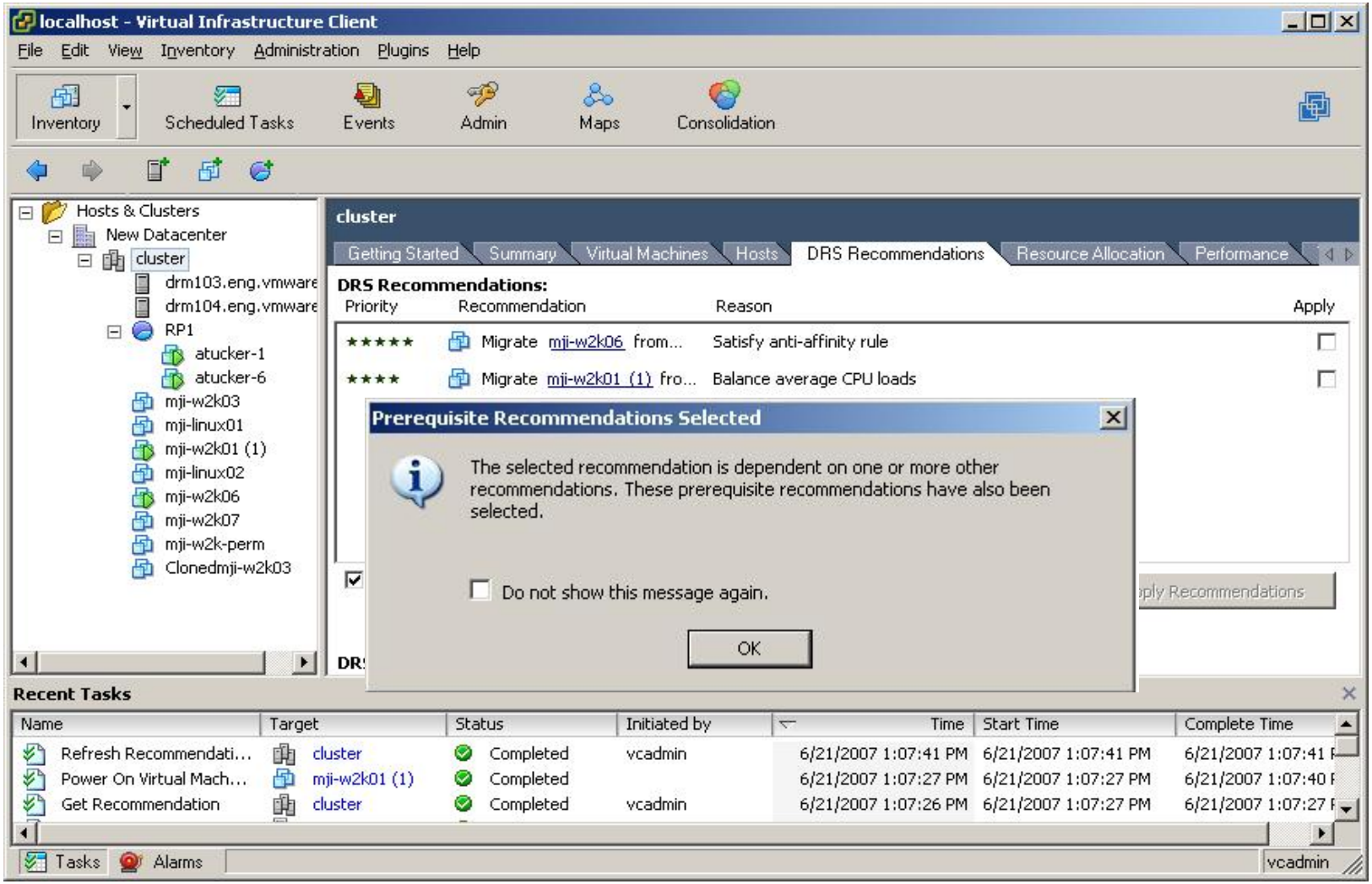
Compound Recommendations



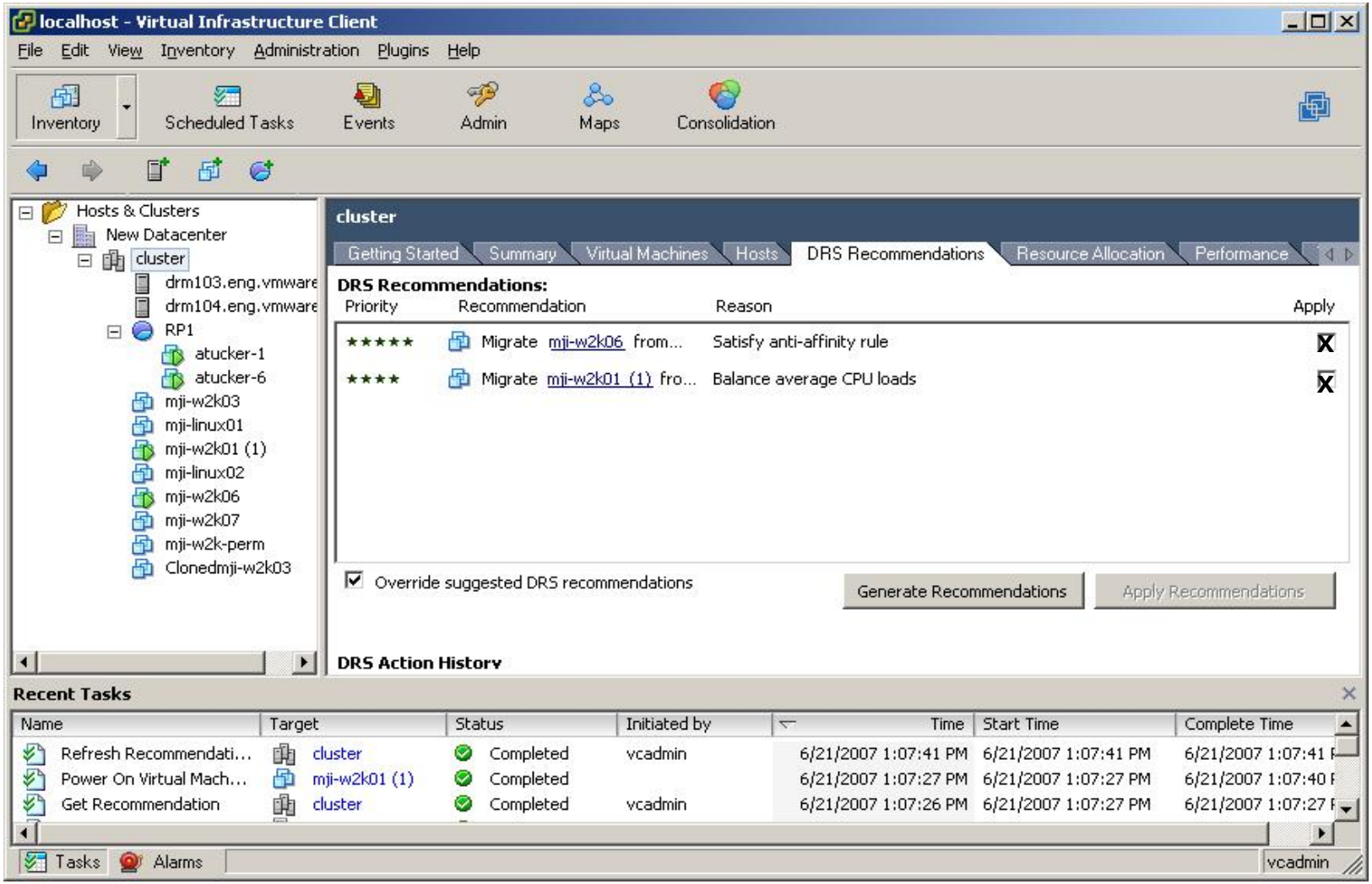
Dependent Recommendations (1)



Dependent Recommendations (1)



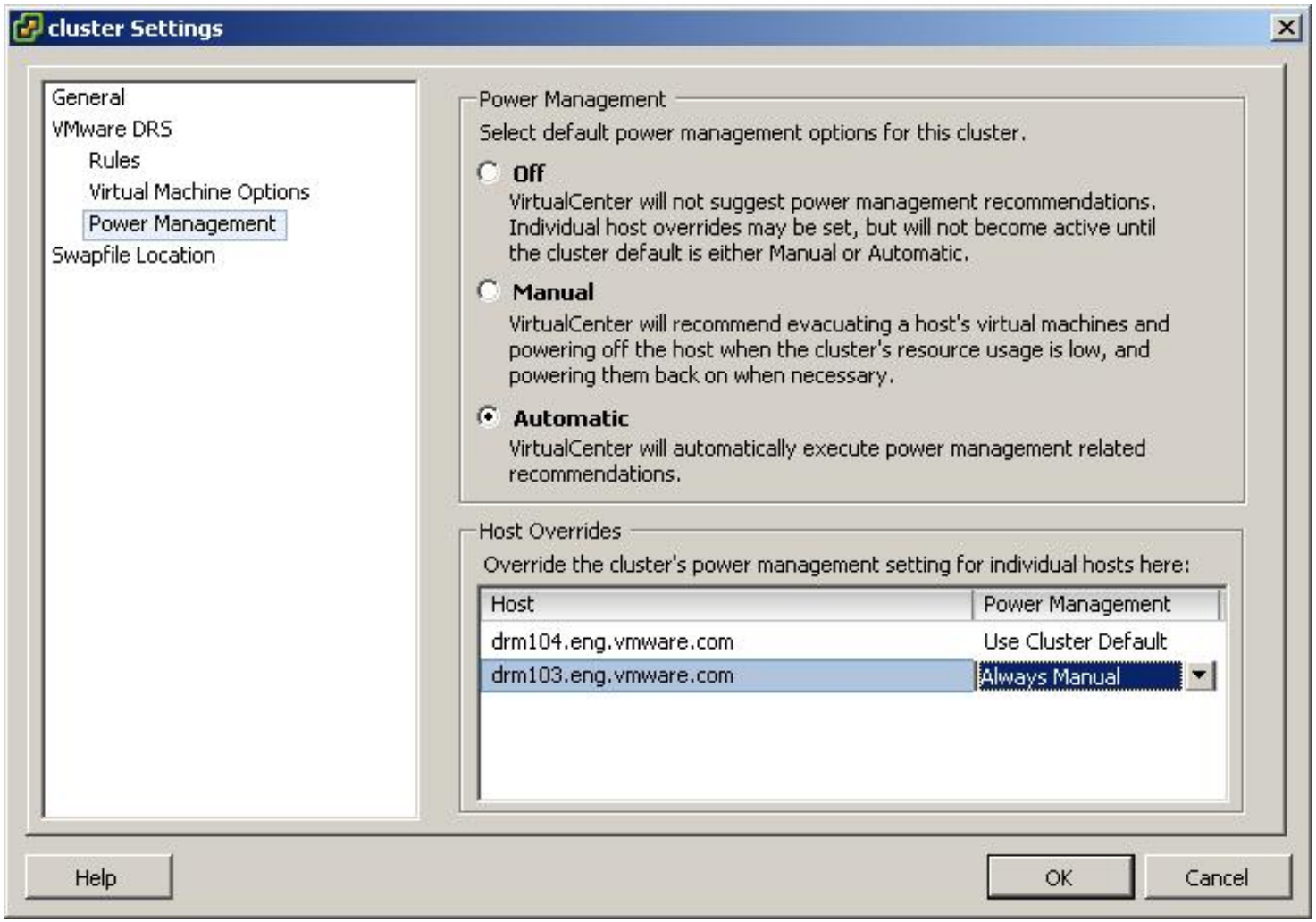
Dependent Recommendations (1)



Outline

- Create and configure a DRS cluster
- Compound and dependent recommendations
- **Enable DPM – power recommendations**
- Power on VMs – host recommendations
- Maintain failover level for High Availability (HA)
- Keep VMs together or separate them – affinity rules

Enable DPM (Distributed Power Management)



DRS/DPM Interactions

- **Conflicting goals ?**
  - DRS balances load to satisfy SLAs
  - DPM optimizes for power consumption
- **High level interaction**
  - DPM determines how many hosts to be powered on
  - DRS balances load among powered-on hosts
- **Detailed interactions**
  - DRS rebalances with DPM-recommended power actions in what-if simulations
  - DPM evaluates potentials of power actions based on DRS rebalancing results
  - Final power actions and VMotions are recommended to user

Power Management Algorithm Overview

- **Goal: try to maintain per-host load/capacity ratio within a target range (centered at 63%), while satisfying all DRS constraints.**
- **After DRS load balancing, divide hosts into highly and lightly loaded groups.**
  - Consider power-on for highly loaded group and power-off for lightly loaded group.
  - Run load balancing simulation for each power-on candidate
    - No VM can move to the new host => do not power it on
  - Run evacuation simulation for each power-off candidate
    - Host cannot be evacuated => do not power it off

Power-on and Power-off Decisions

- **Compute a weighted average load/capacity ratio for each group, i.e., Rh and Rl.**
  - Consider longer load history for power-off decisions (20 minutes) than for power-on decisions (5 minutes) – conservative in powering off and responsive in powering on.
- **Power on for highly loaded group**
  - If Rh > highThreshold, power on additional hosts, until Rh <= highThreshold.
- **Power off for lightly loaded group**
  - If Rl < lowThreshold, power off existing hosts, until Rl >= lowThreshold.
  - Take into account load variation, power operation overhead, and power-performance trade-off in powering off decisions.
- **All parameters are tunable online through advanced options.**

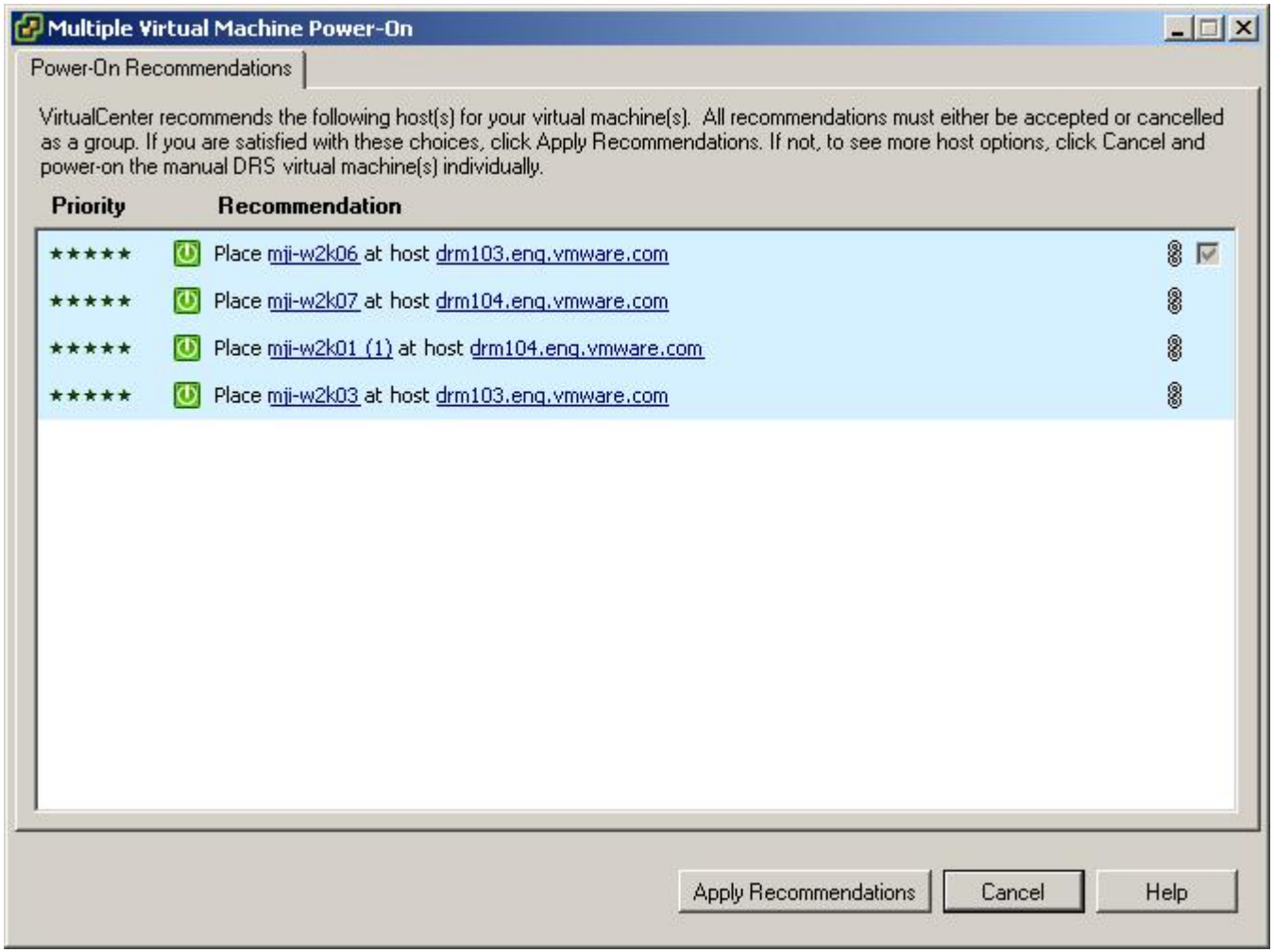
Power Off To Save Energy

Power On To Handle Increased Load

Outline

- Create and configure a DRS cluster
- Compound and dependent recommendations
- Enable DPM – power recommendations
- **Power on VMs – host recommendations**
- Maintain failover level for High Availability (HA)
- Keep VMs together or separate them – affinity rules

Power On Multiple VMs



Power-on With Prerequisite

Host Recommendations for VM Power-On

- **May power on a single VM or multiple VMs simultaneously**
  - In case of multiple power-on, a compound recommendation per DRS cluster
- **May have multiple choices**
- **Each recommendation has at least one VM power-on action, and optionally prerequisite actions**
  - Migration
  - Host power-on (if DPM is enabled)

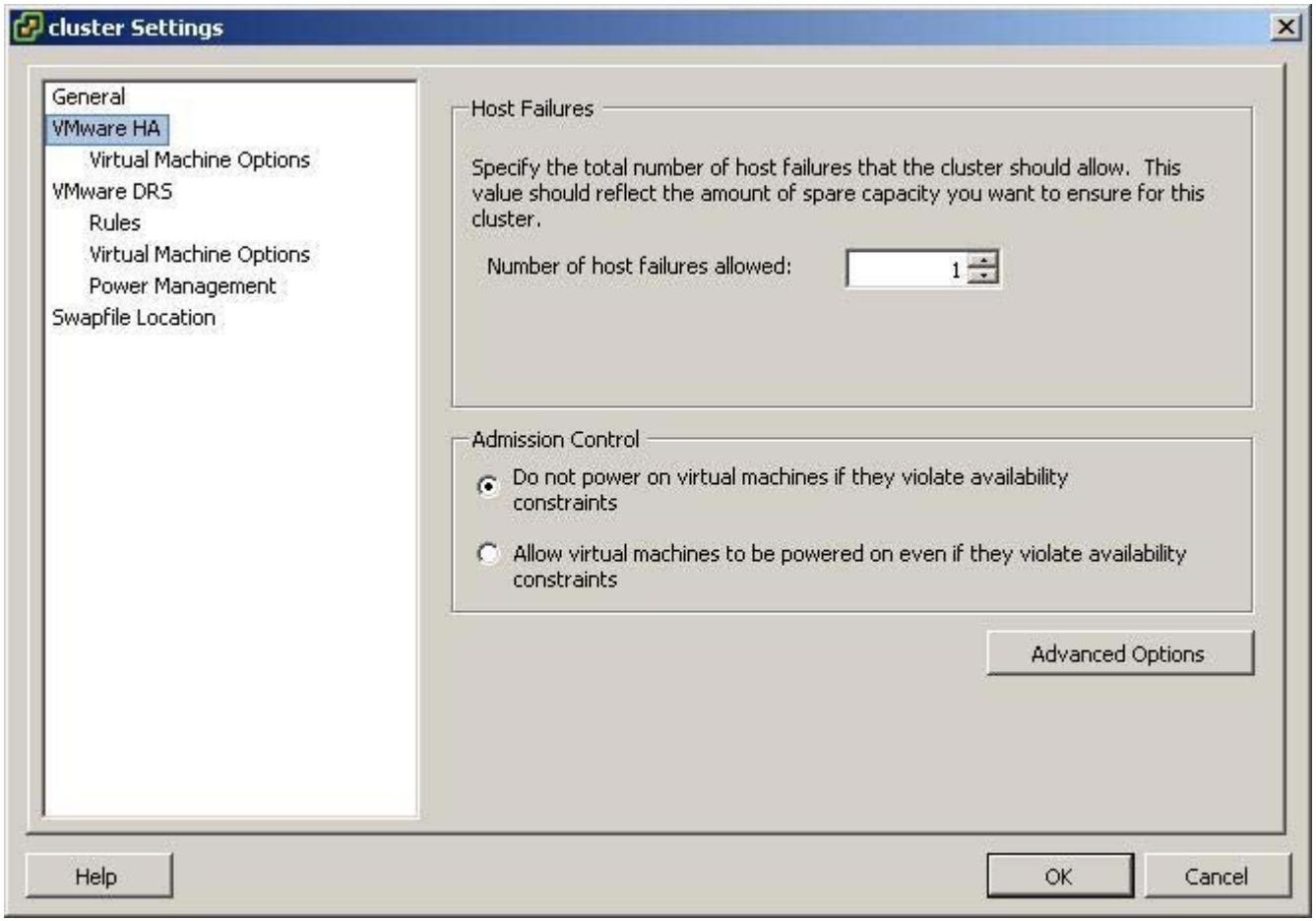
Advantages of Multiple Power-on

- **Better performance**
  - DRS is invoked only once for a multiple power-on, but is invoked N times for N single power-ons.
- **Better placement decision**
  - If powered on in the wrong order, later VMs may fail in single power-on.
  - In multiple power-on, DRS places VMs in descending order of constraints, to avoid failures in later stage.
- **Ideal for Virtual Desktop environments**
  - All VMs need to be powered on at 9am when people come to work

Outline

- Create and configure a DRS cluster
- Compound and dependent recommendations
- Enable DPM – power recommendations
- Power on VMs – host recommendations
- **Maintain failover level for High Availability (HA)**
- Keep VMs together or separate them – affinity rules

Enable HA (High Availability)

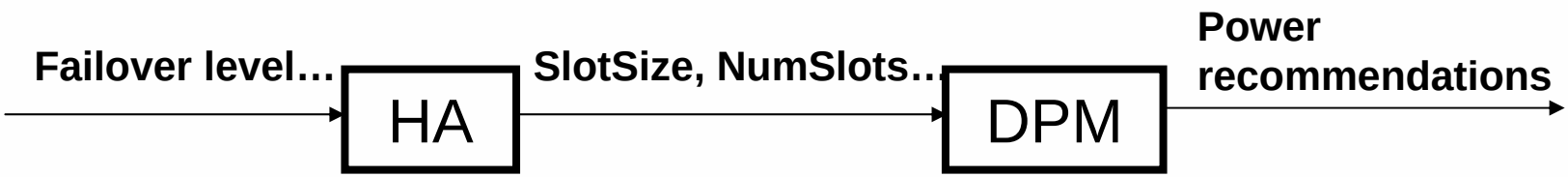


HA/DPM Interactions

- HA does VM admission control to satisfy the constraints:
  - SlotSize = Largest VM reservation
  - NumSlots >= Capacity of N largest hosts / SlotSize
  - NumVms <= ClusterCapacity / SlotSize – NumSlots

DPM does host exit control to satisfy the same constraints:

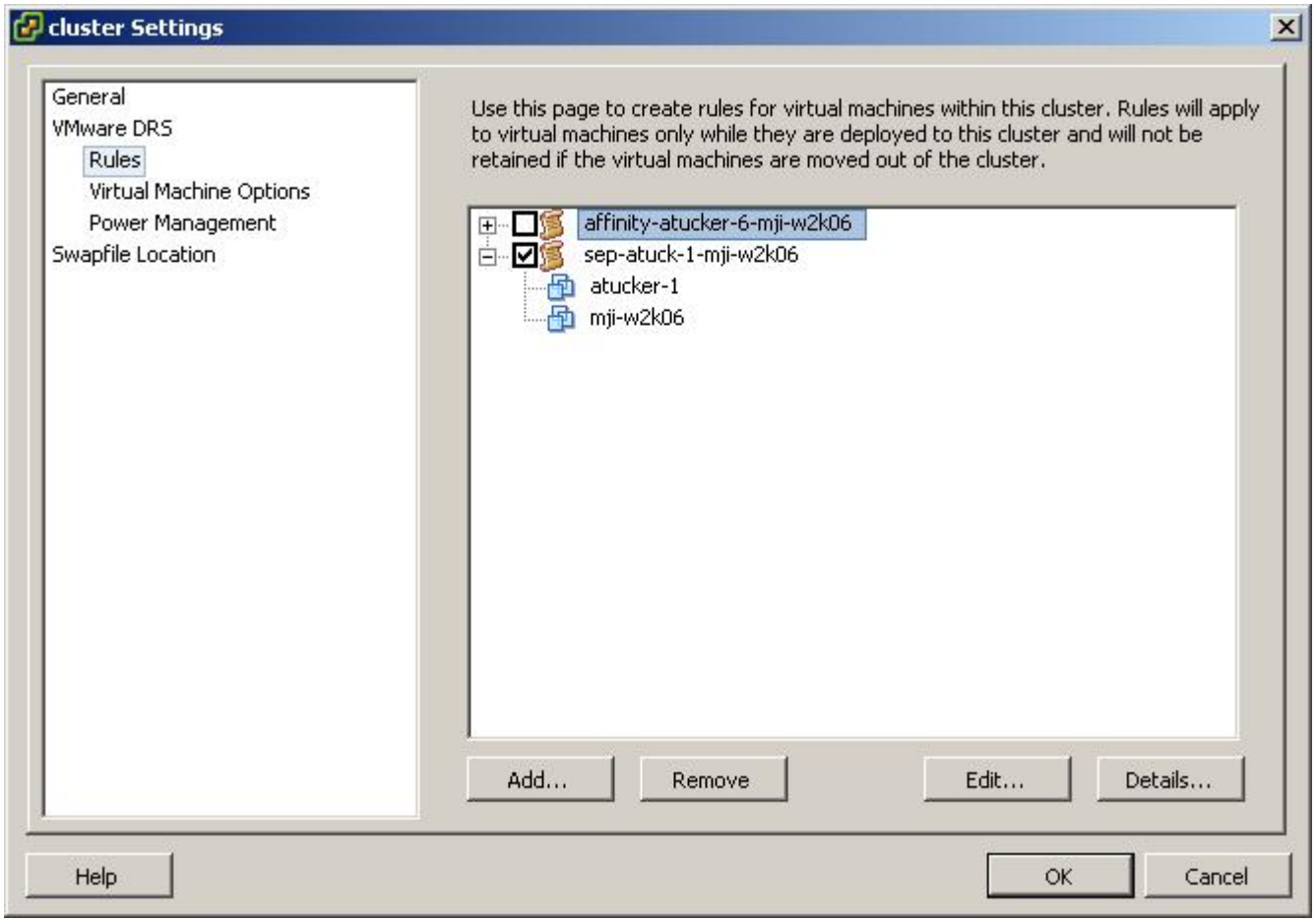
- SlotSize = Largest VM reservation
- NumSlots >= Capacity of N largest hosts / SlotSize
- ClusterCapacity >= (NumVms + NumSlots) * SlotSize



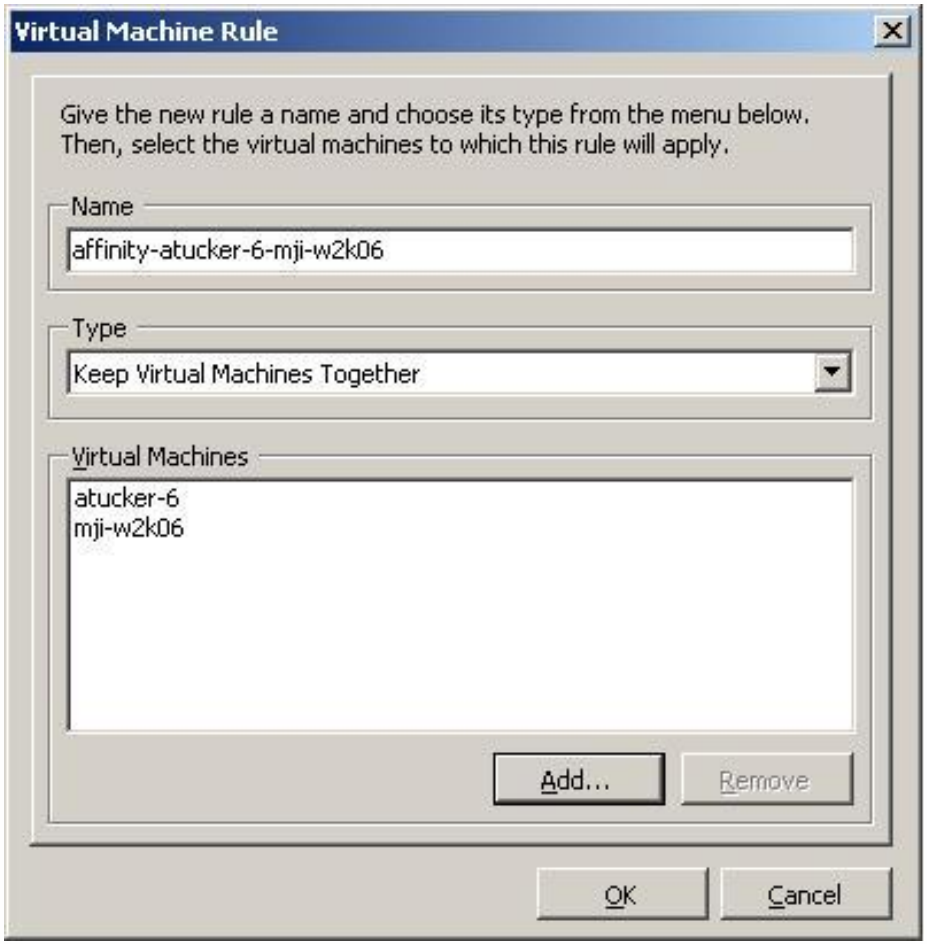
Outline

- Create and configure a DRS cluster
- Compound and dependent recommendations
- Enable DPM – power recommendations
- Power on VMs – host recommendations
- Maintain failover level for High Availability (HA)
- **Keep VMs together or separate them – affinity rules**

Rules Overview



Rules Details



Use Cases and Constraints

- **Affinity rules**
  - VMs communicating a lot with each other
  - VMs sharing a portgroup on a virtual switch (LabManager)
  - Constraint: at most one affinity rule per VM
- **Anti-affinity rules**
  - Redundant VMs for fault tolerance
  - Constraint: at most two VMs per anti-affinity rule (may be lifted in the future)

Enforcing Affinity Rules

- **Each DRS invocation runs in a number of steps:**
  - First, try to enforce SLA, affinity rules and HA requirement.
  - Second, optimize for load balancing and power consumption.
  - Must not violate a new rule in enforcing another.
  - Must not violate a new rule in optimization.
  - Better error reporting is needed when rules cannot be enforced.

Summary

- **DRS control knobs**
  - Migration threshold, automation levels, affinity rules, etc.
- **DRS recommendations**
  - Compound, dependent, initial placement, power, etc.
- **DRS/DPM interactions**
  - DRS balances load among hosts that DPM powers on
- **HA/DPM interactions**
  - HA specifies spare capacity to maintained by DPM

Backup slides

Enforce Affinity Rules

Part of DRS placement algorithm:

```
For each affinity rule {
  If (rule violated) {
    For each host in cluster {
      If (can have all VMs in the rule) {
        Move all VMs in the rule to the host;
          Done with this rule;
        }
      }
    }
}
```

Enforce Anti-Affinity Rules

Part of DRS placement algorithm:

```
For each VM {
  If (violates an anti-affinity rule) {
    If (can move to another host) {
      Move it to another host;
    }
  }
}
```

Questions?

TA24

DRS Deep Dive and Technology Preview of Distributed Power Management

Minwen Ji

VMware

VMWORLD 2007

EMBRACING YOUR VIRTUAL WORLD

BREAKOUT SESSION