

Microsoft HPC

V 1.0 José M. Cámara (checam@ubu.es)



Introduction

Microsoft High Performance Computing Package addresses computing power from a rather different approach.

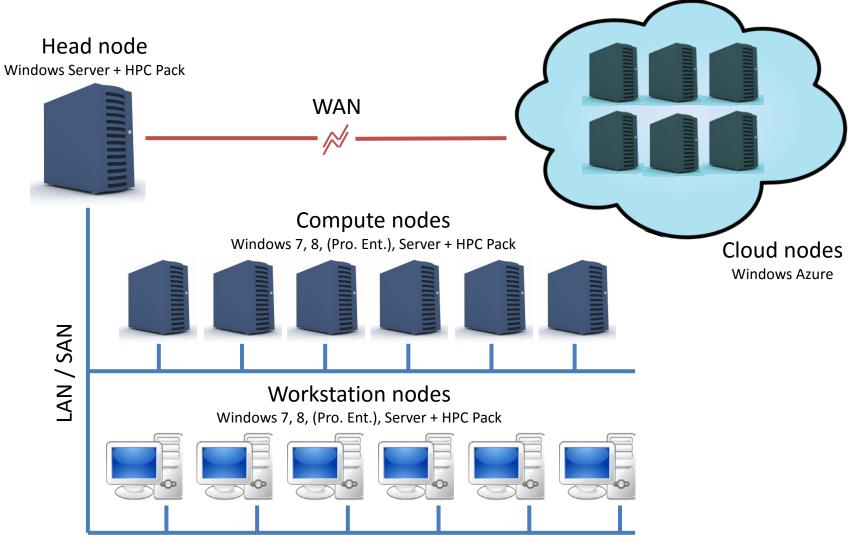
It is mainly focused on commodity clusters.

It is meant to make the most of existing resources thus minimizing investment. The system is highly scalable due to both on premises and cloud resources allocation.

Overall performance is affected by customer's investment on the architectural subsystems: compute nodes hardware & interconnect.

Cluster architecture

UNIVERSIDAD DE BURGOS





- Head node: no particular hardware features are required (double network adapter).
- Compute node: devoted to HPC.
 - Workstation: general purpose computer. It joins the cluster to perform calculation when not used for other purposes. They may be brought online manually or according to a timetable.
 - Cloud node: virtual computer hired to Microsoft.

Operating system

Types of

nodes

- The head node must run Windows Server.
- The rest nodes in premises may be servers but they also admit conventional Microsoft operating systems usually restricted to professional, enterprise or ultimate versions.
- Cloud nodes are integrated under Windows Azure.

Network architecture

- A conventional WAN connection is used for both remote management and connecting cloud resources. Obviously the wider bandwidth, the better.
- On premises resources are connected by means of either LAN or SAN solutions; most vendor systems are supported. Nodes can be linked though private networks, corporate or both.



Cluster deployment

Server deployment			
OS installation.	Head node configuration		Ń
Active domain	Definition of cluster	Node deployment	
configuration.	topology. Configuration of communications. User policies.	OS installation.	
HPC Pack Installation.		Entering the active domain.	
		HPC Pack Installation.	
		Entering the cluster.	



Active domain

For all the cluster operations and services to work safely, all nodes must belong to a common domain or at least to different domains with trust relations established.

The domain is set as a new forest and usually the server is promoted to domain controller, unless other server is in charge.

All nodes and users will set as domain members so they can join the cluster.



HPC Pack

The high performance computing package (HPC Pack) is provided by Microsoft free of charge.

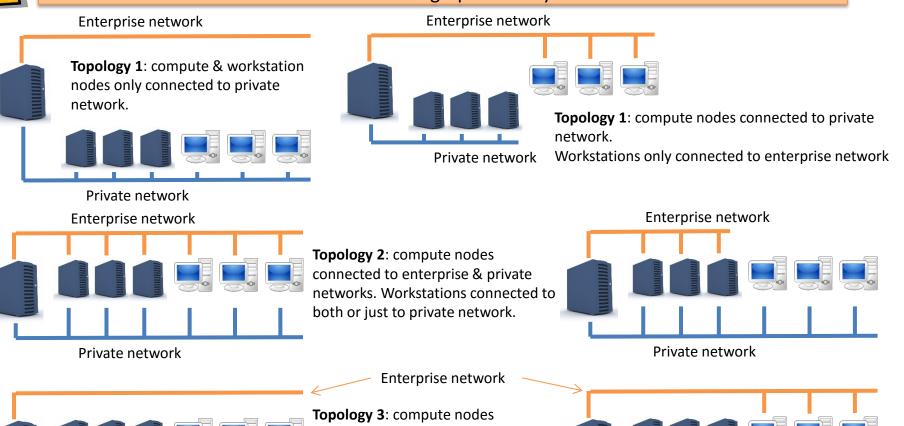
The same package is valid for any type of node to join the cluster.

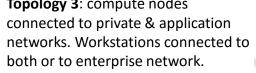
Depending on the type of node selected a different set of tools is installed. In all cases, a group of software tools is made available to the user for cluster management.

Cluster topologies

Microsoft uses this concept to define a series of in premises network connection alternatives. It is not related to the network graph actually.

UNIVERSIDAD DE BURGOS





Private network

Application network (SAN)

Cluster topologies II



UNIVERSIDAD DE BURGOS

Enterprise network

Topology 4: compute nodes connected to enterprise, private & application networks. Workstations connected to all or only to enterprise network.

Private network

Application network (SAN)



Topology 5: no private or application network. All nodes connected to enterprise network



The user is not compelled to match the actual network configuration when setting the topology. Enterprise network may exist on workstations but may not be used for HPC.

Cluster management

UNIVERSIDAD DE BURGOS

> The system administrator is in charge of configuration, supervision and cluster management.

The Cluster Manager is the tool that makes all these operations possible:

- Node management.
- Users management.
- Scheduler configuration.



Node management

Add / remove nodes.

Bring workstations online/offline: only if this operation has previously been set as manual and not according to a certain schedule.

Form node groups to facilitate management. For instance, set certain types of jobs to be allocated to certain node groups.

Monitoring nodes: "node health", properties, network, metrics...



Users management

Add / remove users.

Manage user groups.

Set user roles:

- User: allowed to manage its own jobs.
- Administrator: allowed to manage jobs and resources.
- Job administrator: allowed to manage jobs but not resources.
- Job operator: allowed to manage jobs in a restricted manner (view, cancel, finish, re-queue).



Scheduler management

The scheduler decides what jobs to launch, when and what resources must be assigned to them.

This is done under some policies:

- Queued.
- Balanced.

Backfilling is permitted: smaller jobs can be launched before others ahead of the queue when they are waiting for enough resources and as long as this operation does not delay the awaiting jobs.

Job scheduling policies

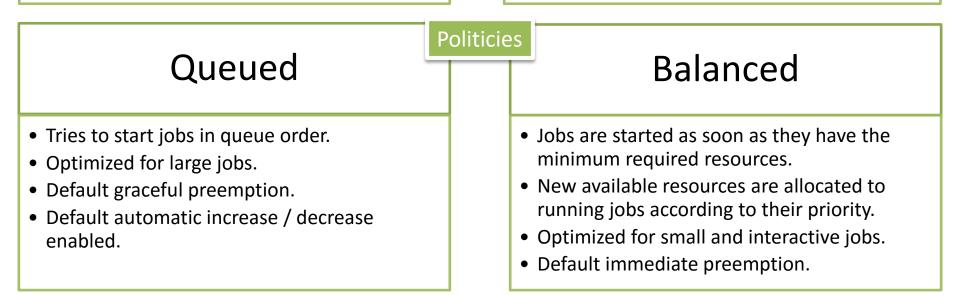
Sub-policies

Preemption: higher priority jobs are allowed to take away resources from lower priority ones already started.

- Graceful: resources are taken away from already finished tasks.
- Immediate: all job's tasks are cancelled at run time.
- Task level: tasks are cancelled individually.

Dynamic resource allocation: resources allocated to a job can be modified during execution.

- Automatic increase: new resource allocation is preferred over starting lower priority jobs.
- Automatic decrease: take away unused resources from jobs with no tasks to be started.





Job Manager

The users submit jobs to the cluster.

Jobs are made up from tasks.

The Job Manager is the tool used to create and submit jobs.

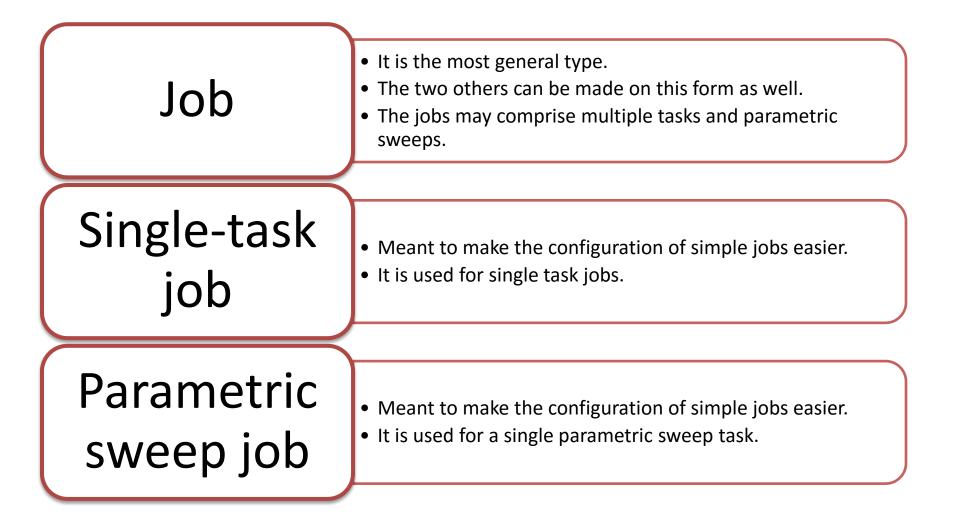


Task types

Basic	• Runs a single instance of a series or parallel application.	
Parametric sweep	 Runs a number of instances of the same tasks. The exact number is determined by a command that takes different values. The command's value itself is meaningful to each instance. 	
Node preparation	 A command or script to be run on each node. It is executed prior to any other task in the job. 	
Node release	 A command or script to be run on each node. It is executed when the node is released from the job. 	
Service	 Runs a command or service on each resource allocated to a job. If an instance of the command exits and the resource is still allocated, another one starts. 	



Job types



Job properties

Job ID	 Numeric ID of the job. Assigned by the Job Manager. 	
Job name	Text name of the job.User assigned.	
Job template	 Name of the job template used to submit the job. Templates define default values and constraints to the jobs and are created by the administrator. 	
Priority	 Numeric value to set the priority of the job. Ranges from 0 to 4000, being 4000 the highest priority. 	
Run time	Maximum time to complete the job.If exceeded the job is cancelled.	
Memory	• The minimum amount of memory on a node to run the job.	
Licenses	• Licenses requested to run the job.	

Apart from the "name", setting the rest is not mandatory.

UNIVERSIDAD

DE BURGOS

There are some more properties. For a complete list refer to: https://technet.microsoft.com/es-es/library/ff919649

Job submission & monitoring.

File View Tasks Options Help	Job properties	1- Job name
Back 🏵 Forward 🛛 Navigation Par	Forward Navigation Part	
Job Management	Select a job template to use for this job. The job template specifies a set of options to use when running a job.	> Actions
All Jobs	Job template: Default 🗸	
Configuring		Job Submission
Active	Send a notification when this job:	
Finished	Starts Completes	New Job
Failed	Send email notifications to:	New Single-Task Job
Canceled		New Parametric Sweep Job
 My Jobs 		🚛 New Job from XML File
	Task properties	Job Actions
Configuring Active	Task name: MPIsingletask	View Job
	Command line:	
Finished	Co <u>m</u> mand line: mpiexec -n 4 c:\mpiapps\MPIapp1.exe	Modify Job
Failed	mprexec -n 4 C. (mprapps (mprappi.exe	Add Task to Job
Canceled		Copy Job
 By Job Template 		🕞 Submit Job
Default		Cancel Job
		(II) Requeue Job
	Working <u>d</u> irectory:	Export Job
	C:\cluster v Browse	
	Standard inp <u>u</u> t:	Task Actions
	C:\cluster\input.txt	🕑 View Task
		Cancel Task
	Standard output:	()) Requeue Task
	C:\cluster\output.txt v Browse	III Requeue Failed Sub-Tasks
	Standard <u>e</u> rror:	i Export Task
	C:\cluster\error.txt V Browse	
	Submit Cancel	
Data updated: 6/1/2015 11:46:45 PM		

Job submission using the Job Manager.

Jobs can be created and submitted logging on to a node and starting the Job Manager.

Choose the type of job and fill the fields in the form.

Set the command line and the output files.

Default values may be left unchanged if not needed.

Jobs can be submitted remotely. A Remote Desktop Connection to a cluster node must be stablished first. For a complete reference on how to submit and monitor jobs remotely refer to: https://technet.microsoft.com/en-us/library/gg315415%28v=ws.10%29.aspx

Queued scheduling.

Scheduling mode:

- Queued Attempt to assign the maximum amount of requested resources to running jobs.
- O Balanced Start as many jobs as possible with the minimum amount of requested resources for each. If additional resources are available on the cluster, grow jobs based on their priority and the Priority Bias setting.

Pre-emption options

- Graceful pre-emption To enable higher priority jobs to start sooner, take resources away from lower priority jobs as their tasks complete.
- Immediate pre-emption To enable higher priority jobs to start sooner, take resources away from lower priority jobs by canceling running jobs.

☑ Task level pre-emption - To enable pre-emption of individual tasks instead of whole jobs.

O No pre-emption - Jobs will continue to run until completion, even if higher priority jobs are waiting for resources.

Adjust resources automatically

- Increase resources automatically (grow) Use available resources to grow higher priority, running jobs to their maximum before starting lower priority jobs.
 - Grow by pre-emption To help grow higher priority, running jobs, use pre-emption to take resources from lower priority, running jobs.
- ✓ Decrease resources automatically (shrink) Automatically release unusued job resources over time when a job holds resources that it cannot use.

More about policy configuration

Main queued scheduling options.

Aims to provide as many resources as possible to running jobs.

Different preemption options can be selected.

It is also possible to decide how allocated resources are adjusted.

Balanced scheduling.

Scheduling mode:

- O Queued Attempt to assign the maximum amount of requested resources to running jobs.
- Balanced Start as many jobs as possible with the minimum amount of requested resources for each. If
 additional resources are available on the cluster, grow jobs based on their priority and the Priority Bias
 setting.

Pre-emption options

- Immediate pre-emption (Recommended) To enable additional jobs to start, take resources away from running jobs by canceling running tasks
- Graceful pre-emption (Advanced) To enable additional jobs to start, take resources away from running jobs as tasks exit
- For most cluster workloads, immediate pre-emption in Balanced mode enables more jobs to start in a time period.

Priority bias

Priority Bias controls how additional resources are allocated to running jobs. A higher bias level allocates more resources to higher priority jobs.

Priority Bias level:

O High bias

Medium bias

🔿 No bias

Rebalancing interval

The job scheduler rebalances resource allocation at a constant time interval. Jobs can grow and shrink in order to start new jobs, fill available resources, and balance resource allocation according to the Priority Bias level.

Seconds between rebalancing:

10 🌲

More about policy configuration Main balanced scheduling options. Tries to start as many jobs as possible even though the maximum required resources are not available.

Different preemption options can be selected.

The bias applied to priority when rebalancing is also selectable.

Rebalancing time is configurable. The goal is to balance flexibility and efficiency.

References

- Cluster network topologies: <u>https://technet.microsoft.com/en-us/library/gg145543.aspx</u>
- Node management: <u>https://technet.microsoft.com/es-es/library/ff919378</u>
- Managing cluster users: <u>https://msdn.microsoft.com/en-us/library/ff919335.aspx</u>
- New cluster user roles: <u>http://blogs.technet.com/b/windowshpc/archive/2013/09/04/hpc-pack-2012-sp1-available.aspx</u>
- Job scheduler: <u>https://technet.microsoft.com/es-es/library/ff919436</u>
- Job manager: <u>https://technet.microsoft.com/es-es/library/ff919691</u>