UNIVERSIDAD DE BURGOS

Microsoft HPC

V 1.0

José M. Cámara

(checam@ubu.es)

UNIVERSIDAD DE BURGOS

Introduction

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

It is mainly focused on commodity clusters.

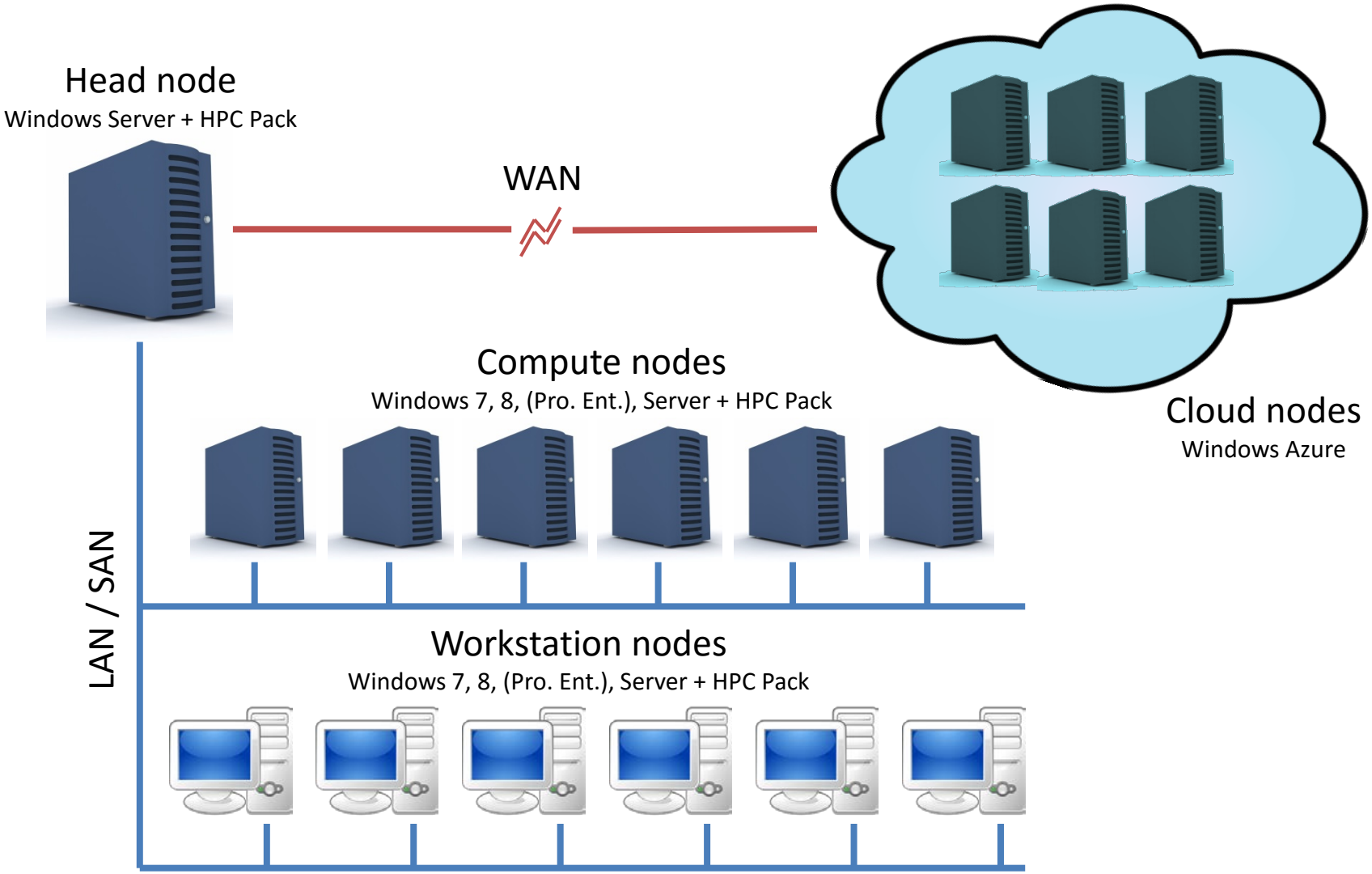
It is meant to make the most of existing resources thus minimizing investment.

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.

UNIVERSIDAD DE BURGOS

Cluster architecture



UNIVERSIDAD DE BURGOS

Cluster architecture II

Types of nodes

- 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

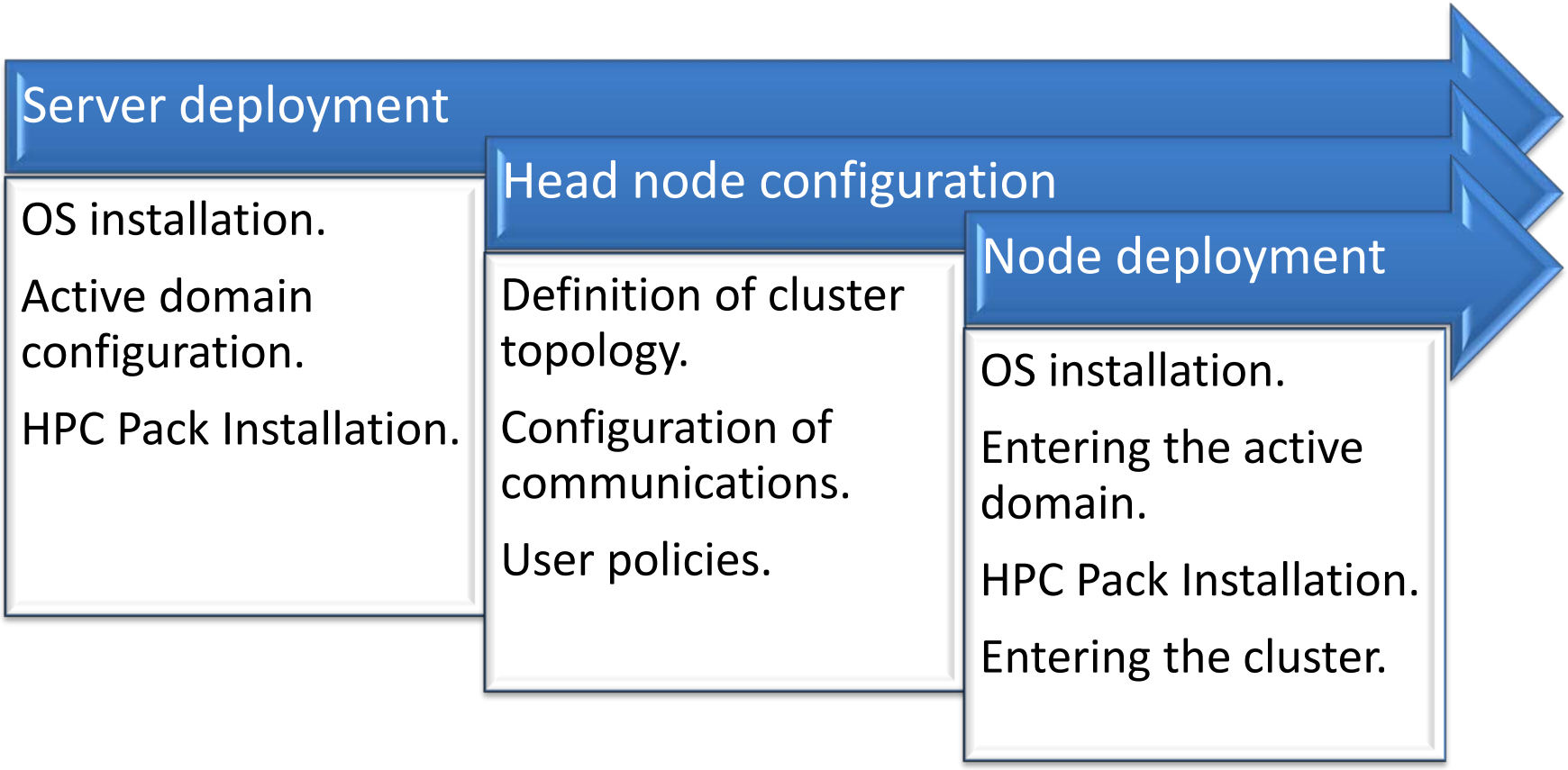
- The head node must run Windows Server.
- The rest nodes in premises may be servers but they also admit conventional Microsoft operating systems usually restricted to professional, enterprise or ultimate versions.
- Cloud nodes are integrated under Windows Azure.

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.

UNIVERSIDAD DE BURGOS

Cluster deployment



UNIVERSIDAD DE BURGOS

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.

UNIVERSIDAD DE BURGOS

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.

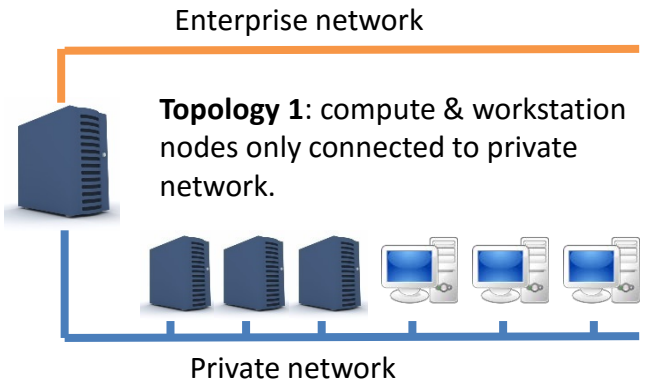
Depending on the type of node selected a different set of tools is installed.

In all cases, a group of software tools is made available to the user for cluster management.

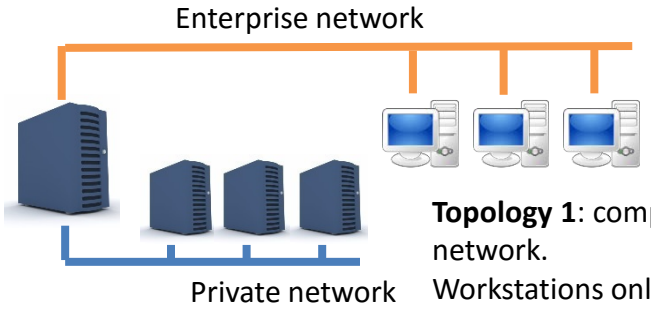
UNIVERSIDAD DE BURGOS

Cluster topologies

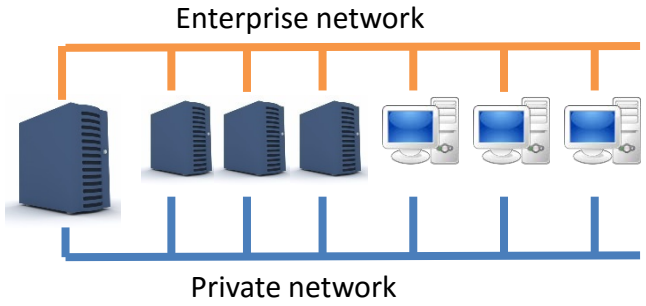
Microsoft uses this concept to define a series of in premises network connection alternatives. It is not related to the network graph actually.



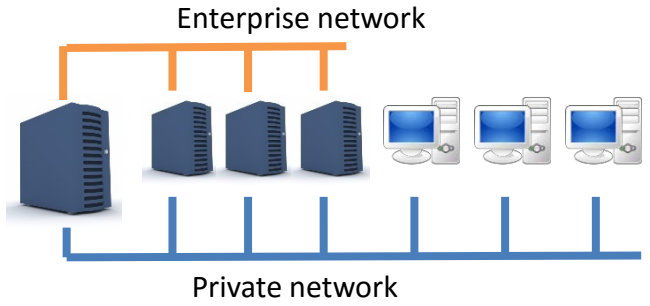
Topology 1: compute & workstation nodes only connected to private network.

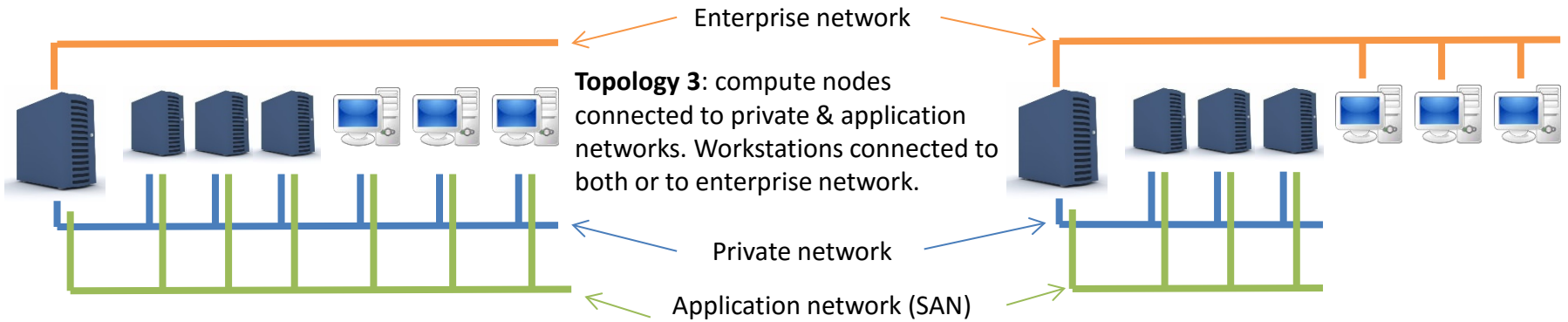


Topology 1: compute nodes connected to private network.
Workstations only connected to enterprise network



Topology 2: compute nodes connected to enterprise & private networks. Workstations connected to both or just to private network.

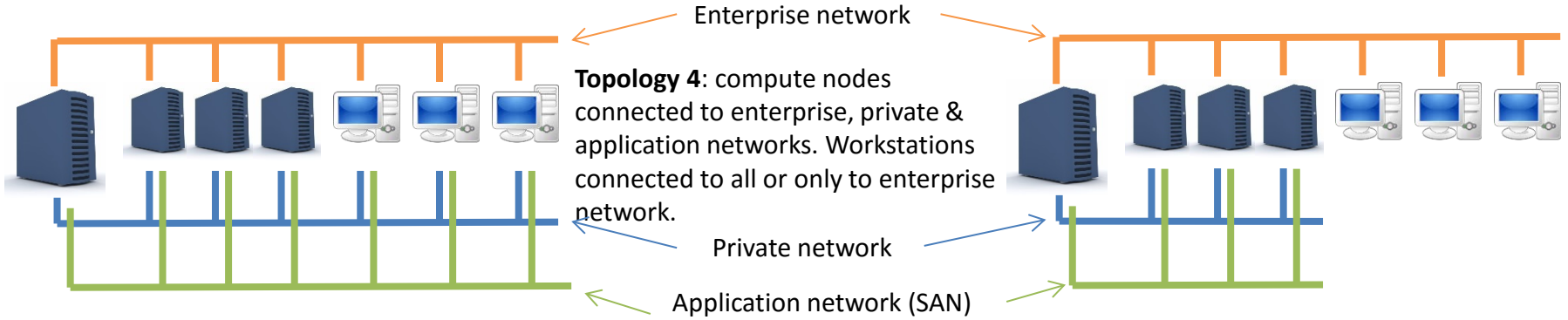




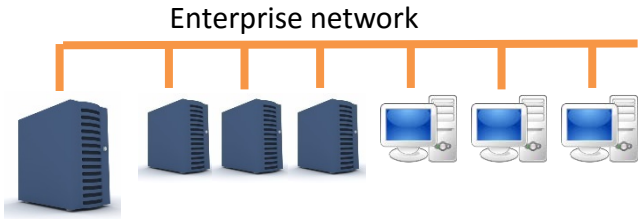
Topology 3: compute nodes connected to private & application networks. Workstations connected to both or to enterprise network.

UNIVERSIDAD DE BURGOS

Cluster topologies II



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

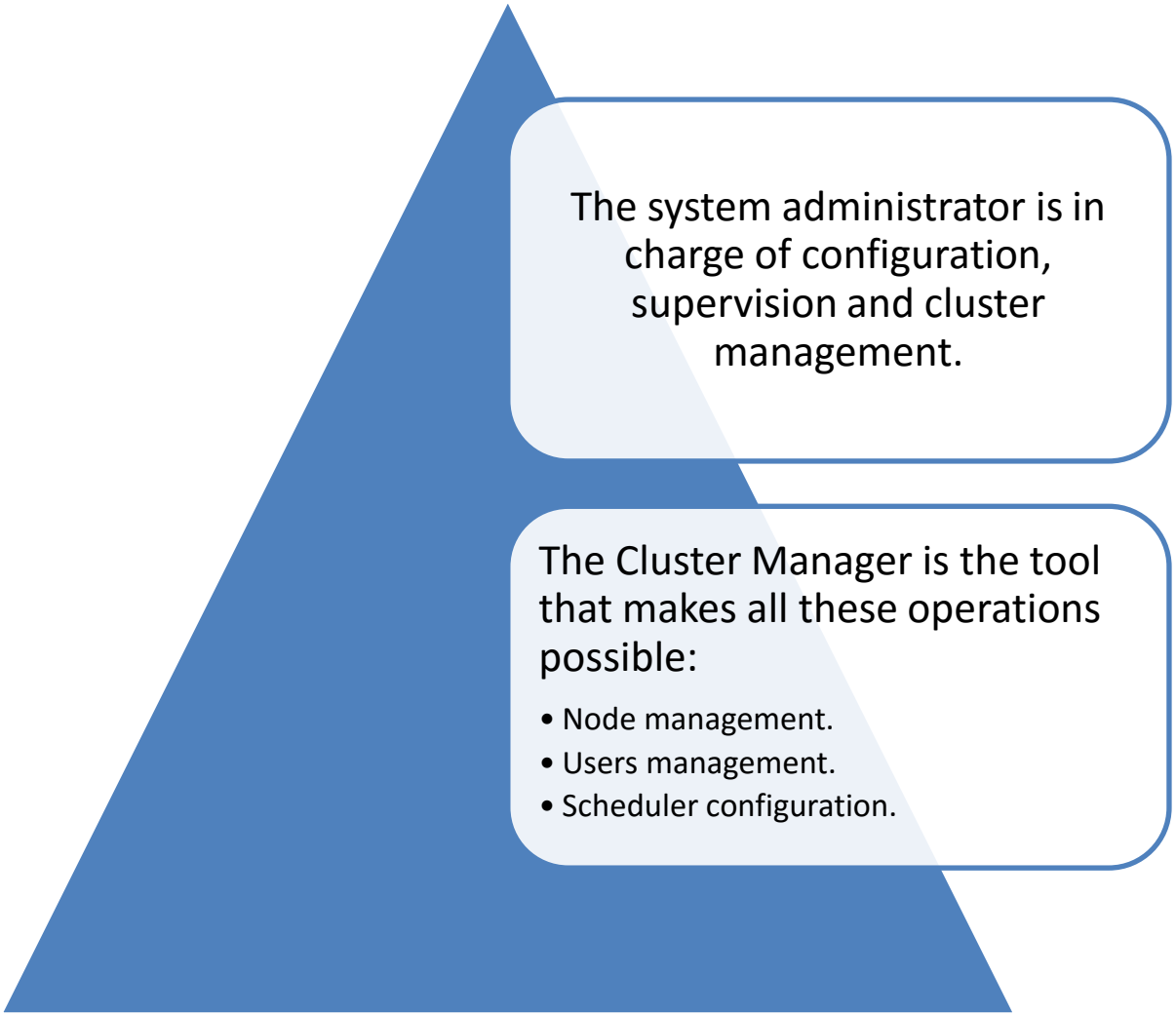


Topology 5: no private or application network. All nodes connected to enterprise network

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.

UNIVERSIDAD DE BURGOS

Cluster management



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.

UNIVERSIDAD DE BURGOS

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…

UNIVERSIDAD DE BURGOS

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).

UNIVERSIDAD DE BURGOS

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.

UNIVERSIDAD DE BURGOS

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.

Politicies

Queued

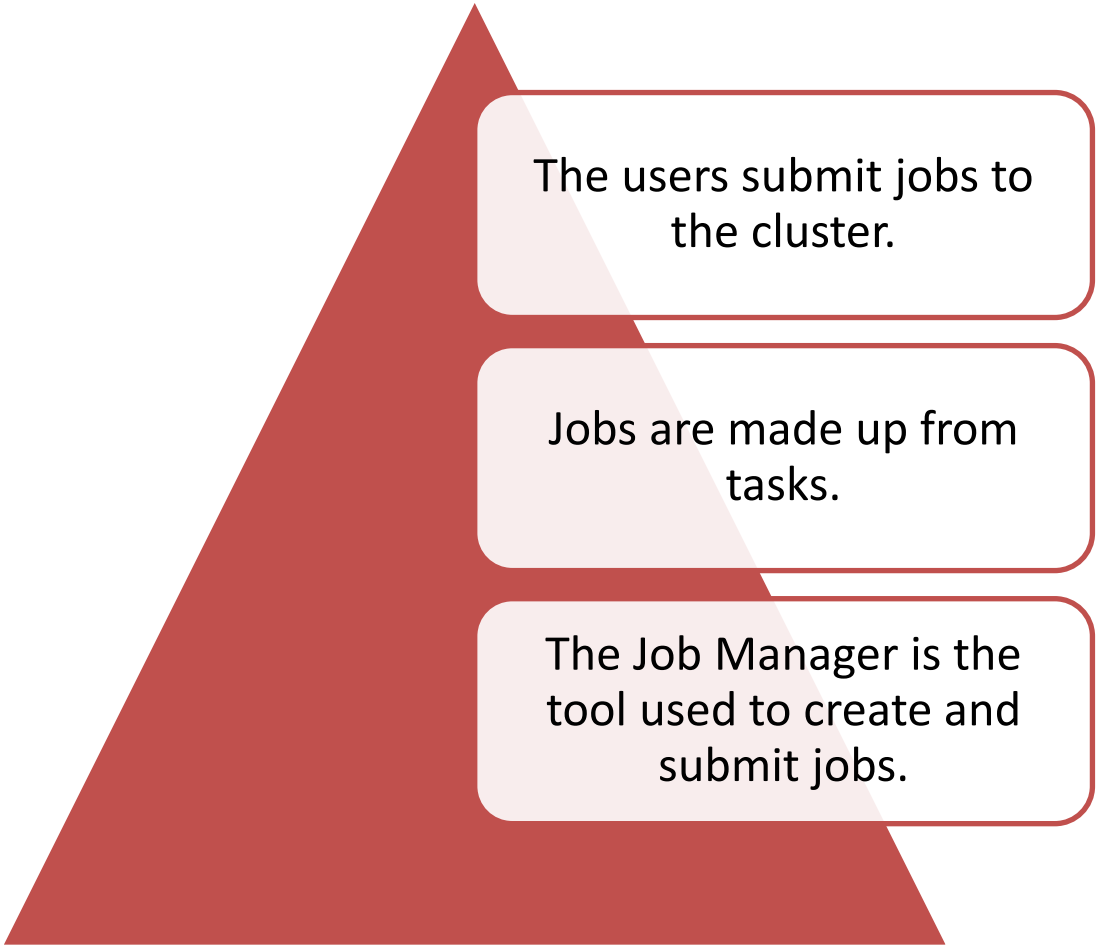
- Tries to start jobs in queue order.
- Optimized for large jobs.
- Default graceful preemption.
- Default automatic increase / decrease enabled.

Balanced

- Jobs are started as soon as they have the minimum required resources.
- New available resources are allocated to running jobs according to their priority.
- Optimized for small and interactive jobs.
- Default immediate preemption.

UNIVERSIDAD DE BURGOS

Job Manager



UNIVERSIDAD DE BURGOS

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.

UNIVERSIDAD DE BURGOS

Job types

Job	• It is the most general type. • The two others can be made on this form as well. • The jobs may comprise multiple tasks and parametric sweeps.
Single-task job	• Meant to make the configuration of simple jobs easier. • It is used for single task jobs.
Parametric sweep job	• Meant to make the configuration of simple jobs easier. • It is used for a single parametric sweep task.

UNIVERSIDAD DE BURGOS

Job properties

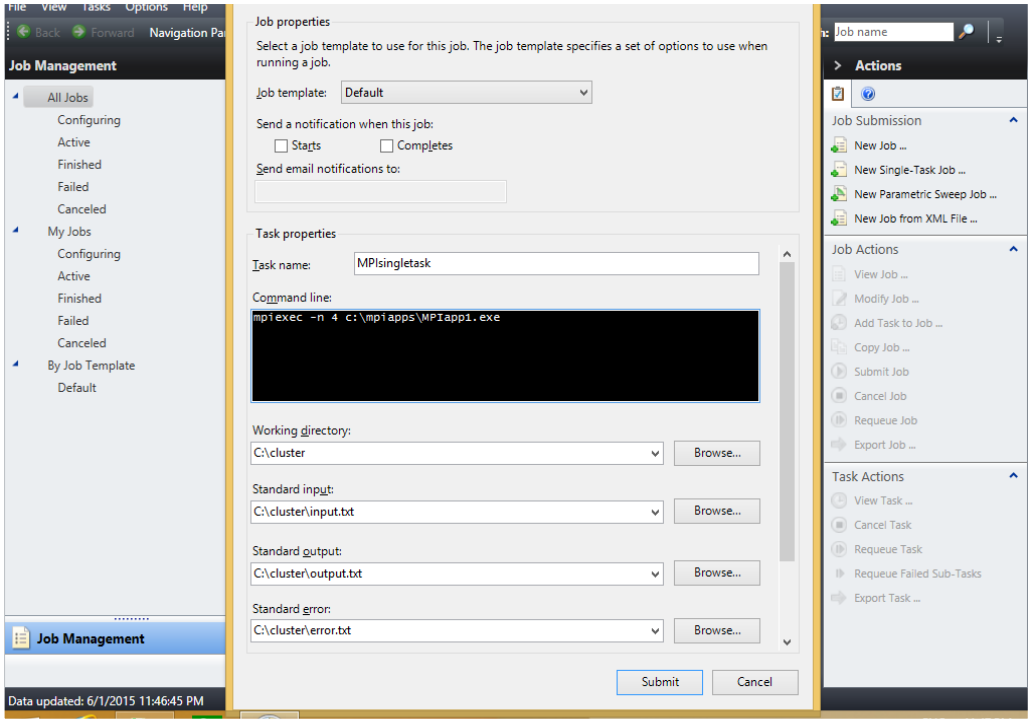
Job ID	• Numeric ID of the job. • Assigned by the Job Manager.
Job name	• Text name of the job. • User assigned.
Job template	• Name of the job template used to submit the job. • Templates define default values and constraints to the jobs and are created by the administrator.
Priority	• Numeric value 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.

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

UNIVERSIDAD DE BURGOS

Job submission & monitoring.



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.

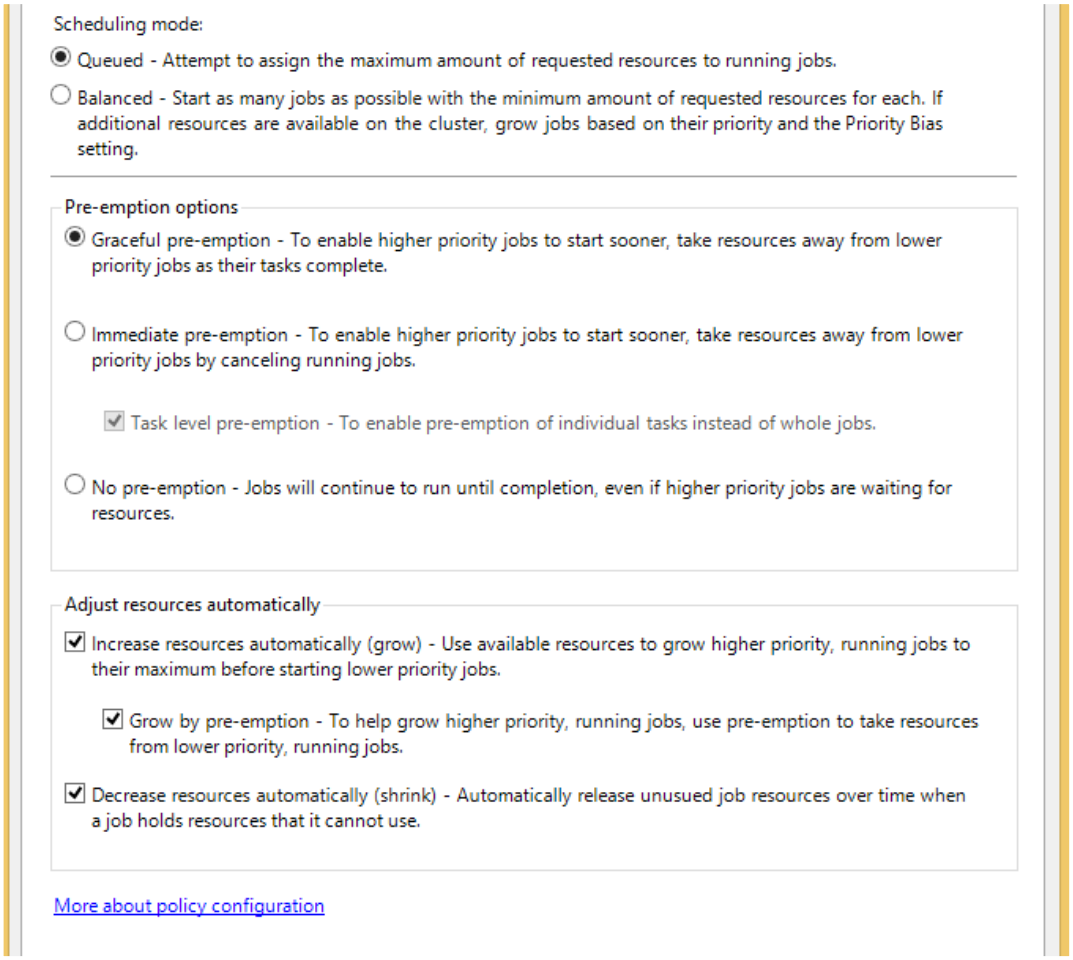
Set the command line and the output files.

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

UNIVERSIDAD DE BURGOS

Queued scheduling.



Main queued scheduling options.

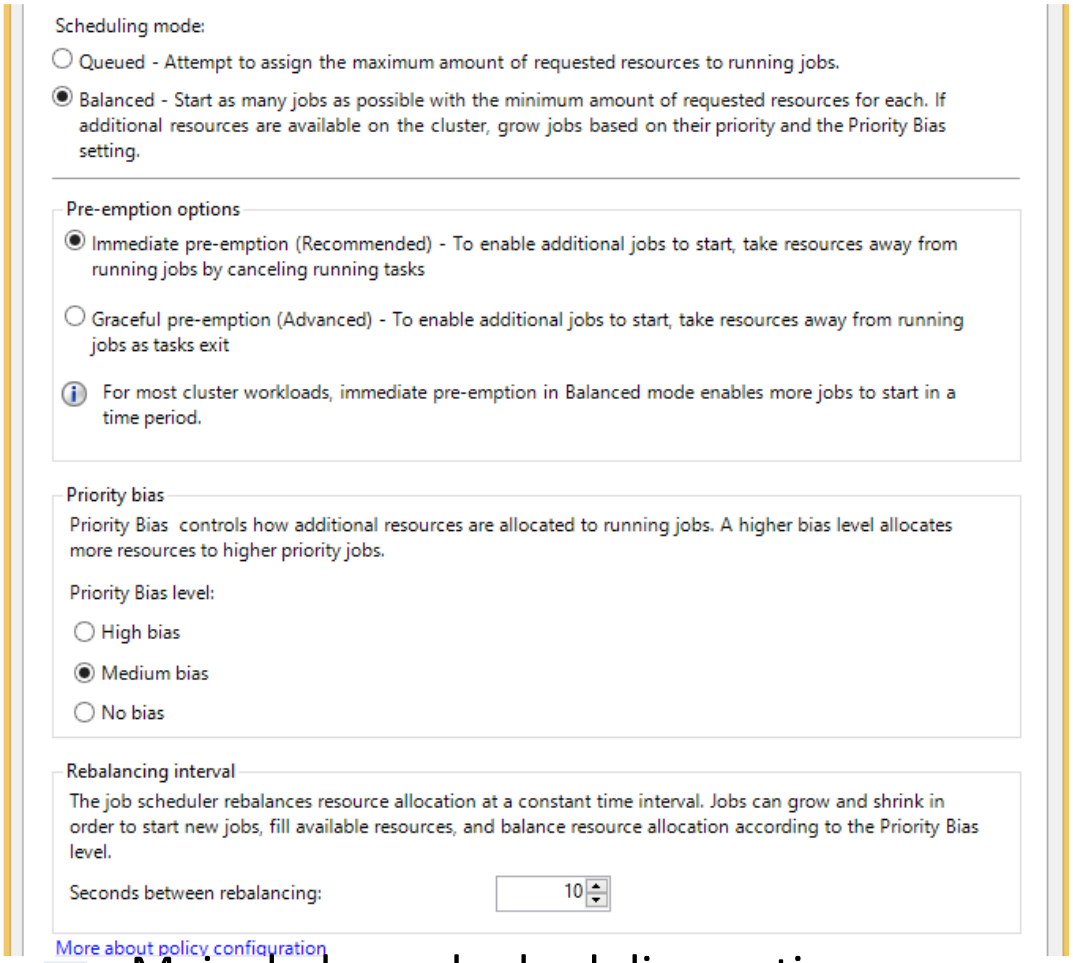
Aims to provide as many resources as possible to running jobs.

Different preemption options can be selected.

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

UNIVERSIDAD DE BURGOS

Balanced scheduling.



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