Servei Meteorològic de Catalunya

PLEC DE PRESCRIPCIONS TÈCNIQUES PER A LA CONTRACTACIÓ DEL SUBMINISTRAMENT I INSTAL·LACIÓ D'UN CLÚSTER DE CÀLCUL D'ALT RENDIMENT (HPC) PER AL SERVEI METEOROLÒGIC DE CATALUNYA

Dr. Roux, 80 1ª Planta
08017 Barcelona
Telèfon (93) 567 60 90
e-mail: contractacio.meteocat@gencat.cat

Dr. Roux, 80 1ª Planta
08017 Barcelona
Telèfon (93) 567 60 90
e-mail: contractacio.meteocat@gencat.cat

1. ABAST DEL CONTRACTE

L'objecte del contracte és el subministrament, instal·lació, configuració i posada en funcionament d'un nou entorn de càlcul d'alt rendiment (HPC) per al Servei Meteorològic de Catalunya (SMC).

Aquesta plataforma dona suport als models de predicció meteorològica i oceanogràfica de l'SMC, que s'executen actualment en un clúster dedicat d'altes prestacions per atendre el càlcul intensiu que requereixen. Les sortides d'aquests models s'utilitzen per elaborar el pronòstic operatiu de l'SMC i activar les alertes de Situació Meteorològica de Perill (SMP). També són la base per confeccionar els productes de predicció automàtica que es difonen en obert i per generar serveis a mida per a clients externs i per a altres entitats de la Generalitat de Catalunya.

El contracte inclou les següents actuacions:

- Subministrament de 8 nodes de càlcul, amb característiques equivalents o superiors a les definides al plec tècnic, per substituir sis dels dotze nodes actuals que han arribat al seu End Of Life (EOL) el 31 de desembre de 2024.
- Incorporació de sis nodes del clúster actual al nou clúster.
- Validació post-instal·lació i integració d'aquests nodes al nou entorn per garantir la coherència operativa.
- Instal·lació i posada en marxa de dos nodes de gestió a l'entorn virtual de l'SMC desplegats sobre la infraestructura VMware vSphere 8, per substituir els actuals que estan en EOL.
    - Creació de les màquines virtuals, instal·lació i configuració del sistema operatiu, del programari HPC i de la xarxa virtual (vSwitch, VLANs).
- La instal·lació i posada en marxa dels servidors al centre de processament de dades de l'SMC, incloent la configuració de la connectivitat de xarxa, emmagatzematge i gestió remota.
- Instal·lació i configuració del sistema operatiu en imatge (Linux HPC), del programari de gestió de cua (SLURM) i de les biblioteques MPI.
- Creació i configuració de volums d'scratch, dades i altres serveis necessaris sobre la cabina NetApp AFF A30 HA (vFilers, LUNs, polítiques de Snapshot, etc.).
- L'elaboració i lliurament de la documentació tècnica completa de la instal·lació realitzada, incloent esquemes, paràmetres de configuració i bones pràctiques d'administració.
- La transferència de coneixement necessària per garantir el manteniment autònom per part de l'Equip de Sistemes TIC de l'SMC.

Aquest abast cobreix tant el component de maquinari com els serveis (instal·lació, configuració, documentació i formació) necessaris per garantir la continuïtat, rendiment, eficiència i seguretat del nou clúster HPC de l'SMC, així com la compatibilitat amb futurs desenvolupaments i projectes científics.

2. SUBMINISTRAMENT I INSTAL·LACIÓ

2.2 SUBMINISTRAMENT

S'han de subministrar un mínim de 8 nodes de càlcul HPC amb el cablejat associat.

Tots els nodes de càlcul subministrats seran iguals amb components idèntics.

Cada node de càlcul haurà de complir els requeriments tècnics mínims descrits a la següent taula. En cas de no fer-ho, l'oferta quedarà desqualificada.

Element	Especificació mínima requerida
Arquitectura de CPU	Intel
Generació de CPU	6a generació
Família de CPU	Xeon
Model de CPU (de referència)	Xeon 6747P
Tipus de nucli	Performance (P-core)
Nombre de cores per CPU	48 cores físics
Freqüència base	2.7 GHz
Velocitat d'interconnexió	24 GT/s
Memòria cau L3 per CPU	288 MB
Memòria cau L3/core	6 MB/core
Nombre de CPU per node	2
Memòria RAM total	512 GB DDR5
Tipus de memòria	RDIMM, DDR5-6400, Dual Rank
Nombre de mòduls de RAM	16 x 32 GB
Canals de memòria actius	8 canals per CPU (total 16 per node)
Xarxa	4 x ports 25GbE SFP28
Emmagatzemament local	Controladora NVME amb 2x M.2 de 480 GB configurats en RAID 1
Gestió remota	Interface out-of-band
Xassís	No pot superar 1U
Subministrament elèctric	Dual, Redundant (1+1), Hot-Plug, Titanium

A continuació es descriu més detalladament les característiques i requisits mínims a complir dels diferents elements.

2.2.1 CPU

Els nodes de càlcul subministrats hauran d'estar equipats únicament amb processadors d'arquitectura Intel de 64 bits (família Xeon o equivalent), per garantir la total compatibilitat i funcionament del model WRF i dels seus components en l'entorn de càlcul.

Cada node es subministrarà amb una configuració de 2 CPUs/Sockets.

Les CPUs seran de la 6a generació de Intel amb un mínim de 48 cores físics de tipus performance (P-core) i un mínim de freqüència base de 2.7 GHz amb tots els cores actius i configurat en server mode. No s'admeten processadors tipus E-core ni híbrids.

La velocitat d'interconnexió entre CPUs (UPI) serà mínim a 24 GT/s i tindrà almenys 4 UPI Links.

La memòria cau L3 total per CPU haurà de mantenir una relació mínima de 6MB per core. Per exemple en el cas d'una CPU de 48 cores, el mínim de memòria cau L3 serà de 288 MB.

2.2.2 Memòria RAM

Cada node de càlcul subministrat haurà d'estar equipat amb un mínim de 512 GB de memòria RAM.

El tipus de memòria serà DDR5 o superior, amb una velocitat mínima efectiva de 6400 MT/s, i ha de garantir una alta capacitat de transferència i baixa latència per a càrregues de treball intensives i paral·leles típiques dels entorns HPC.

Els mòduls de memòria seran de tipus RDIMM Dual Rank o MRDIMM (Multiplexed Rank DIMM) tot garantint la compatibilitat amb les arquitectures de memòria i processador especificades.

La distribució dels mòduls haurà d'omplir tots els canals de memòria disponibles per CPU per garantir el màxim ample de banda i rendiment, minimitzant qualsevol penalització per accés no balancejat.

Tots els mòduls subministrats seran d'idèntiques característiques (mida, velocitat, tipus, tecnologia, etc.). La memòria haurà de ser ECC (Error-Correcting Code) i suportar operació contínua 24/7 a plena càrrega en entorns d'alta temperatura dins dels marges operatius del node.

2.2.3 Xarxa

Cada node de càlcul haurà de disposar d'un mínim de 4 ports de 25GbE SFP28. Com que tots els elements de la part de xarxa han d'estar redundats i no poden presentar un únic punt de fallada, els 4 ports hauran d'estar repartits, com a mínim, en dues targetes de xarxa dual port 25 GbE SFP28.

Les targetes hauran de ser NVIDIA ConnectX-6 Lx Dual Port 25GbE SFP28 o similars per cobrir les següents característiques mínimes:

- Tipus: Ethernet d'alt rendiment amb baixa latència, especialment enfocades per a entorns HPC i data centers.
- Ample de banda: 25 Gbps full-duplex per port.
- Interfície de bus: PCI Express Gen4 x8.
- Factor de forma: OCP 3.0 o PCIe, segons la compatibilitat amb el servidor
- Suport complert per RDMA (Remote Direct Memory Access*)* sobre Ethernet, mitjançant RoCEv2 (RDMA over Converged Ethernet version 2).
- Compatibilitat amb NVMe over Fabrics (NVMe-oF) i accés directe a emmagatzematge remot.
- Offload accelerat de funcions de xarxa (TCP/UDP/IP, SDN, iSER, etc.), reduint la càrrega de la CPU i millorant el rendiment global del sistema.
- Compatibilitat amb S.O. Linux i entorns HPC (OpenMPI, OFED, etc.).
- Qualitat de servei (QoS) avançada per a prioritzar el trànsit crític en entorns de càlcul intensiu.
- Suport per a cables AOC SFP28 amb capacitat d'auto-negociació millorada per a millor rendiment.

2.2.4 Gestió i supervisió remota

Tots els nodes de càlcul hauran d'incloure un sistema d'administració remota (out-of-band) que, com a mínim, permeti:

- Realitzar on/off de l'equip.

- Accés remot complet a la consola.
- Monitoratge de l'entorn (temperatura, consum, etc.).
- Generació d'alertes.
- Detecció de problemes d'hardware i firmware.
- Estat dels leds.
- Accés a la BIOS i actualització en remot de BIOS/firmware.

2.2.5 Xassís, alimentació i format físic

Cada node haurà de disposar de:

- Format 1U optimitzat per alta densitat en rack estàndard de 19" .
- Sistema de ventilació d'alt rendiment i redundància activa, dimensionat per garantir el correcte refredament del node en operacions sostingudes de càlcul intensiu.
- Dues fonts d'alimentació redundants (1+1), tipus Hot-Plug, amb una eficiència mínima 80 Plus Titanium. La potència total haurà de ser suficient per assegurar tolerància total a fallades (full power redundancy) en condicions de màxim consum.
- Rails extensibles (sliding rails). Hauran de permetre l'extracció frontal del node per a operacions de manteniment i disposar, si és possible, de mecanisme de retenció o bloqueig.
- 2 cables de corrent **C13 to C14**, tipus PDU, de mínim 1 metre de longitud i 10 ampers.

2.2.6 Cablejat

Tots els nodes de càlcul subministrats hauran de connectar-se als switchos Nexus mitjançant enllaços a 25 Gbps a través de interfícies SFP28.

Per tant s'han de proveir 4 cables de fibra òptica i 8 transceptors SFP28 (en cas que no estiguin integrats als propis cables de fibra), per cada node de càlcul, amb una velocitat de transferència de dades de 25 Gbps.

A causa de les limitacions físiques del recorregut del cablejat (pas per bastidors i corbes tancades), **no s'admetrà l'ús de cables tipus DAC (Active Direct Attach Copper)**. Per garantir la flexibilitat i fiabilitat de la instal·lació, només s'acceptaran les següents opcions de connexió:

- Cables de fibra òptica multimode tipus OM4 o superior amb transceptors SFP28 SR .
- Cables 25GBase AOC amb transceptors SFP28 integrats.

Per la distribució dels equips als racks habilitats per la instal·lació del clúster al CPD i les distàncies calculades entre els equips, **s'haurà de subministrar un mínim de**:

- **4 cables de 10 metres per node**

Per garantir la compatibilitat, la interoperabilitat i l'òptim rendiment, els cables i transceptors hauran d'estar certificats per l'ús amb els switchos Cisco Nexus 93180YC-FX. També s'ha de garantir la compatibilitat amb les targetes de xarxa dels equips i el funcionament amb RDMA sobre Ethernet (RoCEv2) sense degradació de rendiment ni latència.

D'altra banda, cal tenir en compte també el cablejat per les consoles de gestió remota (punt 2.2.4):

- Cables ethernet categoria 6A mínim. 1 per cadascun dels nodes nous subministrats.
- Mida mínima de 7,5 metres.

2.3 CONDICIONS DE LA GARANTIA/SUPORT DEL MANTENIMENT

Garantia/suport de **5 anys** de maquinari tipus Next Business Day Onsite Service o similar amb les següents característiques:

- Assistència tècnica 24x7: Accés a suport tècnic telefònic i online les 24 hores del dia, 7 dies a la setmana, per a maquinari.
- Servei in situ el següent dia laborable: Si després del diagnòstic remot es determina que cal una reparació, un tècnic de es despla3çarà a les instal·lacions el següent dia laborable per substituir peces o reparar el dispositiu.
- Cobertura de peces i mà d'obra: La substitució de components defectuosos i la mà d'obra estan inclosos durant els 5 anys de contracte.
- Diagnòstic remot: Abans de l'enviament d'un tècnic, es farà un diagnòstic remot per intentar resoldre el problema a distància.
- Gestió de casos i informes: Possibilitat de crear i gestionar casos de suport via portal web i accedir a informes d'estat dels equips.
- Accés a actualitzacions i millores del firmware dels servidors coberts.

3. SERVEIS PROFESIONALS

Els serveis objecte d'aquest plec han de complir les especificacions tècniques que es descriuen a continuació

3.1 MAQUINARI

3.1.1 Nodes de gestió del clúster

Es crearà, instal·larà i configurarà un conjunt de dues màquines virtuals (VMs) a la plataforma VMware vSphere 8 de l'SMC, que actuaran com a nodes de gestió del clúster HPC.

Aquestes màquines assumiran les funcions pròpies de nodes *head* i de serveis centrals del clúster, incloent, però no limitant-se a:

- Gestió de connexions d'usuari (login).
- Gestió del planificador de cues i recursos.
- Administració i distribució d'imatges de sistema operatiu.
- Monitoratge i supervisió del clúster.
- Servei de temps (NTP), resolució de noms (DNS), i altres serveis bàsics necessaris.

Les dues VMs es dimensionaran amb els recursos suficients i adequats, en coordinació amb l'equip tècnic de l'SMC, per garantir un funcionament estable i eficient dels serveis associats.

També s'hi haurà d'instal·lar el sistema operatiu Linux acordat prèviament amb l'equip tècnic, amb la corresponent configuració per a entorn HPC. Igualment, s'hi haurà d'instal·lar, configurar i compilar, si escau, tot el programari de gestió i suport al càlcul detallat a l'apartat 3.2 del present plec.

3.1.2 Nodes de càlcul

3.1.2.1 Nodes de càlcul a subministrar

Els nodes de càlcul del clúster s'instal·laran físicament als racks llogats per l'SMC al Centre de Processament de Dades (CPD) de T-Systems a Cerdanyola del Vallès, on es troben la resta d'elements que integren l'entorn HPC.

La instal·lació del maquinari al CPD s'haurà de dur a terme d'acord amb la normativa i les directrius establertes per T-Systems per als seus centres de dades.

Aquesta normativa serà facilitada al licitador a l'inici de l'execució del projecte, i serà de compliment obligatori durant tot el procés d'instal·lació i integració física. S'haurà de fer:

- Instal·lació física dels servidors
- Connexió de xarxa dels servidors
- Etiquetatge del cablejat

3.1.2.2 Nodes de càlcul existents a integrar

Els nodes de càlcul del clúster HPC actual, Dell R640, només caldrà incorporar-los al nou clúster. Aquesta part inclourà un canvi de configuració lògica de xarxa (reassignació d'adreçament IP i VLAN associada, sense afectació de la infraestructura física ni del maquinari de xarxa existent.

3.1.3 Configuració de xarxa

El clúster disposarà de 3 xarxes pel seu funcionament:

- Xarxa de càlcul: basada en tecnologia 25 Gb Ethernet (RoCE), visible entre els nodes del clúster que permetrà el trànsit generat per les aplicacions MPI.
- Xarxa de dades: basada en tecnologia 25 GB Ethernet (RoCE), per accés dels nodes de càlcul i gestió als volums exportats per NFS de la cabina NetApp i dels diversos serveis del clúster (NTP, DNS, etc.).
- Xarxa de gestió: basada en tecnologia 1GB Ethernet per l'administració remota dels nodes de càlcul (out-of-band).

En els nodes de càlcul subministrats, els ports 25 GbE SFP28 s'organitzaran per proporcionar dues connexions de xarxa redundants i separades evitant un únic punt de fallada. Es configurarà a cada node:

- Un *Link Aggregation Group* (LAG) de dos ports (un de cada NIC) dedicat al trànsit de dades cap a cabines (xarxa de dades del clúster). Cada port anirà punxat a un switch Nexus diferent de l'stack.
- Un altre LAG de dos ports (un de cada NIC) per al trànsit inter-nodes, optimitzat per a la comunicació MPI i el rendiment de càlcul paral·lel (xarxa del clúster) i pel trànsit de serveis del clúster (Slurm, DNS, NTP, etc.). Cada port anirà punxat a un switch Cisco Nexus diferent de l'stack.
- Cada LAG estarà configurat en mode IEEE 802.3ad amb, com a mínim, els següents paràmetres:
    - Mode actiu amb detecció de fallada.
    - Balanceig de càrrega per hash de capa 2+3 (MAC/IP).

- Bonding mode.

Als ESXi, al VMWare vSphere i a les VMs dels nodes de gestió s'hauran de fer les configuracions necessàries per l'accés a les VLANs indicades tot garantint el seu correcte funcionament.

També caldrà configurar als commutadors Cisco Nexus 93180Y tant els ports utilitzats per les connexions LAG com els port individuals, amb la parametrització adient per garantir un rendiment òptim per cada ús, i les VLANs corresponents. I el mateix per l'accés dels ESXi.

3.1.4 Emmagatzemament

L'espai de disc necessari pel clúster HPC es proporcionarà mitjançant una cabina NetApp AFF-30 de l'SMC, equipada amb discos NVMe i SSD. Aquesta cabina està connectada als switchos Nexus i proporciona connectivitat a 25 Gbps per als nodes de càlcul i gestió.

S'hauran de realitzar totes les configuracions necessàries a la cabina NetApp per tal de garantir la integració completa amb el clúster HPC i un rendiment i accés òptim dels nodes i del sistema.

Aquestes tasques inclouran la:

- Creació i configuració de VLANs i interfícies lògiques (LIFs) per a la xarxa de clúster i la xarxa de dades, assegurant l'aïllament necessari i també la redundància i l'alta disponibilitat segons les millors pràctiques de NetApp.
- Creació i configuració de Storage Virtual Machines (SVMs/vFiler) específiques per servir els volums requerits pel clúster HPC.
- Configuració i exportació dels volums necessaris, definits en coordinació amb l'equip tècnic de l'SMC, incloent:
    - Volums *scratch* d'alt rendiment per a càlcul temporal (per entorns).
    - Volum per resultat de compilacions o dades generades pels models.
    - Volum destinat a dades inicials de terreny.
    - Volum per a codis font, llibreries, compiladors (incloent Intel) i dependències, i imatges de sistema i configuracions centralitzades gestionades per Warewulf.
    - Volum per donar suport als arxius de configuració dels models meteorològics, incloent-hi els arxius dels dominis dels models i complements específics com els del model WRF.
    - Volum final que contindrà les dades resultants dels models compilats.

Les configuracions a la cabina – tant de les SVMs com dels volums – hauran d'estar optimitzades per a I/O intensiu, seguint les guies tècniques i bones pràctiques del fabricant (NetApp), incloent:

- Paràmetres de mida de bloc i striping adequats per a HPC.
- Preallocació d'espai per evitar fragmentació i garantir rendiment sostingut.
- Configuració de protocols d'accés (NFS v3/v4, pNFS, NVMe-oF si s'escau) per maximitzar la taxa de transferència i minimitzar la latència.
- Implementació de grups d'interfícies (LACP/ifgrp) per agregació d'ample de banda i redundància, si la infraestructura ho permet.

- Optimització de la configuració NFS (ajust de rsize/wsize, delegació, etc.) segons les necessitats del clúster.
- Qualsevol configuració addicional per millorar l'accés a les dades.

L'objectiu final és garantir que l'accés a disc no esdevingui un coll d'ampolla i que el rendiment global del clúster —tant en càlcul com en gestió— sigui òptim i estable en situacions de càrrega intensiva.

Preservació dels serveis existents

La cabina NetApp AFF-30 ja proporciona serveis d'emmagatzematge per a altres sistemes crítics de l'SMC, com ara LUNs per a bases de dades Oracle i volums NFS i LUNs per a datastores de màquines virtuals.

Per tant, qualsevol configuració o adaptació relacionada amb la integració del clúster HPC haurà de garantir la compatibilitat i la no afectació del funcionament, rendiment ni disponibilitat dels serveis existents.

Caldrà coordinar i consensuar prèviament amb l'equip tècnic de l'SMC qualsevol canvi que pugui tenir impacte sobre aquests serveis, així com adoptar les mesures de seguretat, segmentació i bones pràctiques que estableixi el fabricant NetApp o l'empresa encarregada del suport i manteniment de la cabina de discos.

3.2 PROGRAMARI

El projecte preveu la instal·lació d'un conjunt integral de programari especialitzat per a la gestió i explotació d'un clúster de càlcul d'altes prestacions (HPC) basat en OpenHPC. Aquesta plataforma proporciona una base estandarditzada i robusta, amb un ampli catàleg de paquets precompilats i eines de gestió, que garanteixen la coherència, la seguretat i la facilitat de manteniment de tot l'entorn.

L'objectiu és disposar d'una infraestructura flexible, escalable i altament disponible, capaç de donar resposta a les necessitats de càlcul científic, simulació i processament de dades avançat, optimitzant l'ús dels recursos i facilitant la integració de nous nodes o serveis en el futur.

Si durant el desenvolupament del projecte es detecten mancances, necessitats o funcionalitats no previstes en el present plec, i es plantegen adaptacions, caldrà que siguin aprovades per l'equip tècnic de l'SMC i no implicaran cap cost addicional.

3.2.1 Programari HPC

El sistema disposarà d'eines essencials per a la gestió, el monitoratge i la garantia d'alta disponibilitat del clúster. S'hi inclourà el gestor de mòduls d'entorn Lmod, que facilita el canvi de versions de programari i dependències; el planificador de cues Slurm, per a l'assignació, execució i seguiment de treballs en paral·lel; i la seva interfície web, Slurm-web, que ofereix una visualització intuïtiva i en temps real de l'estat dels treballs i dels recursos disponibles.

Per a la gestió i desplegament centralitzat dels nodes de càlcul s'utilitzarà Warewulf, que permet automatitzar la creació i manteniment d'imatges de sistema i configuracions.

A més, es desplegaran Corosync i Pacemaker per garantir l'alta disponibilitat dels serveis crítics, assegurant la continuïtat operativa davant possibles fallades de components:

Programari	Funció principal
Lmod	Gestió de mòduls d'entorn per a facilitar el canvi de versions de programari i dependències
Slurm	Gestor de cues i recursos per a llançar i monitorar treballs en el clúster
Slurm Web / Slurm Dashboard o similar	Visualització en temps real l'estat dels treballs, historial, durada, erros i ús de recursos
Warewulf	Aprovisionament i administració d'imatges per a nodes de còmput
Corosync/Pacemaker	Gestió de clústers d'alta disponibilitat: Corosync proporciona la comunicació i sincronització entre nodes, mentre que Pacemaker gestiona els recursos i la recuperació automàtica en cas de fallada

3.2.2 Compilació i llibreries

El clúster estarà equipat amb un conjunt ampli de compiladors i llibreries científiques optimitzades per al càlcul d'altes prestacions. Es preveuen compiladors GNU i Intel per a C, C++ i Fortran, així com llibreries bàsiques i avançades per a àlgebra lineal, transformades de Fourier, gestió de formats científics (NetCDF, HDF5), processament de dades meteorològiques i geoespacials, i eines d'anàlisi i visualització. Aquesta selecció garanteix la compatibilitat amb els principals models de simulació i aplicacions científiques, i proporciona la base necessària per a la compilació, execució i optimització de fluxos de treball complexos en entorns HPC:

Programari	Funció principal
Compiladors Intel Parallel Studio versió 18.0.3 (és necessari adquirir llicència amb Intel)	Versió antiga compiladors Intel per compilar WRF DI, Bolam i Moloc
Compiladors Intel Parallel Studio versió actual	Versió actual compiladors Intel per compilar WRF i altres
OpenMPI/MPICH	Llibreries per a computació paral·lela via MPI
IMPI compilat amb icx i ifx	Implementació d'Intel MPI per a comunicació paral·lela eficient en arquitectures Intel
GCC/GFortran	Compiladors per a C, C++, Fortran
NetCDF/HDF5 compilades amb icx, ifx + IMPI	Gestió de formats de dades científiques multidimensionals
OpenBLAS	Llibreria BLAS/LAPACK optimitzada per a càlculs d'àlgebra lineal d'alt rendiment
FFTW	Llibreries per calcular transformades de Fourier
CDO/CDI	Eines per a processament, conversió i anàlisi de dades climàtiques
eccodes	Biblioteca per a lectura, escriptura i manipulació de GRIB/BUFR
ncview	Visualitzador lleuger per a arxius NetCDF
Jasper	Compressió d'imatges, utilitzada per a GRIB2
GDAL	Conversió i manipulació de formats geoespacials
Python	Llenguatge d'scripts, anàlisi de dades, automatització
NCL/Ncarg	Llenguatge i llibreria per a visualització i anàlisi de dades científiques
ecFlow	Gestor de fluxos de treball i dependències per a cadenes de simulació
prun	Llançament paral·lel de processos
pdsh	Execució de comandes en paral·lel a múltiples nodes
Zabbix / Integració Zabbix amb SlurmWeb	Monitoratge de rendiment i visualització de mètriques

4. DESCRIPCIÓ DE LES FASES DEL PROJECTE

4.1 FASE 1: INSTAL·LACIÓ I CONFIGURACIÓ

4.1.1 Fase 1.1: Instal·lació i configuració dels nodes de gestió en l'entorn virtual

Objectiu:

Instal·lar i configurar els dos nodes de gestió en l'entorn virtual VMware vSphere 8 per proporcionar els serveis centrals del clúster HPC.

Tasques:

- Creació i configuració de les VMs i instal·lació del SO.
- Creació de dues VLANs privades específiques:
  - **VLAN de càlcul**: per a la comunicació d'alt rendiment entre nodes de càlcul (*intra-node traffic*).
  - **VLAN de dades**: per a la comunicació amb el sistema de fitxers compartit basat en NetApp AFF i els serveis de clúster.
  - Creació i configuració de la VLAN de dades a la NetApp.
- Assignació de vNICs dedicades als nodes de gestió, una per VLAN.
- Configuració dels switchos Cisco Nexus 93180Y, incloent:
  - Definició de les VLANs.
  - Assignació de ports i etiquetatge (tagged/untagged segons el cas).
  - Polítiques de QoS si escau.
- Integració amb la cabina NetApp
  - Creació i configuració de les VLANs i interficies (LIFs) corresponents a la xarxa de dades dins la NetApp AFF-30.
  - Creació dels vFiler (SVMs) per servir els volums del clúster HPC.
  - Definició i parametrització per l'ús en un HPC dels volums necessaris:
    - Volums *scratch* d'alt rendiment per a càlcul temporal (per cada tipus d'entorn: pro, pre, dev...)
    - Volum *warewulf* per a les imatges de nodes.
    - Volum pels resultats de compilacions o dades generades pels models.
    - Volum per a codi font, llibreries i dependències.
    - Altres volums específics segons les necessitats que es detectin durant l'execució del projecte.
- Instal·lació, configuració i, si s'escau, compilació del programari indicat al punt 3.3. segons els següents criteris:
  - Utilitzar sempre les darreres versions estables que garanteixin la integració i compatibilitat amb la resta del programari i amb el hardware i el sistema operatiu de la infraestructura.
  - Es poden admetre alternatives, canvis i/o programari addicional a l'indicat durant el desenvolupament del projecte sempre i quan siguin acordats amb l'equip tècnic de l'SMC.
- Proves de HA (alta disponibilitat):
  - Testos de tolerància a fallades entre els nodes de gestió.
  - Comprovació del failover automàtic dels serveis crítics:

        - Slurm.
        - Serveis generals HPC
        - Accés als volums HPC NetApp.
    - Validació de que no hi hagi interrupció del servei ni pèrdua de dades en cas de fallada d'un node gestió.

4.1.2 Fase 1.2: Instal·lació i configuració dels models de càlcul

Objectiu:

Instal·lació física dels nous nodes i configuració i integració al cluster HPC.

Tasques:

- Verificació in situ de totes les especificacions del maquinari segons plecs tècnics:
    - Quantitat i tipus de RAM (ex: DDR5 RDIMM/MRDIMM, canals de memòria plens).
    - CPUs, discs locals, NICs i fonts d'alimentació redundants.
    - Estat del firmware i versió de BIOS.
- Instal·lació física i enrackament als racks disponibles
    - Instal·lació dels nodes als racks assignats (1U) amb rails compatibles.
    - Connexió elèctrica.
    - Connexió de NICs als switchos Nexus.
    - Validació de fonts redundants i ventiladors en operació Hot-Plug.
- Configuració inicial de BIOS i firmware
    - Configuració de la BIOS en mode "Performance" (optimització per càlcul).
    - Desactivació de l'Hyper-Threading si no és beneficiós pel perfil de càrrega.
    - Validació o actualització de BIOS i microcodi si és necessari.
    - Activació de funcions com NUMA, Turbo Boost, etc.
- Connexió als switches Nexus d'interconnexió.
- Configuració de HA de connexió :
    - Configuració de LAG als nodes i switchos Nexus.
    - LAG de dades (VLAN de dades).
    - LAG de trànsit inter-nodes (VLAN càlcul).
- Emmagatzematge local
    - Configuració de RAID 1 als discs locals .
    - Validació de coherència entre controladora RAID i SO (via smartctl, mdadm, o eina del fabricant).
    - Assignació de volums temporals si escau (/tmp, /var/tmp, swap...).
- Integració amb els serveis del clúster
    - Alta de nodes al servidor Warewulf (nodes gestió)
    - Assignació d'IP i MAC al fitxer de configuració.
    - Creació i configuració d'imatge per nodes amb OS, Slurm i mòduls.
    - Desplegament automàtic d'imatges.
    - Test de provisionament per PXE i boot correcte des de la imatge.
- Documentació
    - Registre complet per node incloent número de sèrie, MACs, IPs, firmware, BIOS, configuració RAID, etc.

Proves HA:

- Simular la caiguda d'un node de càlcul mentre s'està executant un procés intensiu
- Verificar com afecta aquesta caiguda al sistema Slurm:
    - Com es comporta el planificador davant la fallada.
    - Si el procés es reinicia automàticament o es cancel·la.

  - Si es registra correctament a logs i si el sistema detecta la fallada.
- Entregable: informe detallat amb conclusions i recomanacions.

4.2 FASE 2: COMPILACIÓ MODELS METEOROLÒGICS

Objectiu:

Un cop finalitzada la instal·lació i configuració del clúster, aquesta fase té com a objectiu preparar l'entorn de càlcul perquè sigui plenament operatiu per a l'execució dels models meteorològics. Per fer-ho, l'equip tècnic del Servei Meteorològic compilarà i ajustarà els diferents models i dependències necessàries, optimitzant-los per a la nova infraestructura i verificant-ne el correcte funcionament abans de passar a la fase de validació i proves de rendiment.

4.3 FASE 3: PROVES DE RENDIMENT I ESTABILITAT

Objectiu:

Comparar el rendiment del nou sistema amb el sistema actual mitjançant proves senzilles i objectives.

4.3.1 Prova comparativa de rendiment i funcionalitat d'un model entre el nou clúster i l'actual

Amb l'objectiu de garantir l'excel·lència i la millora substancial del clúster HPC de càlcul nou, es requereix la realització d'una prova comparativa entre el nou clúster HPC instal·lat i l'actual clúster HPC en ús.
Aquesta prova consistirà a executar un model científic de referència, el qual serà compilat i optimitzat expressament per al nou entorn, utilitzant els recursos, compiladors i biblioteques més actuals proporcionats pel nou sistema.
El mateix model, amb la seva configuració i dades associades, està sent executat a l'entorn de producció vigent, on es disposa dels registres de rendiment i resultats.

Cal remarcar que les condicions de la prova no poden ser completament equiparables, ja que el clúster actual no serà modificat en cap aspecte per a aquesta comparativa; en canvi, al nou clúster, la compilació i les execucions es realitzaran aprofitant la nova arquitectura, versions actualitzades de compiladors i llibreries, i possibles millores inherents a l'actualització tecnològica.
L'objectiu d'aquesta prova és demostrar, de manera objectiva, la millora en el rendiment, la fiabilitat i la funcionalitat obtinguda gràcies al desplegament del nou clúster HPC, observant el comportament del model en ambdós entorns.

A partir dels resultats de la comparativa, es validarà el compliment dels requeriments de la licitació quant a la millora substancial del servei de càlcul científic institucional.

4.3.2 Prova de transferència de fitxers i accés a l'emmagatzematge

Amb la finalitat de verificar el correcte accés, rendiment i fiabilitat de la solució d'emmagatzematge implementada al nou clúster HPC, es requerirà la realització d'una prova específica de transferència de fitxers utilitzant l'eina estàndard Iozone Filesystem Benchmark. Aquesta prova tindrà com a objectiu avaluar, de manera objectiva i controlada, la resposta del sistema d'arxius

davant operacions intensives de lectura, escriptura i accés concurrent a l'storage per part dels nodes de còmput del nou entorn.

La metodologia proposada inclou l'execució de la bateria de tests d'iozone en les principals particions i punts de muntatge designats per a dades, emprant una configuració que permeti reflectir tant casos d'ús habituals com escenaris de màxim rendiment. S'analitzaran diferents tipus d'operació (escriptura seqüencial, lectura, relectura, escriptura aleatòria, etc.) i mides de fitxer variables, per tal de garantir una visió àmplia del comportament del sistema sota càrregues diverses.

És important ressaltar que el disseny detallat de la prova, incloent paràmetres, nombre de processos i distribució dels tests, es podrà modular segons les característiques específiques del sistema i els compromisos d'implementació que es determinin en el procés. L'objectiu final és garantir que l'emmagatzematge compleix amb els criteris d'eficiència, escalabilitat i disponibilitat requerits pel servei, demostrant el salt qualitatiu respecte a l'arquitectura prèvia.

4.3.3 Prova de connectivitat entre nodes de càlcul

Per avaluar el rendiment i la latència en les comunicacions entre els nodes de càlcul del nou clúster HPC, es proposa la realització d'una prova de benchmark utilitzant l'eina OSU Micro-Benchmarks (OMB), reconeguda internacionalment per mesurar la qualitat i eficàcia de les operacions MPI en entorns distribuïts. Aquesta eina permet quantificar amb precisió paràmetres clau com la latència punt a punt i l'amplada de banda de la xarxa interna, fonamentals per garantir un funcionament òptim dels processos paral·lels i la coordinació entre nodes.

Es deixarà oberta la possibilitat d'emprar altres eines o benchmarks similars que comparteixin característiques equivalents en quant a rigor i tipologia de proves, per tal d'adaptar-se a possibles requeriments tècnics o preferències d'execució que puguin sorgir.

Pel que fa a la validació dels resultats, aquests s'hauran d'interpretar en base a paràmetres establerts com a referència, incloent la latència mínima assolible i la màxima amplada de banda per rangs de mida de missatges, així com la coherència i estabilitat de les mesures en diferents iteracions i condicions. Aquesta anàlisi servirà per detectar possibles colls d'ampolla, latències anòmales o pèrdues de rendiment, comparant els valors obtinguts amb els estàndards de rendiment esperats per la tecnologia de xarxa implementada i amb dades de sistemes similars.
Si s'escau, els criteris de validació i els procediments d'anàlisi es podran ajustar i detallar conjuntament amb l'empresa adjudicatària per adaptar-los a les especificitats de l'entorn implementat.

4.3.4 Prova d'operativa i estabilitat

Amb l'objectiu de garantir la robustesa, la fiabilitat i el correcte comportament del nou clúster HPC sota condicions d'ús reals, es preveu la realització d'una prova d'estabilitat basada en la simulació de l'operativa habitual de l'SMC.

Aquesta prova consisteix a executar, durant un període aproximat a una setmana, un conjunt de models meteorològics i fluxos de càlcul representatius de l'activitat diària, seguint la planificació esperada i coincidint simultàniament en el temps, tal com succeeix en l'entorn de producció actual.

L'objectiu és monitorar que els processos es llancen en els horaris previstos, que s'executen sense incidències, i que els temps de resposta i d'execució es mantenen dins dels marges esperats. Es farà especial èmfasi en detectar qualsevol tipus d'anomalia relacionada amb l'estabilitat del

sistema: errors d'execució, detencions inesperades, alentiments progressius o degradació en la gestió de la càrrega. S'analitzaran també possibles variacions en les prestacions entre diferents execucions sota càrregues variables, per tal de garantir que el nou entorn pot suportar l'activitat habitual de manera fiable i consistent.

Els resultats d'aquesta prova d'estabilitat permetran validar que el clúster pot assumir la càrrega operativa real, mantenint la qualitat, l'estabilitat i l'eficiència dels serveis de càlcul, i suposaran l'evidència objectiva de què la nova infraestructura compleix els requisits funcionals i de rendiment sol·licitats, esdevenint així un element fonamental per a la validació i acceptació del sistema.

4.3.5 Prova de validació del LAG o Bonding entre les dues interfícies de xarxa

Amb l'objectiu de validar el correcte funcionament de la configuració de bonding o LAG (Link Aggregation Group) entre les interfícies físiques duals de 25 Gb als servidors de càlcul, es proposa dur a terme una prova integral que inclogui la configuració del bonding en mode LACP (IEEE 802.3ad).

L'objectiu és comprovar que les dues interfícies es fusionen correctament en una única interfície lògica, augmentant l'amplada de banda agregada i mantenint alta disponibilitat i tolerància a fallades segons la política de bonding configurada (per exemple, mode LACP 802.3ad o balance-alb). La prova ha de demostrar que, sota càrregues reals o simulades, el trànsit es reparteix entre ambdues interfícies segons el mode establert, i que en cas d'averia d'una de les interfícies el trànsit continua sense interrupcions remarcables.

Metodologia de prova:

- Configuració del bonding en mode LACP (IEEE 802.3ad) entre les dues NICs de 25 Gb del servidor, connectades al switch Nexus.
- Validació de la configuració i l'estat del bonding (cat /proc/net/bonding/<bond_interface>, ethtool, ip link o eines específiques del sistema).
- Execució de tests de transferència de dades simulant trànsit intens i multiconnexió, mitjançant eines per mesurar l'amplada de banda i la latència, com ara **iperf3**, enviant fluxos paral·lels de dades per assegurar l'ús eficient de totes dues interfícies.
- Prova de tolerància a fallades, deshabilitant una de les interfícies mentre es manté la transferència activa, verificant que el bonding fa failover immediat i el trànsit no es veu afectat dràsticament.
- Opcionalment, es poden dur a terme proves de càrrega amb múltiples clients o fluxos per confirmar que el balancejament de càrrega funciona correctament en diversos escenaris i que no es generen colls d'ampolla en una sola interfície.

Validació dels resultats:

- Confirmar que l'amplada de banda agregada (teòrica i pràctica) s'ajusta a l'esperada, idealment propera a la suma de les dues interfícies (fins a 50 Gbps teòrics), segons el tipus i nombre de fluxos paral·lels.
- Verificar que la latència i la càrrega de CPU no es degradin per sobre dels valors habituals.

- Comprovar l'absència de pèrdues de paquets i desconnexions durant la prova, especialment en escenaris de failover.
- Analitzar la distribució del trànsit amb eines d'informació de la targeta de xarxa i del switch per assegurar que el algoritme de hashing o balançament de càrrega està operant correctament.
- Assolir estabilitat i repetibilitat en els resultats amb diferents iteracions de prova.

4.4 FASE 3 – INTEGRACIÓ DELS NODES ANTICS AL NOU CLÚSTER

Observació important:

La fase 3 no podrà ser iniciada fins que el Servei Meteorològic de Catalunya no hagi migrat completament els models actuals del clúster antic al nou clúster, assegurant-ne la coherència i estabilitat del nou entorn.

Tasques:

- Afegir les subxarxes necessàries/vlans als diferents elements de la xarxa CPD, per tal de poder integrar els nodes de càlcul actuals al clúster nou.
- Reconfiguració de la xarxa dels nodes de càlcul, per poder integrar-los al nou clúster.
- Integració al Warewulf dels nodes de càlcul antics a migrar al nou entorn.
- Arrencada i validació del funcionament.

4.5 FASE 4 – VALIDACIÓ FINAL DEL SISTEMA

Objectiu:

Confirmar el funcionament integral del clúster.

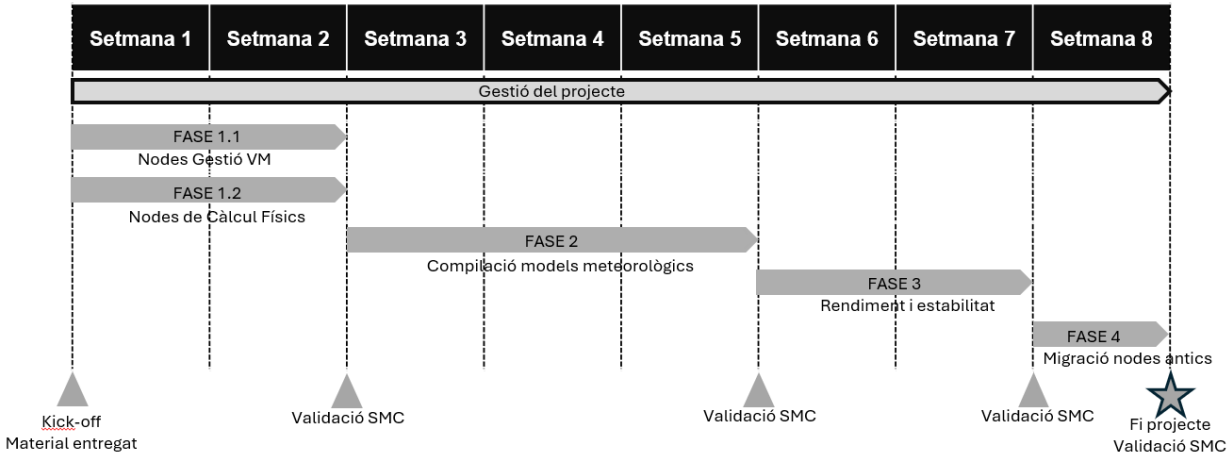
Tasques:

- Confirmació del funcionament del clúster
- Validació de la correcta migració de nodes antics
- Revisió de la documentació entregada
- Formació final i transferència del coneixement

5. CRONOGRAMA DE LES FASES

Cronograma amb les fases del projecte:

La reunió d'inici del projecte (kick-off) es realitzarà una vegada s'hagi fet efectiu el lliurament de tot el material al CPD de l'SMC.



6. SERVEI DE SUPORT POST-INSTAL·LACIÓ

L'empresa adjudicatària haurà de prestar un servei de suport tècnic post-instal·lació durant un període d'1 mes natural amb 10 hores de suport a comptar des de la posada en marxa, per atendre incidències, consultes o dubtes relacionats amb els equips, sistemes i programari instal·lats.

Aquest servei haurà de garantir:

- 10 hores de suport en 1 mes natural.
- El servei inclourà assistència remota (i presencial si cal).
- Compromís de resposta màxim de 4 hores laborables des de la notificació de la incidència/petició.
- La resolució o proposta de solució en un termini màxim de 24 hores laborables, excepte en casos degudament justificats.
- Atenció per telèfon, correu electrònic i/o sistema de gestió de tiquets en horari laboral de dilluns a divendres, de 09:00h a 18:00h.

7. LLOC, TERMINI DE LLIURAMENT I CONDICIONS

7.2 LLOC

Lliurament i instal·lació del subministrament al CPD de l'SMC gestionat per T-Systems i ubicat a Cerdanyola.

7.3 TERMINI D'EXECUCIÓ

Tot el material s'haurà de subministrar i lliurar al CPD de l'SMC ubicat a Cerdanyola del Vallès en el termini màxim **de 6 setmanes naturals** a comptar des de la signatura del contracte.

La instal·lació de tot el sistema s'haurà de finalitzar en el termini màxim de **8 setmanes naturals** a comptar des de la recepció del material i d'acord amb la planificació de l'Equip de Sistemes de l'SMC.

Un cop finalitzada la instal·lació i posada en marxa del sistema, l'empresa adjudicatària haurà de prestar un servei de suport tècnic post instal·lació durant un període de **4 setmanes naturals**, amb un màxim de 10 hores de suport, a comptar des de la posada en marxa, per atendre incidències, consultes o dubtes relacionats amb els equips, sistemes i programari instal·lats.

8. PRESENTACIÓ DE LES OFERTES

Les empreses licitadores han de presentar les seves ofertes incloent-hi el cost del cablejat i enrackat als armaris del CPD de l'SMC.

9. DOCUMENTACIÓ

Les empreses licitadores hauran d'incloure en les ofertes la descripció completa del material a subministrar (element, model, fabricant i garantia).

Pel que fa a l'entrega, els equips o components d'aquest contracte han d'estar degudament documentats, incloent-hi les especificacions tècniques i/o els requeriments del fabricant, la data de fabricació, la informació de garantia, així com amb el test d'acceptació de fàbrica si aquest fos necessari. L'empresa adjudicatària inclourà en l'entrega a l'SMC la documentació original en format digital.

Lliurament digital (preferiblement en PDF, Visio/VSDX) en català abans de 10 dies naturals post proves.

Documentació	Descripció i criteris
Inventari i dades bàsiques del maquinari	Model, nº sèrie, firmware, data fabricació, garantia.
Llistat de Llicències	SupportEdge, SNTC i dates de venciment.
Dossier As-Built	Esquemes rack-layout i cablejat detallat.
Informe d'Instal·lació	Configuracions ONTAP/NX-OS, captures show-tech.
Manual d'Operació i Manteniment	SOPs: creació de volums, snapshots, hardening.
Informe de Proves HA	Mètriques IOPS/latència, logs,..
Sessió Formativa	Presentació PPT, manuals i recursos addicionals (4 h).
Certificats i Garanties	Registre cabina a NetApp ASUP, switches a Cisco SNTC.
Terminis d'entrega	Documents en PDF/Visio abans de 10 dies naturals post proves.

10. FORMA DE PAGAMENT

El pagament es realitzarà contra la presentació de la factura, prèvia certificació de la recepció del material en perfectes condicions per part de l'SMC, la instal·lació i la certificació de que estiguin en funcionament tots els serveis del nou sistema sense cap incidència i amb el rendiment esperat.

11. PENALITZACIONS

En cas d'incórrer en demora per part de l'adjudicatari respecte al compliment dels terminis de lliurament o d'instal·lació, l'SMC podrà imposar penalitats establertes en l'article 193.3 LCSP.

Cap d'Àrea de TIC del
Servei Meteorològic de Catalunya

Dr. Roux, 80 1ª Planta
08017 Barcelona
Telèfon (93) 567 60 90
e-mail: contractacio.meteocat@gencat.cat