

PLEC DE PRESCRIPCIONS TÈCNIQUES

A. OBJECTE

L'objecte d'aquest contracte és l'adquisició per a la Universitat Pompeu Fabra de material de computació d'alt rendiment a emprar al projecte de recerca ALiEN (Autonomous Linguistic Emergence in neural Networks) finançat per fons Consell Europeu d'Investigació en el programa Horizon 2020, núm. 101019291 — ALiEN

L'adquisició compren els equips següents:

- 12 servidors amb capacitat de supercomputació GPU.
- 2 switchos de comunicació Infiniband 200 Gb/s
- 3 targetes HDR 200 Gb/s.
- 2 switchos de comunicació ethernet 10GBase-T.
- 1 armari rack estàndard de 19" i 42U.

B. REQUERIMENTS TÈCNICS

Els licitadors han de presentar una proposta de maquinari que compleixi com a mínim amb els requisits que es resumeixen en l'annex i que a continuació es desenvolupen:

1. Sistema de càlcul

Tots els components presentats han d'estar certificats pel propi fabricant del model en qüestió.

La proposta ha de tenir el mateix model de CPU a tots els servidors, el mateix model de dimm de RAM i la mateixa quantitat de memòria RAM total del server, el mateix model de GPU i els mateixos discs NVMe.

Això si, es poden presentar servidors amb diferents tamanys (U) dependent del nombre de GPUs que hostatgi.

Cadascun dels servidors (node) de càlcul inclòs a l'oferta haurà de tenir, com a mínim, les següents característiques:

1.1 CPU:

Mínim de 2 micro-processadors amb un total mínim de 20 cores per processador a una velocitat mínima de 2,6 Ghz

Per compatibilitat amb l'actual clúster, els processadors han de tenir arquitectura de 64 bits (x86-64), pertanyent a la família Ice Lake SP d'Intel (3ª generació d'Intel).

No seran vàlids entorns que precisin de refrigeració líquida.

1.2. Emmagatzematge:

Cada node tindrà com a mínim 2 discs durs NVMe iguals, de mínim 2,5 Tb, de qualitat Enterprise/Server i configurats en RAID-1.

El RAID s'ha de construir per hardware, no sent vàlides solucions RAID per software.

En cas que la placa base no permeti la creació/gestió de RAIDs NVMe, s'haurà d'afegir una controladora de disc que ho permeti.

El menú de gestió dels RAIDs ha d'aparèixer a la BIOS o en un menú que es mostri en temps d'arrencada.

Solucions on l'eina de gestió estigui fora de l'arrencada quedaran descartades.

1.3. Memòria:

Tenir un mínim de 512GB on els xips de memòria siguin de mínim 32 GB i iguals entre sí.

La memòria, que ha de ser de tipus ECC, ha d'estar equilibrada, es a dir, ha de tenir els mateixos DIMMs de memòria a cada canal de comunicació entre core i memòria (accés simètric).

La memòria ha de ser DDR4 amb freqüència mínima de 3200 Mhz i la més alta que el micro-processador permeti.

1.4 GPU

El servidor han d'incloure un mínim de 4 targetes GPU.

S'han de connectar a un port PCIe 4x.

Cal que tingui tecnologia CUDA per compatibilitat amb el clúster actual, i per tant, es busca una arquitectura Nvidia Ampere.

El model de GPU ha de complir:

- mínim FP64:	5,2 teraFLOPS
- mínim FP64 Tensor Core:	10,3 teraFLOPS
- mínim FP32:	10,3 teraFLOPS
- mínim TF32 Tensor Core:	82 teraFLOPS
- mínim BFLOAT16 Tensor Core:	165 teraFLOPS
- mínim FP16 Tensor Core:	165 teraFLOPS
- mínim INT8 Tensor Core:	TOPS 330
- mínim INT4 Tensor Core:	TOPS 661
- mínim de memòria de GPU:	24 GB HBM2
- mínim ample de banda de memòria de la GPU:	933 GB/s
- interconnexió: PCIe Gen4:	64 GB/s
- Potència màxima de disseny tèrmic (TDP):	165 W

1.5 Servidor

El node ha de ser de màxim 4Us.

Dur dos targetes de xarxa 10GBase-T

Disposar de port IPMI per administració remota. En cas de tenir llicència, haurà d'incorporar una amb validesa perpètua i sense costos addicionals futurs.

Ha de ser "enrackable" en l'armari descrit a l'apartat 3: Armari RACK.

El node comptarà amb almenys una targeta HDR 200Gb/s (connector QSFP56) compatible amb els equips Infiniband descrit en l'apartat 2.1 Infiniband.

Les fonts redundants amb certificació Platinum hauran de suportar la càrrega de totes les GPUs funcionant al 100%. En aquest cas d'avaria d'una font, el servidor haurà de seguir funcionant en el supòsit de tenir totes les GPUs treballant al 100%.

2. Comunicacions

2.1 Infiniband

La UPF actualment disposa de:

- Switch FDR, Lenovo 0724-HCG (basat en el model Mellanox MSX6036F-1BRS).
- 31 nodes connectats a aquest switch amb les següents tipologies de targetes infiniband:

- Mellanox Technologies MT27500 Family [ConnectX-3]
- Mellanox Technologies MT27520 Family [ConnectX-3 Pro]
- Mellanox Technologies MT27640 Family

Per tal de connectar els nous servidors a la xarxa de baixa latència i per compatibilitat amb el switch anteriorment comentat, es requereix l'adquisició de:

2 Switchos Infiniband Mellanox HDR 200 GB/s amb mínim de 40 ports.

Els switchos han de ser compatibles amb el switch FDR descrit anteriorment. S'han de poder estacar (FDR i HDR).

Les targetes infiniband (QSFP+) detallades en l'apartat anterior han de ser compatibles amb el nou model de switch.

En una segona fase, s'estacaran els 2 switchos nous entre si, traient el FDR.

Ha d'incloure tots els cables Mellanox Passive Copper (amb connectors QSFP56) per a la connexió de totes els equips oferts, més tres cables de llargada superior (al voltant de 4 metres).

2.2 Targetes Infiniband

Per tal de millorar les comunicacions del sistema de fitxers Beegfs (nodes de metadades i node de login) i evitar colls d'ampolla en les comunicacions amb aquests nous nodes, es requereix que es subministrin:

3 Targetes IB HDR 200Gb/s (mateix model dels ofertats en l'apartat 1.5 Servidor)

2.3 Ethernet

Per tal de connectar els nous servidors a la xarxa ethernet, es requereix que es lliuri:

2 Switchos 10GBase-T amb les següents característiques:

- 4 ports 40/100GbE (amb connectors QSFP28), els connectors necessaris per fibres multi-mode OM3 (150 metres aproximadament) han d'estar inclosos, aquests connectors han de ser de tipus LC.

No s'acceptaran connectors MTP/MPO amb break-out cables per aquest propòsit.

Aquest equips han d'anar connectats a equips DELL Z9100-ON, els connectors necessaris per connectar els equips d'aquesta licitació amb els equips actuals de la UPF va a càrrec del l'adjudicatari.

- Tot el material ha de ser original del fabricant per garantir la compatibilitat.
- 48 ports 10Gb BASE-T
- Funcionament Actiu/Actiu
- Redundància creuada entre equips (Cross-Chassis Redundancy). Tot node connectat de manera redundada (una connexió a cada commutador) ha de seguir treballant sense cap tall ni delay en el cas de que un dels dos commutadors deixi de funcionar.
- Agregació entre equips a nivell 2 i 3
- Fonts d'alimentació redundades (hot-swap)
- Ventiladors hot-swap.
- Enhanced Object Tracking (Amb HSRP, es pot trackejar un interfície a diferents nivells per decidir si el commutador ha de continuar sent màster)
- Reflexiv Access List
- AutoQoS
- QoS:
 - Hierarchical QoS
 - Aggregate Policing (Ingress-Egress)

- Microflow Policing (Ingress-Egress)
- Remote SPAN (Nivell 3), Encapsulated Remote Switched Port Analyzer
- Capacitat de suportar connexions amb ports a 100G.
- I en el referent als connectors:
 - cada equip ha de tenir 1 connector QSFP28 per fer l'uplink cap als Z9100
 - cada equip ha de tenir 1 connector QSFP28 per fer l'stacking virtual entre els nodes

En resum, caldrà 6 connectors QSFP a 100 Gbps per fer les connexions. 4 per als equips motiu d'aquest plec i 2 per als actuals equips Z9100 (1 per a cada node).

3. Armari RACK

Armari de 19" standard de 42U, per a hostatjar el material ofertat.

S'ha de lliurar muntat i llest per a col·locar al CPD.

S'han d'incloure tantes PDUs com siguin necessàries per a connectar tots els equips de la proposta.

Si el conjunt de servidors oferts supera les 34U, no cal incloure un segon armari extra. La Universitat reubicarà les màquines en altres armaris de la seva propietat.

Les PDUs aniran connectades al sistema de distribució elèctric amb entrada a l'armari de corrent amb mínim 2 connectors de 16A tipus CETAC 2p+T

Portes davanteres i posteriors perforades

Rodes d'alta càrrega amb fre.

Grups de fixació per a instal·lació d'equips.

Els cables de corrent de servidors i switchos han d'estar inclosos en l'oferta i s'han de connectar a aquestes PDUs, fent servir longituds adients però mínimes.

C. ASPECTES RELACIONATS AMB L'EXECUCIÓ DEL CONTRACTE

1. Garantia:

El termini de garantia serà de 3 anys. En aquest termini l'adjudicatari serà responsable de les faltes de conformitat que existeixin en el moment del lliurament i es manifestin dins del termini.

La garantia començarà a comptar a partir de la data de posada en marxa de l'equipament, prèvia acceptació per escrit de la Universitat Pompeu Fabra.

L'empresa adjudicatària haurà de garantir la disponibilitat de peces de recanvi originals o equivalents (directament o a través d'altres agents autoritzats) per tot el cicle de vida previst de l'equip que, com a mínim serà de 10 anys a partir de els equips adquirits es deixin de fabricar.

L'adjudicatari estarà obligat així mateix a prestar en un termini mínim de 3 anys una garantia comercial addicional que compren els treballs de manteniment que garanteixin el correcte funcionament de l'equipament durant el termini de garantia comercial i considerats per mal funcionament que no sigui per un mal ús, incloent els recanvis, mà d'obra, desplaçaments, transport i dietes.

El recanvi de qualsevol element avariats es realitzarà sempre utilitzant material original o certificat pel mateix fabricant.

Finalitzat el període de garantia, el cost dels components es facturarà a preu de mercat.

L'adjudicatari proporcionarà un sistema per a la comunicació i la gestió del seguiment de les incidències, que ha d'explicar de manera explícita en la documentació a presentar amb l'oferta.

2. Interlocució

L'empresa adjudicatària haurà de disposar, un cop formalitzat el contracte, de com a mínim dos perfils d'interlocució, per mantenir contacte durant la fase de preparació de material, lliurament i posada en marxa:

- Un interlocutor comercial
- Un interlocutor tècnic.

Ivan Jiménez Rodà
Cap de la Unitat d'Informàtica del Campus de la Comunicació

Barcelona, 20/9/2021

ANNEX: TAULA D'ESPECIFICACIONS MÍNIMES

Especificacions mínimes servidor	# Equips	#sockets / cpu	velocitat (Ghz)	RAM (Gb)	Dimm RAM (Gb)	Freq (MHz)	Discs (Tb)	# GPU
Servidor	12	2	2,6	512	32	3200	2 x 2,5	4

Tamany màxim de servidor: 4U

Especificacions mínimes per GPU	#GPU	Memòria (GB)	FP64 (teraFLOPS)	FP64 Tensor Core (teraFLOPS)	FP32 (teraFLOPS)	TF32 Tensor Core (teraFLOPS)
gpu (I)	4	24	5,2	10,3	10,3	82

Especificacions mínimes per GPU	BFLOAT16 Tensor Core (teraFLOPS)	FP16 Tensor Core (teraFLOPS)	INT8 Tensor Core (TOPS)	INT4 Tensor Core (TOPS)	Ample Banda (GB/s)	Interconnexió PCIe Gen4 (GB/s)
gpu (II)	165	165	330	661	933	64

Potència màxima de disseny tèrmic (TDP) de GPU: 165 W