Universidad Nacional Autónoma de México
Secretaría de Desarrollo Institucional
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Lineamientos para el acceso y uso de equipo central de cómputo del Laboratorio de Inteligencia Artificial UNAM - HUAWEI

Elaborado por:
Dirección de Sistemas y Servicios Institucionales
24 de mayo de 2021

I. Descripción y usos de equipos centrales de cómputo

A continuación, se describe el equipo central de cómputo instalado en el Laboratorios de Inteligencia Artificial como parte del Espacio de Innovación UNAM – Huawei consiste en los siguientes componentes:

No.	Equipo central	Cantidad
1	Huawei Atlas 800 (Modelo 9000) Servidores de entrenamiento	2
2	Huawei Atlas 800 (Modelo 3010). Servidores de inferencia	2
3	Huawei Thaishan 2280- Servidores ARM	2

1. Dos servidores Huawei Atlas 800 9010- Servidor de entrenamiento

1.1. Factor de forma: Equipo de rack con 2U
1.2. Procesador: Dos Intel Cascade Lake Xeon 6230
1.3. Núcleos totales: 80
1.4. Frecuencia de operación: 3.0 GHz.
1.5. RAM total: 1 TB
1.6. Interfaces de comunicación: PCIe 3.0, 10GE y SAS/SATA,
1.7. Almacenamiento: 12 TB
1.8. Principales usos: Equipo para procesamiento de datos masivos para inteligencia artificial. Ejecución por contenedores en entorno Ubuntu
1.9. Forma de acceso: Remota vía SSH
1.10. Asignación de recursos por contenedor: hasta el 50% de la capacidad total de procesamiento por servidor mínima. Función principal para entrenamiento de algoritmos de inteligencia artificial
1.11. Tiempo total de procesamiento x86 disponible (horas anuales núcleo): 560,640
1.12. Tiempo total de procesamiento GPU disponible (horas anuales núcleo): 140,160
1.13. Tiempo total de procesamiento ARM disponible (horas anuales núcleo): 560,640

2. Dos servidores Huawei Atlas 800 3010- Servidor de inferencia

2.1. Factor de forma: Equipo de rack con 2U
2.2. Procesador: Dos Intel Cascade Lake Xeon 6240
2.3. Núcleos totales: 72
2.4. Frecuencia de operación: 3.0 GHz.
2.5. RAM total: 768 GB
2.6. Interfaces de comunicación: PCIe 3.0, 10GE y SAS/SATA,
2.7. Almacenamiento: 2 TB
2.8. Principales usos: Equipo para procesamiento de datos masivos para inteligencia artificial. Ejecución por contenedores en entorno Ubuntu. Función principal para inferencia de datos que requieran algoritmos para procesamiento de datos masivos en inteligencia artificial
2.9. Forma de acceso: Remota vía SSH
2.10. Asignación de recursos por contenedor: hasta el 50% de la capacidad total de procesamiento por servidor mínima.
2.11. Tiempo total de procesamiento x86 disponible (horas anuales núcleo): 504,576
2.12. Tiempo total de procesamiento GPU disponible (horas anuales núcleo): 140,160
2.13. Tiempo total de procesamiento ARM disponible (horas anuales núcleo): 560,640

3. Dos servidores Huawei Taishan 2280.

3.1. Factor de forma: Equipo de rack con 2U

3.2. Procesador: Huawei Kunpeng 916 ARM v8 de 64 bits
3.3. Núcleos totales: 96
3.4. Frecuencia de operación: 2.4 GHz.
3.5. RAM total: 1 TB.
3.6. Interfaces de comunicación: PCIe 3.0, 10GE y SAS/SATA.
3.7. Almacenamiento: 72 TB.
3.8. Principales usos: Aplicaciones genéricas de informática, almacenamiento o necesidades equilibradas. Aplicaciones de alta carga como análisis de datos masivos, almacenamiento definido por software, nube y aplicaciones nativas ARM.
3.9. Forma de acceso: Remota vía SSH.
3.10. Asignación de recursos por contenedor hasta el 100% de la capacidad total de procesamiento por servidor mínima.
3.11. Tiempo total de procesamiento disponible (horas anuales núcleo): 672,278.

II. Disponibilidad de almacenamiento y procesamiento por equipo de cómputo.

A continuación, se describen las capacidades mínimas y máximas de procesamiento y almacenamiento en cada tipo de equipo central de cómputo dentro el laboratorio:

Servidor modelo	Total equipos	Total CPU x equipo	Total Cores x CPU	Cores totales CPU	CPU horas / Año	Almacenamiento TB
1. Atlas 800 - 9010	2	2	20	80	560640	12
	Total equipos	**Total GPU x equipo**	**Total Cores AI x GPU**	**Cores totales GPU AI**	**GPU AI / Año**	**RAM TB**
	2	4	2	16	**140160**	1
	Total equipos	**Total GPU x equipo**	**Total cores ARM x GPU**	**Cores totales GPU ARM**	**GPU ARM / año**	
	2	4	8	64	560640	

Servidor modelo	Total equipos	Total CPU x equipo	Total Cores x CPU	Cores totales CPU	CPU horas / Año	Almacenamiento TB
2. Atlas 800 - 3010	2	2	18	72	504576	2
	Total equipos	**Total GPU x equipo**	**Total Cores AI x GPU**	**Cores totales GPU AI**	**GPU AI / Año**	**RAM TB**
	2	4	2	16	**140160**	0.768
	Total equipos	**Total GPU x equipo**	**Total cores ARM x GPU**	**Cores totales GPU ARM**	**GPU ARM / año**	
	2	4	8	64	560640	

Servidor modelo	Total equipos	Total CPU x equipo	Total Cores x CPU	Cores totales CPU	CPU horas / Año	Almacenamiento TB
3. Taishan 2280	2	1	48	96	**672768**	72
	Total equipos	Total GPU x equipo	Total Cores AI x GPU	Cores totales GPU AI	GPU AI / Año	**RAM TB**
	0	0	0	0	**0**	1
	Total equipos	Total GPU x equipo	Total cores ARM x GPU	Cores totales GPU ARM	GPU ARM / año	
	0	0	0	0	0	

III. Usuarios, aplicaciones y datos

Se consideran los siguientes tipos de usuarios, aplicaciones y datos para el uso del equipo central de cómputo en el Laboratorio de Inteligencia Artificial:

1. **Usuarios por convocatoria**: Son quienes tienen asignados recursos en horas de procesamiento y espacio de almacenamiento por el Grupo Especial de Innovación y a raíz de la propuesta de proyecto de investigación que hayan presentado en la respectiva convocatoria publicada. La disponibilidad de los recursos asignados será hasta el agotamiento de estos o el vencimiento del período de ejecución permitido.

2. **Usuarios por colaboración**: Son quienes tienen acceso a horas de procesamiento y almacenamiento por pertenecer al Grupo Especial de Innovación o realizar una solicitud expresa. La disponibilidad de los recursos será en función del máximo permitido por el remanente de los recursos empleados por los usuarios por convocatoria. El uso de los recursos asignados será sin fines de lucro.

3. **Usuarios académicos**. Son quienes están adscritos a la Licenciatura en Ciencia de Datos (estudiantes, profesores e investigadores) de la Universidad Nacional Autónoma de México y que utilizarán los recursos del laboratorio exclusivamente para los objetivos docentes de la Licenciatura.

4. **Administradores**: Son las instancias que gestionan la infraestructura central del Laboratorio, de acuerdo con lo estipulado por el Grupo Especial de Innovación y sus resoluciones para la asignación de recursos a los diversos tipos de usuarios. La administración de la infraestructura central es responsabilidad tanto de la DGTIC como del Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS).

5. **Aplicaciones**: Se distinguen los siguientes tipos de aplicaciones dentro del equipo central de cómputo del Laboratorio de Inteligencia Artificial.

    5.1. **Propietarias**: Desarrolladas por un usuario del laboratorio, sus equipos de trabajo y/o colaboradores. Pueden ser compartidas con otros usuarios del laboratorio de mutuo acuerdo.
    5.2. **Comunales**: Desarrolladas en conjunto entre dos o más usuarios del laboratorio y/o sus equipos de trabajo. También se consideran aplicaciones comunales a los códigos y aplicaciones que se obtengan de la red de desarrolladores Huawei.
    5.3. **Comerciales**: Adquiridas por el usuario para el desarrollo de su proyecto dentro del laboratorio. Pueden ser compartidas con otros usuarios en función del tipo de licenciamiento contratado.

IV. Límites de asignación y ejecución

1. La utilización de la infraestructura central del laboratorio considera como valores máximos asignables por proyecto las horas de procesamiento y cuotas de almacenamiento en cada categoría de equipo, como se resumen en la siguiente tabla:

No.	Equipo central	Horas anuales máximas de procesamiento disponibles	Capacidad máxima de almacenamiento disponible
1	Huawei Atlas 800 (Modelo 9000) Servidores de entrenamiento	**140,160**	**8 TB**
2	Huawei Atlas 800 (Modelo 3010). Servidores de inferencia	**140,160**	**1 TB**
3	Huawei Thaishan 2280- Servidores	**672,768**	**30TB**

No.	Equipo central	Horas anuales máximas de procesamiento disponibles	Capacidad máxima de almacenamiento disponible
	ARM		

2. Para los Usuarios por Convocatoria, el máximo de recursos disponibles en conjunto para todos los proyectos será del 80% de la infraestructura de cómputo central instalada en el Laboratorio de inteligencia Artificial UNAM – Huawei, lo que incluye horas de procesamiento y capacidad de almacenamiento.
3. Para los usuarios por colaboración, el mínimo de acceso a recursos en conjunto será del 10% de la infraestructura de cómputo central instalada en el Laboratorio de inteligencia Artificial UNAM – Huawei y el máximo será del 20%, si los usuarios por convocatoria no estuvieran usando hasta un 30% de la capacidad total.
4. Para los usuarios académicos, el mínimo de acceso a recursos en conjunto será del 10% de la infraestructura de cómputo central instalada en el Laboratorio de inteligencia Artificial UNAM – Huawei y el máximo será del 30%, si los usuarios por convocatoria no estuvieran usando hasta un 30% de la capacidad total
5. La asignación de recursos a los usuarios por convocatoria dependerá de la aprobación al proyecto presentado ante el Grupo Especial de Innovación y en ningún caso excederá el máximo permitido por la convocatoria dentro de la categoría asignada al proyecto.
6. La asignación de recursos a los usuarios por colaboración dependerá de la autorización por el Grupo Especial de innovación y en función de los recursos disponibles. Esta asignación podrá autorizarla el Grupo Especial de Innovación al usuario en colaboración que así lo solicite, por hasta el máximo de capacidad disponible en función del criterio IV.3
7. La asignación de recursos a los usuarios académicos estará autorizada por el Grupo Especial de Innovación siempre que el uso se apegue a los objetivos académicos de la Licenciatura.
8. La utilización de los recursos se apegará a los criterios de colas de trabajos descritas en la Sección V

V. Colas de trabajos

1. Las colas de trabajos en la infraestructura central de cómputo del Laboratorio de Inteligencia Artificial UNAM - Huawei tienen dos objetivos principales
   1.1. Asegurar que cada usuario por convocatoria tiene acceso al menos a una parte proporcional de los recursos que tiene asignados sin demasiada espera en la cola
   1.2. Asegurar que los usuarios por colaboración tienen acceso al máximo de reserva para este perfil de usuario.

2. Las colas de trabajos se establecerán bajo el principio de reducir los tiempos de espera y optimizar al máximo la utilización de la infraestructura

3. Para el cálculo de la asignación de trabajos en las colas se considerará lo siguiente:
   3.1. CP: Promedio de núcleos de procesamiento de uso constante, resultado de dividir el número de horas asignado al proyecto entre el total de horas de vigencia del proyecto
   3.2. CN: Promedio de núcleos de procesamiento requeridos para el tipo de trabajo en ejecución
   3.3. Si CP => CN, el proyecto requerirá de recursos para ejecutar al menos 1 trabajo de forma constante-
   3.4. Si CN > CP, el proyecto requerirá recursos para ejecutar al menos 1 trabajo ocasionalmente.

4. Tipos de colas de trabajos
   4.1. Cola de uso intensivo: Para los usuarios por convocatoria que emplean intensamente el equipo por la naturaleza de su investigación y el total de recursos que les fueron asignados.

4.1.1. Es la cola de más alta prioridad, donde el sistema de gestión procurará concluir los trabajos a la brevedad y corresponde a la situación donde CP => CN
4.1.2. Los trabajos o contenedores que requieran de más procesamiento estarán accediendo de manera concurrente con trabajos de demandas similares de recursos.
4.1.3. Los trabajos tendrán menor tiempo de espera
4.1.4. Los trabajos tendrán un máximo de ejecución de 220 horas para evitar largos períodos de procesamiento sin revisión de resultados parciales
4.1.5. La suma de todos los CP no deberá exceder el 80% de la capacidad de los equipos en el laboratorio

4.2. Cola de alto rendimiento y baja prioridad: Para los usuarios por convocatoria con baja demanda, donde se tendrá acceso a los recursos disponibles hasta el máximo establecido por los límites de asignación y ejecución
4.2.1. Es la cola en nivel siguiente inferior de prioridad, y corresponde a la situación CN > CP
4.2.2. Los trabajos en esta cola podrán extenderse más en el tiempo, ya que no requieren un uso intenso de procesamiento
4.2.3. Se evitará que estos trabajos bloqueen el desarrollo de los aquellos que requieren un uso más intenso.
4.2.4. Los trabajos podrán utilizar hasta el máximo de recursos disponibles dentro de los límites que permita la ejecución de la cola de uso intensivo.
4.2.5. Se establecerá un límite de utilización por trabajo de 72 horas, así como recurso para la priorización dinámica, bajo el principio de evitar el bloqueo de recursos por un solo proyecto

4.3. Cola residual: Para los usuarios por colaboración y usuarios académicos, en donde accederán hasta el máximo de recursos de procesamiento que los límites de la sección IV permiten
4.3.1. Son trabajos de baja prioridad para uso de recursos superiores al 10% del total en caso de estar disponible y hasta un 50% del total.
4.3.2. En caso de que un trabajo de las colas de uso intensivo o de alto rendimiento requieran liberar trabajos que hayan excedido el tiempo de ejecución, podrá utilizar espacio de la cola residual
4.3.3. Se consideran trabajos que pueden tener muchos recursos disponibles, pero su ejecución dependerá de la realización de los trabajos en las otras colas asociadas a los usuarios por convocatoria.