UNIVERSIDAD
Panamericana

FACULTAD DE INGENIERÍA

"Evaluación de plataformas de supercómputo en la nube para la investigación científica: caso de estudio en la Facultad de Ingeniería de la Universidad Panamericana campus Aguascalientes"

Tesis que presenta

Ing. Alfredo Márquez Martínez

Para obtener el grado de

Maestro en Ciencias

CON RECONOCIMIENTO DE VALIDEZ OFICIAL DE ESTUDIOS DE LA SECRETARÍA DE EDUCACIÓN PÚBLICA, SEGÚN ACUERDO CON EL No. 20171659 DE FECHA 12 DE MAYO 2017

Director de tesis

Juan Carlos García Sánchez

Codirector de tesis

Pedro Manuel Rodrigo Cruz

Universidad Panamericana Campus Aguascalientes
Facultad de Ingeniería
Aguascalientes, Ags., septiembre 2023

UNIVERSIDAD
Panamericana

Aguascalientes, Ags., 23 de agosto del 2023

LIBERACIÓN DE DOCUMENTO RECEPCIONAL

Por medio de la presente, certificamos en nuestra calidad de asesores de tesina, que el trabajo de **Alfredo Márquez Martínez** que lleva como título**: EVALUACIÓN DE PLATAFORMAS DE SUPERCÓMPUTO EN LA NUBE PARA LA INVESTIGACIÓN CIENTÍFICA: CASO DE ESTUDIO EN LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD PANAMERICANA CAMPUS AGUASCALIENTES** cumple con los requisitos establecidos por el reglamento vigente de la Facultad de Ingeniería para presentarse como documento recepcional de titulación del programa de Maestría en Ciencias.

Atentamente,

M.C. Miguel Angel Díaz Melchor

Revisor

Josemaría Escrivá de Balaguer 101, Fracc. Villas Bonaterra, Aguascalientes 20296, México. Tel. +52 (449)9106200
www.up.edu.mx

Dedicatoria

Dedico esta tesis a Dios que en su infinita misericordia y providencia divina nos acompaña siempre y ha bendecido este camino de esfuerzo y dedicación para entregar un pequeño fruto a las gracias que nos da siempre por medio del Espíritu Santo, junto con nuestra madre la Virgen de Guadalupe que siempre intercede por nosotros.

A mi esposa Mónica y mis hijos Héctor Alfredo, Sofía del Rocío y Fátima les quiero dedicar esta tesis como una muestra de mi agradecimiento por todo su apoyo, paciencia y amor incondicional a lo largo de este camino. Gracias por estar siempre a mi lado, animándome y motivándome para seguir adelante incluso cuando las cosas se pusieron difíciles y que siempre han creído en mí.

Esta tesis es un logro compartido y quiero que sepan que, sin ustedes, esto no habría sido posible. Espero que esta tesis sea una muestra de mi gratitud y amor por ustedes.

Con todo mi amor y gratitud,
Alfredo Márquez Martínez

Agradecimientos

- A nuestro Señor Jesucristo, a los dones del Espíritu Santo que se hacen presente a cada instante y a nuestra madre la Virgen de Guadalupe que siempre intercede.

- A mi esposa Mónica y nuestros hijos Héctor Alfredo, Sofía del Rocío y Fátima que me apoyaron y soportaron con paciencia el largo camino para la elaboración de esta tesis, así como el curso para elaborarla que fue durante pandemia.

- Al Maestro Juan Carlos García Sánchez quien fungió como asesor técnico de tesis y proyecto, así como a los Doctores Josué Ortiz Medina quien nos acompañó en el Seminario de Investigación y Tesis y Pedro Manuel Rodrigo Cruz quien fue mi asesor de tesis.

Contenido

Lista de Figuras

Lista de Tablas

Resumen

La presente tesis aborda la evaluación y las etapas para una futura implementación de supercómputo en la nube para la investigación científica para la Facultad de Ingeniería de la Universidad Panamericana Campus Aguascalientes.

La creciente cantidad de datos y la complejidad de los procesos científicos han llevado a los investigadores a buscar soluciones más eficientes y escalables y lo que se presenta es una opción viable para que lo obtengan con el cómputo en la nube.

En esta investigación se evaluaron diferentes plataformas de supercómputo en la nube, incluyendo Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform. Se propusieron soluciones para la implementación de supercómputo en la nube en diversas áreas de investigación.

Se analizó el rendimiento y escalabilidad de las soluciones de supercómputo en la nube, comparando los resultados con el cómputo local. Los resultados demostraron que las soluciones de supercómputo en la nube ofrecen un alto rendimiento y escalabilidad a un costo menor en comparación con las soluciones locales.

Se concluye que la implementación de supercómputo en la nube es una solución eficaz para la investigación científica, ya que ofrece una mayor eficiencia y escalabilidad a un menor costo. Se recomienda a los investigadores que consideren la implementación de supercómputo en la nube para optimizar sus procesos y mejorar la calidad de sus investigaciones.

Capítulo 1 – Introducción.

1.1 – Contexto.

La presente tesis busca estudiar si las tecnologías de Cómputo en la nube están lo suficientemente maduras para que puedan ser utilizadas en el ámbito de la investigación académica universitaria, recomendando, de ser el caso, una propuesta, planeación, estrategia de implantación y los primeros pasos o tareas para lograr el objetivo en el caso concreto de la Facultad de Ingeniería de la Universidad Panamericana campus Aguascalientes.

En ese contexto, se inició el estudio haciendo una revisión bibliográfica respecto del estado del arte del Cómputo en la nube, identificando los modelos de servicio que soporta, su arquitectura, estándares y otros componentes relevantes.

Además, se estudió la viabilidad de utilizar los servicios de cómputo en la nube identificados en el ámbito universitario de investigación para la Facultad de Ingeniería de la Universidad Panamericana de Aguascalientes.

En la actualidad existe un reto globalizado sobre el alto volumen de información estructurada y no estructurada que generan las organizaciones en todos los rubros por lo que se contempla la necesidad de establecer el primer laboratorio virtual de investigación sobre súper cómputo que permitan obtener avances importantes en esta materia que conduzcan a mejores decisiones y movimientos estratégicos así como cubrir una necesidad importante de una institución educativa concreta de preparar a sus alumnos e investigadores en las nuevas tendencias tanto del mercado laboral como son los temas de súper cómputo.

La computación en la nube ha transformado la forma en que muchas organizaciones gestionan sus actividades, lo que representa beneficios donde se incluyen: ahorro de costos, agilidad, eficiencia, consolidación de recursos y nuevas oportunidades de negocio; pero lo que se busca en esta tesis es que pueda servir como marco de referencia para un programa de implementación de cómputo en la nube para el ámbito de la investigación enfocado a la consecución de estrategias para el desarrollo de la misma.

El paradigma de la computación en la nube subyace en su propia funcionalidad, la cual busca optimizar la funcionalidad de los servicios existentes de tecnologías de la información TI y habilitar otras funciones (Luis Felipe Ortiz Clavijo, 2018).

En este caso lo que se busca es optimizar la investigación mediante herramientas tecnológicas del súper-cómputo en la nube.

Además, el cómputo en la nube; representa una convergencia de dos grandes tendencias en las tecnologías de la información:

- Poder TI: donde la potencia de los servidores es usada con mayor eficiencia a través de configuraciones de hardware y software altamente escalables.

- Inteligencia: Al poder ser usada como una herramienta de valor agregado, gracias al procesado paralelo, uso de analítica y que responden en tiempo real a los requerimientos del usuario.

Existen diversas definiciones de lo que se conoce como 'nube', entre las que se encuentra la siguiente: "Una 'nube' es un tipo de centro de datos distribuida, la cual proporciona infraestructura de TI como servicios. Dentro de las características esenciales del cómputo en la nube es que dispone de recursos en forma masiva, los cuales son proporcionados a los usuarios de manera sencilla, dinámica, flexible y elástica, permitiendo además un monitoreo en tiempo real de los servicios recibidos. (Manuel Yrigoyen Quintanilla, 2011)

Cómputo en la nube tiene en general cuatro características (Figura 1):

- Autoservicio bajo demanda.
- Acceso de red asegurado.
- Medición, monitoreo y monetización o cobro por uso.
- Elasticidad y pool de recursos.

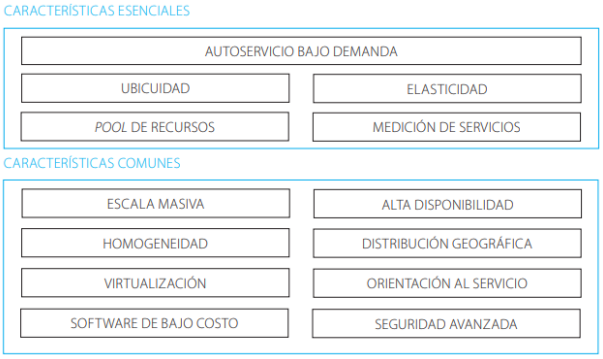

Figura 1 – Características del cómputo en la nube (Manuel Yrigoyen Quintanilla, 2011).

1.2 – Antecedentes.

El aprendizaje automático (o machine learning) puede ayudar a mejorar la eficiencia de la ciencia y la investigación. Esta técnica consiste en acumular grandes cantidades de datos e información para que un algoritmo extraiga patrones y vaya aprendiendo para ofrecer soluciones a problemas concretos. Es lo que ha hecho el Joint Research Centre de Ispra, en Italia, uno de los

grandes nodos de investigación de la Comisión Europea para el ámbito de la agricultura. (Comunicación, 2022).

Existen diversos centros de supercómputo tradicionales o con equipo físico de alto desempeño como el Laboratorio Nacional de Supercómputo del Sureste de México (LNS) o el Laboratorio Nacional de Cómputo de Alto Desempeño (LANCAD) de la UNAM, pero en lo que a la nube concierne, no existen laboratorios de supercómputo para investigadores, ya que lo que encontramos actualmente son varias soluciones existentes por proveedores de entornos de cómputo en la nube públicos como son Amazon AWS con HPC (High Power Compute), Google Cloud (Google Cloud Data Science), Microsoft Azure IA e IBM (IBM Cloud HPC).

De acuerdo con el documento científico: Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model (Patel & Kushwaha, 2020) hace mención de la existencia de dos modelos de cargas a clústeres de cómputo de nube propuestos que son K-Means y Gaussian Mixture que consiste en agrupación de datos en el clúster y el otro en métodos probabilísticos de cargas de trabajo para cómputo de nube que pueden funcionar para generar un entorno de pruebas adecuado para las cargas de trabajo de investigación de supercómputo en la nube. (Patel & Kushwaha, 2020)

Actualmente, basado en el documento científico ya citado, existen varios modelos de servicios facilitados por el cómputo en la nube dados por los proveedores antes mencionados que pueden otorgar una amplia gama de oportunidades para desarrollar un ambiente adecuado, seguro y confiable para los investigadores de la Facultad de Ingeniería de la Universidad Panamericana (Figura 2).

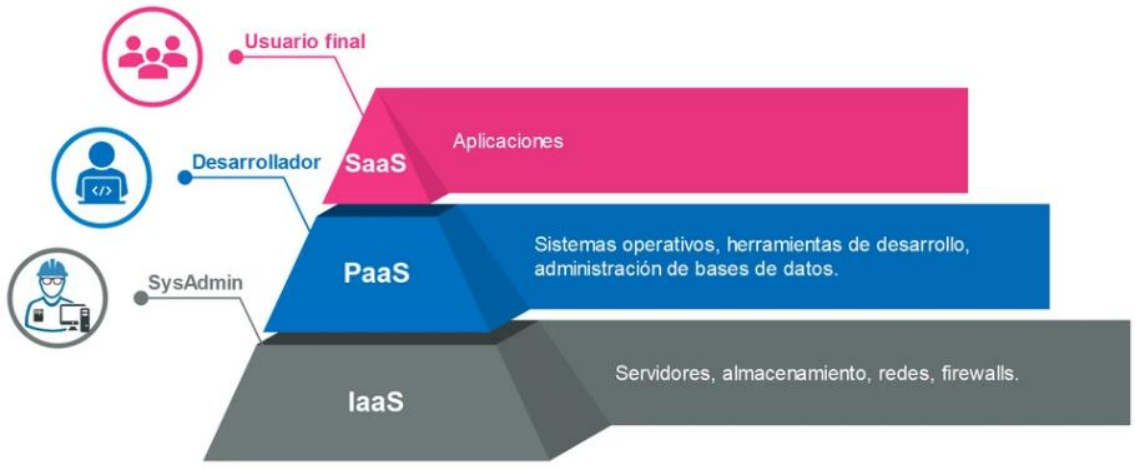

Figura 2 – Modelos de servicio de cómputo en la nube: SaaS (Software como servicio), PaaS (Plataformas como servicio) e IaaS (Infraestructura como servicio). (Lunazco Roger & Chavez Jaime Tomas, 2022)

1.3 – Justificación.

El proyecto se basa en la investigación de un clúster virtual basado en cómputo en la nube para generar los ambientes y entornos de procesamiento necesarios para la investigación sobre temas relacionados en súper cómputo, así como estudio y práctica de las nuevas tendencias en software y el código de calidad. El tiempo planeado para considerar la implementación de este laboratorio virtual es de un año, donde el desarrollo de la fase de implementación queda fuera de los objetivos de la presente investigación, pero estableciendo documentos necesarios para definir un periodo de sustentabilidad de este bajo un periodo pertinente de renovación y ampliación.

Uno de los problemas fundamentales a los que se enfrenta la investigación universidades es aumentar la productividad de su comunidad investigadora. El objetivo es utilizar el cómputo en la nube como una herramienta para ayudar a mejorar la productividad de actividades de investigación. Esto se logra identificando y abordando los principales factores que influyen en la mejora de productividad e identificando las necesidades de los investigadores desde la perspectiva del cómputo en la nube y aplicaciones informáticas.

Para obtener las necesidades y los factores que afectan la productividad de los investigadores universitarios, se realizaron entrevistas informales a investigadores de la Facultad de Ingeniería para determinarlos.

Dentro de estos resultados se encontraron diversos factores a considerar dentro del planteamiento de esta iniciativa: Independencia de perfiles, seguridad, confiabilidad, multiplataforma y escalabilidad.

1.4 – Objetivos e hipótesis.

El objetivo general de la investigación es la propuesta y planeación de un laboratorio o Clúster Virtual basado en cómputo de nube a bajo costo para la Facultad de Ingeniería de la Universidad Panamericana campus Aguascalientes para fomentar la investigación con las capacidades que este ambiente provee para temas relacionados con súper cómputo tanto para investigadores de la facultad de ingeniería como para los alumnos de la carrera de Ingeniería en Inteligencia Artificial. Dentro de este proyecto se estudiará involucrar a las tecnologías líderes de cómputo en la nube como son Google, Microsoft e IBM.

Para poder llevar a cabo el presente proyecto de generar un clúster de supercómputo en la nube se requiere de un análisis adecuado de los tres grandes competidores de nube con los que la Universidad Panamericana tiene relación como es Amazon, Google y Microsoft para revisar las diferentes capacidades de cómputo de alta densidad que generarían el ambiente de supercómputo, relacionados con métodos de desarrollo para ciencia de datos como ejemplo está

Google Cloud con Google Collaboratory (ambientes de Python), y un repositorio de proyectos de investigación en dónde se pueda registrar los avances y códigos generados para un futuro reúso para investigadores, de esta manera se puede resolver con cargas establecidas y probadas con los modelos citados con anterioridad proveer de un ambiente de nube con capacidad de cómputo suficiente para los investigadores y sus necesidades.

La hipótesis de investigación es que la planeación de un Clúster Virtual basado en cómputo de nube para la Facultad de Ingeniería de la Universidad Panamericana campus Aguascalientes arrojará resultados competitivos en cuanto al costo de la solución y cualitativamente representará ventajas respecto al modo de investigación tradicional que ayuden a potenciar el rendimiento de las tareas de investigación.

Capítulo 2. Estado del arte.

El campo del supercómputo ha experimentado un rápido avance en los últimos años que están transformando la investigación con supercomputadoras (Universitaria, 2013):

- **Computación cuántica:** La computación cuántica es una tecnología emergente que utiliza principios cuánticos para procesar información. Las computadoras cuánticas prometen resolver problemas complejos mucho más rápido que las computadoras clásicas. Aunque aún en etapa experimental, la computación cuántica tiene el potencial de transformar la investigación en áreas como la física, la química, la biología y la criptografía.

- **Computación heterogénea:** La computación heterogénea utiliza múltiples tipos de procesadores en una sola máquina, lo que permite la optimización de la carga de trabajo para tareas específicas. Por ejemplo, una computadora heterogénea podría utilizar procesadores de CPU, GPU y FPGA para diferentes tareas, lo que resultaría en una mejor eficiencia energética y un mayor rendimiento.

- **Cómputo en la nube:** La computación en la nube ha revolucionado el campo de la supercomputación al permitir el acceso remoto a recursos de supercomputación a través de Internet. Esto significa que los investigadores pueden ejecutar simulaciones y análisis de datos en supercomputadoras de clase mundial sin tener que invertir en hardware costoso o tener una infraestructura local de supercomputadoras.

- **Inteligencia Artificial:** La inteligencia artificial y el aprendizaje automático están impulsando el desarrollo de nuevas herramientas de simulación y análisis de datos en la supercomputación. Por ejemplo, la inteligencia artificial se puede utilizar para la clasificación de imágenes, el reconocimiento de voz y la optimización de algoritmos de simulación.

- **Interconexión de alta velocidad:** La interconexión de alta velocidad es esencial para la comunicación entre los componentes de la supercomputadora y para la transferencia de grandes cantidades de datos entre la memoria y el procesador. La interconexión de alta velocidad se ha vuelto cada vez más importante a medida que las aplicaciones de supercomputación se vuelven más exigentes en términos de ancho de banda y latencia.

2.1 – Supercómputo como herramienta en la investigación científica.

El supercómputo es la tecnología informática más avanzada de cálculo numérico que existe actualmente para desarrollar investigaciones complejas de alto nivel de especialización; es la única herramienta que le permite al investigador llevar a cabo, con certeza y velocidad, billones de cálculos matemáticos para estudiar problemas de gran magnitud.

Supercómputo se refiere a las computadoras más poderosas que existen, aquellas que actualmente se usan para hacer predicciones de clima o para hacer cálculos que serían imposibles en las computadoras de escritorio, estos cálculos son normalmente para el campo de la ciencia.

El uso del supercómputo con fines científicos permite abordar una gran cantidad de problemas científicos que de otro modo serían difíciles o imposibles de resolver. Por ejemplo, la predicción confiable del clima a escalas de tiempo cada vez mayores sería imposible sin el uso de las supercomputadoras. Otros ejemplos son el estudio de nuevos fármacos en el tratamiento de enfermedades; el estudio de las corrientes de aire en un nuevo diseño aerodinámico de un avión o de un auto; el análisis de los datos de los mayores aceleradores de partículas para buscar nuevas partículas; el estudio de nuevas moléculas, por mencionar solo algunos. (Villaseñor Cendejas, n.d.)

Podemos afirmar entonces que prácticamente en todas las áreas del conocimiento se requiere o se requerirá en el futuro próximo, del uso del supercómputo, lo cual nos permitirá aumentar la capacidad de análisis y estudios a una mayor profundidad.

2.2 – Supercómputo en la nube como servicio.

El servicio de supercómputo en la nube es para que investigadores no tengan que construir o administrar un sistema desde cero o en cuánto tenga disponibilidad de uso con que puedes reducir el costo de implementación al ya no necesitar una supercomputadora físicamente instalada en el centro de datos de la institución para una amplia variedad de cargas de trabajo.

La tecnología de supercómputo en la nube pasa rápidamente los datos a través de múltiples procesadores, un proceso clave de supercomputación en el que muchos chips actúan como una colmena para realizar una gran tarea, por ejemplo, cálculos complejos.

Contar con altos niveles de rendimiento computacional en la nube elimina la latencia que se produce al transmitir datos, al tiempo que proporciona las capacidades de cómputo necesarias para conducir análisis a fondo y capturar un alto volumen de datos, incluyendo imágenes y videos en alta resolución; sin embargo, en la mayoría de los casos, los investigadores sólo necesitan ver partes específicas de los datos. La disección y procesamiento de los datos evita la latencia y eleva la eficiencia y la velocidad.

Las supercomputadoras en nube son una parte esencial del avance científico, gracias a su facilidad para realizar cálculos numéricos para desarrollar investigación con alto nivel de especialización. La supercomputadora permite a los investigadores reducir los cálculos para sus trabajos científicos a semanas, lo que en una computadora normal llevarían años.

2.2.1– Ventajas del supercómputo en la nube.

El supercómputo en la nube, ofrece varias ventajas en comparación con la computación tradicional (Villaseñor Cendejas, n.d.). Algunas de las ventajas del supercómputo en la nube son:

1. **Infraestructura potente y flexible** para dar servicio a cargas de trabajo escalables.
2. **Cada equipo puede acceder a su propio clúster**, escalable y adecuado a sus necesidades, lo que permite reducir los tiempos de las colas de grandes cargas de trabajo y mitigar la limitación de recursos de computación.
3. **Pagar únicamente por lo que se utiliza** ya que se puede acceder a configuraciones de máquinas personalizadas para controlar los costos con descuentos por compromiso de uso y uso continuado. Además, puedes ahorrar hasta un 80 % con las máquinas virtuales interrumpibles.
4. **Innovación con recursos de alto rendimiento bajo demanda**: Se puede diseñar tu propio servidor de cómputo de alto rendimiento con los últimos procesadores de Intel y AMD, GPUs NVIDIA y funciones de almacenamiento de objetos y archivos de baja latencia y alto rendimiento.
5. **Virtualización**: La virtualización es una tecnología que permite crear servicios de TI útiles, con recursos que están tradicionalmente limitados al hardware. Gracias a que distribuye las funciones de una máquina física entre varios usuarios o entornos, posibilita el uso de toda la capacidad de la máquina. Los recursos se dividen según las necesidades, desde el entorno físico hasta los numerosos entornos virtuales. Los usuarios interactúan con la informática y la ponen en funcionamiento dentro del entorno virtual (generalmente denominado máquina Guest o máquina virtual). La máquina virtual funciona como un archivo de datos único; por eso, tal como ocurre con cualquier archivo digital, es posible trasladarla de una computadora a otra, abrirla en cualquiera de ellas, y tener la tranquilidad de que funcionará de la misma forma. (RedHat, n.d.)
6. **Disponibilidad**. En este sentido, la virtualización juega un papel fundamental, ya que el proveedor puede hacer uso de esta tecnología para diseñar una infraestructura

redundante que le permita ofrecer un servicio constante de acuerdo con las especificaciones del investigador.

7 **Acceso desde cualquier punto geográfico**. El uso de las aplicaciones diseñadas sobre el paradigma del cómputo en la nube puede ser accesible desde cualquier equipo de cómputo en el mundo que esté conectado a Internet.

8 **Escalabilidad y seguridad.** Las actualizaciones y nuevas funcionalidades son instaladas prácticamente de manera inmediata.

2.2.2– Desventajas del supercómputo en la nube.

1 **Percepción**. La percepción de inseguridad que genera una tecnología que pone la información (sensible en muchos casos), en servidores fuera de la organización, dejando como responsable de los datos al proveedor de servicio.

2 **Falta de control sobre recursos y los datos.** Al tener toda la infraestructura e incluso la aplicación corriendo sobre servidores que se encuentran en la nube, es decir, del lado del proveedor, el investigador carece por completo de control sobre los recursos e incluso sobre su información, una vez que ésta es subida a la nube.

3 **Dependencia**. En una solución basada en cómputo en la nube, el investigador se vuelve dependiente no sólo del proveedor del servicio, sino también de su conexión a Internet, debido a que el usuario debe estar permanentemente conectado para poder alcanzar al sistema que se encuentra en la nube.

4 **Propiedad intelectual.** La propiedad intelectual de los archivos que depositamos en la nube y en otros bancos virtuales varía según el programa que se utilice. La gran mayoría de las cláusulas están detalladas en las condiciones de servicio de cada programa, puntos para tener en cuenta si se va a utilizar alguno de estos servicios de almacenamiento. Microsoft en sus condiciones de servicio apuntan a que la propiedad pertenece al usuario y no se apropian de ella. Caso diferente es el de Google, quien se adueña del contenido a través de una licencia mundial para usar, alojar, almacenar, reproducir, modificar o crear obras derivadas, entre otras acciones y que seguirán siendo vigentes, aunque el usuario deje de utilizar esos servicios. Es decir, legalmente a quién le pertenecen los datos almacenados y los datos generados. Jurídicamente debe protegerse al investigador. (Francisco Naranjo, n.d.)

5 **Integración**. No en todos los entornos resulta fácil o práctica la integración de recursos disponibles a través de infraestructuras de cómputo en la nube con sistemas desarrollados de una manera tradicional, por lo que este aspecto debe ser tomado en cuenta por el investigador para ver qué tan viable resulta implementar una solución basada en la nube dentro de la universidad, por ejemplo (Tabla 1).

Ventajas	Desventajas
1. Mejorar accesibilidad al servicio 2. Respaldo y recuperación 3. Competitividad incrementada a través de acceso a recursos 4. Escalabilidad 5. Incrementar flexibilidad 6. Colaboración (recurso compartidos) 7. Inversión menor / costo por adelantado 8. Menores costos operacionales 9. Menores costos de personal de TI	1. No hay una adecuación a necesidades (confiabilidad, disponibilidad, accesibilidad, robustez, resiliencia, recuperabilidad) 2. Integridad (funcionalidad de software) 3.Mantenibilidad alcance limitado de solución 4. Riesgos contingentes (alto impacto en operación de servicios) 5. Interrupciones mayores de servicios 6. Necesidad de soporte a una operación de negocio flexible 7. Riesgos de seguridad (servicio, datos, autenticación y autorización, de-negación de ataques de servicio) 8. Estrategias de gestión de riesgo (cumplimiento y uso)

Tabla 1 – Ventajas y desventajas del cómputo en la nube. (Villaseñor Cendejas, n.d.)

2.3 – Soluciones existentes de supercómputo en la nube.

El National Institute of Standards and Technology, NIST (Instituto Nacional de Estándares y Tecnologías) proporciona la definición más amplia y adoptada acerca del cómputo en la nube.

Esta definición identifica cinco características esenciales, así como tres modelos de servicio y cuatro modelos de despliegue.

De acuerdo con la definición del NIST:

- El cómputo en la nube es un modelo para crear acceso conveniente, ubicuo y bajo demanda, vía internet, a un conjunto compartido de recursos de cómputo configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios), los cuales pueden ser rápidamente asignados y provistos con un mínimo de gestión administrativa e interacción con el proveedor. Este modelo promueve la disponibilidad; tiene cinco características esenciales, tres modelos de servicio y cuatro modelos de despliegue. (Barnard & Delgado, n.d.)

Cómputo en la nube es una metáfora para describir la web como un lugar donde la informática ha sido preinstalada y está disponible como un servicio, donde los datos, las aplicaciones, los sistemas operativos, el almacenamiento y la capacidad de procesamiento son todos disponible en la web y lista para ser compartida entre los usuarios. (Alemami et al., 2023)

También lo podemos considerar como una colección de centros de datos conectados a internet para ofrecer sus servicios, y estos centros de datos se basan en la virtualización de su infraestructura, para proveer servicios como plataforma de software, sistema operativo, desarrollo de aplicaciones, gestión de bases de datos, software de gestión de sistemas y aplicaciones, Internet y redes, investigación y big data, etc. (Alemami et al., 2023)

Hay varios líderes en el mercado de supercómputo en la nube, cada uno con sus propias fortalezas y enfoques. Aquí hay una lista de algunos de los principales proveedores de supercómputo en la nube (Figura 3):

- **Amazon Web Services (AWS):** AWS es uno de los principales proveedores de servicios en la nube y también ofrece servicios de supercómputo en la nube, como Amazon EC2, que permite a los usuarios lanzar y escalar instancias de supercómputo.

- **Microsoft Azure:** Azure es otra plataforma de computación en la nube popular que ofrece una amplia gama de servicios, incluyendo opciones de supercómputo en la nube. Azure Batch, por ejemplo, permite a los usuarios ejecutar aplicaciones de alta carga de trabajo en paralelo y escalarlas automáticamente.
- **Google Cloud Platform (GCP):** GCP es otra plataforma de computación en la nube que ofrece servicios de supercómputo en la nube, como Compute Engine, que proporciona acceso a máquinas virtuales de alta capacidad y escalabilidad.

Figura 3 – Cuadro Gartner sobre líderes de infraestructura en la nube (Gartner, n.d.)

La idea básica detrás de la nube es que todo lo que pueda hacerse en los sistemas informáticos en una organización, desde el almacenamiento y la colaboración hasta el procesamiento y la comunicación, se pueden desplazar hacia la nube.

Esencialmente, el cómputo en la nube es un servicio o conjunto de servicios prestados por medio de internet, bajo demanda del usuario y desde una ubicación remota, en lugar de residir en un equipo de escritorio, una laptop o los servidores de la organización. Así, las organizaciones contratan a un proveedor de servicios que ofrezca almacenamiento, procesamiento y aplicaciones a través de la web.

Los recursos del cómputo en la nube están disponibles bajo demanda para acceder a información, aplicaciones y procesamiento, independientemente de la ubicación y de los dispositivos de acceso.

El cómputo en la nube ofrece flexibilidad y comodidad porque los usuarios pueden trabajar cuando y donde quieran sin importar de dónde vienen los datos que ven en pantalla, siempre y cuando haya acceso a internet. Además, para ese propósito el cómputo en la nube permite a los proveedores utilizar centros de datos distantes. (Barnard & Delgado, n.d.)

2.3.1 - Soluciones de supercómputo en la nube actuales.

Existen diversos modelos de servicio y se dividen en tres categorías principales: Nube pública, la nube se hace disponible a través de un acuerdo medido para el público en general, Nube privada, Centros de datos interno a una organización que no se hace disponible al público en general y Nube hibrida, una combinación de desarrollos de nube privadas y públicas. Estas tres categorías también pueden describirse como micro características de modelos de nube (Barnard & Delgado, n.d.)(Tabla 2).

Modelo	Micro Características
Publica	• Flexible • Usuarios distribuidos • Elástica • Libertad para autoservicio • Pagas conforme se usa • Segura • Medida
Privada	• Internaliza procesos de negocio • Usuarios restringidos • Escalable • Accesible • Elástica • Compartida
Hibrida	• Elástica • Por demanda • Localidades y equipo virtual del usuario • Combinación de servicios restringidos y abiertos.

Tabla 2 - Características de Modelos de Nube (Barnard & Delgado, n.d.)

Microsoft (DatacenterDynamics, n.d.) ha informado sobre la creación de su nueva supercomputadora en la nube que está ejecutando para el grupo de investigación de IA OpenAI. Microsoft anunció que invirtió 10 mil millones de dólares en el grupo, lo que esencialmente eliminó su estado sin fines de lucro a medida que avanza hacia las operaciones comerciales. (NY Times, n.d.)

Construyó el sistema en colaboración con OpenAI, y que la supercomputadora tenía 285.000 núcleos de CPU, 10.000 GPU y 400 gigabits por segundo de conectividad de red para cada uno de los servidores de GPU.

Amazon, por su parte, definitivamente ha orientado su infraestructura hacia el proverbial "99 por ciento". Aun así, Amazon está tratando de acercarse cada vez más al 1 por ciento con sus instancias de cómputo en clúster, que pueden ejecutar chips Intel por sí mismos o en combinación con GPU NVIDIA (Amazon Web Services, n.d.).

Aquí hay un vistazo a las especificaciones de Amazon para una instancia llamada clúster Compute Eight Extra Large:

- 60,5 GB de memoria
- Dos procesadores Intel Xeon E5-2670 Sandy Bridge, ocho núcleos cada uno
- 3370GB de almacenamiento
- Interconexión de 10 Gigabit Ethernet

La instancia de Clúster Compute "Cuádruple extragrande" utiliza dos procesadores Nehalem de cuatro núcleos, y viene con aproximadamente la mitad de memoria y almacenamiento. Una instancia de Clúster GPU también usa dos procesadores Nehalem de cuatro núcleos, pero con el impulso adicional de dos GPU NVIDIA Tesla M2050. Las tres instancias de clúster de Amazon utilizan 10 Gigabit Ethernet, a diferencia de Gigabit Ethernet para las instancias EC2 estándar.

En la lista más reciente de las 500 supercomputadoras más rápidas del mundo, los clústeres de gama más alta utilizaron principalmente InfiniBand o una de las numerosas interconexiones personalizadas o patentadas diseñadas específicamente para la informática de alto rendimiento.

Ethernet en realidad representa 224 de los 500 sistemas principales, y 210 utilizan velocidades de un solo Gigabit por segundo en lugar de 10 Gigabit. Eso coloca a Ethernet justo por delante de InfiniBand en general, pero InfiniBand domina en la cima, con dos de los cinco sistemas más rápidos y cinco de los 10 más rápidos.

El clúster de mayor rango que utiliza una conexión Ethernet de 10 Gigabit es en realidad uno construido por Amazon en su propia nube con el propósito de demostrar su poder. El clúster de 17.000 núcleos de Amazon con procesadores Intel Xeon alcanzó velocidades de 240 teraflops para ocupar el puesto 42 en todo el mundo . Amazon ejecutó todo el clúster en una sola región de centro de datos, lo que sin duda aceleraría las cosas en comparación con los clústeres creados por el cliente que se extraen de los centros de datos de Amazon en varios continentes para garantizar que se satisfagan las necesidades de capacidad.

Google cloud maneja la solución de cómputo de alto rendimiento (HPC), con una infraestructura potente, ya que las máquinas virtuales de Compute Engine cuentan con las CPUs más recientes de las CPU, admiten la migración en tiempo real, pueden usar

objetos de alto rendimiento, bloqueos y almacenamiento de archivos, y están desarrolladas para tener un alto rendimiento y poca latencia de red entre máquinas virtuales.

Permite escalar fácilmente tus cargas de trabajo por lotes pasando por ejecutar tus cargas de trabajo de HPC en contenedores de forma flexible con GKE hasta desplegar entornos de clúster de HPC que se escalan automáticamente con el kit de herramientas de HPC de Google Cloud, las herramientas y los servicios de HPC facilitan la ejecución de las cargas de trabajo más difíciles. Además, Google colabora con una amplia variedad de desarrolladores de aplicaciones, gestores de cargas de trabajo, proveedores de almacenamiento e integradores de sistemas para asegurarse de que sus aplicaciones se ejecuten desde el primer momento en Google Cloud (Figura 4).

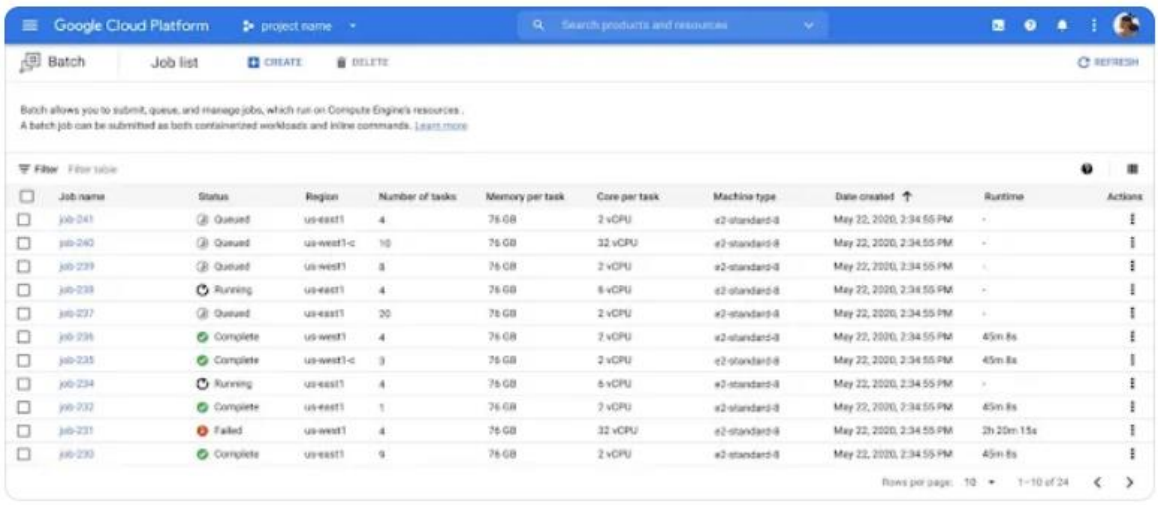

Figura 4 - Interfaz de la lista de tareas por lotes de la consola de Google Cloud

2.3.2 – Ejemplo de Soluciones de supercómputo actuales en la nube para la investigación.

Actualmente, The National Center for Biotechnology Information (NCBI) es un repositorio de datos de secuenciación de alto rendimiento con más de 36 petabytes a través de nubes públicas, como Google Cloud Plataforma (GCP) y AWS. (ncbi.nlm.nih.gov, n.d.). Una solución es configurar suficiente computación recursos locales (internos) dentro de un laboratorio u organización. Sin embargo, si los recursos computacionales aumentan, lograr tanto los costos de configuración como los de mantenimiento suelen ser difíciles.

Por el contrario, los bajos recursos informáticos dificultan el manejo de análisis de datos a gran escala dentro de un período de investigación. Además, varios pasos de análisis de datos requieren diferentes memorias y CPU; por lo tanto, el número total de recursos

informáticos y componentes cambiar diariamente. Recientemente, en los Estados Unidos, el AnVIL (el Instituto Nacional de Salud Investigación Nacional del Genoma Humano (NHGRI) ha estado promoviendo una plataforma en la nube
ejecutándose en el GCP diseñado para administrar y almacenar a gran escala genómica para permitir el análisis a escala de población (https://anvilproject.org/). Como un esfuerzo de colaboración internacional, investigadores del Consorcio Internacional del Genoma del Cáncer desarrolló una interfaz unificada para buscar y acceder a datos a usuarios autorizados de una nube comercial, AWS y una nube académica, Cancer Genome Collaboratory (dcc.icgc.org, n.d.). (El-Kassabi et al., 2023)

2.4 - Estructura del supercómputo.

La estructura de una supercomputadora se conforma por un clúster de computadoras, conformado por más de dos servidores en donde cada servidor es llamado nodo. Estos equipos son con características de alto rendimiento y a veces de alta disponibilidad de cómputo o accesibilidad. Se puede afirmar que la unión de cada nodo hace la fuerza de la supercomputadora, ya que por la arquitectura actual de los nodos de cómputo se define el poder de un equipo por el número de núcleos (Cores en inglés), la memoria interna disponible, su velocidad de interconexión, su capacidad de almacenamiento externo, su arquitectura, sistema operativo y software instalado (Figura 5).

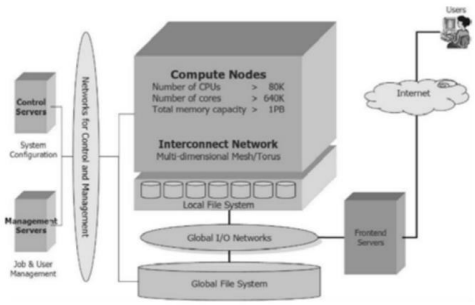

Figura 5 - Estructura de los componentes que son parte del sistema de supercómputo. (El-Kassabi et al., 2023)

Además, existen maneras de realizar pruebas de eficiencia o estrés para estos nodos por ejemplo con una prueba llamada Linpack y se mide su poder de cómputo por la cantidad de operaciones con números decimales que puede realizar en un segundo, llamados flops (Floating Point Operations Per Second).

La capacidad de procesamiento se puede considerar desde unos cientos de núcleos hasta millones. Por ejemplo, la máquina actual más poderosa tiene un poco más de 10.6 millones de núcleos y puede realizar un poco más de 93,000 Tflops.

Es importante destacar que el uso del supercómputo es para resolver problemas muy complejos o difíciles, que requieren mucho tiempo para hallar su solución en un laboratorio o computadora normal; y que son de interés para la ciencia, tecnología, salud, la economía o alguna área del conocimiento.

Para esto además de tener acceso a la supercomputadora se requiere: de software adecuado, el cual puede ser una solución generada por la comunidad de investigadores de un área del conocimiento, el cual sea de distribución libre (software libre) o se tenga que pagar por él (software comercial); plantear correctamente el problema, preparar los datos suficientes y necesarios al equipo de supercómputo; y finalmente utilizar un equipo en el cual el software procese esos datos en un tiempo razonable, de no ser así podrían pasar años, décadas o siglos para hallar la solución buscada y en general después de ese tiempo se podría perder el interés por los resultados, y esto en el caso que la supercomputadora en ese tiempo no falle o se quede sin energía, por ejemplo.

2.4.1 – Algunos ejemplos del uso de supercómputo para la investigación.

Un ejemplo interesante del uso de una supercomputadora corresponde a la investigación de los efectos de diferentes sustancias químicas diseñadas (fármacos) que sean candidatos para ayudar a curar o prevenir una cierta enfermedad o comportamiento de algún microorganismo en un ser vivo, como puede ser un animal o vegetal. Es posible mediante software de química para calcular la interacción entre el fármaco candidato y la célula o tejido de interés y en un tiempo corto (segundos, minutos u horas) tener información de su efecto en las condiciones programadas, sin necesidad de hacer pruebas en un laboratorio y esperar los resultados (días, semanas, meses o años). Entonces el uso de la supercomputadora mediante un programa de "simulación" permite descartar los candidatos con poco o ningún efecto y quedarse con los candidatos con mayores posibilidades de éxito. Luego estos podrán ser probados en un laboratorio y en consecuencia al ser menos, el tiempo para hallar el fármaco buscado se reducirá mucho. (Fuente: Saberes y Ciencias, n.d.)

Otro ejemplo corresponde al estudio de señales eléctricas del cerebro relacionadas con eventos de ataques epilépticos. Para esto se procesan registros de individuos que padecen de la enfermedad y se buscan patrones que permitan medir con anticipación la probabilidad que en los próximos digamos 10 segundos se vaya a producir un ataque epiléptico severo. Esta predicción permite suministrar de forma automática al paciente un medicamento adecuado luego de reconocerse el patrón previo al evento epiléptico y en consecuencia ayudar a reducir los efectos del ataque. De igual forma se puede plantear una estrategia para eventos cardiacos o de otro tipo. En estos casos la tarea primaria es hallar los patrones (tarea pesada) y luego mediante un método rápido hacer la identificación (tarea ligera), la primera puede requerir supercómputo y la segunda la puede hacer por ejemplo un teléfono celular. (Fuente: Saberes y Ciencias, n.d.)

Para situaciones genéricas planteadas en los ejemplos anteriores se ha desarrollado software y de igual manera para otros escenarios (y en caso de no haber un software específico, si ya se tiene un modelo de solución se puede crear uno adecuado de forma incremental) para plataformas de supercómputo de diferentes tallas y así resolver algunos de los problemas que la ciencia y tecnología demandan, ayudando las supercomputadoras a reducir los tiempos para hallar respuestas y en consecuencia poner a disposición de los profesionales y personas soluciones a múltiples problemas.

2.4.2 – Clúster de supercómputo.

La arquitectura de un clúster de supercómputo típicamente consta de nodos de cómputo, nodos de almacenamiento y nodos de gestión, todos conectados a través de una red de alta velocidad. Aquí hay una descripción general de cada uno de estos componentes (Figura 6):

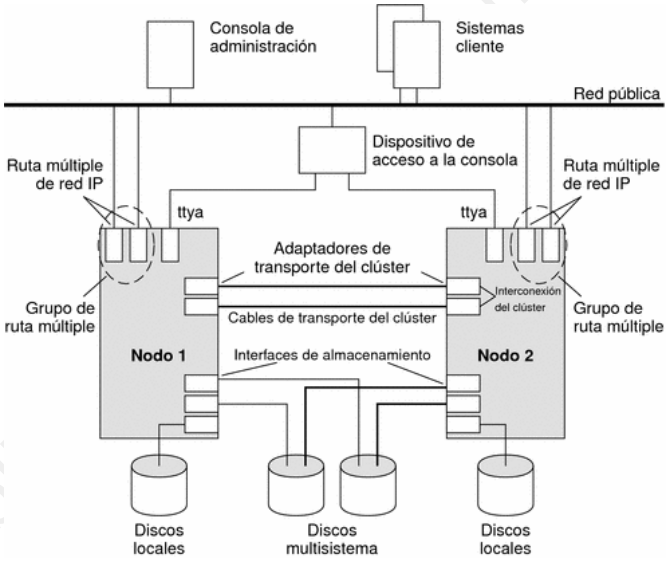

Figura 6 – Esquema de arquitectura de clúster para supercómputo con 2 nodos como parte del clúster. (Villaseñor Cendejas, n.d.)

- **Nodos de cómputo:** son las computadoras que realizan el trabajo de procesamiento de datos en el clúster. Estos nodos suelen tener una gran cantidad de procesadores de alto rendimiento, como procesadores multinúcleo o procesadores gráficos (GPU), que pueden procesar múltiples tareas de manera simultánea. La cantidad de nodos de cómputo en un clúster puede variar desde unos pocos hasta varios miles, dependiendo del tamaño del clúster.

- **Nodos de almacenamiento:** son nodos que proporcionan almacenamiento masivo y de alta velocidad para los datos en el clúster. Estos nodos suelen tener discos duros de alta capacidad y velocidad, como sistemas de almacenamiento en red (NAS) o sistemas de almacenamiento en disco (SAN). Los nodos de almacenamiento también

pueden tener un software especializado para el acceso y la gestión de los datos, como sistemas de archivos paralelos (por ejemplo, Lustre o GPFS).

- **Nodos de gestión:** son nodos que proporcionan servicios de administración y coordinación para el clúster. Estos nodos son responsables de monitorear y gestionar el estado de los nodos de cómputo y de almacenamiento, programar trabajos en el clúster y coordinar la comunicación entre los nodos. Los nodos de gestión también pueden proporcionar servicios de seguridad, como autenticación y cifrado de datos.

- **Red de interconexión:** es la red de alta velocidad que conecta todos los nodos en el clúster. La red de interconexión es crucial para el rendimiento del clúster, ya que permite la comunicación y el intercambio de datos entre los nodos de manera eficiente. Las tecnologías de red de alta velocidad utilizadas en los clústeres incluyen InfiniBand y/o Ethernet de alta velocidad.

A partir de un conjunto de varios (decenas, centenas, miles) servidores individuales (nodos de cálculo), que están interconectados usando redes de comunicación de alta velocidad, que cuentan con dispositivos con una gran capacidad de almacenamiento masivo de datos y que por lo general se emplean en problemas que requieren de cómputo numérico intensivo.

2.4.3 – Campos de la investigación científica aplicables al entorno de supercómputo.

Hay dos grandes campos donde trabajar: simulación y búsqueda de patrones.

- En simulación se usa para crear modelos virtuales de un problema y correr varias pruebas con diferentes variables para observar el comportamiento de un fenómeno, como, por ejemplo, el clima, una reacción nuclear, la división e integración de proteínas o el tráfico vehicular.
- La búsqueda de patrones sirve para decodificar mensajes encriptados, o para interpretar señales radio magnéticas que bien podrían indicar la existencia de una inteligencia extraterrestre, o para encontrar discrepancias en el comportamiento de mercados de valores.

El uso del supercómputo con fines científicos permite abordar una gran cantidad de problemas científicos que de otro modo serían difíciles o imposibles de resolver. Por ejemplo, la predicción confiable del clima a escalas de tiempo cada vez mayores sería imposible sin el uso de las supercomputadoras. Otros ejemplos son la simulación de la creación y evolución de las galaxias que componen a nuestro Universo; el estudio de nuevos fármacos en el tratamiento de enfermedades; el estudio de las corrientes de aire en un nuevo diseño aerodinámico de un avión o de un auto; la búsqueda de números

primos que tienen más de 100 millones de dígitos; el análisis de los datos de los mayores aceleradores de partículas para buscar nuevas partículas.

Podemos afirmar que en prácticamente todas las áreas del conocimiento se requiere actualmente, o se requerirá en el futuro próximo, del uso de la supercomputación, lo cual nos permitirá aumentar substancialmente nuestro conocimiento de la naturaleza y sin duda incrementar la duración y la calidad de la vida humana.

Algunas aplicaciones relevantes son:

- **Investigación médica** modelando el cerebro y otros órganos y sistemas del cuerpo humano, secuenciando y estudiando el doblamiento de proteínas para el desarrollo de nuevos fármacos,
- **Seguridad** mediante el análisis masivo de datos y la obtención de patrones de éstos, generando, validando y descifrando métodos criptográficos,
- **Diseño urbano** optimizando la red y el flujo para el tránsito vehicular,
- **Estudio de desastres naturales** simulando terremotos, inundaciones y tsunamis,
- **Modelos aerodinámicos** de automóviles
- **Industria aeronáutica** diseñando y optimizando aeronaves y sus componentes,
- **Simulación de procesos** en distintos ámbitos como: la simulación de accidentes vehiculares, procesos subatómicos, movimientos de partículas, fenómenos astrofísicos y cosmológicos, yacimientos fracturados, entre muchos otros.

Capítulo 3. Análisis de necesidades.

Antes de implementar el ambiente de cómputo en la nube para la investigación, es importante realizar un análisis de necesidades para garantizar que la implementación sea efectiva y adecuada para las necesidades de los investigadores.

3.1. – Hallazgos de las necesidades de los investigadores.

Durante el análisis de viabilidad y factibilidad del proyecto, uno de los puntos a considerar fue el tener una entrevista o sesión en línea (ya que se realizó en medio de la pandemia COVID-19) con algunos de los investigadores con los que cuenta la Facultad de Ingeniería de la Universidad Panamericana Campus Bonaterra para poder obtener hallazgos que permitan determinar las necesidades fundamentales a cubrir y solventar mediante la implementación del clúster virtual para investigadores mediante supercómputo en la nube.

3.1.1 – Profesores Investigadores entrevistados.

Actualmente el total de investigadores con los que cuenta la Facultad de Ingeniería de la Universidad Panamericana Campus Bonaterra es de 16 investigadores de diversas áreas y líneas de investigación. (Panamericana, n.d.)

Se realizó una entrevista con un grupo de investigadores del campus a los vía Google Meet ya que se realizó durante pandemia, para obtener cuales son los puntos que se debe considerar para implementar esta solución en la nube. Para la entrevista se hizo la invitación a un total de 6 investigadores lo que significa un total del 50% de la plantilla de investigadores.

Los participantes fueron:

- Dra. Claudia Nallely Sánchez Gómez – Inteligencia Artificial
- Dr. Héctor Eduardo Gilardi Velázquez – Sistemas Físicos
- Dr. Josué Ortiz-Medina – Ingeniería Bioelectrónica
- MC. Ricardo Espinosa Loera – Inteligencia Artificial
- Dra. María Teresa Orvañanos Guerrero – Sistemas Mecatrónicos
- Dr. José Sebastián Gutierrez Calderón – Sistemas Mecatrónicos

3.1.2 – Hallazgos de necesidades.

A continuación, se analiza los requerimientos que surgieron a través de la entrevista realizada como se puede ver en la Figura 7:

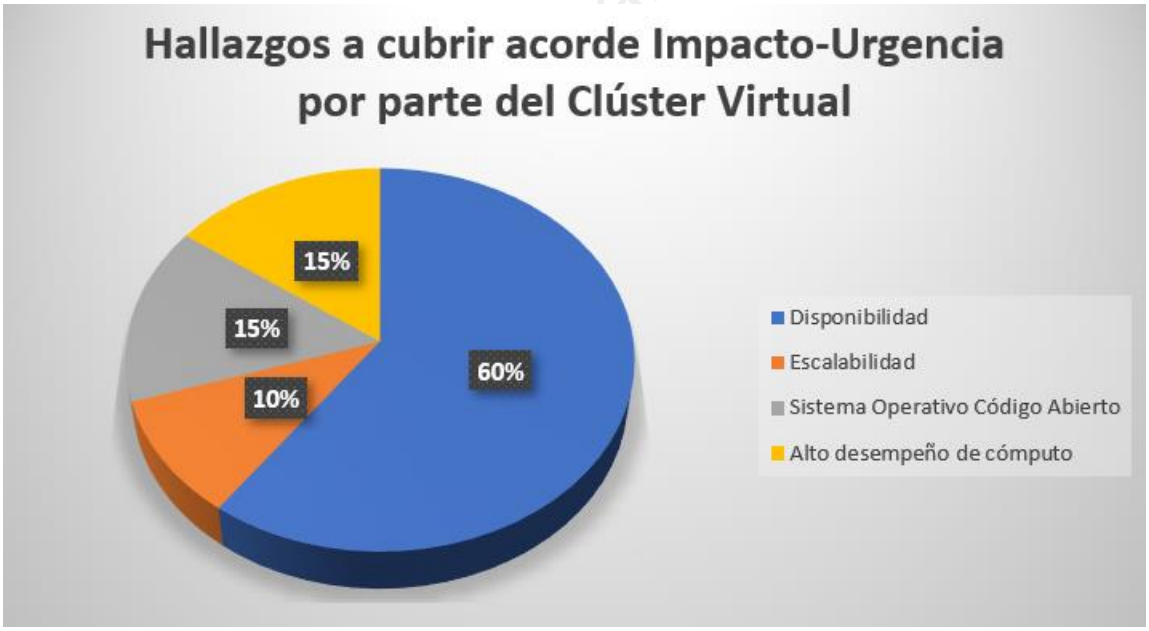

Figura 7 – Porcentajes de los hallazgos que los investigadores ven con mayor Impacto-Urgencia por solventar por medio del clúster virtual para investigadores.

- **Disponibilidad**: Los investigadores hicieron mucho hincapié que actualmente con cómputo tradicional necesitan esperar turno de algún servidor para hacer uso de él, además de que tiene que ser dentro del Campus para poder acceder a este. Además de que requieren que esté disponible la mayor parte del tiempo y que no tengan límites de tiempo de uso. ***Porcentaje de impacto-necesidad: 60%***

- **Escalabilidad:** Otro de los puntos que los investigadores del Campus refirieron es el hecho de que en caso de que se requiera incrementar las capacidades de las

máquinas virtuales, o replicar una máquina virtual para tener una copia exacta para pruebas y producción, que se pueda realizar sin inconvenientes, así como poder reducir las capacidades una vez que ya no se requieran como tal. ***Porcentaje de impacto-necesidad: 10%***

- **Sistema Operativo Libre o de código abierto:** De preferencia, los investigadores mencionaron que prefieren trabajar con sistemas operativos libres, como Linux, dado que la mayoría de las librerías o herramientas de investigación están soportados únicamente para este tipo de sistemas además de que en tema de licenciamiento suele incrementar costos que impactan a la sustentabilidad de la investigación. ***Porcentaje de impacto-necesidad: 15%***

- **Alto Rendimiento de cómputo:** Por último comentaban que es necesario contar con altos niveles de cómputo tanto de procesamiento como de almacenamiento pero hicieron mucho énfasis específicamente en términos de GPUs ya que muchas de las investigaciones están basadas en análisis de imágenes, pixeles, y mapas celulares que requieren un alto poder de manejo por parte de los ambientes de máquinas virtuales, y que actualmente en ambientes de cómputo tradicionales son un cuello de botella muy grande porque no renderizan en tiempo y forma como ellos quisieran. ***Porcentaje de impacto-necesidad: 15%.***

3.1.3 – Puntos a considerar a partir de los hallazgos.

- **Análisis de costos**: La implementación de la computación en la nube puede ser costosa. Por lo tanto, es importante realizar un análisis de costos para evaluar si la implementación de la computación en la nube es más rentable que la infraestructura de cómputo local. Esto incluye el costo de los recursos informáticos, la energía, el mantenimiento, la seguridad y la capacitación.

- **Selección de proveedor de servicios en la nube:** Hay una variedad de proveedores de servicios en la nube disponibles. Es importante seleccionar un proveedor que tenga experiencia en la gestión de recursos informáticos para la investigación y que tenga una infraestructura de alta disponibilidad y seguridad.

- **Evaluación de seguridad:** La seguridad es una consideración crítica en la implementación de la computación en la nube para la investigación. Es importante evaluar las políticas de seguridad del proveedor de la nube y asegurarse de que se cumplan los requisitos de seguridad de los investigadores.

- **Desarrollo de políticas y procedimientos:** Es importante desarrollar políticas y procedimientos claros para el uso de la computación en la nube para la investigación.

Esto incluye políticas de acceso a datos, políticas de seguridad y procedimientos para la gestión de recursos informáticos.

- **Capacitación y soporte:** Es importante proporcionar capacitación y soporte a los investigadores para que puedan utilizar la infraestructura de la nube de manera efectiva. Esto incluye capacitación en el uso de la infraestructura, así como soporte técnico y de servicio al cliente.

3.2. – Análisis de cada uno de los hallazgos.

3.2.1 – Disponibilidad

- **Disponibilidad en nube:** Alta disponibilidad es un protocolo de diseño del sistema y su implementación asociada que asegura un cierto grado absoluto de continuidad operacional durante un período de medición dado. Para los investigadores esto es de suma importancia ya que si confiadamente deja en ejecución un programa, algoritmo, prueba de eficiencia o cualquier procedimiento que requiera de cómputo de alto rendimiento no debe de presentar corte alguno de energía, suministro o conexión.

- La disponibilidad de un sistema es la proporción de tiempo en la que puede atender a sus usuarios. Se mide en 9s, por ejemplo 5 **9s** significa 99.999% de uptime o tiempo funcionando. El cálculo 99.999% uptime implica 1 - 0.99999 = 0.00001 de downtime (tiempo sin funcionar) dónde 1 = 100% y 0.99999 = 99.999%. Es decir que en un año sólo podemos tener un downtime de 0.00001 * 365 * 24 * 60 = 5 minutos y 15.36 segundos. En el mejor de los casos, 5 minutos es lo que tardo en leer una alerta en el celular y encender la computadora. (Ojeda, n.d.)

- Con este acuerdo el usuario final puede esperar que el servicio no esté disponible durante los siguientes períodos de tiempo:

✓ Diario: 0.9 segundos
✓ Semanal: 6.0 segundos
✓ Mensual: 26,3 segundos
✓ Anual: 5 minutos y 15,6 segundos

- Google Cloud por ejemplo establece el 99.9% (3 9s) de disponibilidad para interconexión dedicada (Google Cloud, n.d.-b) mientras que Microsoft con Azure y AWS ofrecen el 99.99% (4 9s). (AWS, n.d.; JowsNunez, n.d.)

- Esto en un ambiente físico sería prácticamente imposible y muy costoso ya que requeriría de una alta inversión de conectividad tanto eléctrica para tener un par de

acometidas, plantas de emergencia y baterías de respaldo (UPS) de alta gama para mantener en funcionamiento un clúster de servidores para investigación.

- Disponibilidad también se refiere a la habilidad de la comunidad de investigadores para acceder al clúster, ejecutar nuevos trabajos, actualizar o alterar trabajos existentes o recoger los resultados de trabajos previos. Si un investigador bajo su perfil de usuario no puede acceder al sistema se dice que está no disponible es decir tiene tiempo de inactividad o caída de sistema o servidor.

- La alta disponibilidad y el costo en el cómputo en la nube están correlacionados de la siguiente manera:

✓ **Mayor disponibilidad, mayor costo:** La alta disponibilidad en la nube implica tener redundancia de servidores y la capacidad de mantener los servicios en funcionamiento incluso en caso de fallas. Esto requiere una infraestructura y recursos adicionales, lo que a menudo se traduce en mayores costos para el proveedor del servicio en la nube. Por lo tanto, los servicios en la nube que ofrecen alta disponibilidad tienden a tener precios más altos.

✓ **SLA y costos:** Los acuerdos de nivel de servicio (SLA, por sus siglas en inglés) son contratos entre los proveedores de servicios en la nube y los clientes que definen los niveles de disponibilidad garantizados. Los proveedores de servicios en la nube que ofrecen una alta disponibilidad suelen tener SLAs más estrictos, lo que implica que deben cumplir con ciertos estándares de disponibilidad. Para lograr esto, deben invertir en infraestructura y tecnología más robustas, lo cual se refleja en los costos.

- Dentro de la alta disponibilidad también debemos de considerar el tema de copias de seguridad o respaldos. Se puede definir que una copia de seguridad es un respaldo de algo que en un tiempo determinado es posible recuperarlo. Gracias a las copias de seguridad es posible recuperar objetos, servidores e incluso infraestructuras enteras en caso de desastre o perdida. (Daniel Romero Sanchez, n.d.)

- Normalmente las copias de seguridad se pueden realizar en local y/o en remoto. Hay una regla de buenas prácticas que se denomina 3-2-1 y consiste en:

✓ Realizar 3 copias de los datos cada día, siempre que sea posible.
✓ Guardar las copias en 2 soportes distintos. Estos pueden ser cintas, discos, USB, etc.
✓ Almacénalas 1 un lugar distinto de donde se encuentran los datos. En el caso de nube estamos hablando de una región distinta.

- Existen diferentes formas de realizar las copias de seguridad:

- ✓ Copia completa de los datos. Se dispone de todos los datos y es más sencillo llevar un control de versiones. Tiene el inconveniente que es muy lenta.
- ✓ Copia de seguridad incremental. A partir de una copia completa se pueden ir guardando los últimos cambios producidos en los datos. Tiene la desventaja que para restaurar datos serán necesarias varias copias de seguridad donde estén almacenados todos los datos.
- ✓ Copia de seguridad diferencia. A partir de una copia completa, se realizan Backups de todos los cambios. Si disponen de copias de todos los cambios desde la última copia completa. Para la recuperación se necesita la copia completa y la diferencial.

- Un snapshot es más común encontrarlo en ambientes de virtualización o de nube y es una instantánea de los metadatos de un sistema en un tiempo determinado. Esta instantánea puede ser de una máquina virtual, un volumen, o una base de datos. Se realizan en un medio simular al original. Los snapshot se guardan y se restauran rápidamente siendo excelentes para el control de versiones.

- Los snapshot son complementarios a las copias de seguridad ya que es capaz de disponer de la copia en un momento en el tiempo. Son instantáneas rápidas de realizar y de restaurar, pero pueden corromperse y perderse además de que utilizan el mismo lugar de almacenamiento del sistema, pero utilizan menos almacenamiento.

- Mientras que las copias de seguridad son una copia completa de los datos capaz de replicarse, pueden almacenarse en localizaciones y medios distintos, pero utilizan gran capacidad de almacenamiento y permiten restauraciones parciales de los datos.

3.2.2 – Escalabilidad

- La escalabilidad es la capacidad que tiene la infraestructura de cómputo en la nube para crecer al mismo tiempo que se incrementa la demanda de las soluciones que se ejecutan sobre ella. De esta manera siempre podemos estar seguros de que todas las peticiones de los usuarios están siempre resueltas.

- Dentro del tema de escalabilidad unas de las inquietudes de los investigadores entrevistados fue el que el ambiente de investigación en nube no quede corto en cuanto a las capacidades de procesamiento.

- Otras de las inquietudes de la escalabilidad es el hecho de que se corra el riesgo de sobredimensionar la arquitectura de cómputo.

- La ventaja fundamental del cómputo en la nube es que la escalabilidad es parte importante de su concepto como trabajo:
    - Se aplica a todos los servicios de nube: servidores, almacenamiento, GPUs, conectividad, etc.
    - Ajusta los costos de infraestructura a la realidad del consumo.
    - Elimina el sobredimensionamiento de recursos.
    - Evita problemas a los usuarios y clientes a la hora de utilizar los servicios.
    - Permite reducir costos y mantiene la infraestructura ordenada.

3.2.3 – Poder de cómputo

- Los investigadores requieren un gran poder de procesamiento ya que la mayoría de los proyectos de investigación están relacionados con análisis de imágenes, operaciones de grandes cantidades de iteraciones, machine learning, procesamiento de lenguaje natural o ciencia de datos.
- Ante este reto lo que se debe de solventar es el uso de GPUs de alta capacidad, una gran cantidad de núcleos en paralelo, nodos de cómputo y memoria RAM suficiente para el procesamiento adecuado de las operaciones que serán ejecutadas.

3.2.4 – Sistema Operativo de código abierto

- Los investigadores comentaron que preferían el uso de software de código abierto, específicamente Linux.
- El software de investigación abierto o software de investigación de Código Abierto se refiere al uso y al desarrollo de software para el análisis, la simulación y la visualización cuyo código fuente completo está disponible.
- Los investigadores prefieren los sistemas operativos de código abierto ya que tienen comandos que el sistema operativo Windows simplemente no ofrece. Como Linux se puede personalizar, hay menos limitaciones que con el sistema operativo Windows. Existen "hacks" para alterar el sistema operativo Windows, pero hacer esto infringe la licencia de usos aceptables de Windows. Por lo general, las actualizaciones a un sistema operativo sobrescriben los cambios personalizados no aprobados, así que no sirve de nada implementar cambios "hackeados" en el código del sistema operativo.

Capítulo 4 – Solución propuesta.

Es necesario trabajar en un ambiente que sea un clúster de alta disponibilidad y de alto rendimiento para obtener una configuración de equipos diseñado para proporcionar capacidades de cálculo mucho mayores que la que proporcionan los equipos individuales mientras que esté garantizado para el funcionamiento ininterrumpido de ciertas aplicaciones.

Dentro de lo que se propone en el presente proyecto es contar con un clúster de alta disponibilidad que no es más que un conjunto de dos o más máquinas que se caracterizan por mantener una serie de servicios compartidos y por estar constantemente monitorizándose.

4.1 – Características de la solución propuesta.

- ***Alta disponibilidad de infraestructura:*** Si se produce un fallo de hardware en alguna de las máquinas del clúster, el software de alta disponibilidad es capaz de arrancar automáticamente los servicios en cualquiera de las otras máquinas del clúster. Y cuando la máquina que ha fallado se recupera, los servicios son nuevamente migrados a la máquina original. Adicionalmente se cuenta con los recursos de disponibilidad de la nube que proporciona un ambiente similar, pero en los servidores físicos del fabricante dónde está almacenado nuestro clúster virtual en la nube.

- ***Alta disponibilidad de aplicación:*** Si se produce un fallo del hardware o de las aplicaciones de alguna de las máquinas del clúster, el software de alta disponibilidad es capaz de arrancar automáticamente los servicios que han fallado en cualquiera de las otras máquinas del clúster. Y cuando la máquina que ha fallado se recupera, los servicios son nuevamente migrados a la máquina original. Esta capacidad de recuperación automática de servicios nos garantiza la integridad de la información, ya que no hay pérdida de datos, y además evita molestias a los usuarios, que no tienen por qué notar que se ha producido un problema.
- ***Alto Desempeño:*** Un clúster de alto rendimiento es un conjunto de servidores que está diseñado para dar altas prestaciones en cuanto a capacidad de cálculo. Los motivos para utilizar un clúster de alto rendimiento son: El tamaño del problema por resolver y El precio de la máquina necesaria para resolverlo.

Los clústeres en la nube son los que se están realizando en algunas entidades con instancias gratuitas con alto desempeño en RAM, CPU y GPU que consiguen competir en capacidad de cálculo con supercomputadoras de alto nivel.

4.1.1 – Alto desempeño en GPUs (Procesamiento de gráficos).

La tecnología de procesadores gráficos tuvo un gran avance hace aproximadamente 10 años y aunque su origen fue la aceleración de tareas referentes al procesamiento gráfico, dado su buena prestación en el cálculo de números en punto flotante se comenzaron a utilizar en la resolución de problemas de propósito general.

Se comenzó a desarrollar el área GPGPU (General Purpose Graphical Processor Unit) incorporando más funciones, mayor poder de cómputo a las tarjetas y desarrollando herramientas que permitan manipular de forma sencilla los distintos dispositivos.

En la actualidad, las GPUs se componen de multiprocesadores orientados al cálculo de punto flotante en simple y doble precisión que disponen de una memoria local independiente. Las tarjetas son interconectadas a los nodos a través del bus PCI y generando, de esa forma, nodos híbridos con CPUs y GPUs.

- **Procesamiento en paralelo:** es el que se puede realizar mediante varias CPU o unidades de procesamiento de gráficos (GPU). Las GPU, diseñadas originalmente para gráficos independientes, son capaces de realizar diferentes operaciones aritméticas por medio de una matriz de datos (como pixeles de pantalla) de forma simultánea.

La capacidad para trabajar en varios planos de datos al mismo tiempo hace que las GPU sean la elección natural para el procesamiento en paralelo en tareas de aplicaciones de aprendizaje automático (AA), como el reconocimiento de objetos en videos.
Lo que se busca que es que este clúster de cómputo de alto rendimiento para investigadores tenga una capacidad suficiente tanto de CPU y GPU para este trabajo en paralelo a gran escala y permita ejecutar aplicaciones de alto densidad gráfica para el manejo adecuado de pixeles, imágenes telescópicas a gran escala, gráficas moleculares, etc. A nivel del dispositivo GPU, se proveen hilos de ejecución (uno por procesador), administrados en forma eficiente por el dispositivo, que son agrupados en bloques. Los hilos del mismo bloque comparten la memoria disponible en el multiprocesador y todos comparten la memoria global del dispositivo, formando lo que se conoce como una grilla. No hay un orden fijo de ejecución entre bloques, se ejecutan paralelamente si hay suficientes multiprocesadores disponibles en la tarjeta o, si no, en tiempo compartido.

Para superar los límites de la supercomputación, se necesitan diferentes arquitecturas de sistemas. Para que el procesamiento en paralelo pueda llevarse a cabo, la mayoría de los sistemas de alto rendimiento integran varios procesadores y módulos de memoria a través de interconexiones con un ancho de banda enorme. Ciertos sistemas de alto rendimiento combinan varias CPU y GPU, lo que se conoce como computación heterogénea (Figura 7).

Figura 8 - Comparación entre un CPU convencional y un GPU (Patel & Kushwaha, 2020)

4.1.2 – Clúster de procesos auto escalable.

Con los clústeres Auto Escalables lo que se busca es que pueden escalar automáticamente según sea necesario para satisfacer las demandas de recursos de todas las tareas y servicios en su clúster, incluida la escala hacia y desde cero.

Con esto se mejora la fiabilidad, la escalabilidad y el costo de ejecutar cargas de trabajo ya que al momento de que el clúster deja de utilizar los recursos que se incrementaron, regresa al esquema anterior para tener eficiencia no sólo de recursos de cómputo sino también económicos al no pagar por algo que ya no se está utilizando.
Permite a su vez tener históricos de cuándo se requirió esta auto escalación para saber en qué periodos de tiempo se requirió este tipo de acción.
Hay que hacer énfasis que actualmente esta característica se encuentra sólo en contenedores Kubernetes de clústeres de servidores.

4.1.3– Sistema Operativo UBUNTU.

Dentro de la comunidad científica hay muchas tareas que requieren de un sistema operativo que poder ajustar a las necesidades propias de la tarea que se quiera desarrollar. GNU/Linux permite esa personalización mejor que ningún otro sistema operativo. Ya sea porque requiere de ciertas necesidades o porque se quiere incorporar ciertas herramientas y distribuir ese sistema con esas herramientas en muchos equipos. GNU/Linux por su potencia, eficacia, desarrollo y seguridad es una buena alternativa. Existen en particular muchas distribuciones basadas en Linux Ubuntu ya establecidas para ambientes científicos como: Poseidón Linux, Bio-Linux entre otros (Figura 8).

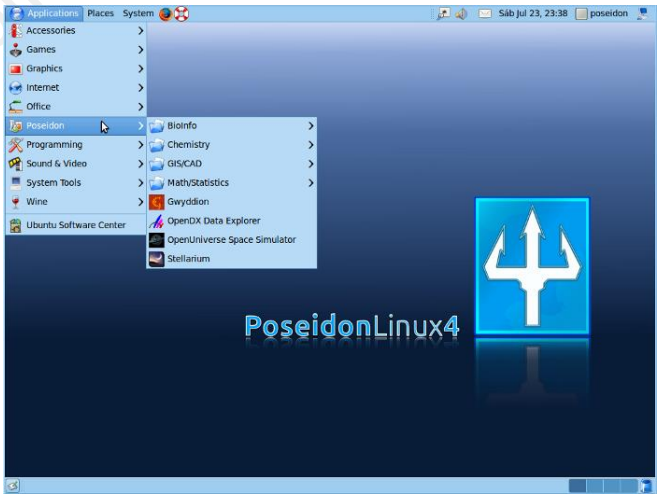

Figura 9 - Pantalla de inicio de la distribución de UBUNTU – Poseidón Linux (Linux, n.d.)

4.1.4 – Topología de red: Hub and Spoke - VPN.

La capacidad de conectarse a través de métodos más seguros, como la utilización de redes privadas virtuales (VPN) en una infraestructura pública, ha aumentado la capacidad de los proveedores de servicios en la nube para proteger los datos de los clientes públicos y empresariales a través de una arquitectura de nube privada.

Una red privada virtual tradicional puede soportar una pequeña escala de conexiones seguras, en contraste, la nube es de una escala impresionante. Por lo tanto, VPN no puede permitirse el ajuste de flexibilidad en la infraestructura de la nube. Las corporaciones y los proveedores de servicios pueden acceder a los canales de comunicación públicos mediante el uso de la arquitectura VPN a través de IPsec, SSL o PPTP para reducir los costos asociados.

Para que los usuarios confíen en una seguridad mecanismo en la nube, una infraestructura de nube privada tiene que ser adquirido que desplegaría y garantizaría la seguridad de los datos en esta plataforma

Aunque las VPN convencionales se pueden aplicar a la nube, sólo una escala relativamente pequeña de conexiones puede ser gestionado por el Administrador. Para ello se propone la topología **Hub-and-Spoke** con Puerta de Salida VPN que son capaces de gestionar la gran escala de la nube. (Alvarado et al., 2013). Esta configuración de red en estrella tipo hub-and-spoke usa los siguientes elementos arquitectónicos (Microsoft, n.d.):

- **Conectividad de red virtual.** Esta arquitectura conecta redes virtuales mediante conexiones de emparejamiento o grupos conectados. Las conexiones de emparejamiento y los grupos conectados son conexiones no transitivas de baja latencia entre las redes virtuales. Las redes virtuales emparejadas o conectadas pueden intercambiar tráfico a través de la red troncal de sin necesidad de un enrutador.
- **Firewall**. Existe una instancia de Firewall administrada en su propia subred.
- **Puerta de enlace de VPN Gateway.** Una puerta de enlace de red virtual permite que una red virtual se conecte a un dispositivo de red privada virtual (VPN). La puerta de enlace proporciona conectividad de red entre entornos locales. Para obtener más información, consulte Conexión de una red local a una red virtual de Microsoft Azure y Extensión de una red local mediante VPN.
- **Dispositivo VPN.** Un dispositivo o servicio VPN proporciona conectividad externa a la red entre entornos locales. El dispositivo VPN puede ser un dispositivo de hardware o una solución de software.

4.2 – Diagrama de flujo del sistema.

En el presente diagrama se propone el flujo de sesión (Figura 9), aplicación de seguridad y topología de red propuesta para la entrega del servicio en la nube seleccionada(Wicaksono et al., 2023), dónde:

- MV = Máquina Virtual
- Hub and Spoke = Es la topología de acceso a la red bajo la seguridad de una puerta de acceso de VPN (Red virtual privada).

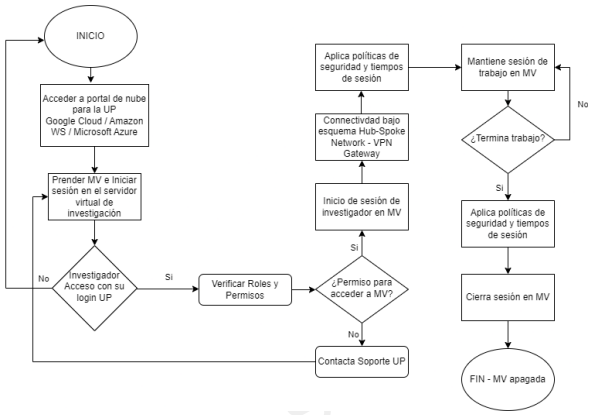

Figura 10 – Diagrama de flujo del sistema.(Kummar Maurya et al., 2023)

- En la aplicación del sistema de diagrama de flujo, se divide en dos, a saber, desde el lado del administrador, que proporcionará acceso al usuario, y el usuario mismo que accederá a la biblioteca de servidores que contiene el clúster.

- La misma figura describe cómo se ejecuta el proceso del sistema. El administrador abre la plataforma en la nube para activar la MV y generar una clave para ingresar a la MV, el siguiente paso es abrir el portal de la plataforma de nube y que se pueda acceder a la MV bajo el recurso de trabajo autorizado dependiendo del grupo de seguridad o rol establecido, luego configure una red privada virtual (VPN) para hacer un túnel entre los servidores de la nube y los usuarios pueden conectarse con una conexión estable. Al ejecutar una conexión VPN, la instancia mantendrá la conectividad entre la MV y el usuario a través de una red segura.

- Si el usuario desea acceder a la MV, puede hacerlo a través del portal, o mediante conexión SSH (Secure Shell) o en su defecto por Conexión de Escritorio Remoto si es una MV con sistema operativo Microsoft.

- La MV puede mantener la conexión entre los usuarios si hay una conexión a Internet estable. El sistema de la MV en la nube que se ejecuta no sobrecargará el dispositivo del usuario. Cuando se completa la actividad del usuario en la MV, el usuario puede salir de la MV y el administrador cerrará la sesión. (Wicaksono et al., 2023)

4.3 – Fases para la implementación.

La propuesta es implementar un clúster en la nube para que los investigadores de la Universidad Panamericana, pero a su vez los estudiantes puedan tener un entorno de poder de cómputo para los proyectos de investigación, así como áreas específicas de trabajo en 4 fases:

Fase 1: Investigación y elección del líder en cómputo de nube para implementar el laboratorio al menor costo posible para la institución: Google Cloud, Microsoft Azure o IBM Cloud y revisar sustentabilidad.

Fase 2: Establecer los entornos de cómputo para las áreas de investigación basado en las prioridades de investigadores, profesores de la facultad de Ingeniería y el mercado actual basado en cuatro rubros de las tendencias tecnológicas actuales para implementar en el laboratorio virtual para Inteligencia Artificial:

- Procesamiento de imágenes
- Procesamiento de lenguaje natural
- Machine Learning
- Ciencia de datos

Fase 3: Elaboración del entorno virtual y los recursos de cómputo necesarios
Fase 4: Pruebas del laboratorio virtual con investigadores y liberación de entorno.

Dentro de estas fases, existen puntos específicos en los que se debe de poner mayor atención para el desarrollo de estas a las que llamaremos iteraciones:

Iteración 1: Análisis y toma de decisión sobre proveedores de servicios de nube:
Se debe de considerar los 3 grandes entornos de nube con los que la Universidad Panamericana cuenta ya sea en el uso de herramientas propias del proveedor o intereses de uso de estas:

- Microsoft Azure
- Google Cloud
- Amazon Web Services (AWS)

Después de un análisis de los 3 proveedores, se pudo determinar que todos cuentan con un conjunto completo de recursos de proceso, red y almacenamiento integrados con servicios de orquestación de cargas de trabajo para aplicaciones HPC. Cuentan con toda la infraestructura de los Centros de Datos con los que cuentan a nivel mundial y con soluciones y servicios de aplicaciones optimizados creados específicamente para supercómputo.

Se revisaron los siguientes aspectos:

a. Rendimiento optimizado con control de costos: Cuenta con una gama completa de características de CPU, GPU entre otras e interconexión rápida para reducir el tiempo de realización de los trabajos de días a minutos. Podemos encontrar servidores Linux y Windows para ambientes habilitados para GPU y CPU respectivamente.
b. Cuentan con almacenamiento de gran escala tanto en discos administrados y no administrados, de mecanismos mecánicos o de estado sólido dependiendo de las necesidades de la aplicación o de la base de datos o uso del servidor que permiten aceptar grandes demandas de datos tanto de lectura como de escritura.
c. La conectividad de red y la alta disponibilidad es un valor agregado a este entorno de supercómputo.

Iteración 2: Elección de la arquitectura por proveedor adecuada para hacer el análisis de costos:

d. Solicitar los costos definidos en un entorno de cómputo en clúster para cada uno de los proveedores y realizar un análisis de costos entre plataformas mediante revisión de cotizaciones con los descuentos que se puedan otorgar a una institución educativa.
e. Elaboración de una Matriz de impacto entre implementación y costos de las plataformas bajo revisión.
f. Resolución de plataforma a elegir basado en costos.

Iteración 3: Análisis de cargas y almacenamiento para revisar las opciones de cómputo:

g. Análisis de cargas de GPU para entornos de investigación
h. Análisis de cargas de CPU para entornos de investigación
i. Análisis de cargas de Memoria RAM para entornos de investigación
j. Resolución de plataforma a elegir basado en los resultados de un mejor rendimiento.

Iteración 4: Creación del clúster virtual (Proceso):

k. Creación del tenant o subscripción para el clúster virtual
l. Generación de usuarios administradores
m. Generación de usuarios para investigadores
n. Definición de servidores con carga GPU
o. Creación de clúster de servidores Linux con GPU seleccionados
    i. Se debe de crear un recurso desde la plantilla diseñada para clústeres dependiendo del proveedor.
    ii. Cuántos nodos se deben asignar al clúster (2 o 3 servidores)
    iii. Asignar el tipo de máquina a asignar (tener en cuenta con GPU y que sea de la capa gratuita)

iv. Asignar la seguridad que debe de tener (preferentemente asignar un certificado SSL para el clúster)
v. Una vez finalizado el proceso corresponderá comenzar la generación de los usuarios locales para que cada investigador tenga su perfil.
vi. Se debe de configurar el administrador del clúster para que detecte los nodos con los que va a comenzar a trabajar.

p. Instalación de librerías y drivers necesarios para los servidores Linux
q. Instalación y puesta a punto de aplicaciones para investigadores
r. Generación de servidor ya instalado y configurado para crear plantillas para futuro uso.
s. Generación de servidores plantillas
t. Todos los servidores estarán conectados a unidades virtuales de disco compartidos con una entidad de almacenamiento central que permita tener acceso a la documentación de cada uno de los investigadores ya sea desde el servidor o sus sesiones locales de dominio en LDAP (Microsoft u Open Source).

Iteración 5: Roles y seguridad

u. Establecer los roles dentro del entorno que serán permitidos para el uso adecuado del clúster de investigación virtual que en la mayoría de los proveedores de cómputo en la nube se les llama "Roles IAM" (Identity and Access Management – Administración de identidades y accesos).
v. Estos roles pueden ser desde los establecidos de manera orgánica por la plataforma o realizar roles e identidades "customizadas" para lo que se busca, por ejemplo:
    - Rol de investigador
    - Rol de estudiante
    - Rol de jefe de investigadores
    - Etc.
w. Dentro de la seguridad dichos roles ya cuentan con un nivel de seguridad inherido por la plataforma, además de que los proveedores de cómputo en si cuentan con los sistemas más avanzados en materia de seguridad, agregando además el hecho de que existe relación de confianza con el Directorio Activo (Active Directory) del dominio **UP.EDU.MX.LOCAL** para que utilicen los grupos de seguridad establecidos.
x. Además, se establecen políticas de seguridad de uso, para que limite el uso de lo que se puede hacer y no dentro del clúster de investigación virtual.

Iteración 6: Pruebas y accesos a los investigadores:

Todos los servidores, servicios, accesos que serán creados bajo el dominio y tenant de up.edu.mx para que los investigadores puedan acceder a los equipos mediante su cuenta de dominio o ID que normalmente la Universidad les provee

y deberán ser probados de manera adecuada para evitar conflictos en la puesta a producción del entorno. Todos los servidores, servicios, accesos que serán creados bajo el dominio y tenant de up.edu.mx para que los investigadores puedan acceder a los equipos mediante su cuenta de dominio o ID que normalmente la Universidad les provee y deberán ser probados de manera adecuada para evitar conflictos en la puesta a producción del entorno.

Iteración 7: Dimensionamiento presupuestal del ambiente:

Dentro del presupuesto se debe de dimensionar y monitorear los siguientes puntos:

- Recursos de cómputo:
  - Total de clústeres o máquinas virtuales.
  - Tamaños (SKUs) autorizados para no exceder los costos.
  - Tiempo de utilización.
  - Rendimiento y poder de uso de las máquinas virtuales.
- Almacenamiento:
  - Respaldos
  - Tipo de almacenamiento ya que por ejemplo un almacenamiento Hot (caliente) que son datos o archivos en constante uso el costo es mayor a un tipo de almacenamiento para archivos en reposo o Cool ya que el almacenamiento es menor porque no tiene operaciones sobre los datos.
- Taza de transferencia de datos:
  - Envío y entrada de datos diarios, semanales y mensuales.
- Servicios adicionales:
  - Contratos de licenciamiento y soporte por parte del proveedor (Google, Microsoft o Amazon).

4.4 – Análisis de costos y ahorro.

El Cómputo en la nube ha sido, y sigue siendo, toda una evolución a nivel tecnológico para los entornos de cómputo y más aún para entidades educativas y de investigación. La posibilidad de desvincularse de la adquisición y mantenimiento de equipos físicos como servidores, o la disponibilidad de sus recursos desde cualquier lugar son solo algunos de los beneficios de la computación en la nube. El uso de soluciones de cómputo en la nube supone un cambio en la forma de gestionar los recursos tecnológicos de las empresas, así como una nueva forma de contratarlos. En este sentido, el precio también es una nueva variable para tener cuenta. Vamos a analizar cuáles son los aspectos clave del costo de un servicio en la nube para investigación (Figura 10).

Figura 11 – Ventajas y desventajas del cómputo en la nube en sus diferentes ámbitos.(Barnard & Delgado, n.d.)

4.4.1– Pago por uso.

La principal novedad que se encontrará todo negocio que dé el salto a una solución en la nube frente a un recurso informático tradicional es la **modalidad de pago por uso**. El precio de un servicio de Cómputo en la nube no es un desembolso inicial único ni una cuota fija mensual. En su lugar, se utiliza una **tarificación que tiene en cuenta los recursos reales consumidos** de las soluciones contratadas (Figura 11).

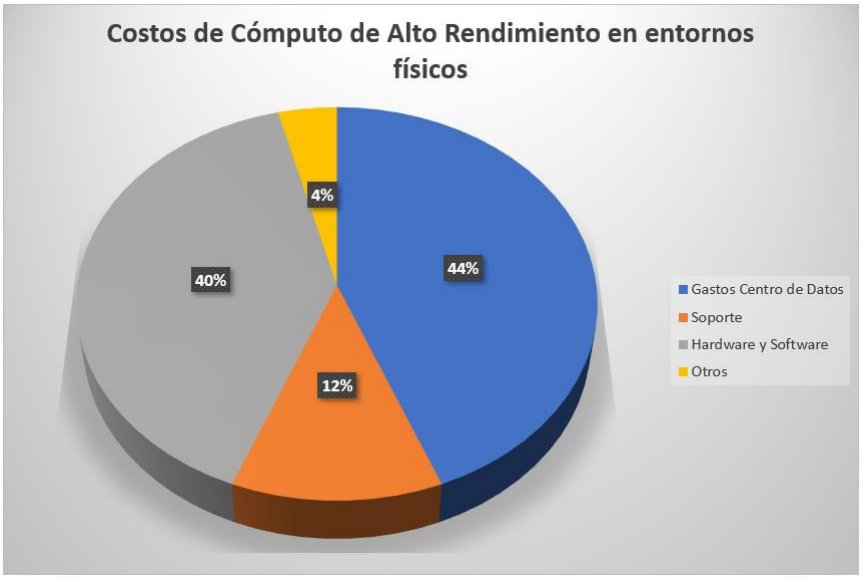

Figura 12 – Aspectos generales de los porcentajes relacionados con los costos del cómputo tradicional(Wang, 2023)

Por ejemplo, si contratamos un servidor virtual en la nube para alojar una tienda on-line, seguramente tengamos más visitas en los meses de Navidad o en días de rebajas. Durante dichos períodos, aumentará el consumo de recursos de servidor para evitar que nuestra web funcione de forma incorrecta. En consecuencia, nuestra cuota aumentará. Por el contrario, en aquellos meses en el que el consumo sea menor, pagaremos menos.

4.4.2 – ¿Qué costos desaparecen con el Cómputo en la nube?

La adopción del servicio de cómputo en la nube frente al uso de recursos informáticos tradicionales supone, como hemos visto anteriormente, un cambio en la modalidad de pago. Esto implica la desaparición de costos típicos asociados a soluciones hechas en casa, que podemos resumir en:

- **Mantenimiento.** Desaparecen las tareas de conservación y actualización de los recursos tecnológicos (por ejemplo, un servidor). En el Cómputo en la nube, nuestro proveedor del servicio será el encargado de llevar a cabo el mantenimiento de los recursos e infraestructura que hayamos contratado.
- **Adquisición de hardware**. En el entorno de la nube no tendremos que comprar un nuevo servidor si queremos tener acceso a más recursos o potencia. En lugar de ello, podremos escalar los recursos contratados en nuestra solución de Cómputo en la nube.
- **Infraestructura asociada**. Costos de electricidad, seguridad o climatización asociados habitualmente al mantenimiento de servidores también desaparecen, dado que será nuestro proveedor quien se encargue de los mismos.

4.4.3 – Ahorros que se obtienen mediante el cómputo en la nube a una solución informática tradicional.

En general, la modalidad de pago por uso y los costos que desaparecen hacen que el precio del Cómputo en la nube sea más económico para la mayoría de los casos.

Por el contrario, uno de los temores más comunes sobre el esquema del pago por uso del Cómputo en la nube es la dificultad de prever el consumo de recursos y, por ende, el costo asociado a los mismos. (Figura 12)

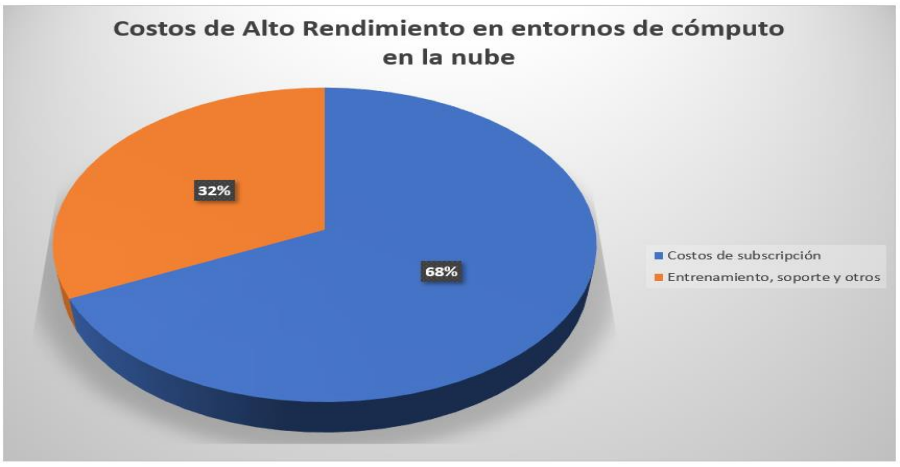

Figura 13 - Aspectos generales de los porcentajes relacionados con los costos del cómputo en la nube (Wang, 2023)

4.4.4 – Costos operativos de cómputo en la nube vs servicios en sitio: Estimación de ahorros.

Se estima que el costo principal en la empresa es el costo de personal de operación, y en el centro de datos, este costo es del orden del 15%. En una empresa promedio el radio típico de personal de TI es de 1:100. La automatización es parcial, y el error humano es causa de un gran porcentaje de problemas. En el centro de datos en la nube, la automatización es un tema mandatorio para lograr escalación, y es un principio de diseño fundamental. En un centro de datos para la nube el radio de miembros de personal es de 1:1000. La automatización y las técnicas de recuperación del cómputo en la nube resuelven la mayoría de los problemas que surgen.

Las diferencias principales de costos entre los centros de datos tradicional y el servicio de nube son:

1 Las grandes economías de escala, donde el tamaño de los centros de datos (algunos de 100,000 servidores) presentan una oportunidad de explotar la economía de escala al disminuir costos por volumen, lo cual no sucede en las empresas que deben invertir grandes montos iniciales;

2 Aumento proporcional, donde la optimización de espacio y numero de dispositivos, en la nube se explota el uso eficiente de carga de trabajo distribuidos en un gran número de servidores de bajo costo.

Para la operación de los centros de datos se requieren diversos rubros donde se incluye el costo operativo de TI y los costos de personal de TI empresarial son del orden del 15% del total de costos de operación y en el caso de la nube es del orden del 5%. Por lo que el monto estimado de ahorro podría ser de hasta un 10% del monto total dedicado a la operación de TI, considerando que al mover la operación a la nube se llega al 5% partiendo del 15% de costo de personal. Podemos observar las diferencias en precios entre los 3 competidores principales del cómputo en la nube (Tabla 3):

AWS vs. Azure vs. Google		
Provider	**Storage**	**Pricing**
Amazon S3	S3 Standard Storage	First 50 TB / Month $0.023 per GB Next 450 TB / Month $0.022 per GB Over 500 TB / Month $0.021 per GB
	S3 Standard-Infrequent Access (S3 Standard-IA) Storage	All storage / Month $0.0125 per GB
	S3 One Zone-Infrequent Access (S3 One Zone-IA) Storage	All storage / Month $0.01 per GB
Amazon EBS	Amazon EBS General Purpose SSD (gp2) Volumes	$0.10 per GB-month of provisioned storage
	Amazon EBS Provisioned IOPS SSD (io1) Volumes	$0.125 per GB-month of provisioned storage $0.065 per provisioned IOPS-month
	Amazon EBS Throughput Optimized HDD (st1) Volumes	$0.045 per GB-month of provisioned storage
Amazon Glacier	S3 Glacier Storage	All storage / Month $0.004 per GB
	S3 Glacier Deep Archive Storage	All storage / Month $0.00099 per GB
Google Cloud Storage	Multi-Regional	$0.026 - $0.036 per GB/month
	Regional	$0.02 - $0.035 per GB/month
	Nearline	$0.01 - $0.02 per GB/month
	Coldline	$0.004 - $0.014 per GB/month
Microsoft Azure	Block Blobs	$0.002/GB per month
	Azure Data Lake Storage	$0.001/GB per month
	Managed Disks	$1.54 per month
	Files	$0.060/GB per month

Tabla 3 - Descripción general de las principales opciones de precios en la nube para AWS, Azure y GCP (Jay Chapel, n.d.)

Una de las principales inquietudes sobre la implementación de un nuevo modelo de infraestructura en la nube es su costo, comparado contra los gastos de un modelo tradicional en sitio. La gran mayoría de las empresas está en proceso de invertir en infraestructuras virtuales porque son conscientes de que es la base para su transformación digital. (IT cloud services, n.d.)

Presentamos una comparación de costos anualizados (Figura 13) sobre un caso práctico de dos servidores ubicados en la nube contra su ubicación tradicional en sitio, comprendiendo su instalación, acondicionamiento de espacio físico y servicios de administración y mantenimiento.

	On premise		Nube	
Infraestructura	**Set-up** (costo único)	**Servicios** (anualizado)	**Set-up** (costo único)	**Servicios** (anualizado)
Infraestructura (2 servidores)	$ 22,300	-	-	$ 11,400
Equipamiento (UPS, switches, routers)	$ 2,800	-	-	-
Administración	-	2,300	-	-
Total	**$ 25,100**	**$ 2,300**	**$ 0**	**$ 11,400**
Total infraestructura 1er. año	**$ 27,400**		**$ 11,400**	

* Precios en dólares americanos

Figura 14 – Comparativa precios (USD) de implementación de súper cómputo en la nube vs. Súper Cómputo tradicional (IT cloud services, n.d.)

Dentro del mantenimiento de la plataforma deben considerarse los costos de consumo eléctrico, climatización, renovación de licencias, sueldos de personal y demás gastos adicionales, pero esto es únicamente para el cómputo tradicional, ya que en el tema de cómputo en la nube esos costos están implícitos en el pago mensual del uso de la infraestructura (Figura 14) (IT cloud services, n.d.)

Mantenimiento	*On premise* Servicios (anualizado)	Nube Servicios (anualizado)
Consumo eléctrico (2 servidores)	$ 3,516	-
Aire acondicionado Data Center	$ 4,220	-
Mantenimiento	$743	-
Total mantenimiento 1er. año	**$ 8,479**	**$ 0**

* Precios en dólares americanos

Figura 15 - Comparativa precios (USD) de mantenimiento de súper cómputo en la nube vs. Súper Cómputo tradicional (IT cloud services, n.d.)

El costo de implementar un clúster de supercómputo en en la nube (Amazon Web Services, Microsoft Azure o Google Cloud Platform) dependerá de varios factores, como el tamaño y la complejidad del clúster, la región que se utiliza, el tipo de instancia que se elige, y el tipo de almacenamiento y transferencia de datos que se requiere.

- **Instancias**:
    - Amazon Elastic Compute Cloud (EC2) g3.4xlarge (GPU)
    - Microsoft Azure N6 optimizado a GPU
    - Google Cloud g2-standard-32

  Estas instancias han sido elegidas porque:
        - Potencializan el poder de cómputo
        - Optimizadas al uso de GPUs
        - Mayor rendimiento y escalabilidad
- Otros costos por considerar son:
    - Costo de almacenamiento.

Es importante destacar que se ofrecen diferentes planes de precios y herramientas para optimizar el costo de la implementación de un clúster de supercómputo en la nube. Es recomendable que los usuarios utilicen estas herramientas para controlar los costos y optimizar la implementación. En resumen, el costo de implementar un clúster de supercómputo en la nube

dependerá de los factores mencionados anteriormente, pero podría oscilar entre cientos y miles de dólares al mes.

Aquí se mencionan los costos específicos para este caso de estudio mostrando los costos principales de los tres principales proveedores de supercómputo en la nube con GPUs que podrían involucrarse en la implementación de un clúster de supercómputo en la nube (Tabla 4).

	Amazon Web Services (AWS)	Microsoft Azure	Google Cloud Platform
Tipo instancia	g3.4xlarge (GPU)	N6 (GPU)	g2-standard-32
vCPUs	16	6	32
RAM	122	56	128
GPU	Tesla M60 de NVIDIA	NVIDIA TESLA L4	NVIDIA TESLA L4
Servidores	4	4	4
Almacenamiento (USD)	$ 200.00	$ 180.00	$ 200.00
Total por mes (USD)	**$ 5,096.84**	**$ 2,805.94**	**$ 3,964.86**
Total al año (USD)	**$ 61,162.08**	**$ 33,671.28**	**$ 47,578.32**

Tabla 4 – Comparativa de precios con la infraestructura analizada para su implementación (Precios en dólares americanos) (Google Cloud, n.d.-a)

A partir de los datos arrojados en la Tabla 5, podemos dar como conclusión que la mejor opción para implementar el clúster de supercómputo de investigación para la Facultad de Ingeniería en la nube es con Google Cloud Platform dado los siguientes puntos a considerar:

1. **Costo-Beneficio**: De los 3 competidores, es el segundo lugar en costo mensual y anual, pero considerando que Google Cloud Platform es el que mayor capacidad de cómputo tiene de las 3 por lo que se tiene un mayor impacto de procesamiento en esta solución.
2. **Capacidad de cómputo:** Google Cloud Platform presenta servidores virtuales en la nube con el doble de capacidad de núcleos o procesadores virtuales (virtual CPUS) así como el más alto en el rango de memoria RAM, lo cual nos da como resultado tener más del 50% de capacidad de cómputo sobre las demás opciones.
3. **Capacidad de GPUs:** Uno de los puntos a considerar por los investigadores de la Facultad de Ingeniería fue el cuidar la capacidad de procesamiento de imágenes (GPUs) por lo que cuenta junto con Microsoft Azure con la mejor tecnología establecida hoy en día, pero sin olvidar el punto anterior de la capacidad de cómputo.

4.5 – Análisis de viabilidad económica.

La viabilidad económica del supercómputo en la nube en comparación con el cómputo tradicional puede variar dependiendo de varios factores, como el tipo de aplicación, el tamaño de los datos y los requisitos de rendimiento. A continuación, se presentan algunos puntos a considerar:

- **Costos iniciales:** Inversión inicial para la implementación de supercómputo en la nube contra un cómputo tradicional.
- **Escalabilidad:** En caso de ser requerido, la capacidad de incrementar o reducir la capacidad de cómputo según sea necesario en ambos ambientes.
- **Costos operativos:** Se refiere a costos adicionales a los que ya realiza la facultad para poder dar el servicio requerido.
- **Modelo de precios:** Si solo se realiza el pago por lo utilizado (ya sea al día, mes o año) o si es un modelo tradicional de inversión inicial.

Es importante enfatizar que no se considera dentro de los costos el licenciamiento de sistemas operativos dado que se tiene en plan utilizar Sistemas Operativos de libre uso como Linux Ubuntu. En la presente tabla (Tabla 5) podemos revisar los siguientes datos dónde podemos asegurar que la viabilidad económica para implementar el ambiente en nube tiene un mayor impacto tanto en lo económico como en la operación:

	Google Cloud Platform	Servidores y almacenamiento tradicional
Inversión inicial (MXN)	$ 0	$ 1,028,000.00
Costos mensuales (MXN)	$ 68,592.08	$ 0 con respecto a servidores
Escalabilidad	100%	25%
Costos operativos mens.	$ 0	$ 150,000.00
Modelo de precios	Pago por uso	Pago inicial, soportes y mantenimientos mensuales, pólizas
Total por mes (MXN)	$ 68,592.08	$ 235,666.67
Total al año (MXN)	**$ 823,104.94**	**$ 2,828,000.00**

Tabla 5 – Comparativa de precios entre super cómputo de nube vs. Cómputo tradicional (Google Cloud, n.d.-a)

A considerar los costos mensuales de operación presentados en la Figura 15:

Figura 16 – Costos de operación de un centro de datos tradicional contra cómputo en la nube. (Amin, n.d.)

Podemos presentar como la viabilidad económica de implementar el clúster de super cómputo en la nube para investigadores de la facultad muy por encima de tratar de implementar un ambiente similar en un centro de datos tradicional (Data Center) dónde:

- **La inversión inicial** es mucho menor en cómputo en la nube ya que los costos son mensuales conforme a uso.

- Mientras que en el cómputo tradicional necesitas hacer la inversión de compra de equipamiento tanto de servidores como de centro de datos (no considerado en la tabla).

- **La escalabilidad** en el cómputo de nube es inmediata e inclusive se propone que las 4 instancias se configuren con Auto escalabilidad ya establecida para que en caso de que necesite incrementar capacidad de cómputo lo pueda ser y regrese al modelo inicial cuándo ya no sea requerido, y eso se realiza en un ambiente productivo y al instante.
  - Por otro lado, en un ambiente tradicional se requiere que el servidor físico tenga la capacidad de crecimiento, que existan las piezas y además realizar el apagado total del servidor, en horario fuera de trabajo para hacer el cambio y normalmente por personal calificado que tiene un costo adicional.

- **Costos operativos** para el ambiente de super cómputo en la nube se pueden considerar nulos dado que solo se utiliza la conexión actual de internet, así como el personal actual de la Universidad en el departamento de TI para dar soporte al ambiente. No se necesita instalar nada especial o adicional para el funcionamiento de este.
  - En un ambiente tradicional, es necesario acondicionar un espacio en el SITE de ubicado en la Biblioteca de la Universidad para colocar un rack adicional en caso de que no exista espacio para instalar los 4 servidores nuevos, nuevo cableado, modificar el sistema de enfriamiento ya que estos 4 servidores estarán dando una mayor carga de calor al SITE así como de energía, adquirir una planta de emergencia y UPS para dar alta disponibilidad al servicio para lograr algo similar a lo que se puede obtener en el cómputo de nube, etc. Por lo tanto, los costos de operación son considerablemente más altos.

- El modelo de precios o pagos en la nube es de pago por uso, es decir, si el equipo de investigadores no requiere que los equipos estén en funcionamiento todo el tiempo, estos pueden ser apagados y no generarán ningún costo. Por el contrario, si están encendidos todo el tiempo, pero no existe actividad el consumo es más bajo.
  - En el esquema de cómputo tradicional no existe el pago por uso, ya que para ello se realizó una inversión inicial fuerte y hay que consumirlo como tal.

Se puede observar que la diferencia de económica entre el super cómputo en la nube y el esquema tradicional es muy grande con ahorros considerables que podrá traer grandes beneficios a la Facultad de Ingeniería y a la Universidad misma.

Además de los aspectos económicos, el supercómputo en la nube puede ofrecer beneficios adicionales, como la disponibilidad y accesibilidad global, la posibilidad de colaboración en línea y la capacidad de integrarse con otros servicios en la nube, lo que puede mejorar la eficiencia y productividad en ciertos casos.

4.6 – Curva de adopción y aprendizaje.

La curva de aprendizaje del uso de cómputo en la nube se refiere al proceso de adquirir habilidades y conocimientos necesarios para aprovechar eficientemente los servicios y recursos

que ofrece la computación en la nube. La curva de aprendizaje puede variar en duración y complejidad dependiendo de la experiencia previa del usuario y en este caso de la Facultad de Ingeniería así como de los investigadores con tecnologías de la información y la complejidad de los servicios en la nube que se desean utilizar.

A continuación, se presenta una descripción general de los principales puntos de la curva de aprendizaje del cómputo en la nube:

- **Comprensión de conceptos básicos:** Al comenzar, es importante familiarizarse con los conceptos fundamentales del cómputo en la nube, como qué es la nube, sus modelos de servicio (SaaS, PaaS, IaaS), los beneficios que ofrece, y cómo se comparan con los enfoques tradicionales de infraestructura de TI.

- **Selección de proveedor y servicios:** Existen múltiples proveedores de servicios en la nube, como Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), entre otros. Cada proveedor ofrece una amplia gama de servicios, y es necesario aprender a seleccionar los que mejor se adapten a las necesidades específicas del proyecto o aplicación.

- **Uso de la consola de administración:** Cada proveedor ofrece una interfaz de consola o panel de control para administrar los recursos en la nube. Aprender a navegar y utilizar estas interfaces es crucial para desplegar, configurar y gestionar recursos de manera efectiva.

- **Provisionamiento y configuración de recursos:** A medida que se gana experiencia, se aprende cómo aprovisionar y configurar máquinas virtuales, bases de datos, almacenamiento, redes, y otros servicios que componen la infraestructura en la nube necesaria para consolidar el clúster virtual para investigación.

- **Seguridad y cumplimiento**: Es importante entender las mejores prácticas de seguridad para proteger los datos y sistemas alojados en la nube, así como el cumplimiento de regulaciones y normativas relacionadas con la privacidad y la protección de datos y cumplir con el acceso adecuado con respecto a los perfiles tanto de administración como de los investigadores.

- **Monitoreo y optimización:** Aprender a monitorear el rendimiento y la utilización de los recursos en la nube es crucial para asegurarse de que se esté utilizando eficientemente y para identificar posibles problemas o áreas de mejora.

- **Automatización y orquestación:** Con el tiempo, se puede avanzar hacia la automatización y orquestación de tareas en la nube utilizando herramientas como scripts, plantillas de infraestructura, y contenedores.

Es importante tener en cuenta que el aprendizaje del cómputo en la nube es un proceso continuo, ya que la tecnología y los servicios en la nube están en constante evolución. La práctica y la experiencia práctica desempeñan un papel fundamental en la mejora de las habilidades y la comprensión de cómo aprovechar al máximo los beneficios de la computación en la nube como podemos ver en la Figura 16.

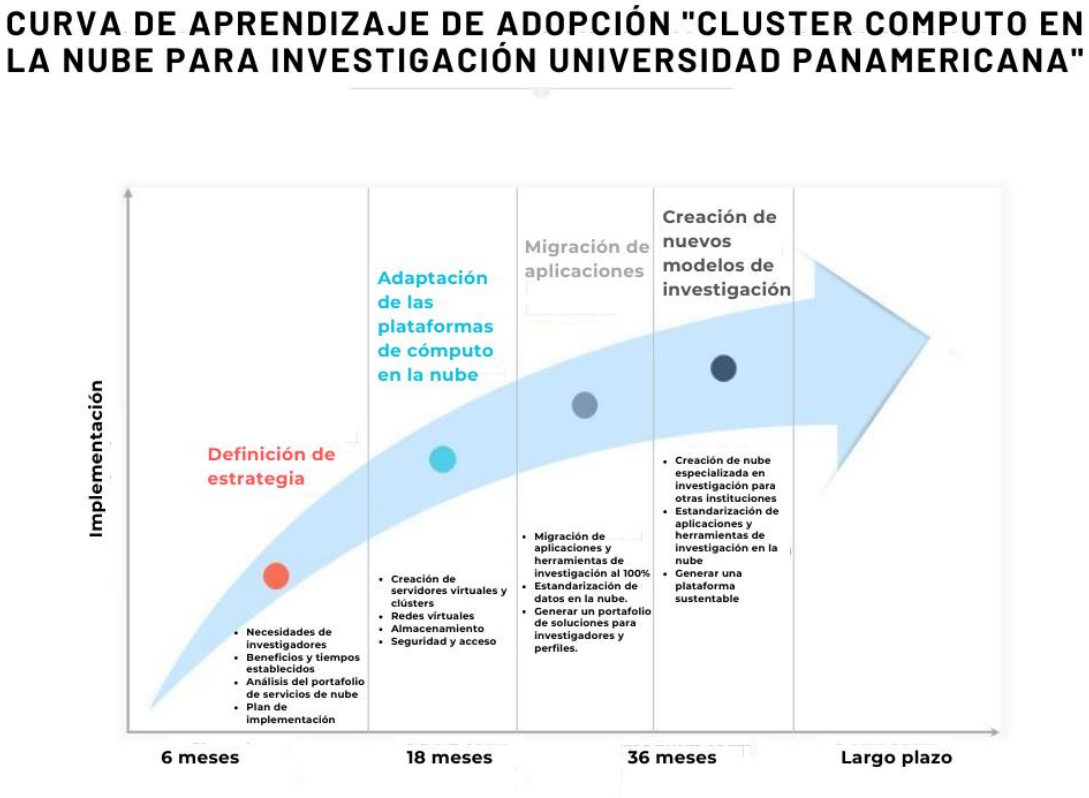

Figura 17 – Curva de aprendizaje y adopción del clúster virtual para investigación de la Universidad Panamericana.

Capítulo 5 – Conclusiones.

Esta tesis lo que se busca demostrar que el uso de la tecnología, en este caso el cómputo de alto rendimiento o supercómputo puede tener un impacto en los proyectos de investigación de la Universidad Panamericana, incentivar a más estudiantes a enfocarse a la investigación mediante entornos de trabajo tecnológicos modernos y facilitar a los equipos de investigadores o poder concluir sus investigaciones. A través de la aplicación de un clúster virtual de cómputo de alto rendimiento en la nube, se busca mitigar los problemas actuales que los lleva a los investigadores para poder realizar sus trabajos.

Los resultados arrojan un claro incremento en la capacidad de cómputo, así como en el rendimiento a gran escala, con una alta disponibilidad en el uso de servidores con entornos de código abierto y que pueden ser escalables a las necesidades de cada uno de los investigadores y que inclusive puedan continuar con sus trabajos indistintamente se encuentren en el campus o no. Además, damos respuesta a los ahorros que se obtienen al utilizar nodos de cómputo administrados por proveedores de nube como Microsoft, Amazon o Google ya que esta baja se da a que no existe necesidad de compra o uso de hardware de supercómputo o alto rendimiento y sus costos adyacentes y que repercute notablemente en la administración de presupuestos enfocados al pago de uso en el cómputo en la nube, fomentar el uso de tecnologías verdes y el pago solo a lo que se ha utilizado. Esto supone una disminución en los tiempos de proyectos de investigación y el incremento de cantidad de proyectos en ejecución.

Se define que la computación en clúster agrega y coordina una colección de máquinas virtuales que trabajan en conjunto para realizar una tarea recordando que los clústeres suelen tener un solo nodo principal o controlador, una cantidad de nodos de procesamiento y, posiblemente, algunos nodos especializados. El nodo principal administra muchas tareas, incluidas las siguientes:

- Registrar nodos de procesamiento en el sistema
- Asignar trabajos a nodos particulares
- Supervisar los nodos y los trabajos que se ejecutan en ellos

Lo que se busca es mitigar las actuales necesidades de los investigadores del campus que fueron entrevistados que son temas como accesibilidad a los equipos en cualquier horario y lugar, uso de código abierto y que si se requiere más capacidad de cómputo o GPU puede ser realizado sin mayor problema para poder continuar o concluir con los trabajos de investigación. Por último, esta investigación arroja datos sobre el tipo de infraestructura de nube sugerida para la creación de clústeres de cómputo, conectividad necesaria, bajo costo en licenciamiento al hacer uso de plataformas de código libre incluso llegando a los casos en que se utilicen servicios a costo cero, sin dejar de lado temas como seguridad y roles manejados desde los proveedores de servicios como por los propios administradores de la nube.

Lo más importante a considerar es que después del análisis realizado en este trabajo de investigación y lo presentado en la Tabla 5, la mejor opción en cuanto a rendimiento, pero sobre todo costo vs capacidad de cómputo es Google Cloud Platform, además de considerar que cuentan con múltiples de apoyos para la educación y la investigación.

Referencias

Alemami, Y., Al-Ghonmein, A. M., Al-Moghrabi, K. G., & Mohamed, M. A. (2023). Cloud data security and various cryptographic algorithms. *International Journal of Electrical and Computer Engineering*, *13*(2), 1867–1879. https://doi.org/10.11591/ijece.v13i2.pp1867-1879

Alvarado, M. D., Agrawal, R., & Baker, Y. (2013). Security mechanisms utilized in a secured cloud infrastructure. *Conference Proceedings - IEEE SOUTHEASTCON*, *April 2013*, 3–8. https://doi.org/10.1109/SECON.2013.6567430

Amazon Web Services. (n.d.). *Amazon Gravition*. https://aws.amazon.com/es/blogs/aws-spanish/comparando-el-desempeno-de-las-instancias-amazon-ec2-aws-graviton2-intel-x64-y-amd-epyc/

Amin, J. C. P. (n.d.). *Introducción a la Computación en la Nube*. https://www.easynube.co.uk/introduccion-a-la-computacion-en-la-nube/

AWS. (n.d.). *AWS Alta disponibilidad*. https://aws.amazon.com/es/blogs/aws-spanish/continuidad-de-negocio-en-aws-disponibilidad-y-resiliencia/#:~:text=Una región de AWS puede,de disponibilidad para algunos servicios.

Barnard, A., & Delgado, A. (n.d.). *Introducción al cómputo en la nube 8*.

Comunicación, C. (2022). *La nube multiplica la computación para propulsar la ciencia*. https://www.csic.es/es/actualidad-del-csic/la-nube-multiplica-la-computacion-para-propulsar-la-ciencia

Daniel Romero Sanchez. (n.d.). *Copias de Seguridad*. https://www.dbigcloud.com/backups/390-diferencias-entre-backup-y-snapshots.html

dcc.icgc.org. (n.d.). *dcc.icgc.org*. https://dcc.icgc.org/icgc-in-the-cloud

El-Kassabi, H. T., Serhani, M. A., Masud, M. M., Shuaib, K., Khalil, K., Nagasaki, M., Sekiya, Y. Y., Asakura, A., Teraoka, R., Otokozawa, R., Hashimoto, H., Kawaguchi, T., Fukazawa, K., Inadomi, Y., Murata, K. T., Ohkawa, Y., Yamaguchi, I., Mizuhara, T., Tokunaga, K., … Matsuda, F. (2023). Design and implementation of a hybrid cloud system for large-scale human genomic research. *Journal of Cloud Computing*, *12*(1). https://doi.org/10.1038/s41439-023-00231-2

Francisco Naranjo. (n.d.). *Propiedad Intelectual en la nube*. https://comunica-web.com/blog/marketing-digital/a-quien-pertenecen-los-documentos-que-subimos-a-internet/

Fuente: Saberes y Ciencias. (n.d.). *No Title*. https://lns.buap.mx/?q=noticias/11272018-1524/¿para-qué-usar-una-supercomputadora#:~:text=Un ejemplo interesante del uso,ser un animal o vegetal

Gartner. (n.d.). *No Title*. https://www.gartner.com/reviews/market/cloud-infrastructure-and-platform-services

Google Cloud. (n.d.). *Calculadora Google*. https://cloud.google.com/products/calculator#id=6acec8f7-be44-490f-a593-9e3a33037ecb

IT cloud services. (n.d.). *¿Cómo optimizar al máximo tus procesos manteniendo un nivel de flexibilidad a la medida de tus necesidades?* https://itcloudservices.mx/infraestructura-en-la-nube/

Jay Chapel, P. (n.d.). *Una revisión de los costes de almacenamiento en la nube*. https://www.datacenterdynamics.com/es/opinion/una-revision-de-los-costes-de-almacenamiento-en-la-nube/

JowsNunez. (n.d.). *Azure Alta Disponibilidad*. https://github.com/josejesusguzman/acordeon-az900-innovaccion/blob/main/res/formulario_sla.md

Kummar Maurya, S., Malik, S., & Kumar, N. (2023). Virtual machine tree task scheduling for load balancing in cloud computing. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(1), 388. https://doi.org/10.11591/ijeecs.v30.i1.pp388-393

Linux. (n.d.). *Poseidon Linux*. https://distrowatch.com/table.php?distribution=poseidon

Luis Felipe Ortiz Clavijo. (2018). Computación en la Nube: Estudio de Herramientas Orientadas a la Industria 4.0. *Computación En La Nube: Estudio de Herramientas Orientadas a La Industria 4.0*. https://www.redalyc.org/journal/6139/613964507007/html/

Lunazco Roger, C., & Chavez Jaime Tomas, C. (2022). *UNIVERSIDAD PERUANA DE LAS AMÉRICAS ESCUELA PROFESIONAL DE INGENIERIA DE COMPUTACION Y SISTEMAS TRABAJO DE INVESTIGACIÓN IMPLEMENTACIÓN DE UN SERVICIO CLOUD COMPUTING PARA MEJORAR LOS PROCESOS EN LA EMPRESA RCL, AÑO 2022 PARA OPTAR EL TITULO PROFESIONAL D*. 0–2. http://repositorio.ulasamericas.edu.pe/bitstream/handle/upa/2419/1.Trabajo de Investigacion final_Calixto Lunazco Roger_G7_PAP.pdf?sequence=2&isAllowed=y

Manuel Yrigoyen Quintanilla, C. T. P. (2011). *SERVICIOS EDUCATIVOS MEDIANTE LA UTILIZACIÓN DE TECNOLOGÍAS DE CLOUD COMPUTING*. 9–38.

Microsoft. (n.d.). *Topología de red en estrella tipo hub-and-spoke en Azure*. https://learn.microsoft.com/es-mx/azure/architecture/reference-architectures/hybrid-networking/hub-spoke?tabs=cli#workflow

ncbi.nlm.nih.gov. (n.d.). *ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/sra/docs/sra-nube/

NY Times. (n.d.). *Microsoft IA investment*. https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html

Ojeda, G. (n.d.). *Alta Disponibilidad 9s*. https://www.linkedin.com/pulse/disponibilidad-9s-y-slas-guillermo-ojeda/?originalSubdomain=es

Panamericana, U. (n.d.). *Investigadores UP*. https://research-eng-ags.up.edu.mx/investigadores-researches

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, *171*(2019), 158–167. https://doi.org/10.1016/j.procs.2020.04.017

RedHat. (n.d.). *Virtualización*. https://www.redhat.com/es/topics/virtualization/what-is-virtualization

Universitaria, R. D. (2013). *SUPERCÓMPUTO : APLICACIONES Y*. 1–9.

Villaseñor Cendejas, L. M. (n.d.). *Supercómputo como herramienta*. https://saberesyciencias.com.mx/2016/12/04/el-supercomputo-como-herramienta-en-la-investigacion-cientifica/

Wang, L. (2023). Providing Compliance in Critical Computing Systems. In *Springer Series in Reliability Engineering*. https://doi.org/10.1007/978-3-031-02063-6_10

Wicaksono, A. B., Munadi, R., & Sussi. (2023). Cloud server design for heavy workload gaming computing with Google cloud platform. *International Journal of Electrical and Computer Engineering*, *13*(2), 2197–2205. https://doi.org/10.11591/ijece.v13i2.pp2197-2205