UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSTGRADO

MAESTRÍA EN TELECOMUNICACIONES

"ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO."

Trabajo de Investigación previo a la obtención del Título de Magíster en Telecomunicaciones.

AUTOR(A):

ALEXANDRA NATALY CULQUI MEDINA.

DIRECTOR:

MSC. CARLOS ALBERTO VÁSQUEZ AYALA

IBARRA - ECUADOR

2021

APROBACIÓN DEL TUTOR

Yo, **Carlos Alberto Vásquez Ayala**, certifico que la estudiante **Alexandra Nataly Culqui Medina** con Cédula Nro. 100292537-6, ha elaborado bajo mi tutoría la sustentación del trabajo de grado titulado: "ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO".

Este trabajo se sujeta a las normas y metodologías dispuestas en el reglamento del título a obtener, por lo tanto, autorizo la presentación a la sustentación para la calificación respectiva.

Ibarra, 28 de junio de 2021

CARLOS ALBERTO VASQUEZ AYALA

Firmado digitalmente por CARLOS ALBERTO VASQUEZ AYALA
Fecha: 2021.07.05 14:42:52 -05'00'

Msc. Carlos Alberto Vásquez Ayala.

Tutor

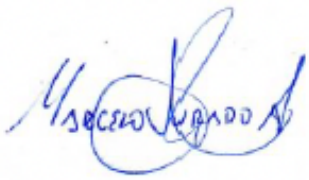
C.I.: 100242498-2

APROBACIÓN DEL TRIBUNAL

El presente trabajo de grado titulado "ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO", constituye requisito previo para la obtención del título de Magister en Telecomunicaciones del Instituto de Posgrado de la Universidad Técnica del Norte.

Autora: Alexandra Nataly Culqui Medina

Trabajo de grado, aprobado en nombre de la Universidad Técnica del Norte, por el siguiente jurado: Msc. Edwin Marcelo Jurado Ávila, Msc. Carlos Alberto Vásquez Ayala, Msc. Freddy Mauricio Tapia León, a los 7 días del mes de agosto de 2021.



Msc. Edwin Marcelo Jurado Ávila

Presidente del Tribunal

CARLOS ALBERTO VASQUEZ AYALA
Firmado digitalmente por CARLOS ALBERTO VASQUEZ AYALA
Fecha: 2021.08.11 14:21:17 -05'00'

Msc. Carlos Alberto Vásquez Ayala

Tutor

PRBDDY MAURICIO TAPIA LBON

Msc. Freddy Mauricio Tapia León.

Asesor

UNIVERSIDAD TÉCNICA DEL NORTE

BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DEL CONTACTO			
CÉDULA DE IDENTIDAD:	100292537-6		
APELLIDOS Y NOMBRES:	CULQUI MEDINA ALEXANDRA NATALY		
DIRECCIÓN:	Chorlavi		
EMAIL:	nathalymedina@hotmail.com		
TELÉFONO FIJO:	062932-450	**TELÉFONO MÓVIL:**	0980438045

DATOS DE LA OBRA	
TÍTULO:	ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO.
AUTORA:	CULQUI MEDINA ALEXANDRA NATALY
FECHA:	07 DE AGOSTO DE 2021
PROGRAMA DE POSGRADO:	MAESTRÍA EN TELECOMUNICACIONES.
TÍTULO POR EL QUE OPTA:	MAGISTER EN TELECOMUNICACIONES.
DIRECTOR:	MSC. CARLOS ALBERTO VÁSQUEZ AYALA

2. CONSTANCIAS

La autora Alexandra Nataly Culqui Medina, manifiesta que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de esta y saldrá en defensa de la Universidad en caso de reclamación por parte de los terceros.

Ibarra, a los 12 días del mes de agosto del año 2021.

ALEXANDRA NATALY CULQUI MEDINA

Firmado digitalmente por ALEXANDRA NATALY CULQUI MEDINA
Fecha: 2021.08.12 12:28:30 -05'00'

Alexandra Nataly Culqui Medina

C.I.: 1002925376

DEDICATORIA

El presente trabajo de investigación lo dedico a mis bellos hijos quienes son mi fuente de inspiración y fortaleza para cada día esforzarme y ser mejor persona y profesional. A mi amado esposo, quién estuvo a mi lado en cada paso brindándome palabras de aliento y el apoyo necesario para culminar con éxito este gran escalón profesional y académico. A mis padres, a mis suegros y mi familia quienes me han apoyado también en esta gran etapa, con su tiempo y apoyo incondicional.

RECONOCIMIENTO

Mis más sinceros sentimientos de gratitud al Msc. Carlos Vásquez por ser una guía para el desarrollo y culminación con éxito de este proyecto de investigación.

Al Msc. Freddy Tapia por su actitud siempre predispuesta y de interés en colaborar con sus conocimientos a la presente investigación.

A un gran amigo el Ing. Fabián Jimenez, ex Administrador del servicio de LSF (Load Share Facility) y de Aplicaciones Científicas del Supercomputador Quinde I de la ex Empresa Pública Siembra E.P., quién con su gran profesionalismo y pasión por el mundo de la Supercomputación, ha estado siempre dispuesto en colaborar en cada etapa de esta investigación, en especial durante el caso práctico referente a una temática muy importante respecto al COVID19; trabajo que hoy se encuentra publicada en Springer Link como “Natural Products as Potential Inhibitors for SARS-CoV-2 Papain-Like Protease: An in Silico Study”.

A la ex Empresa Pública Siembra E.P, quién me brindó todas las facilidades técnicas y administrativas para desarrollar el presente trabajo, y de esta forma contribuir con un granito de arena al Servicio de Supercomputación, el cual anhelo que este sueño de fomentar la investigación científica a través del Supercomputador Quinde I continúe y se fortalezca en beneficio de la comunidad académica – científica.

UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSGRADO

PROGRAMA DE MAESTRÍA EN TELECOMUNICACIONES

"ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO"

Autor: Alexandra Nataly Culqui Medina

Tutor: Msc. Carlos Alberto Vásquez Ayala

Año: 2021

RESUMEN

El presente trabajo de investigación comprende realizar un Análisis de Rendimiento de la Red de Altas prestaciones InfiniBand sobre una arquitectura de computación paralela del Supercomputador Quinde I, mediante el paso de mensajes sobre la interfaz MPI, utilizando la aplicación b_eff o también llamado Benchmark b_eff; el análisis se realizará sobre ciertos escenarios seleccionados basados en el HPC Challenge Benchmark y en la interconexión de los nodos de cómputo; donde se define los parámetros base para analizar el rendimiento de la red sobre ambientes paralelos. Una vez obtenidos los datos a través de la aplicación b_eff, se realizará un análisis de los escenarios y la comparación de los datos con otra infraestructura de Altas Prestaciones. Adicional, se realizará un análisis referencial de los datos obtenidos del Benchmark b_eff, con los datos del Test de Linpack. Parte de la presente investigación es aplicar los parámetros utilizados en el Análisis de Rendimiento de la Red de Altas prestaciones sobre un caso práctico con el fin de analizar el comportamiento de los procesos de cómputo en un ambiente real dentro de un proyecto de investigación científica. Al finalizar el análisis de los escenarios, comparación de los datos y la aplicación del caso práctico, se presenta una guía de buenas prácticas para la ejecución de procesos de cómputo en arquitecturas paralelas con el fin de obtener el mejor Rendimiento de la Red de Altas prestaciones InfiniBand en cualquier proyecto de investigación que se desarrolle en el Supercomputador Quinde I.

Palabras clave: *b_eff, Benchmark, Supercomputador, InfiniBand.*

UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSGRADO

PROGRAMA DE MAESTRÍA EN TELECOMUNICACIONES

"ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO"

Autor: Alexandra Nataly Culqui Medina

Tutor: Msc. Carlos Alberto Vásquez Ayala

Año: 2021

ABSTRACT

This research work includes a Performance Analysis of the InfiniBand High Performance Network on a parallel computing architecture of the Quinde I Supercomputer, by passing messages on the MPI interface, using the b_eff application or also called Benchmark b_eff; the analysis will be carried out on certain scenarios based on the HPC Challenge Benchmark and on the interconnection of the computing nodes; where the base parameters are defined to analyze the performance of the network on parallel environments. Once the data has been obtained through the b_eff application, an analysis of the scenarios will be carried out and the data will be compared with another High Performance infrastructure. Additionally, a referential analysis will be performed on the data obtained from the b_eff Benchmark, with the data from the Linpack Test. Part of the present investigation is to apply the parameters used in the Performance Analysis of the High Performance Network on a practical case in order to analyze the behavior of the computing processes in a real environment within a scientific research project. At the end of the analysis of the scenarios, comparison of the data and the application of the practical case, a guide of good practices is presented for the execution of computing processes in parallel architectures in order to obtain the best Performance of the High Performance Network InfiniBand in any research project that takes place on the Quinde I Supercomputer.

Keywords: *b_eff, Benchmark, Supercomputer, InfiniBand.*

PRESENTACIÓN

El presente trabajo “ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO”, en el Supercomputador Quinde I de la Empresa Pública SIEMBRA E.P., el cual se ha llevado a cabo con el propósito de fomentar la investigación científica de alto impacto, democratizar el uso de herramientas de cómputo de altas prestaciones, facilitando el procesamiento en paralelo de los trabajos de investigadores y estudiantes de Supercomputación.

El proyecto se basa en el Análisis de Rendimiento de la Red de Altas Prestaciones en aplicaciones que usan paralelismo de procesos o procesamiento en paralelo, con el objetivo de obtener tiempos de respuesta considerablemente cortos en comparación a equipos de cómputo de procesamiento serial, esta rapidez de procesamiento se logra a través de una red que facilita el paso de información entre procesos, a velocidades inigualables al contrario de los equipos domésticos, por lo tanto esto implica disponer de un óptimo grado de latencia y buen rendimiento de la Red de Altas prestaciones.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

CAPÍTULO I

EL PROBLEMA

Introducción

En la actualidad Ecuador ha impulsado políticas que aporten al campo de la investigación como se menciona en el Plan Nacional del Buen Vivir 2017-2021 (Secretaría Nacional de Planificación y Desarrollo, SENPLADES) dentro del Objetivo 5: "Impulsar la productividad y competitividad para el crecimiento económico sustentable de manera redistributiva y solidaria", política 5.3 "Promover la investigación, la formación, la capacitación, el desarrollo y la transferencia tecnológica, la innovación y el emprendimiento, en articulación con las necesidades sociales, para impulsar el cambio de la matriz productiva.". (Senplades, 2017).

Como parte del Plan Nacional del Buen Vivir el Gobierno ecuatoriano ha impulsado la formación de docentes e investigadores como estrategia de largo plazo para el cambio de la matriz productiva y energética. Como complemento se han implementado varias universidades como la Universidad de YachayTech en Urcuquí; la Universidad Regional Amazónica (IKIAM) en Tena; Universidad de las Artes (UNIARTES) en Guayaquil; y, Universidad Nacional de la Educación (UNAE) en Azogues (Senescyt, 2013). (Senplades, 2017).

Considerando las estrategias y políticas del país, la Empresa Pública Siembra E.P anteriormente llamada Empresa Pública Yachay E.P. entidad que cambio de denominación a partir del 10 de diciembre de 2019 por Decreto Ejecutivo Nro. 945; este organismo público es el encargado de liderar el proyecto Ciudad del Conocimiento, mismo que tiene como objetivo el cambio en la matriz productiva nacional, así como fomentar el desarrollo del conocimiento. Además, de impulsar el

Desarrollo de un ecosistema de docencia, investigación, innovación y producción en el país. (Constitucional., 2019)

La Empresa Pública Siembra E.P. implementa en el año 2017 el primer Supercomputador llamado “Quinde I” como herramienta tecnológica esencial para la comunidad científica del país, con el fin de impulsar varios proyectos de investigación que requieran una gran capacidad de procesamiento computacional o supercomputación. (Empresa Pública Siembra E.P., 2020)

En este contexto se destaca la Computación de Alto Rendimiento o HPC (High performance Computing por sus siglas en inglés), siendo este un instrumento muy importante en el campo de la investigación, el cual permite el procesamiento de datos a mayor velocidad y eficiencia; de esta forma abordar problemas u operaciones muy complejas que requieren de mayor capacidad computacional dentro de distintas áreas como la física, química, matemática, biología, medicina, inteligencia artificial, etc. En base a varios autores de bibliografías muestra el concepto de HPC como una herramienta que permite realizar una gran cantidad de operaciones computacionales en el menor tiempo posible, de la forma más eficiente posible. (Ocampo Yahuarcani & Campos Baca, 2017)

Problema de Investigación

El Supercomputador “Quinde I” cuenta con una de las infraestructuras con mayor capacidad a nivel regional, la cual dispone de una capacidad de procesamiento

de 231,9 Tflops[1] con una arquitectura de computación paralela Power 8 MIMD[2] NUMA (del inglés Multiple Instruction, Multiple Data, en español "múltiples instrucciones, múltiples datos") NUMA[3] (del inglés Non-Uniform Memory Access, en español "acceso a memoria no uniforme"), con una Red de 100 Gbit/s InfiniBand[4] EDR (Enhanced Data Rate por sus siglas en inglés) que brinda el servicio de Supercomputación a la comunidad académica, científica e industria del país. (Empresa Pública Siembra E.P., 2020)

Ecuador ha ido dando pasos pequeños en cuanto a Supercomputación en los últimos años, donde los principales actores han sido las Universidades públicas y privadas quienes han realizado varias investigaciones de aporte científico para el país, a pesar de ya disponer de clústeres de computación de diferentes capacidades, existe un gran vacío en cuanto a conocimientos de computación paralela sobre Redes de Altas Prestaciones, lo cual ha perjudicado el normal desarrollo y ejecución de los procesos de cómputo de varias investigaciones.

El Supercomputador “Quinde I” ha ejecutado más de 30 proyectos de investigación de diferentes áreas de la ciencia y tecnología. Varios proyectos de investigación se han ejecutado basados en parámetros y variables de otros trabajos de investigación y benchmarks[5], que permiten la ejecución normal de los procesos de cómputo, ya que esto da un aporte muy importante al investigador para mejorar los

[1] Flops. - Las operaciones de coma flotante por segundo, también conocidas como FLOPS, son una forma de medir el rendimiento que tiene un ordenador o una tarjeta gráfica.
[2] MIMD. - Múltiples tareas heterogéneas pueden ser ejecutadas al mismo tiempo, y cada procesador opera independientemente con ocasionales sincronizaciones con otros.
[3] NUMA (non-uniform memory access). - Los computadores se comunican explícitamente usando pase de mensajes entre ellos.
[4] La Arquitectura InfiniBand (IBA) es un estándar que define un subsistema de conmutación de alta velocidad para conectar con nodos de procesamiento dentro de un área de red.
[5] Benchmark. – Es una metodología que permite determinar el rendimiento de un equipo de supercomputación, medir su capacidad antes ciertas variables y escenarios propuestos.

tiempos de ejecución de sus jobs[6] o tareas de cómputo o a su vez crear nuevos escenarios experimentales de cualquier proyecto de investigación.

Ahora bien, el investigador debe contar con todos los recursos necesarios que le permitan ejecutar de manera eficiente sus jobs en una infraestructura de computación paralela sobre una Red de Altas Prestaciones. El investigador utiliza parámetros preestablecidos en documentos técnicos o de benchmarks realizados en otros supercomputadores, datos que sirven de referencia para la ejecución de los Jobs de los trabajos de investigación; esto permite mejorar sus ambientes de trabajo sobre cualquier aplicación científica o aplicación HPC. (HPC-AI Advisory Council, 2020)

En este contexto, el investigador o usuario no cuenta con la información acerca del Rendimiento real o del comportamiento de la Red de Altas Prestaciones InfiniBand sobre una arquitectura de computación paralela del Supercomputador Quinde I; para la ejecución eficiente de los Jobs o tareas de cómputo de los proyectos de investigación. Por lo tanto, es necesario tomar en cuenta las siguientes consideraciones, que influyen en la eficiencia de los Jobs:

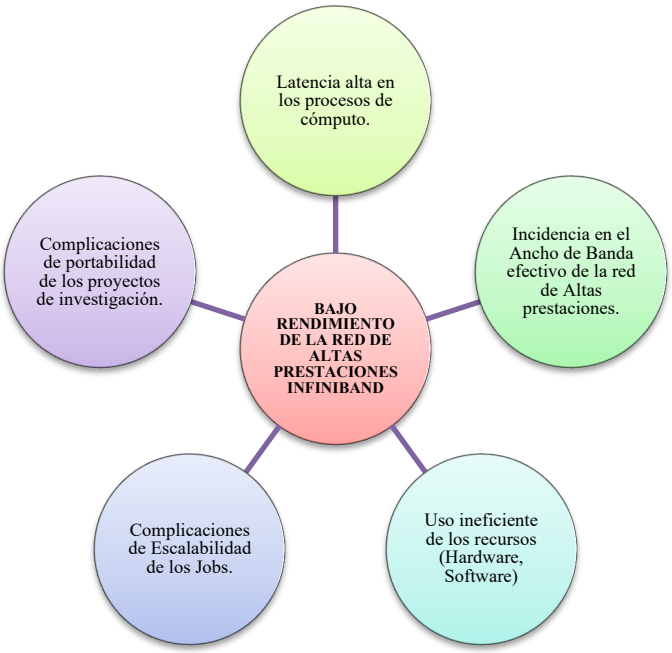
- Escalabilidad: Es un parámetro muy crítico a nivel del sistema, que influye directamente en el rendimiento de la red de altas prestaciones sobre una infraestructura de computación paralela, ya que los Jobs se adaptan a los recursos disponibles que presta el Supercomputador “Quinde I” y presenta resultados con base a lo configurado en cada Job de los proyectos de Investigación.

[6] Job en español trabajo. - es una o varias tareas lógicas de cómputo que se ejecutan en una infraestructura de computación para obtener un resultado específico del algoritmo o programa ejecutado.

- Eficiencia: Para la ejecución de los Jobs, estos requieren de los recursos que proporciona el Supercomputador “Quinde I” en función de la complejidad del proyecto, esto puede representar un mal uso de los recursos de hardware como de software, si estos no se configuran adecuadamente en la ejecución de los Jobs. Esto puede incurrir en un mayor consumo de los recursos y ralentizar los procesos de cómputo, si no se configura con los parámetros específicos basados en la funcionalidad de la red InfiniBand sobre una arquitectura de computación paralela.
- Portabilidad: Es importante para el investigador poder ejecutar sus procesos de cómputo en diferentes entornos, arquitecturas e infraestructuras de supercomputación, que le permita portar su información y adaptarla a cualquier medio con el fin de garantizar la eficiencia de sus procesos de cómputo y por ende obtener resultados precisos.
- Compartición de recursos: Cada usuario investigador realiza un requerimiento inicial de los recursos necesarios para sus trabajos de investigación; sin embargo, es importante considerar este factor al momento de ejecutar los Jobs, que un equipo o gestor externo se encarga de balancear la carga de procesamiento en cada nodo de cómputo, lo cual puede influenciar en el porcentaje de eficiencia de la red de altas prestaciones al momento de correr un Job.

En la Figura 1 se muestra el problema principal y las consecuencias que representa el Bajo Rendimiento de la Red de Altas prestaciones en una infraestructura de Computación paralela.

Figura 1. Árbol de Problemas y consecuencias del trabajo de investigación.



Nota. - Elaborado por el autor

En resumen, el Supercomputador "Quinde I", se ha convertido en una herramienta vital para el desarrollo de la investigación científica del Ecuador, por lo cual es de suma importancia realizar documentos guías, manuales y benchmarks, que permitan medir la capacidad y rendimiento de este tipo de infraestructuras de altas prestaciones, y sobre todo presentar el rendimiento real de la red InfiniBand en comparación a lo datos teóricos y de esta forma tener mejor comprensión de las infraestructuras de Altas Prestaciones, y así alcanzar porcentajes razonables de máximo rendimiento de la Red que afectan directamente en los Jobs.

Formulación del problema

Incidencia en los tiempos de ejecución de los procesos de cómputo corridos sobre una infraestructura de computación paralela utilizando una red de Altas prestaciones del Supercomputador “Quinde I” de la Empresa Pública Siembra E.P. localizado en el cantón de Urcuquí provincia de Imbabura.

Objetivos de la Investigación

Objetivo general

Realizar un análisis de Rendimiento de la Red de Altas Prestaciones en una infraestructura de computación paralela, a través de una aplicación HPC, con el fin de elaborar una guía de buenas prácticas para la ejecución de procesos de cómputo en el Supercomputador Quinde I de la Empresa Pública Siembra E.P.

Objetivos específicos

- Describir las características y funcionalidades principales de una infraestructura de computación paralela y de la Red de Altas prestaciones InfiniBand.
- Describir las características y funcionalidades principales de tres aplicaciones HPC de software libre, con el fin de determinar el software idóneo para el análisis del rendimiento de la Red de Altas prestaciones InfiniBand.
- Analizar y definir los parámetros y variables necesarias para crear los escenarios de prueba del rendimiento de la Red de Altas prestaciones InfiniBand.

- Compilar y Ejecutar la aplicación de cómputo por medio de tareas MPI[7] (Messaging Paralell Interface) para obtener los datos del rendimiento de la Red de Altas prestaciones en computación paralela.
- Recopilar y analizar los datos obtenidos como tiempos de respuesta, ancho de banda y latencia de los jobs o procesos de cómputo ejecutados, por medio de métodos estadísticos.
- Comparar los resultados obtenidos con otros trabajos de investigación en otro supercomputador.
- Validar los datos obtenidos con el Test de linpack[8] entregado por el fabricante de la infraestructura de altas prestaciones.
- Presentar los resultados del análisis de los datos obtenidos a través de dashboards o herramientas de software libre.
- Presentar un caso práctico en una de las investigaciones ya realizadas en el Supercomputador "Quinde I" con el fin de verificar los resultados obtenidos del análisis.

Justificación de la Investigación.

Con base a lo publicado en el sitio web de la Empresa Pública Siembra E.P. *"La supercomputación, High Performance Computing, HPC, computación de altas prestaciones y en los últimos años denominada informática de alto rendimiento o informática de gama alta, es una herramienta estratégica fundamental para el Ecuador, que ha sido implementada por la Empresa Pública Siembra E.P, y que*

[7] MPI. – Es un estándar de interfaz de paso de mensajes portátil, eficiente y flexible utilizado para escribir programas independientes del proveedor.
[8] Test de linpack. – Es una prueba de referencia para medir la eficiencia de sistemas informáticos dedicados esencialmente al cálculo científico y técnico, es decir a procesos de tratamiento numérico.

potenciará el progreso científico, el desarrollo y la innovación industrial, la seguridad nacional y permitirá afrontar los retos sociales del país de mejor manera a través del modelamiento y la simulación computacional." (Empresa Pública Siembra E.P., 2020)

"El Servicio Nacional de Supercomputación de Siembra EP, cuenta con un Modelo de Gestión; el mismo que ha sido estructurado en tres fases, que comprende al final la triple hélice: ACADEMIA, INDUSTRIA Y ESTADO, para fomentar la Gestión del conocimiento técnico, científico y coadyuvar al desarrollo industrial y productivo del país. El Modelo de Gestión cuenta con el aval del SENESCYT y del Comité de Acceso Quinde 1." (Empresa Pública Siembra E.P., 2020).

En el campo de la investigación, la supercomputación tiene una gran relevancia en diferentes áreas de la ciencia, tecnología, física, matemática, salud, etc.; el Supercomputador "Quinde I", actualmente cuenta con varios proyectos en distintas líneas de investigación, que están transformando la academia, industria y estado, por lo tanto; los investigadores requieren contar con datos precisos como guía para la ejecución eficiente de los procesos de cómputo sobre una Red de Altas prestaciones InfiniBand en una infraestructura de computación paralela.

Esta temática se encuentra dentro de la línea de investigación de Innovación tecnológica y productos de Telecomunicación de la Universidad Técnica del Norte, que permitirá coadyuvar al campo de la investigación, academia e industria de la zona norte del país.

Este proyecto permite coadyuvar al Servicio de Supercomputación que brinda la Empresa Pública Siembra E.P. a toda la comunidad académica – científica, aportando al desarrollo del país. Además, permite impulsar uno de los objetivos del

Plan Nacional del Buen Vivir 2017-2021 (Secretaría Nacional de Planificación y Desarrollo, SENPLADES) dentro del Objetivo 5: "*Impulsar la productividad y competitividad para el crecimiento económico sustentable de manera redistributiva y solidaria",* política 5.3 "*Promover la investigación, la formación, la capacitación, el desarrollo y la transferencia tecnológica, la innovación y el emprendimiento, en articulación con las necesidades sociales, para impulsar el cambio de la matriz productiva.".* (Senplades, 2017).

Con base a los antecedentes expuestos, el desarrollo de este trabajo es de vital importancia, ya que permitirá presentar a la comunidad académica – científica, datos precisos del rendimiento de la Red de Altas prestaciones InfiniBand sobre una infraestructura de computación paralela, y una guía de buenas prácticas para la ejecución eficiente de los procesos de cómputo sobre el Supercomputador "Quinde I." Sobre todo, este trabajo permitirá impulsar proyectos de investigación futuros que aporten al desarrollo científico y social del país, convirtiendo a Ecuador un referente en el campo de la investigación a nivel regional.

Pregunta de investigación

¿Cómo incide el tipo de Infraestructura de Computación Paralela en el rendimiento de la Red de Altas prestaciones InfiniBand durante los procesos de cómputo?

Hipótesis

La Red de Altas prestaciones InfiniBand presenta un mejor rendimiento en una Infraestructura de Computación Paralela durante la ejecución de los procesos de cómputo en el Supercomputador Quinde I.

Preguntas directrices

- ¿Qué factores influyen en el rendimiento de la Red de Altas prestaciones InfiniBand?
- ¿Cuáles son las principales ventajas de una Infraestructura de computación Paralela en un Supercomputador?
- ¿Existen investigaciones o trabajos relacionados utilizando infraestructuras de Computación Paralela que permitan brindar un marco de referencia al presente análisis?
- ¿Cuál es el nivel de conocimientos de una Red de Altas prestaciones InfiniBand sobre una Infraestructura de computación paralela?
- ¿Es posible mejorar los tiempos de ejecución de los procesos de cómputo de los proyectos de investigación desarrollados en el Supercomputador “Quinde I”?
- ¿Es posible establecer parámetros referenciales para nuevas investigaciones o trabajos en redes de Altas prestaciones sobre una infraestructura de computación paralela?

Variables e indicadores

Independiente: Infraestructura de Computación Paralela.

Dependiente: Rendimiento de la red de Altas prestaciones InfiniBand.

A continuación, se detalla los indicadores de las variables a ser aplicadas en el presente trabajo de investigación:

Variable Independiente: Infraestructura de Computación Paralela. En la Tabla 1 se muestra los indicadores de la variable Independiente.

Tabla 1. Indicadores de la variable Independiente.

CONCEPTUALIZACIÓN	DIMENSIONES	INDICADORES	ITEMS BÁSICOS	TÉCNICA O INSTRUMENTO
Procesamiento paralelo: Muchos problemas son más fáciles de modelar usando paradigmas paralelos, ya sea por la estructura que se usa para su resolución o porque el problema es intrínsecamente paralelo. Es decir, si desde el principio se puede pensar en los mecanismos paralelos/concurrentes para resolver un problema, eso puede facilitar la implantación del modelo computacional. Esto podría permitir obtener mejores soluciones para los problemas a resolver, en tiempos razonables de ejecución. Así, este enfoque permite el surgimiento de modelos de cálculos diferentes, a los modelos secuenciales. **(Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)**	Paralelismo	Grado de paralelismo	¿Cuál es el grado de paralelismo utilizado?	Análisis de archivos
	Librerías	Factibilidad de librerías	¿Qué tipo de librerías se puede utilizar para el paso de mensajes en una infraestructura de computación paralela?	Análisis de librerías.
	Carga de trabajo	Cantidad de operaciones ejecutadas.	¿ Cuál es la cantidad de operaciones de cómputo ejecutadas en el Supercomputador?	Análisis de archivos.

	Procesadores	Cantidad de procesadores.	¿Cuál es la cantidad de procesadores utilizados?	Análisis de recursos

Nota: Realizado por el Autor

Variable Dependiente: Rendimiento de la Red de Altas prestaciones InfiniBand. En la Tabla 2 se muestra los indicadores de la variable dependiente.

Tabla 2. Indicadores de la variable Dependiente.

CONCEPTUALIZACIÓN	DIMENSIONES	INDICADORES	ITEMS BÁSICOS	TÉCNICA O INSTRUMENTO
Red de Altas Prestaciones InfiniBand. A través del gran avance de tecnologías de redes de interconexión, que cada vez son capaces de proporcionar un ancho de banda y latencia ultra baja para sistemas computacionales mucho más exigentes en el campo de la Supercomputación, se habla de Redes de Altas prestaciones que tienen un mejor rendimiento para grandes procesos de cómputo.	Ancho de Banda	Capacidad de Ancho de Banda	¿Qué capacidad de Ancho de Banda tiene la Red de Altas prestaciones?	Análisis de hoja técnica y monitoreo de Ancho de banda
	Latencia de red	Tiempo que tarda en el envío de un paquete de su origen al destino.	¿Qué tiempo tarda el paquete en transmitirse en una Red de Altas prestaciones?	Análisis de archivos.
En el caso del Supercomputador Quinde I de la Empresa Pública Siembra E.P utiliza una Red de Altas prestaciones con tecnología InfiniBand, que permite la interconexión de los nodos de cómputo y sistemas de almacenamiento tanto de entrada y salida, a mayor capacidad; que facilita el paso de datos y mensajes	Escalabilidad	Grado de Escalabilidad	¿Qué grado de escalabilidad presenta la Red de Altas prestaciones en un Sistema de Supercomputación?	Análisis de archivos.

sin la participación de la CPU con acceso directo a memoria remota (RDMA) y descargas de envío / recepción de mensajes, que son administradas y realizadas por los adaptadores InfiniBand, que proporciona mayor ancho de banda, menor latencia y mejor escalabilidad. **(Infiniband Trade Association, 2014)**

Nota: Elaborado por el Autor.

CAPÍTULO II

MARCO REFERENCIAL

Marco Teórico

Antecedentes Investigativos

Un Supercomputador es un computador con capacidades y características de procesamiento y cálculo muy altas, en comparación a los equipos tradicionales. Este tipo de equipos fue diseñado para procesar grandes cantidades de información en poco tiempo, reduciendo los tiempos de respuesta de un job (Ocampo Yahuarcani & Campos Baca, 2017).

En Ecuador se implementó el primer Supercomputador llamado "Quinde I", el cual está ubicado en la provincia de Imbabura en el cantón Urcuquí, perteneciente a la Empresa Pública Siembra E.P, el cual nació en el año 2014 como parte del Plan maestro para la construcción de la primera Ciudad del Conocimiento "Yachay"; el equipo inicio sus funciones en el mes de noviembre del año 2016. Esta infraestructura se encuentra montada sobre un Data Center en contendedores como se muestra en la Figura 2 la cual cuenta con prestaciones técnicas de un TIER 3[9]. Hasta la presente fecha se han ejecutado alrededor de 30 proyectos de investigación en diferentes áreas como inteligencia artificial, nanotecnología, biomedicina, análisis matemáticos, análisis financieros, etc. (Empresa Pública Siembra E.P., 2020)

[9] Tier 3: Es un centro de datos que cuenta con las especificaciones de un Tier 1 + Tier 2 + Equipos de alimentación eléctrica dual y varios enlaces de salida, garantizando una disponibilidad del 99.982%.

Figura 2. Fotografía del Data Center de la Empresa Pública Siembra E.P., donde se encuentra alojado el Supercomputador Quinde I.

Nota: Recuperado del sitio web https://hpc.yachay.gob.ec

Con base al estudio realizado según (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004), se debe considerar varios aspectos para este tipo de infraestructuras como es el uso de recursos, carga de trabajo, latencia y entre otros aspectos que pueden influir en la degradación del rendimiento del sistema de Supercomputación, al momento de ejecutar los procesos de cómputo.

Siendo el principal problema el factor Rendimiento en una Infraestructura de computación paralela, por lo que muchos investigadores requieren contar con los recursos e información necesaria para la ejecución eficiente de sus procesos de cómputo sobre una Red de Altas prestaciones con tecnología InfiniBand con la que cuenta el Supercomputador Quinde I de la Empresa Pública Siembra E.P.

Mediante el Análisis de Rendimiento de la Red de Altas prestaciones, permitirá presentar a la comunidad académica – científica, datos reales del rendimiento de la Red de Altas prestaciones InfiniBand y una guía de buenas prácticas para la ejecución eficiente de sus procesos de cómputo en una infraestructura de computación paralela con la tecnología que dispone el Supercomputador Quinde I.

Además, este trabajo permitirá impulsar proyectos de investigación futuros que aporten al desarrollo científico del país, convirtiendo a Ecuador un referente en el campo de la investigación a nivel regional.

Fundamentación legal.

A continuación, se resalta los fundamentos legales que sustentan la presente investigación:

Reglamento de Régimen Académico:

"Art. 3.- Objetivos. - Los objetivos del régimen académico son:

Literal d. articular la formación académica y profesional, la investigación científica, tecnológica y social, y la vinculación con la colectividad en un marco de calidad, innovación y pertinencia.

Literal h. Impulsar el conocimiento de carácter multi, inter y trans disciplinarios en la formación de grado y postgrado, la investigación y la vinculación con la colectividad.

Literal j. Desarrollar la educación superior bajo la perspectiva del bien público social, aportando a la democratización del conocimiento para la garantía de derechos y la reducción de inequidades."

Plan de la Sociedad de la Información y el Conocimiento (2018-2021)

Programas del Plan

Proyecto 4. Promoción de uso y apoyo a la formación de profesionales en tecnologías emergentes. El presente proyecto busca impulsar la adopción de tecnologías emergentes que apuntalen el desarrollo de la sociedad de la información y del conocimiento, y difundir las recomendaciones internacionales relacionadas con la implementación de estas nuevas tecnologías. Es por eso que en el presente proyecto se unificará esfuerzos con la academia y se enfocará a trabajar en campañas de difusión de tecnologías emergentes, así como se coordinará con la academia la formación de profesionales en estas temáticas. Con este proyecto se aborda el objetivo de la Política de: "Apoyar al trabajo conjunto entre academia, sector público y privado para la investigación, innovación y transferencia de conocimiento a través de las Líneas de Investigación que tienen una orientación sobre el impacto social y productivo para la mejora de la matriz productiva del país" y "Fomentar el acercamiento entre la oferta y la demanda del sector TIC, a través de eventos de intercambio de experiencias y mejores prácticas".

Acciones para alcanzar el objetivo.

Las acciones a realizar en este proyecto son las siguientes:

- Implementación de campañas de difusión, publicaciones en el Observatorio de las TIC y de la Sociedad de la Información, webinars, seminarios, concursos, etc.
- Reuniones con las autoridades de la academia y el ente rector de la educación superior, con el fin de coordinar la inclusión de "Tecnologías emergentes" como parte de la formación de los futuros profesionales, y de las áreas y líneas de investigación de desarrollo e innovación (I+D+i) en los proyectos auspiciados por estas entidades.

Plan Nacional de Desarrollo 2017-2021

Este proyecto de investigación se encuentra alineado al Objetivo 5, política 5.6 del Plan Nacional de Desarrollo:

Objetivo 5: Impulsar la productividad y competitividad para el crecimiento económico sustentable de manera redistributiva y solidaria

Política 5.3. Promover la investigación, la formación, la capacitación, el desarrollo y la transferencia tecnológica, la innovación y el emprendimiento, la protección de la propiedad intelectual, para impulsar el cambio de la matriz productiva.

Plan Estratégico Institucional del MINTEL

OEI 1: INCREMENTAR LA APROPIACIÓN DE LAS TIC EN LA POBLACIÓN PARA EL DESARROLLO SOCIAL E INCLUSIVO DEL PAÍS

a) Desarrollar competencias digitales en la población

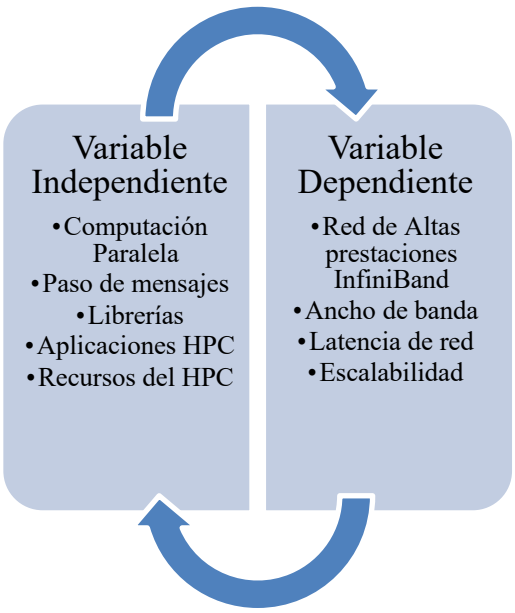
b) Potenciar el acceso y asequibilidad a las TIC.

c) Impulsar el desarrollo de servicios digitales

Esquema del Marco Teórico de la Investigación

A continuación, se muestra en la Figura 3 el esquema del Marco metodológico de la presente investigación:

Figura 3. Esquema del Marco Teórico de la Investigación



Nota: Elaborado por el Autor

Los temas a desarrollar en esta sección son los siguientes:

- Introducción a la Supercomputación o HPC
- Computación Paralela
- Conceptos básicos de supercomputación
- Características del Supercomputador Quinde I.
- Paso de mensajes
- Librería MPI
- Recursos del Supercomputador Quinde I
- Aplicaciones de cómputo
- Red de Altas prestaciones InfiniBand
- Arquitectura de la Red de Altas prestaciones

Introducción a la Supercomputación o HPC

La Supercomputación o High Performance Computing (Computación de Altas Prestaciones) comúnmente es conocida como la práctica de agregar mayor potencia al procesamiento de cómputo, que permite tener mejor rendimiento en los procesos que se ejecuten entre varios computadores, esto con el fin de resolver grandes problemas en cualquier campo de la ciencia, ingeniería o negocios en el menor tiempo posible. (insideHPC, s.f.)

De acuerdo al Dr. Fabián García Nocetti, de la UNAM de México, menciona: "La computación de alto rendimiento (HPC) es el uso de procesamiento paralelo para ejecutar aplicaciones avanzadas de manera eficiente, confiable y rápida. El término se

aplica en especial en sistemas que operan arriba de un Teraflops. El término se usa a veces como sinónimo de super cómputo.” (García Nocetti, 2014)

Los recursos que se utiliza dentro de la Supercomputación está el uso de procesadores con sus núcleos (CPU), aceleradores gráficos (GPU), memorias, bus de comunicaciones de alta velocidad, sistemas operativos, unidad de almacenamiento y las aplicaciones informáticas o aplicaciones de cómputo para una arquitectura de computación paralela o serial. (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

Computación paralela

El procesamiento paralelo permite ejecutar varios procesos al mismo tiempo de forma concurrente. Existe varios tipos de procesamiento paralelo como:

- El que ejecuta procesos independientes simultáneamente, los cuales son controlados por el sistema operativo (usando tiempo compartido, multiprogramación y multiprocesamiento).
- El que descompone los programas en tareas (controladas por el sistema operativo, los compiladores, los lenguajes de programación, etc.), algunas de las cuales pueden ser ejecutadas en paralelo.
- El que se basa en usar técnicas de encauzamiento para introducir paralelismo a nivel de instrucciones, lo que implica dividirlas en pasos sucesivos que pueden ser ejecutados en paralelo, cada uno procesando datos diferentes.

Con base a las necesidades que se presentan en el campo de la ciencia e ingeniería, la computación paralela ha estado en constantes cambios y sujetos a varios factores que han permitido la evolución y transformación de las infraestructuras de Supercomputación, que son importantes mencionar a continuación: (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

- Requerimiento de mayor potencia de cálculo, que muchas veces se ve limitado por el tipo de tecnología con el que cuenta la infraestructura. Hoy en día el avance tecnológico ha permitido usar múltiples procesadores juntos para ejecutar varias tareas de manera distribuida y de esta forma aumentar la potencia de cálculo.
- Mejor relación de costo/rendimiento, siendo este un factor muy importante a nivel económico, ya que muchas veces el investigador se ve limitado por los costos que representa la capacidad de cálculo que requiere los trabajos de investigación, por lo que es importante segmentar adecuadamente los elementos de cálculo de acuerdo al requerimiento de potencia de procesamiento.
- Potencia expresiva de los modelos de procesamiento paralelo: este factor se refiere al uso de mecanismos paralelos/concurrentes para resolver un problema, que desde el inicio el investigador implemente un modelo computacional diferente a los convencionales de manera secuencial, y de esta forma mejorar los tiempos de ejecución de los procesos de cálculo.

La computación paralela es bastante utilizada a nivel mundial en el avance y desarrollo de trabajos de investigación de diferentes campos como la bioquímica,

inteligencia artificial, finanzas, biotecnología, física, matemáticas, medicina, etc.; que ha exigido mayor procesamiento computacional y por ende disponer de redes de comunicación que soporten una gran transferencia de información de manera eficiente y rápida, llamadas redes de Altas prestaciones (Ocampo Yahuarcani & Campos Baca, 2017).

Conceptos básicos de Supercomputación

A continuación, se presenta varios conceptos básicos que comúnmente son utilizados en el campo de la Supercomputación:

CPU: "Unidad Central de Procesamiento", o procesador, es la Unidad que permite el procesamiento de forma secuencial de las operaciones de un computador. Es un componente electrónico, en el que se realiza los procesos lógicos, el cual dispone núcleos para la ejecución de tareas (Ocampo Yahuarcani & Campos Baca, 2017).

Flops.- Del acrónimo de floating point operations per second, se traduce como "operaciones o cálculos de coma flotante por segundo", el cual es una medida de rendimiento de las computadoras expresado en un número entero multiplicado por un exponente, que permite determinar la velocidad de un computador para el procesamiento de datos. La fórmula para calcular los flops de un procesador (CPU): 1 Gigaflops = (Velocidad de la CPU en GHz) x (Nº núcleos) x (Instrucciones por ciclo), donde GHz es gigahercio. (Ocampo Yahuarcani & Campos Baca, 2017)

Algoritmos: Es un conjunto de operaciones realizadas de forma sistemática a través de un proceso definido, que en el campo de la supercomputación este es ejecutado con el fin de obtener un resultado específico (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004).

Job o tarea de cómputo.- En español trabajo, es una o varias tareas lógicas de cómputo que se ejecutan en una infraestructura de computación para obtener un resultado específico del algoritmo o programa ejecutado. (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

Aplicación HCP o aplicación paralela: Es aquella aplicación que permite ejecutar una o más jobs sobre un sistema de cómputo de Altas prestaciones, con el fin de obtener un resultado o resolver algún problema específico. (Jorba Esteve, 2006)

GPU: "Unidad de Procesamiento Gráfico", es la Unidad que permite la aceleración de los procesos gráficos u operaciones de coma flotante, con el fin de coadyuvar con la carga de trabajo del procesador central (Ocampo Yahuarcani & Campos Baca, 2017).

Procesador: Es aquel que puede ser uno o más dentro de un equipo de cómputo, que permiten leer las instrucciones y ejecutar las tareas al mismo tiempo o de forma distribuida (Ocampo Yahuarcani & Campos Baca, 2017).

Benchmark: Es una prueba de rendimiento o de evaluación comparativa de un equipo computacional, con el fin de medir la capacidad real o el rendimiento de

uno de sus componentes o de toda la infraestructura de Altas prestaciones. (EcuRed, s.f.)

HPL: Es la implementación portable del High-Performance Linpack Benchmark, el cual permite medir la potencia de cálculo de punto flotante de una infraestructura de Altas Prestaciones de memoria distribuida con el fin de proporcionar datos en el TOP500 de Supercomputador a nivel mundial. (top500, 2020)

Test de Linpack o Linpack Benchmark: Es parte del HPL, el cual permite medir la tasa de ejecución de punto flotante de una infraestructura de Altas Prestaciones. Se determina ejecutando un programa de computadora que resuelve un sistema denso de ecuaciones lineales. (top500, 2020)

Centro de Datos o Data Center: Es un conjunto de elementos o componentes que permiten el normal funcionamiento y las condiciones adecuadas para mantener una infraestructura HPC, entre estos elementos se destacan: (Ocampo Yahuarcani & Campos Baca, 2017)

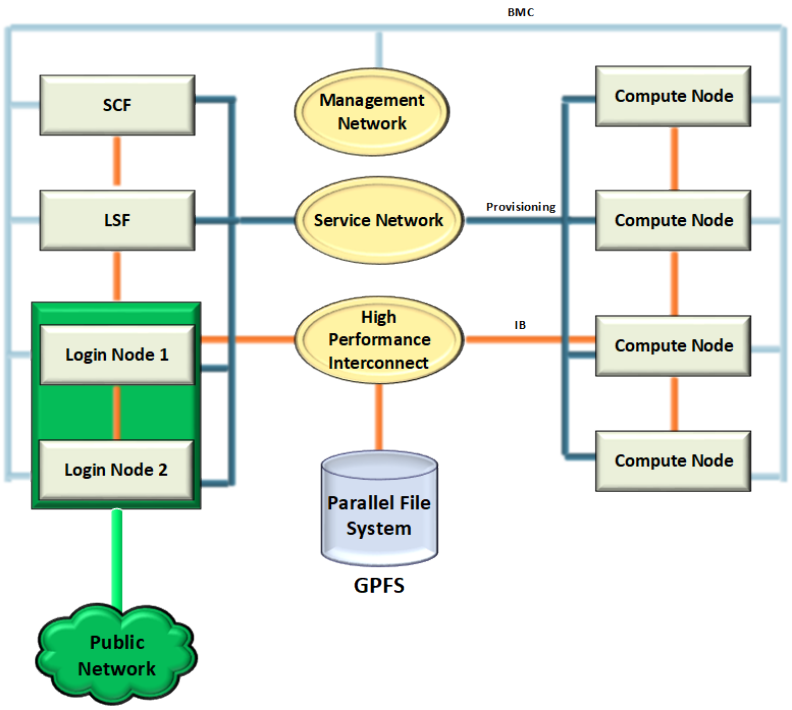
- Servidores de grandes capacidades de procesamiento.
- Equipos de almacenamiento.
- Sistema de refrigeración.
- Sistema eléctrico para acceso y respaldo de energía.
- Sistema eléctrico de soporte.
- Software de monitoreo y gestión.

- Personal encargado de la gestión y acceso.

Arquitectura General del Supercomputador Quinde I

En la figura 4 se muestra la arquitectura general del Supercomputador Quinde I de la Empresa Pública Siembra E.P.:

Figura 4. Arquitectura del Supercomputador Quinde I.



Nota: Tomado de la página web de la (Empresa Pública Siembra E.P., 2020)

El Supercomputador Quinde I dispone de una arquitectura compuesta por varios componentes o elementos que se describen a continuación, basado en la figura 4:

Compute Node o Nodo de cómputo: es un equipo de cómputo que forma parte del clúster HPC, que tiene características avanzadas para el procesamiento de datos (Empresa Pública Siembra E.P., 2020).

Login node: son equipos que permiten el acceso al usuario científico a la consola de línea de comandos del Supercomputador Quinde I. Cuenta con dos login node, con el fin de balancear la cantidad de accesos de los usuarios. (Empresa Pública Siembra E.P., 2020)

LSF: Plataforma LSF (Load Sharing Facility) es un conjunto de productos de administración de recursos distribuidos que permite: (Empresa Pública Siembra E.P., 2020)

- Conectar los servidores dentro de un clúster (or "Grid")
- Monitorear la carga de los sistemas.
- Distribuir la agenda y balancear la carga de trabajo.
- Controlar el acceso y carga basado en políticas.
- Analizar la carga de trabajo.
- Proveer un acceso transparente a todos los recursos disponibles (procesadores, aplicaciones, etc.)

SCF: Spectrum Cluster Fundation, es el equipo que administra los clústeres de cómputo. Permite administrar todos los equipos de cómputo como uno solo, automatizar la implementación del sistema operativo y los componentes de software, así como el aprovisionamiento y el mantenimiento de los equipos. Además,

proporciona el monitoreo centralizado con alertas y acciones personalizables de toda la infraestructura de supercomputación. (Empresa Pública Siembra E.P., 2020)

NSD: Network Shared Disk, es un protocolo creado por IBM Spectrum Scale que permite implementar un interfaz de nivel de bloque sobre la red que es llamado NSD, evitando la complejidad de configuración de una red SAN. Los equipos NSD brindar acceso a los datos del usuario que están conectados a la red SAN. (IBM, 2020)

GPFS: General Parallel File System es un sistema de ficheros distribuido de alto rendimiento desarrollado por IBM, el cual permite el acceso a alta velocidad a las aplicaciones de cómputo que pueden estar ubicadas en los nodos de cómputo y pueden verse como archivos compartidos por grupos de proyectos de investigación. (Wikipedia, 2019)

Public Network: es la red pública por la que acceden los usuarios o investigadores a la interfaz del Supercomputador, a través de una IP pública a un determinado puerto por conexión SSH. (Empresa Pública Siembra E.P., 2020)

IB o Red InfiniBand: es el bus de comunicaciones de Alta velocidad, baja latencia y de baja sobrecarga de CPU, o también denominado Red de Altas Prestaciones, que permite la interconexión de los nodos de cómputo del Supercomputador Quinde I a una velocidad de 100Gbps. (Infiniband Trade Association, 2014)

Red BMC: es una red de comunicaciones que permite realizar la administración de los nodos de cómputo y nodos de gestión, con una velocidad de 1Gpbs. (Empresa Pública Siembra E.P., 2020).

Red de provisioning o aprovisionamiento: permite realizar el aprovisionamiento de los nodos de cómputo que pertenecen al clúster de supercomputación a una velocidad de 10Gbps. (Empresa Pública Siembra E.P., 2020).

Con base a los datos obtenidos de los informes técnicos de la Empresa Pública Siembra E.P., en la Tabla 3, se detalla un resumen de las características técnicas principales del Supercomputador Quinde I:

Tabla 3. Características del Supercomputador Quinde I

CARACTERÍSTICAS DEL SUPERCOMPUTADOR "QUINDE I"	
FABRICANTE	IBM (INTERNATIONAL BUSINESS MACHINES)
ARQUITECTURA	MIMD NUMA (Multiple Instruction Multiple data, Non-Uniform Memory Access)
NODOS DE ADMINISTRACIÓN	7 (2 Login nodes, 1 LSF, 1 SCF, 3 NSD)
NODOS DE CÓMPUTO	84
RMAX (TEST DE LINPACK - HPL)	231,9 TFLOPS
RPEAK	488,9 TFLOPS
CPU CORES	1640
GPU-CORES POR NODO	9984 cuda cores
SISTEMA DE ARCHIVOS	GPFS
CAPACIDAD DE ALMACENAMIENTO	350 TB
MEMORIA RAM	10,5 TB
RED DE ALTAS PRESTACIONES	MELLANOX INFINIBAND EDR 100 Gbps

RED DE PROVISIONAMIENTO	10 Gbps
RED DE ADMINISTRACIÓN	BMC (1Gbps)
ARQUITECTURA DEL PROCESADOR (NODOS DE CÓMPUTO)	Nodos S822LC con procesadores Power 8 RISC (Reduced instruction set computer)[10]
DATOS TÉCNICOS DE LOS NODOS DE CÓMPUTO	(20) procesadoress físicos Power 8 de 3.4 GHz, 512KB Cache L2, 96 Cache L3 compartida entre procesadoress, veinte (20) procesadoress activos. (128) GB de memoria de RAM física, (128) GB de memoria activa (2) tarjetas NVIDIA KEPPLER GPU K80 (9984 GPU cuda cores, 24 GB DDR5) (1) Tarjeta de Red 10 Gbps de 2 puertos (1) Tarjeta de Red 1 Gbps de 4 puertos (1) Tarjeta de Red InfiniBand EDR de 2 puertos habilitada para CAPI (2) discos duros de 960 GB SSD SAS SFF (02) Fuentes de poder 1300W (4) ventiladores cooling fan configurados en redundancia N + 1

Nota: Tomado de la página web de la (Empresa Pública Siembra E.P., 2020)

Arquitectura Paralela: MIMD NUMA (Multiple Instruction Multiple data, Non-Uniform Memory Access)

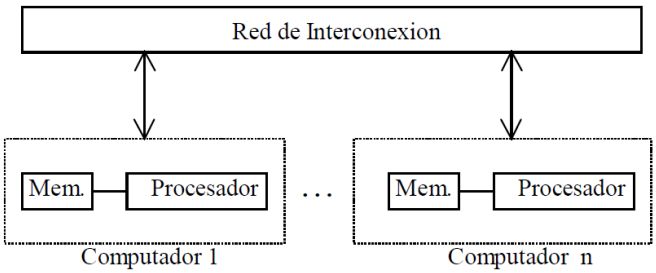
El Supercomputador Quinde I, cuenta con una arquitectura MIMD NUMA (Multiple Instruction Multiple data, Non-Uniform Memory Access), en el que múltiples tareas pueden ejecutarse al mismo tiempo utilizando una memoria distribuida de forma asíncrona, en el que los procesadores se comunican por pase de mensajes, donde se puede resaltar que el rendimiento de las máquinas está muy ligado al rendimiento de las comunicaciones de las máquinas y el paralelismo a utilizar debe ser totalmente claro por parte de los investigadores o usuarios del servicio de

[10] RISC: El objetivo de diseñar máquinas con esta arquitectura es posibilitar la segmentación y el paralelismo en la ejecución de instrucciones y reducir los accesos a memoria.

Supercomputación de la Empresa Pública Siembra E.P. (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

Esta arquitectura permite un control distribuido de los datos, se replica la memoria, procesadores y los controladores. La red de interconexión es utilizada para que los procesadores se comuniquen a través de pase de mensajes, en el que los programas se dividen en tareas y son ejecutadas de forma concurrente en espacios de memorias locales como se muestra en la Figura 5.

Figura 5. Red de Interconexión para pase de mensajes



Nota: Tomado de (Jorba Esteve, 2006)

Cada procesador ejecuta un conjunto separado de instrucciones sobre sus propios datos locales, la memoria está distribuida entre los procesadores del sistema, donde posee su propio programa y los datos asociados. Una red de interconexión conecta los procesadores (y sus memorias locales), mediante enlaces (links) de comunicación, usados para el intercambio de mensajes entre los procesadores. (Jorba Esteve, 2006)

Por lo tanto, la red de interconexión se convierte en el elemento principal de la arquitectura para un buen rendimiento de todo el sistema. Los nodos de cómputo del Supercomputador Quinde I están conectados a través de una red de Altas

prestaciones llamada InfiniBand a una velocidad de 100Gbps, que permite que las Entrada/Salidas de los nodos estén conectadas directamente a través de la red, presentando la ventaja de crecimiento y flexibilidad de toda la infraestructura. Una desventaja que presenta este tipo de arquitectura es el ineficiente uso de la memoria y los problemas de sincronización. (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

Generalidades del Paralelismo

Es importante resaltar ciertos aspectos del Paralelismo, que permita una mejor comprensión del presente trabajo de investigación, y, sobre todo permita realizar el análisis del comportamiento de la Red de Altas Prestaciones sobre una infraestructura de computación paralela con la que está compuesta el Supercomputador Quinde I de la Empresa Pública Siembra E.P.

Metodología de Programación

Existe una gran variedad de metodologías de programación, que no están definida exactamente para aplicaciones paralelas. El Supercomputador Quinde I cuenta con una arquitectura de memoria distribuida, por lo tanto, como parte de la presente investigación se describirá uno de los métodos más comunes para este tipo de arquitecturas como es la Interfaz de Paso de Mensajes o MPI (Message Passing Interface).

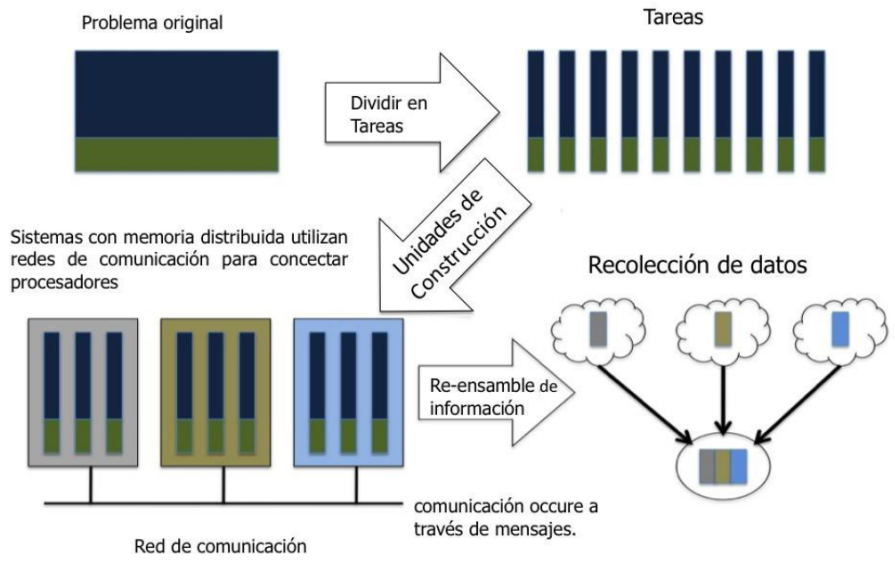
Paso de mensajes

El Paso de mensajes es un modelo muy utilizado en la programación paralela, que permite el intercambio de mensajes transmitidos a través de la red de interconexión, con el fin de ejecutar tareas de forma sincronizada entre varios nodos de cómputo.

Una de las ventajas del paso de mensajes es que permite una conexión directa con la red de comunicaciones interna o red cableada. En el caso del Supercomputador Quinde I, esta infraestructura dispone de una red de interconexión llamada Red InfiniBand entre los nodos de cómputo, lo cual le brinda versatilidad, ya que le permite ser un modelo más flexible para adaptar el hardware y obtener mejores prestaciones. Además, este modelo es muy útil y completo para expresar algoritmos paralelos y presenta mejores prestaciones al programador para expresar la asociación de datos con procesos y de esta forma mantener un control de la localidad, y permitiendo a compiladores, y caches optimizar su rendimiento. (Jorba Esteve, 2006)

Es significativo resaltar que el Supercomputador Quinde I dispone de una arquitectura con memoria distribuida, por lo tanto, la metodología de paso de mensajes permite agrupar tareas independientes de forma concurrente a través de diversos procesadores conectados entre sí, siendo de vital importancia la sincronización y comunicación entre ellos durante el procesamiento paralelo como se muestra en la . figura 6. (Acosta Berlinghieri, 2009)

Figura 6. División de tareas a través del paso de mensajes.



Nota: Recuperado de (Acosta Berlinghieri, 2009)

Uno de los principales problemas de este modelo es que toda la responsabilidad recae sobre el programador para el manejo y control para la distribución de los datos. Por tal razón, el programador debe hacer uso de librerías que le permitan implementar este modelo de programación, en el que usa una Interfaz de Programación de Aplicación o llamada API[11] por sus siglas en inglés. (Red Hat, s.f.)

Entre las API que soportan el modelo de Paso de mensajes está MPI que es una de las más utilizadas en ambientes de computación paralela.

MPI de sus siglas en inglés Message Passing Interface o Interfaz de paso de mensajes, es un estándar que permite la implementación de librerías de paso de mensajes, siendo el principal objetivo el permitir la interacción entre la eficiencia y

[11] Una API es un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones. API significa interfaz de programación de aplicaciones.

portabilidad, que proporciona una librería de funciones para C, C++ o Fortran que son empleadas para comunicar datos entre procesos. MPI es portable, fue diseñada y optimizada principalmente para trabajar sobre arquitecturas de memoria distribuida. A partir de esta especificación se desarrollan librerías de software libre como lo es OpenMPI, la cual será sujeto de estudio en el presente informe de investigación. (Jorba Esteve, 2006)

OpenMPI (Message Passing Interface)

Es una API de código abierto utilizada para programación paralela y/o distribuida, que se basa en el estándar MPI, la cual presenta las principales características (open-mpi, 2020):

- Está conforme al estándar MPI 3.1.
- Seguridad de hilos y concurrencia.
- Permite la distribución de procesos de forma dinámica.
- Alto rendimiento en todas las plataformas.
- Tolerancia a fallos: capacidad de recuperarse de forma transparente de los fallos de los componentes (errores en el envío o recepción de mensajes, fallo de un procesador o nodo).
- Soporta redes heterogéneas: permite la ejecución de programas en redes cuyos ordenadores presenten distinto número de nodos y de procesadores.
- Es una única librería que es compatible con todas las redes.
- Sistemas compatibles de 32 y 64 bits.
- Portable: funciona en los sistemas operativos Linux, OS-X, Solaris.

- Modificable por los instaladores y usuarios finales: presenta opciones de configuración durante la instalación de la API, la compilación de programas y su ejecución.
- Licencia de código abierto basada en licencia BSD.

Open MPI es una de las API más utilizadas y la más adecuada para aplicar el modelo de Paso de mensajes en infraestructuras de computación paralela. Además, permite mayor control de las tareas y mayor eficiencia en la realización de operaciones complejas. (open-mpi, 2020)

Desafío HPC o HPC Challenge

El HPC Challenge consiste en siete retos o benchmarks que permiten medir o evaluar el rendimiento de una infraestructura de Altas Prestaciones en diferentes aspectos del mismo: (ICL UT, n.d.)

- **HPL. -** High Performance Linpack, es el test más usado en sistemas científicos y de ingeniería, el cual permite medir el rango a nivel de punto flotante la ejecución para resolver sistemas de ecuaciones lineales complejos, y de esta forma determinar la eficiencia de los procesadores. Estos valores que refleja el test son utilizados para establecer el Top 500 de las infraestructuras HPC a nivel mundial.
- **DGEMM. -** Mide la tasa de ejecución en punto flotante de la multiplicación de matrices reales de doble precisión.

- **STREAM. -** es un programa para realizar un benchmark simple sintético, que mide el ancho de banda (GB/s) de la mqemoria y la tasa correspondiente de cómputo para un kernel vectorial simple.
- **PTRANS (parallel matrix transpose). -** Ejercita las comunicaciones donde pares de procesadores se comunican entre sí simultáneamente. Es una prueba muy útil para la capacidad total de las comunicaciones de la red.
- **RandomAccess. -** mide la tasa de las actualizaciones de la memoria randómica integrada (GUPS).
- **FFT. -** mide la tasa de punto flotante de la ejecución de la Transformada de Fourier Discreta compleja unidimensional de doble precisión.
- **Ancho de banda de Comunicación y latencia. -** es un test que permite medir la latencia y el ancho de banda de un número de patrones de comunicación simultáneos; basado en b_eff (effective bandwidth benchmark).

Con base a los retos propuestos por "HPC Challenge", se considera el reto "**Ancho de banda de Comunicación y latencia**" como parte del objetivo del presente estudio, el cual permitirá realizar el Análisis de rendimiento de la Red de Altas Prestaciones.

Aplicaciones HPC para análisis de rendimiento de la red de Altas prestaciones.

Como parte del presente estudio se seleccionó tres aplicaciones de las más conocidas para realizar el Análisis de rendimiento de redes de altas prestaciones sobre una infraestructura de computación paralela, de esta forma determinar una

herramienta idónea para este tipo de análisis que se ajusten a la arquitectura del Supercomputador Quinde I.

A continuación, se realiza una descripción general de cada una de las aplicaciones seleccionadas:

Iperf: es una herramienta para mediciones activas del ancho de banda máximo en redes IP, que permite testear el ancho de banda, pérdidas y otros parámetros. (iperf, 2020)

- Permite ajustar varios parámetros relacionados con la sincronización, buffers y protocolos (TCP, UDP, SCTP con IPv4 e IPv6).
- Iperf fue desarrollado originalmente por NLANR / DAST.
- Fue liberado para licencia BSD.
- Soporta varias plataformas como: Linux, Windows. MacOSX, OpenBSD, NetBSD, VxWorks, Solaris, etc.
- Cliente y servidor pueden tener múltiples conexiones simultáneas (opción -P).
- Corre en el servidor como un demonio (opción -D).

TAU (Tuning and Analysis Utilities) (Universidad de Oregon, 2020)**:** es un kit de herramientas portátil de creación de perfiles y seguimiento para el análisis del rendimiento de programas paralelos escritos en Fortran, C, C ++, UPC, Java, Python.

- Es una herramienta para afinación y análisis.

- Permite recopilar información de rendimiento a través de la instrumentación de funciones, métodos, bloques básicos y declaraciones, así como un muestreo basado en eventos.
- Desarrollado por la Universidad de Oregon.
- Herramienta de Análisis de rendimiento Escalable y flexible.
- Instrumentación automática a través de la base de datos del programa.
- Puede generar trazas de eventos que se pueden mostrar con las herramientas de visualización de trazas Vampir, Paraver o JumpShot.

Effective Bandwith (B_eff) (Rolf Rabenseifner, 2020)**:** El ancho de banda efectivo b_eff mide el ancho de banda acumulado de la red de comunicación de un sistema informático paralelo y / o distribuido.

- Utiliza varios tamaños de mensajes, patrones de comunicación y métodos.
- El algoritmo usa un promedio para tener en cuenta que en aplicaciones reales los mensajes cortos y largos se transfieren en diferentes valores de ancho de banda.
- Utiliza varios métodos de programación para medir el ancho de banda efectivo independientemente de los métodos MPI que estén optimizados en una plataforma determinada.

Comparación de aplicaciones para el Análisis de Rendimiento

En la tabla 4 se muestra una comparación de las aplicaciones antes descritas con el fin de determinar la más idonea para realizar el Análisis de Rendimiento de la Red de Altas prestaciones del Supercomputador Quinde I:

Tabla 4. Aplicaciones para el Análisis de rendimiento de la Red de Altas prestaciones.

COMPARACIÓN DE APALICACIONES DE TES DE RENDIMIENTO DE REDES DE ALTAS PRESTACIONES			
CARACTERÍSTICA	**IPERF**	**TAU**	**B_EFF**
FUNCIONAMIENTO	Crea flujos de datos TCP y UDP y medir el rendimiento de la red.	TAU (Tuning and Analysis Utilities) es capaz de recopilar información de rendimiento a través de la instrumentación de funciones, métodos, bloques básicos y declaraciones, así como un muestreo basado en eventos.	Mide el ancho de banda efectivo donde el número de procesos MPI multiplicado por el ancho de banda asintótico multiplicado por la proporción del área bajo la curva "ancho de banda sobre longitudes de mensaje" y el área bajo la curva de ancho de banda asintótico constante en el mismo diagrama.
Funcionamiento en sistemas de computación parela y/o distribuida	IPerf mide el paso de mensajes, pero unidireccional.	Mide el paso de mensajes de forma paralela.	Mide el paso de mensajes en arquitecturas paralelas.
Soporta MPI	No	Si	Si
LENGUAJES QUE SOPORTA	Windows, Linux, Android, MacOS X, FreeBSD, OpenBSD, NetBSD, VxWorks, Solaris, etc.	C++ language	Linux, Unix
Arquitectura RISC	Es compleja la aplicación en Arquitecturas RISC.	No optimizado para este tipo de arquitecturas	Optimizado para este tipo de arquitecturas
Software libre	Licencia BSD	Si	Si
Usado en Benchmark de Infraestructuras de Altas Prestaciones	No es muy utilizado.	No es muy utilizado	Es parte de uno de los Benchmarks del HPC Challenge

Nota: Elaborado por el Autor.

Con base a las mejores prácticas detalladas en la página del Consejo Asesor de HPC o “HPC-AI Advisory Council”, la cual se encarga de potenciar el uso de sistemas

HPC y sistemas de Inteligencia Artificial en el campo de la investigación a nivel mundial; además de presentar a los usuarios herramientas necesarias para el desarrollo de la computación paralela (HPC-AI Advisory Council, 2020), y la comparación realizada en la tabla 4. se ha seleccionado la aplicación "b_eff", la cual permitirá determinar el comportamiento de la Red de Altas prestaciones InfiniBand sobre una infraestructura de computación paralela, a través del paso de mensajes sobre memoria distribuida.

Descripción de B_eff

Es una aplicación de software libre que permite medir el ancho de banda efectivo de la red de comunicación de un sistema informático paralelo y / o distribuido, que utiliza diferentes tamaños de mensajes y métodos de comunicación para validar el comportamiento de la Red de Altas prestaciones (Gerrit Schulz , 2020).

El algoritmo utiliza una media para tener en cuenta que los mensajes cortos y largos se transfieren con diferentes valores de ancho de banda en aplicaciones reales. La prueba genera varias tablas de salida, una de ellas es el anillo aleatorio. (HPC-AI Advisory Council, 2020)

Random Ring Bandwidth. - crea un anillo de comunicación aleatorio, el rango i se comunica con el rango j y el rango k (selecciona el rango aleatorio j, k).

Características (Gerrit Schulz , 2020)**:**

- Versión 3.6 + bugfix 3.6.0.1

- Desarrollada en lenguaje C++ y paralelizada con la librería MPI.
- El algoritmo usa un promedio para tener en cuenta que en aplicaciones reales los mensajes cortos y largos resultan en diferentes valores de ancho de banda.
- El promedio de todos los patrones cartesianos y el promedio de todos los patrones aleatorios se calculan en la escala logarítmica.
- Utiliza varios métodos de programación para medir el ancho de banda efectivo independientemente que los métodos MPI estén optimizados en una plataforma determinada. Los métodos usados son: MPI_Sendrecv, MPI_Alltoallv y non-blocking MPI_Irecv and MPI_Isend with MPI_Waitall.

Definición del ancho de banda efectivo (b_eff) (Gerrit Schulz , 2020)**:**

Ecuación 1. Fórmula del ancho de banda efectivo.

$$b_{eff} = \text{logavg}\Big(\text{logavg}_{\text{ring patterns}}\ \big(\text{sum}_L\ (\max_{mthd}\ (\max_{rep}\ (\ b(\text{ring pat.},L,mthd,rep)\ ))\big)/21\ ),\ \text{logavg}_{\text{random patterns}}\big(\text{sum}_L\ (\max_{mthd}\ (\max_{rep}\ (\ b(\text{random pat.},L,mthd,rep)\ ))\big)/21\ )\Big)$$

Nota: Tomado de (Gerrit Schulz , 2020)

Con:

- b (pat,L,mthd,rep) = L * (total de número de mensajes de un patrón "pat") * looplength / (tiempo máximo en cada proceso para ejecutar el patrón de comunicación de tiempo looplength)
- Cada medición se repite en 3 tiempos (rep=1...3). Se utiliza el ancho de banda máximo de todas las repeticiones (mirar maxmthd en la fórmula anterior).

- Cada patrón es programado con tres métodos. Es usado el ancho de banda máximo para todos los métodos (maxmthd).
- La medición se realiza para diferentes tamaños de mensaje: La longitud del mensaje L tiene los siguientes 21 valores:

  con L = 1B, 2B, 4B, ... 2kB, 4kB, 4kB*(a**1), 4kB*(a**2), ... 4kB*(a**8) y 4kB*(a**8) = Lmax and Lmax = (memoria por procesador) / 128
- El looplength es dinámicamente reducido para alcanzar el tiempo de ejecución de cada loop entre 2.5 y 5msec.
  El looplength para la primera iteración es calculada con algún PRE_MSG_LOOPS. El mínimo looplength es 1.
  Se calcula el promedio del ancho de banda de todos los tamaños de mensajes (sumL(...)/21).
- Se utiliza un conjunto de patrones ring y patrones aleatorios.
- El promedio de todos los patrones ring y el promedio de todos los patrones aleatorios se calcula en escala logarítmica (patrones logavgring y patrones logavgrandom).
- Finalmente, el ancho de banda efectivo es el promedio logarítmico de estos dos valores (logavg(logavgring patterns, logavgrandom patterns).

Detalles del algoritmo (Gerrit Schulz , 2020)**:**

La comunicación se programa con varios métodos de programación. Esto permite medir el ancho de banda efectivo independientemente de qué métodos MPI

estén optimizados en una plataforma determinada. Se utiliza el ancho de banda máximo a través de los siguientes métodos:

- MPI_Sendrecv: Envía y recibe un mensaje
- MPI_Alltoallv: Recopila datos de todos los miembros de un grupo y los dispersa. Cada proceso envía datos distintos a cada uno de los receptores.
- non-blocking MPI_Irecv: Inicia una operación de recepción y devuelve un identificador a la operación de comunicación solicitada.
- MPI_Isend con MPI_Waitall: Inicia una operación de envío de modo estándar y devuelve un identificador a la operación de comunicación solicitada. MPI_Waitall completa varias operaciones pendientes.

Para producir una medición equilibrada en cualquier topología de red, b_eff utiliza diferentes patrones de comunicación.

La comunicación se implementa sobre la interfaz MPI y para cada medición se utiliza el ancho de banda máximo. El ancho de banda se traza sobre la longitud del mensaje y los valores de longitud del mensaje utilizados se trazan equidistantes en la abscisa, es decir, a lo largo de dos escalas logarítmicas, una de 1 byte a 4 kbyte (12 intervalos) y la siguiente de 4 kbyte a L (8 intervalos). (Gerrit Schulz , 2020)

La latencia / ancho de banda mide la latencia (tiempo necesario para enviar un mensaje de 8 bytes de un nodo a otro) y el ancho de banda (tamaño del mensaje dividido por el tiempo que se tarda en transmitir un mensaje de 2.000.000 bytes) de la

comunicación de red utilizando rutinas MPI básicas. La medición se realiza durante la comunicación no simultánea (ping-pong benchmark) y simultánea (patrón random y natural ring) y, por lo tanto, cubre dos niveles extremos de contención (sin contención y contención causada por cada proceso que se comunica con un vecino elegido al azar en paralelo) que podría ocurrir en una aplicación real. (icl, s.f.)

El benchmark b_eff, la latencia y el ancho de banda se miden principalmente con tres patrones de comunicación (ping-pong, anillo aleatorio, anillo natural) y dos tamaños de mensaje (8 bytes para latencia y 2,000,000 bytes para mediciones de ancho de banda) y estos diferentes resultados se reportan de forma independiente. La memoria intermedia siempre se reutiliza en un ciclo de mediciones. El objetivo de b_eff es calcular un valor de ancho de banda promedio que representa varios patrones de anillo (ordenados de forma secuencial y aleatoria) y 21 tamaños de mensajes diferentes. Se prohíbe la reutilización de la memoria.

Se ejecuta en dos procesos. Desde el proceso del cliente, se envía un mensaje (ping) al proceso del servidor y luego se devuelve al cliente (pong). Se utiliza el envío y la recepción de bloqueo estándar de MPI. Los patrones de ping-pong se hacen en un bucle. Para lograr el tiempo de comunicación de un mensaje, el tiempo total de comunicación se mide en el proceso del cliente y se divide por el doble de la longitud del bucle. El punto de referencia en hpcc utiliza mensajes de 8 bytes y una longitud de bucle = 8 para comparar la latencia de la comunicación. El punto de referencia se repite 5 veces y se informa la latencia más corta. Para medir el ancho de banda de comunicación, se repiten dos veces 2.000.000 de mensajes de bytes con una longitud de bucle 1. (icl, s.f.)

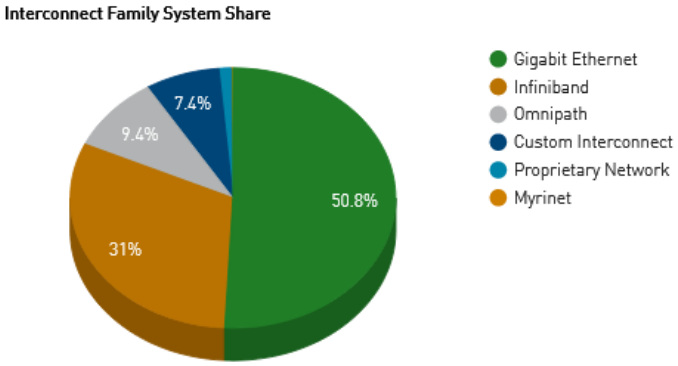
Top de Supercomputadores más potentes a nivel mundial.

La tabla TOP500 muestra los 500 sistemas de cómputo más potentes a nivel mundial. La página muestra información varios datos de cada sistema de cómputo como (top500, 2020):

- Nworld. - Posición dentro del ranking del TOP500.
- Manufacturer. - Fabricante o vendedor.
- Computer .- Tipo indicado por el fabricante o vendedor.
- Installation Site - Cliente
- Location – Ubicación y país
- Year – Año de instalación / última actualización importante.
- Field of Application. - Campo de aplicación.
- #Proc.- Numero de procesadores
- Rmax.- Rendimiento máximo de LINPACK alcanzado.
- Rpeak.- Rendimiento máximo teórico.
- Nmax.- Tamaño del problema para alcanzar Rmax
- N1/2.- Tamaño del problema para lograr la mitad de Rmax.

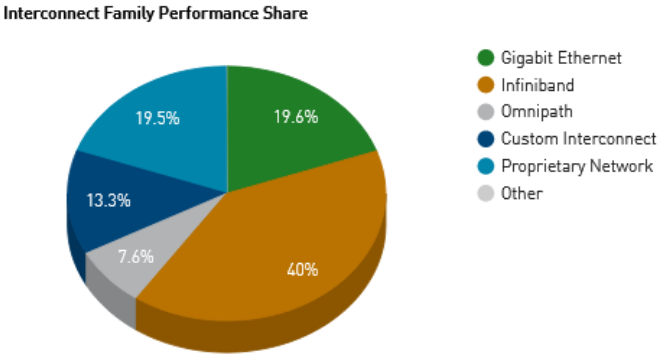
En el mes de noviembre de 2020, la página del TOP 500 presenta las gráficas acerca de la categoría de las familias de interconexión más utilizadas en infraestructuras de Altas prestaciones, que se puede observar en la figuras 7 y 8 las tendencias a nivel del uso de varias tecnologías de red utilizada por los más potentes supercomputadores a nivel mundial.

Figura 7. Familia de Interconexión del Sistema Compartido



Nota: Tomado de (top500, 2020)

Figura 8. Familia de Interconexión del Rendimiento compartido.



Nota: Tomado de (top500, 2020)

Figura 9. Familia de Interconexión del Rendimiento compartido.

	Interconnect Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	Gigabit Ethernet	254	50.8	475,356,880	1,012,892,944	17,601,296
2	Infiniband	155	31	971,927,068	1,477,791,334	20,231,602
3	Omnipath	47	9.4	184,605,368	291,945,234	4,518,896
4	Custom Interconnect	37	7.4	321,955,166	481,808,575	21,673,420
5	Proprietary Network	6	1.2	472,942,300	575,003,229	8,351,312
6	Myrinet	1	0.2	1,975,070	6,594,560	89,600

Nota: Tomado de (top500, 2020)

Con base a las Figuras 7,8 y 9 mostradas anteirormente, se puede observar que el uso de redes InfiniBand ocupan el segundo lugar dentro del TOP500 a nivel mundial, con un porcentaje del 40% de la Familia de Interconexión de Rendimiento compartido. Por lo tanto, estas estadísticas es parte de la motivación del presente trabajo de investigación, ya que permitirá presentar datos reales sobre el rendimiento

de una red InfiniBand sobre ciertos escenarios en una infraestructura de computación paralela.

Red de Altas prestaciones InfiniBand

El mundo de la informática ha sufrido grandes cambios tecnológicos que ha exigido a las redes de comunicaciones disponer de mayores prestaciones que se adapten a nuevas necesidades y arquitecturas de computación. Hoy en día existe una gran exigencia a nivel de procesamiento de cómputo y alto rendimiento por las nuevas tecnologías como: Cloud Computing, Big data y Supercomputación; las redes altas prestaciones mediante un software adecuado permiten la comunicación y control de los procesos de manera eficiente y de esta forma obtener un rendimiento aceptable de las soluciones tecnológicas. Siendo InfiniBand una tecnología de interconexión entre sistemas de procesamiento y dispositivos de E/S, como una arquitectura independiente del sistema operativo y de la plataforma, permitiendo manejar de mejor manera los procesos de cómputo.

InfiniBand

InfiniBand es una nueva y potente tecnología diseñada para soportar la conectividad E/S para infraestructuras de computación. InfiniBand es compatible con los principales proveedores de servidores OEM[12] como un medio para expandirse más allá y crear el estándar de interconexión E/S de próxima generación de servidores.

[12] OEM (del inglés, Original Equipment Manufacturer) o fabricante de equipos originales que confecciona piezas, un subsistema o software que se utilizan en los productos de otras empresas. Algunos ejemplos son los sistemas operativos y los microprocesadores en equipos.

InfiniBand es el único en proporcionar tanto una solución de backplane “in the box”, una interconexión externa y “Ancho de banda out the box”, por lo que proporciona conectividad de una manera previamente reservada solo para interconexiones de redes tradicionales.

Arquitectura de InfiniBand

InfiniBand define una Red de Área de Sistema (System Area Network, SAN) que permite interconectar ordenadores, sistema de E/S y dispositivos de E/S, para la comunicación entre ellos o transacciones de E/S, de forma simultánea con gran ancho de banda y baja latencia. (EcuRed, s.f.)

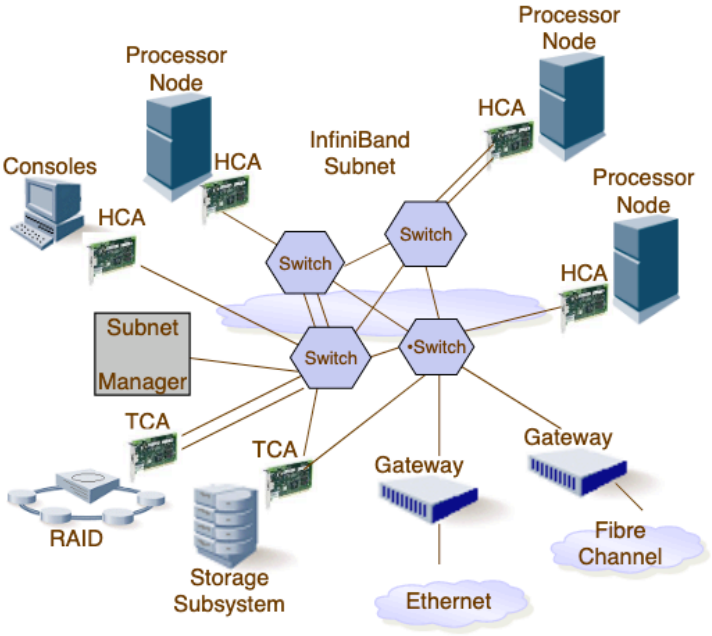
En la Figura 10, se puede identificar una red de área de sistema, que interconecta Nodos procesadores, subsistema RAID, los Discos, switches y routers. Cada Nodo procesador dispone del CPU y de las tarjetas HCA[13] (Host Channel Adapter o Adaptador de Canal de Host) interconectado a través de los switches sobre una red InfiniBand. (Villar Ortiz, 2004)

La arquitectura definida por InfiniBand es independiente del sistema operativo y de la plataforma. Siendo sus principales elementos de la arquitectura InfiniBand el HCA, switches, SM[14] (Subnet Managers) y el Gateway o router. (Villar Ortiz, 2004)

[13] HCA (Host Channel Adapter) es Dispositivo donde termina un enlace IB y ejecuta funciones a nivel de transporte y soporta la interfaz de verbos

[14] El Subnet Manager (SM) es una entidad de software que configura su subred local y asegura su funcionamiento continuo. Establece rutas primarias y secundarias entre cada punto final para que las decisiones de reenvío del flujo de tráfico estén pre programadas y los datos lleguen en el menor tiempo posible.

Figura 10. Red de área de sistema con InfiniBand.



Nota: Tomado de (HPC-AI Advisory Council, 2020).

InfiniBand permite resolver ciertas limitaciones que presentan los actuales buses PCI (por ejemplo, cuellos de botella, fiabilidad, escalabilidad, etc.). Un sistema InfiniBand puede variar desde un pequeño servidor formado por un procesador y unos cuantos dispositivos de E/S conectados, hasta un supercomputador masivamente paralelo con miles de procesadores y dispositivos de E/S conectados mediante internet a otras plataformas de procesamiento y/o sistema de E/S. (EcuRed, s.f.)

Niveles en la arquitectura de InfiniBand

El funcionamiento de InfiniBand se describe mediante la interacción de una serie de niveles. El protocolo que gobierna cada nivel es independiente del resto de niveles. (EcuRed, s.f.)

Nivel Físico. - El nivel físico especifica cómo se trasladan los bits al cable. InfiniBand utiliza el sistema de codificación 8B/10B, con lo que, de cada 10 bits enviados solamente 8 son de datos. InfiniBand especifica tres tipos de medios físicos: par trenzado, fibra óptica o circuito en placa. El nivel físico también especifica los símbolos que se usaran para indicar principio y final de paquete, datos, espacio entre paquetes, el protocolo de señalización a utilizar, etc.

Nivel de Enlace. - El nivel de enlace describe los formatos de paquete a usar y los protocolos para las operaciones con ellos. También son tareas de este nivel el control de flujo y el encaminamiento de los paquetes dentro de la misma subred.

Nivel de Red. - El nivel de red describe el protocolo para encaminar un paquete entre distintas subredes. Este nivel define una Global Route Header (GRH) que debe estar presente en los paquetes que tengan que ser encaminados entre dos o más subredes. La GRH identifica los puertos origen y destino usando direcciones globales (GID) en el formato de una dirección IPv6.

Nivel de Transporte. - Los protocolos de enlace y de red llevan un paquete al destino deseado. La parte de nivel de transporte del paquete se encarga de hacer que se entregue el paquete en la cola (QP) apropiada, indicándole además como procesar los datos contenidos en el paquete. El nivel de transporte es el responsable de segmentar una operación en varios paquetes si los datos a transmitir exceden el tamaño máximo de paquete (MTU). El QP en el extremo final re ensambla los paquetes para formar la secuencia de datos que se quiso enviar.

Características de InfiniBand

InfiniBand presenta las principales características: (EcuRed, s.f.)

Pares de colas. - Los pares de colas (queue pairs, o QP) son la interfaz virtual que el hardware proporciona a un productor de información en InfiniBand, y el puerto de comunicación virtual que proporciona para el consumidor de dicha información. De esta forma, la comunicación tiene lugar entre un QP fuente y un QP destino. (EcuRed, s.f.)

Tipos de servicio:

- **Servicio orientado a conexión frente a no orientado a conexión:** El servicio no orientado a conexión suele llamarse también datagrama. En un tipo de servicio orientado a conexión cada QP fuente está asociado con un único QP destino, y viceversa. En un tipo de servicio no orientado a conexión está permitido que un QP envíe/reciba paquetes a/desde cualquier otro QP en cualquier nodo.
- **Servicio con confirmación frente a sin confirmación:** En un servicio confirmado, cuando un QP recibe un paquete debe confirmar al QP origen que lo ha recibido correctamente. Estos mensajes de confirmación pueden ir integrados en otro paquete con información, o en un mensaje ACK o NAK (negative acknowledged) propio. El servicio confirmado se dice que es fiable, pues el protocolo de transporte garantiza un envío sin errores y con entrega en orden para el posterior re ensamblado de los paquetes en un mensaje de nivel

superior. Por contra, el servicio sin confirmación se dice que es no fiable pues el protocolo de transporte no asegura que la información llegue a su destino.

- **Servicio de transporte de InfiniBand frente a otro tipo de transporte:** El servicio de transporte de InfiniBand permite transmitir paquetes en bruto encapsulando paquetes de otros protocolos de transporte, como por ejemplo IPv6, o de otros tipos de redes.

Claves. - Las claves (keys) proporcionan un cierto nivel de aislamiento y protección del tráfico. Se insertan en los paquetes. Las aplicaciones solo podrán acceder a los paquetes que contengan claves para los que ellas estén habilitadas. Los diferentes tipos de claves son: (EcuRed, s.f.)

- **Management Key (M Key):** Esta clave se usa para tareas de gestión y se puede asignar una distinta a cada puerto. Una vez hecha la asignación, todo el tráfico de control con ese puerto deberá llevar esa clave insertada en los paquetes.
- **Baseboard Management Key (B Key):** Permite que actúe el gestor de subred en placa. Esta clave la contienen cierto tipo de paquetes de gestión de la subred.
- **Partition Key (P Key):** Permite la división lógica de la subred en distintas zonas. Cada adaptador contiene una tabla de claves de partición que define las particiones para las que ese adaptador está habilitado. Hay un gestor de particiones (Partition Manager, PM) único, que se encarga de gestionar las claves de las particiones.

- **Queue Key (Q Key):** Permite controlar el derecho de acceso a las colas para los servicios sin conexión (datagrama). De esta forma, dos nodos que no hayan establecido previamente unas conexiones pueden intercambiar información de forma que esta clave identifique unívocamente a los interlocutores.
- **Memory Keys (L Key y R Key):** Permite el uso de direcciones de memoria virtuales y dota al consumidor de un mecanismo para controlar el acceso a dicha memoria. El consumidor le especifica al adaptador una zona de memoria y recibe de este una L Key y otra R Key. El consumidor usa la L Key en las gestiones locales de memoria, y pasa la R Key a los consumidores remotos para que la usen en las operaciones remotas de DMA (RDMA[15]).

Las claves no proporcionan seguridad por si solas pues dichas claves están disponibles en la cabecera de los paquetes que circulan por la red. (Villar Ortiz, 2004)

Canales virtuales. - Los canales virtuales (VL) constituyen un mecanismo para crear múltiples enlaces virtuales con un único enlace físico. Un canal virtual representa un conjunto de buffers de transmisión y recepción en un puerto.

Control de la tasa de inyección. - InfiniBand define enlaces serie punto a punto full-dúplex funcionando a una frecuencia de 2,5 GHz. La velocidad de transmisión que se obtiene es 2,5 Gb/seg, que se denomina 1X. Sin embargo,

[15] Remote Direct Memory Access (RDMA) o su traducción al castellano acceso remoto directo a memoria consiste en el acceso directo desde la memoria principal de un ordenador en la de otro sin cooperación del sistema operativo. Esto permite la realización de sistemas de alto rendimiento, así como comunicaciones de baja-latencia lo cual es muy importante en sistemas MPP.

InfiniBand permite alcanzar mayores velocidades usando varios de esos enlaces en paralelo, como 10 Gb/seg (4X) y 30 Gb/seg (12X).

Multicast. - Multicast es un paradigma de comunicación uno-muchos/muchos-muchos diseñado para simplificar y mejorar la comunicación entre un conjunto de nodos finales. InfiniBand permite la comunicación multidestino.

Hardware InfiniBand

InfiniBand ofrece múltiples niveles de rendimiento de enlace, que actualmente alcanza velocidades hasta 400 Gb/s. Cada una de estas velocidades de enlace también proporciona comunicación de baja latencia dentro de la estructura, lo que permite un mayor rendimiento agregado que otros protocolos. Esto posiciona de manera única a InfiniBand como la interconexión de E/S ideal para centros de datos. (Infiniband Trade Association, 2014)

Existe una gran variedad de chipset de última generación para la comunicación InfiniBand, que ofrece Mellanox el proveedor líder de estas soluciones ahora perteneciente a Nvidia, en diversas HCAs dependiendo del slot de conexión, la memoria de la tarjeta y las prestaciones puesto que existe varias versiones como se muestra en la Figura 11:

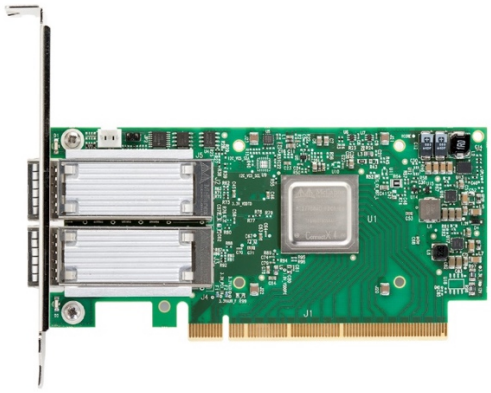
Figura 11. Velocidad de datos de InfiniBand.

Name	Abbreviation	Raw Signaling Rate	Applied Encoding	Effective Data Rate	Aggregated (4x) Throughput
Single Data Rate	SDR	2.5 Gb/s	8b/10b	2 Gb/s	**8 Gb/s**
Double Data Rate	DDR	5 Gb/s	8b/10b	4 Gb/s	**16 Gb/s**
Quad Data Rate	QDR	10 Gb/s	8b/10b	8 Gb/s	**32 Gb/s**
Fourteen Data Rate	FDR	14.1 Gb/s	64b/66b	13.64 Gb/s	**54.5 Gb/s**
Enhanced Data Rate	EDR	25.8 Gb/s	64b/66b	25 Gb/s	**100 Gb/s**
High Data Rate	HDR	51.6 Gb/s	64b/66b	50 Gb/s	**200 Gb/s**
Next Data Rate	NDR	TBD	TBD	TBD	**TBD**

Nota: Tomado de (Infiniband Trade Association, 2014)

El Supercomputador Quinde I de la Empresa Pública Siembra E.P, cuenta con tarjetas HCA ConnectX®-4 VPI adapter card, EDR IB (100Gb/s) y 100GbE, puerto dual QSFP28, PCIe3.0 x16, tall bracket. (Mellanox Technologies, s.f.)

Figura 12. Tarjeta HCA ConnectX 4 MCX456A-ECAT



Nota: Tomado de (Mellanox Technologies, s.f.)

Beneficios: (Mellanox Technologies, s.f.)

- Latencia 0.6 µs, dependiendo bastante del software y el firmware de la infraestructura.

- Silicio de mayor rendimiento para aplicaciones que requieren un gran ancho de banda, baja latencia y alta tasa de mensajes
- Rendimiento de almacenamiento, red y clústeres de clase mundial
- Interconexión inteligente para plataformas de almacenamiento y computación x86, Power, Arm y basadas en GPU
- Rendimiento de vanguardia en redes superpuestas virtualizadas (VXLAN y NVGRE)
- Consolidación de E/S eficiente, que reduce los costos y la complejidad del centro de datos
- Aceleración de virtualización
- Eficiencia energética
- Escalabilidad a decenas de miles de nodos.
- Calidad de Servicio, Canales de Entrada/ Salida Independientes a nivel del Adaptador. Tiene lineas virtuales a nivel de la capa de enlace.

Características: (Mellanox Technologies, s.f.)

InfiniBand

- Puertos EDR InfiniBand 100 Gb/s
- Cumple con la Especificación IBTA 1.3
- RDMA, Semántica de Envío/Recepción.
- Control de Congestión basado en Hardware.
- Operaciones Atómicas.
- 16 millones de canales de Entrada/Salida
- 256 a 4Kbyte MTU, 2Gbyte mensajes

- 8 líneas Virtuales + VL15
- Mellanox Multi-Host (4 hosts)

Características mejoradas:

- 150M mensajes/segundo
- Transporte confiable basado en Hardware.
- Operaciones de transporte de baja carga de CPU (CPU offloading of transport operations)
- Aplicaciones de baja carga.
- Offloads en Operaciones Colectivas.
- Offloads de Operaciones Colectivas Vectoriales.
- Aceleración de comunicación Mellanox PeerDirect RDMA (también conocido como GPUDirect®)
- Codificación 64/66
- Transporte Extended Reliable Connected (XRC)
- Dynamically Connected transport (DCT)
- Operaciones Atómicas Mejoradas.
- Compatibilidad con mapeo de memoria avanzado, que permite el registro en modo de usuario y la reasignación de memoria (UMR)
- Paginación bajo demanda (ODP)
- Registro de acceso directo a la memoria RDMA

CPU Offloads (Baja carga de CPU)

- RDMA over Converged Ethernet (RoCE)

- TCP/UDP/IP stateless offload
- LSO, LRO, checksum offload
- RSS (Se puede hacer en paquetes encapsulados), TSS, HDS, VLAN inserción / extracción, dirección de flujo de recepción
- Coalescencia de interrupción inteligente.

Storage Offloads:

- T10 DIF Signature Handover
- Operación de handover de firma a la velocidad del cable, para el tráfico de entrada y salida.

Overlay Networks

- Stateless offloads para redes overlay y protocolos de tunelización
- Offload de hardware de encapsulación y desencapsulación de redes overlay NVGRE

Hardware-Based I/O Virtualization

- Single Root IOV
- Multifunción por Puerto.
- Dirección de traducción y protección
- Varias colas por máquina virtual.
- QoS Mejorada por vNICs
- Soporta VMware NetQueue

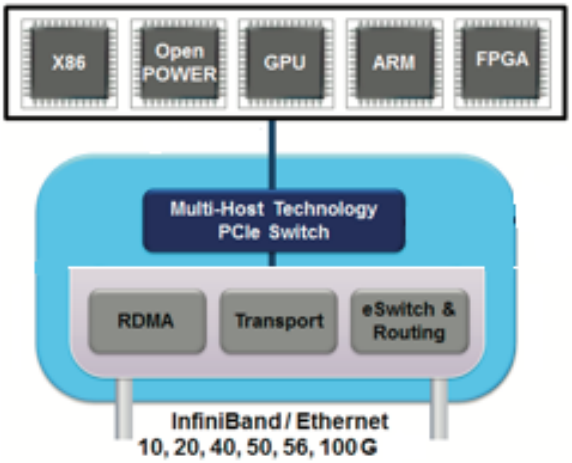
Soporte de protocolos

- OpenMPI, IBM PE, OSU MPI (MVAPICH/2), Intel MPI,
- Platform MPI, UPC, Mellanox SHMEM
- TCP/UDP, EoIB, IPoIB, SDP, RDS, MPLS, VXLAN, NVGRE, GENEVE
- SRP, iSER, NFS RDMA, SMB Direct
- uDAPL

Arquitectura de Tarjeta ConnectX-4

El Adaptador ConnectX-4 de Mellanox ofrece un rendimiento de 10, 20, 25, 40, 50, 56 y 100 Gb/s compatible con los protocolos estándar InfiniBand y Ethernet, y la flexibilidad para conectar cualquier arquitectura de CPU: x86, GPU, POWER, ARM, FPGA y más. Con un rendimiento de clase mundial a 150 millones de mensajes por segundo, una latencia muy baja y motores de aceleración inteligente como RDMA, GPUDirect y SR-IOV. ConnectX-4 permite plataformas de almacenamiento y computación más eficientes como se muestra en la Figura 13.

Figura 13. Arquitectura Adaptador ConnectX-4.



Nota: Tomado de (Announcing the Mellanox ConnectX-5 100G Infiniband Adapter, 2016)

RDMA

Remote direct memory access (RDMA) o su traducción al castellano acceso remoto directo a memoria se refiere al acceso directo desde la memoria principal de un ordenador en la de otro sin intervención del sistema operativo, permitiendo alta rendimiento en sistemas de cómputo y de baja latencia. (HPC-AI Advisory Council, 2020)

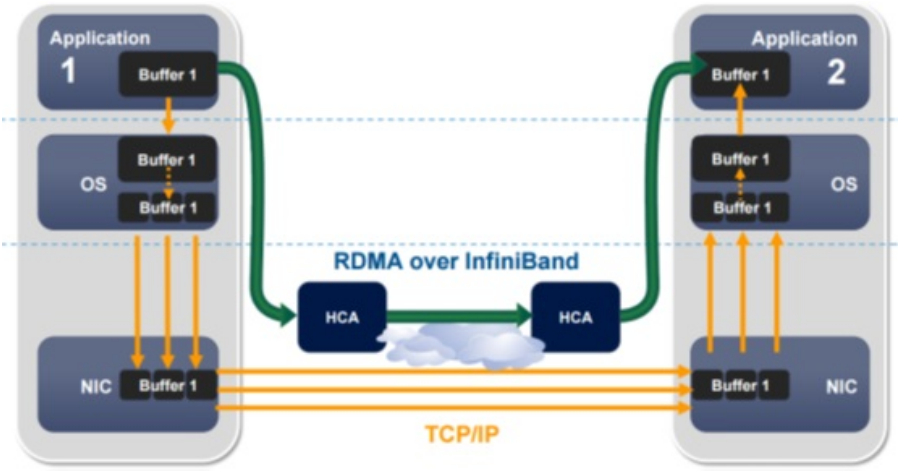
InfiniBand permite el acceso remoto (RDMA) a una región de memoria previamente registrada en el adaptador con el fin de permitir la transferencia eficiente de datos entre servidores y almacenamiento, sin la participación del CPU en la ruta de datos. (HPC-AI Advisory Council, 2020)

RDMA se ha utilizado ampliamente en computación de alto rendimiento (HPC), inteligencia artificial (AI), computación en la nube y de borde, servicios financieros y de telecomunicaciones, y más. RDMA permite un alto rendimiento y escalabilidad, una latencia ultra baja y una sobrecarga de CPU baja para las

transferencias de datos de aplicaciones con uso intensivo de datos y computación. RDMA es la tecnología clave para una transferencia de datos eficiente y escalable entre los servidores del centro de datos, el almacenamiento y la plataforma de borde. Los programadores altamente capacitados en RDMA son ahora extremadamente deseados por los empleadores de la industria. (HPC-AI Advisory Council, 2020)

InfiniBand crea un canal conectando directamente una aplicación en su espacio virtual de direcciones a una aplicación en otro espacio virtual de direcciones. Las dos aplicaciones pueden estar en espacios de direcciones físicos separados, alojados en diferentes servidores como se muestra en la Figura 13. (Grun, 2010)

Figura 14. RDMA sobre InfiniBand.



Nota: Tomado de (Grun, 2010)

Es conveniente dar un nombre a los puntos finales del canal; los llamaremos pares de cola (QP); cada QP consta de una cola de envío y una cola de recepción, y cada QP representa un extremo de un canal. Si una aplicación requiere más de una conexión, se crean más QP. Los QP son la estructura mediante la cual una aplicación accede al servicio de mensajería de InfiniBand. Para evitar involucrar al sistema

operativo, las aplicaciones en cada extremo del canal deben tener acceso directo a estos QP. Esto se logra mapeando los QP directamente en el espacio de direcciones virtuales de cada aplicación. Por lo tanto, la aplicación en cada extremo de la conexión tiene acceso virtual directo al canal que la conecta a la aplicación (o almacenamiento) en el otro extremo del canal. Ésta es la noción de E / S de canal. (Grun, 2010)

InfiniBand crea canales privados y protegidos entre dos espacios de direcciones virtuales separados, proporciona un canal endpoint, llamado QP, a las aplicaciones en cada extremo del canal, y proporciona un medio para que una aplicación local transfiera mensajes directamente entre aplicaciones que residen en esos espacios de direcciones virtuales separados, lo cual se refiere a un canal de E/S. (Grun, 2010)

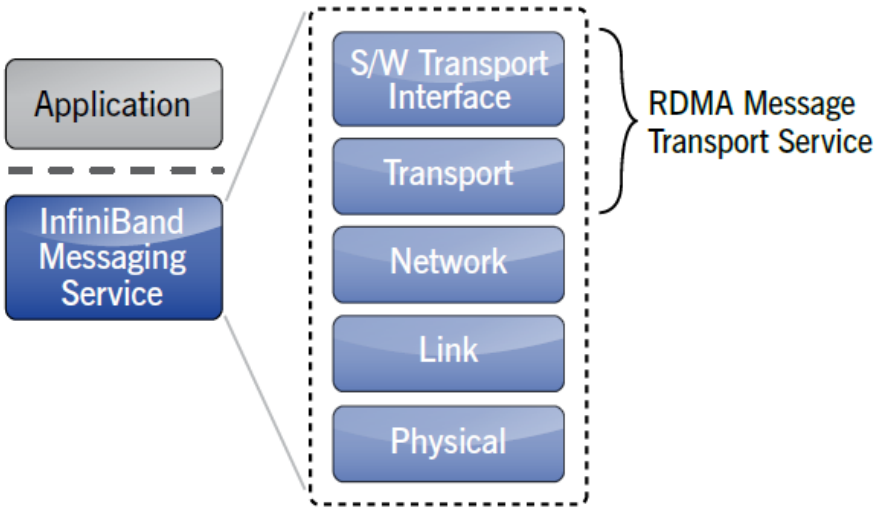
InfiniBand proporciona dos semánticas de transferencia; una semántica de canal a veces llamada SEND / RECEIVE y un par de semánticas de memoria llamadas RDMA READ y RDMA WRITE. Cuando se utiliza la semántica del canal, el mensaje se recibe en una estructura de datos proporcionada por la aplicación en el lado receptor. Esta estructura de datos se publicó previamente en su cola de recepción. Por tanto, el lado emisor no tiene visibilidad de los búferes o estructuras de datos del lado receptor; en cambio, simplemente ENVÍA el mensaje y la aplicación receptora RECIBE el mensaje. (Grun, 2010)

La semántica de la memoria es algo diferente; en este caso, la aplicación del lado receptor registra un búfer en su espacio de memoria virtual. Pasa el control de

ese búfer al lado emisor, que luego usa las operaciones RDMA READ o RDMA WRITE para leer o escribir los datos en ese búfer. (Grun, 2010)

Como parte de la Arquitectura InfiniBand y ubicada justo encima de la capa de transporte se encuentra la interfaz de transporte de software. La interfaz de transporte de software contiene los QP; los pares de colas son la estructura mediante la cual se accede al servicio de transporte de mensajes RDMA. La interfaz de transporte de software también define todos los métodos y mecanismos que una aplicación necesita para aprovechar al máximo el servicio de transporte de mensajes RDMA. Por ejemplo, la interfaz de transporte de software describe los métodos que utilizan las aplicaciones para establecer un canal entre ellas. Una implementación de la interfaz de transporte de software incluye las API y las bibliotecas que necesita una aplicación para crear y controlar el canal y utilizar los QP para transferir mensajes como se muestra en la Figura 14. (Grun, 2010)

Figura 15. Stack de Comunicación de InfiniBand similar al modelo OSI.



Nota: Tomado de (Grun, 2010)

Debajo de las cubiertas del servicio de mensajería, todo esto aún requiere una pila de red completa, tal como la encontraría en cualquier red tradicional.

Incluye la capa de transporte InfiniBand para proporcionar confiabilidad y garantías de entrega (similar al transporte TCP en una red IP), una capa de red (como la capa IP en una red tradicional) y capas de enlace y físicas (cables y conmutadores). Pero es un tipo especial de pila de red porque tiene características que facilitan el transporte de mensajes directamente entre la memoria virtual de las aplicaciones, incluso si las aplicaciones son "remotas" entre sí. Por lo tanto, la combinación de la capa de transporte de InfiniBand junto con la interfaz de transporte del software se considera mejor como una memoria remota directa. Acceso al servicio de transporte de mensajes (RDMA). La pila completa en conjunto, incluida la interfaz de transporte de software, comprende el servicio de mensajería InfiniBand. (Grun, 2010)

La arquitectura InfiniBand proporciona mecanismos simples definidos en la interfaz de transporte de software para colocar una solicitud para realizar una transferencia de mensajes en una cola. Esta cola es el QP, que representa el endpoint del canal. La solicitud se denomina Solicitud de trabajo (WR) y representa una cantidad única de trabajo que la aplicación desea realizar. Un WR típico, por ejemplo, describe un mensaje que la aplicación desea ser transportado a otra aplicación.

Librería Mellanox

La biblioteca MellanoX Messaging (MXM) proporciona mejoras a las bibliotecas de comunicaciones paralelas utilizando completamente la infraestructura

de red proporcionada por el hardware de Mellanox HCA / switch. Esto incluye una variedad de mejoras que aprovechan el hardware de las redes de Mellanox que incluye: (Mellanox Technologies, s.f.)

- Soporte de transporte múltiple, incluidos RC, XRC y UD
- Gestión adecuada de los recursos de HCA y las estructuras de memoria.
- Registro de memoria eficiente
- Semántica de comunicación unilateral.
- Gestión de conexiones
- Recibir coincidencia de etiquetas laterales
- Comunicación de memoria compartida intra-nodo

Estas mejoras aumentan significativamente la escalabilidad y el rendimiento de las comunicaciones de mensajes en la red, aliviando los cuellos de botella dentro de las bibliotecas de comunicaciones paralelas. (Mellanox Technologies, s.f.)

CAPÍTULO III

MARCO METODOLÓGICO

Para explicar la metodología usada en el presente proyecto de titulación, se definirá los parámetros que son usados; así: descripción de área de estudio, enfoque y tipo de investigación, métodos y procedimiento de investigación.

Descripción de área de estudio

Uno de los principales problemas en un Supercomputador con una arquitectura de computación paralela, es la degradación del rendimiento. En un ambiente ideal, el rendimiento crece linealmente al crecer el número de procesadores, si se duplica el número de procesadores en un sistema, el rendimiento debería igualmente duplicarse. Esto no sucede en la realidad, ya que a partir de un cierto número de procesadores los excesivos intercambios de mensajes de control y de transferencia de datos entre los procesadores, provocan un efecto de saturación en el sistema, lo que genera automáticamente una degradación del rendimiento siendo evidente la relación entre la optimización del rendimiento y la minimización de los intercambios entre los procesadores. Existen otros aspectos que juegan un papel importante para optimizar el rendimiento en un sistema Distribuido/Paralelo, como es el equilibrio de la carga de trabajo, y la minimización del número de operaciones de entradas/salidas y entre otros. (Aguilar Castro & Leiss, Introducción a la Computación Paralela, 2004)

En este capítulo se estudia la metodología o reglas a seguir para analizar el rendimiento de la Red de Altas Prestaciones al momento de ejecutar los jobs en arquitecturas de computación paralela.

Este trabajo comprende seleccionar una aplicación HPC, compilarla y ejecutar mensajes MPI sobre ciertos escenarios que permitan analizar el rendimiento de la Red de Altas Prestaciones InfiniBand en una infraestructura de computación paralela. Una vez obtenido los datos se realizará una comparación con otro equipo de computación paralela que tengan una arquitectura de memoria distribuida; lo cual puede dar un punto de partida para realizar el análisis del rendimiento de la red InfiniBand.

A través de esta comparación se pretende obtener datos reales del rendimiento y comportamiento de la red en arquitecturas de computación paralela ante ciertos parámetros, y definir una guía de buenas prácticas con base a los parámetros considerados en los escenarios de prueba de la red de altas prestaciones, que sea utilizado por el usuario investigador en los proyectos a desarrollarse en el Supercomputador Quinde I.

Una vez realizado la comparación y validación de los datos se presentará los resultados en herramientas de visualización de software libre que muestre de manera entendible los gráficos estadísticos.

Diseño de la Investigación

La presente investigación es de tipo documental y experimental, donde se realizará un análisis del rendimiento de la red de altas prestaciones InfiniBand del Supercomputador Quinde I, basado en parámetros preestablecidos en benchmarking

de infraestructuras de Supercomputadores; además, de realizar una comparación del comportamiento de la red en varios escenarios en dicha infraestructura. A continuación, se describe los tipos de investigación utilizados:

Investigación Documental: Para la construcción del análisis de Rendimiento de la Red de Altas Prestaciones se basará en las variables dependientes e independientes identificadas en el capítulo anterior. Se usará fuentes como documentos, artículos, revistas y entre otros medios relacionados con test de Rendimiento o benchmarks realizados en infraestructuras de computación de Alto Rendimiento, con el fin de emitir criterios válidos y fundamentados con bases teóricas.

Investigación Experimental: Se considera este tipo de investigación ya que se realizará pruebas con escenarios controlados utilizando una aplicación de computación paralela sobre una interfaz de paso de mensajes MPI, con el fin de analizar el comportamiento de la red de Altas Prestaciones InfiniBand, de esta forma comparar con otras infraestructuras similares, y describir la causa – efecto que produce esta investigación en el Supercomputador Quinde I de la Empresa Pública Siembra E.P.

Como base documental y experimental para ejecutar el análisis de Rendimiento de la Red de Altas Prestaciones se usará las mejores prácticas planteadas por la HPC-AI Advisory Council, quien tiene la misión principal de cerrar la brecha entre la computación de alto rendimiento (HPC) y el uso de la Inteligencia Artificial (IA), y sobre todo dar a conocer el potencial y beneficios que tiene el HPC y la Inteligencia Artificial en el campo de la investigación, educación, innovación y fabricación de productos. Además, de brindar a los usuarios la experiencia necesaria para operar sistemas HPC e IA y dar a los diseñadores de aplicaciones las herramientas

necesarias para incursionar en el mundo de la computación paralela. (HPC-AI Advisory Council, 2020).

Se tomará en cuenta las reglas indicadas por el HPC Challenge Benchmark, el cual plantea como referencia para seguir ciertos pasos y parámetros que permitirán medir el rendimiento de una infraestructura de altas prestaciones y comparar los resultados en diferentes ámbitos. Como propósito de la presente investigación se considera analizar los datos del Ancho de banda y latencia de comunicaciones.

Estrategia Técnica y procedimiento de investigación

Considerando el significado del concepto de Benchmark conocido como **"punto de referencia",** es un término muy utilizado en el campo empresarial para mejorar el rendimiento de las empresas líderes en cualquier línea de negocio; mientras que en el área de las tecnologías aplicar un Benchmark es realizar un análisis o comparación del rendimiento de varios componentes del hardware o software de una infraestructura o sistema.

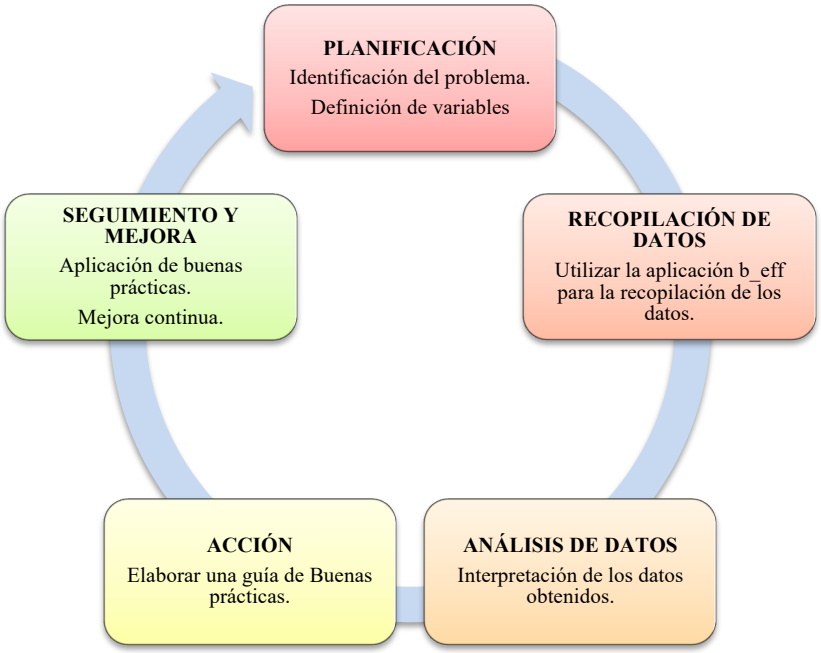
En el mundo de la Supercomputación se utiliza ciertos procedimientos y reglas para analizar el rendimiento de la red, la memoria RAM, el procesador o de cualquier componente o del todo de una infraestructura de supercomputación; con base a la calificación obtenida se posiciona a este tipo de infraestructuras en el TOP 500 o simplemente medir el rendimiento de este. Además, a través de un Benchmark se presenta al campo de la investigación datos reales del comportamiento de un Supercomputador bajo ciertos escenarios de prueba, lo cual brinda una guía de buenas prácticas para la ejecución de los diferentes proyectos de investigación.

Considerando que este proyecto es una investigación experimental se utilizará el procedimiento del Benchmarking definido por HPC -AI Advisory Council para realizar el análisis del rendimiento de la red InfiniBand sobre una infraestructura de computación paralela, utilizando el software Effective Bandwitdth benchmark (b_eff), el cual establece ciertas reglas y pasos a seguir para cumplir con los parámetros de medición del rendimiento y comparación de infraestructuras de Computación de Alto Rendimiento.

Se considera este tipo de investigación experimental ya que se realizará pruebas con escenarios controlados utilizando la aplicación b_eff sobre una infraestructura de computación paralela utilizando mensajes MPI; esto con el fin de analizar el comportamiento particular de la red InfiniBand del Supercomputador Quinde I, y comparar con otra infraestructura similar; además, describir la causa – efecto que produce esta investigación en un caso práctico de unos de los proyectos de investigación del Supercomputador Quinde I de la Empresa Pública Siembra E.P.

Como se muestra en la Figura 16 se considera las siguientes etapas para esta investigación, basado en el origen del concepto de un Benchmark, y con relación al HPC Challenge Benchmark:

Figura 16. Etapas de un Benchmark



Nota. - Elaborado por el autor.

1. **Planificación:** en esta fase se define las variables dependientes e independientes de la presente investigación y los recursos a utilizar. Además, de identificar la aplicación que se utilizará para el análisis de rendimiento de la Red InfiniBand sobre una infraestructura de computación paralela y los escenarios a utilizar para cada test. Como se detalló en el capítulo del Marco Referencial, se describió las características de la Red de Altas prestaciones InfiniBand, del Supercomputador Quinde I y otros aspectos importantes para el desarrollo de la presente investigación. Además, en el mismo capítulo se definió utilizar la aplicación b_eff para el análisis de rendimiento de la Red de Altas Prestaciones utilizando mensajes MPI.

2. **Recopilación de Datos:** Una vez identificada la aplicación para el análisis de Rendimiento de la Red de Altas prestaciones, la cual se determinó anteriormente con base a la comparativa mostrada en la tabla 4, se utilizará la aplicación b_eff, donde es necesario ajustar los parámetros necesarios en cada test para recopilar los datos de cada log generado. La aplicación b_eff arroja varios datos referentes al ancho de banda y latencia que permitirán analizar el rendimiento de la Red de Altas Prestaciones en varios escenarios; por lo tanto, basado al objeto de la presente investigación y con el fin de comparar los datos con otra infraestructura similar, se analizará los siguientes datos:

Ancho de banda efectivo (b_eff): El ancho de banda efectivo es el número de procesos MPI multiplicado por el ancho de banda asintótico multiplicado por la proporción del área bajo la curva "ancho de banda sobre longitudes de mensaje" y el área bajo la curva de ancho de banda asintótico constante en el mismo diagrama. Para medir el ancho de banda, se aplican varios patrones de comunicación. Los patrones se basan en anillos y en distribuciones aleatorias. Se calcula el promedio logarítmico en todos los patrones de anillo y en todos los patrones aleatorios y b_eff es el promedio logarítmico de estos dos valores.

Ancho de banda ping-pong: El Ancho de banda es aquel que se obtiene a través del tamaño del mensaje dividido para el tiempo que toma en transmitir un mensaje de 2,000,000 de bytes usando rutinas básicas de MPI, de acuerdo con los parámetros establecidos en b_eff.

Latencia ping-pong: Es el tiempo requerido para enviar un mensaje de 8 bytes desde un nodo a otro usando rutinas básicas de MPI, de acuerdo a los parámetros establecidos en b_eff.

La medición del ancho de banda y de la latencia se realiza durante la comunicación no simultánea (benchmark de ping-pong) y la comunicación simultánea (patrón random y natural ring). Para el caso de la presente investigación se analizará para una comunicación no simultánea, usando solo los parámetros obtenidos por medio del benchmark ping pong para análisis el rendimiento de la Red de Altas Prestaciones y de esta forma comparar con otra infraestructura.

3. **Análisis de datos:** Se realizará un análisis de los datos obtenidos a través de la ejecución de la aplicación b_eff, configurando los parámetros de cada escenario de prueba, y de esta forma interpretar los datos del ancho de banda efectivo, ancho de banda ping-pong y de la latencia ping-pong.

4. **Acción:** Se detallará un guía de buenas prácticas para la ejecución de proyectos de investigación sobre una red de Altas Prestaciones InfiniBand en una arquitectura de computación paralela. Además, de presentar los resultados obtenidos sobre gráficas estadísticas de los datos, que permitan conocer el comportamiento de la red sobre ciertos escenarios de prueba.

5. **Seguimiento y mejora:** En esta última etapa, es parte de la aplicación de las buenas prácticas en proyectos futuros del Supercomputador Quinde I,

con el fin de dar seguimiento de la ejecución de cada trabajo de investigación. De esta forma, mantener un constante monitoreo del Rendimiento de la Red de Altas prestaciones; la idea es mantener un ciclo de mejora continua del uso y funcionamiento de la infraestructura, con la ejecución de escenarios con diferentes aplicaciones de investigación según la necesidad.

Entre los instrumentos de investigación a utilizar están:

- Archivos de logs de cada escenario.
- Nodos de cómputo del Supercomputador Quinde I.
- Recursos de los nodos de cómputo a utilizar como memoria RAM, cantidad de procesadores, almacenamiento, etc.)
- Aplicación para conexión mediante el protocolo SSH a la interfaz de línea de comandos del Supercomputador.
- Computador para acceder al servicio SSH del servicio de supercomputación.
- Herramientas para presentar los gráficos estadísticos de los datos obtenidos.

Reglas para realizar un Benchmark

Se considera una línea base para la ejecución de cada archivo en un sistema de cómputo o un archivo optimizado para cada sistema de cómputo. A continuación, se describe las reglas según el HPC Challenge Benchmark para correr un test a nivel de red utilizando la aplicación b_eff:

1. Compilar y cargar opciones

Se permiten los indicadores de compilador o cargador que son compatibles y documentados por el proveedor. Estos incluyen portabilidad, optimización e invocación de preprocesador.

2. Bibliotecas

Se permite la vinculación a versiones optimizadas de las siguientes bibliotecas, las cuales deben poder ser utilizadas por terceros:

- BLAS[16]
- FFT[17]
- MPI

El uso aceptable de dichas bibliotecas está sujeto a las siguientes reglas:

- Todas las bibliotecas utilizadas se divulgarán con la presentación de resultados. Cada biblioteca se identificará por nombre de biblioteca, revisión e institución que proporcione el código fuente.
- No se permiten bibliotecas que no estén disponibles en general, a menos que la organización informante las ponga a disposición en un plazo de 6 meses.
- Las llamadas a las subrutinas de la biblioteca deben tener la misma sintaxis y semántica que en el código de referencia publicado. No se

[16] Basic Linear Algebra Subprograms (BLAS), en español Subprogramas Básicos de Álgebra Lineal, es una especificación que define un conjunto de rutinas de bajo nivel para realizar operaciones comunes de álgebra lineal tales como la suma de vectores, multiplicación escalar, producto escalar, combinaciones lineales y multiplicación de matrices.
[17] FFT es una biblioteca de subrutinas C para calcular la transformada discreta de Fourier (DFT) en una o más dimensiones, de tamaño de entrada arbitrario, y de datos tanto reales como complejos.

permiten modificaciones de código para acomodar varios formatos de llamadas a bibliotecas.

3. Herramientas de software

Todas las herramientas utilizadas para construir y ejecutar el benchmark (incluidos preprocesadores, compiladores, enlazadores estáticos y dinámicos, sistemas operativos) deben estar disponibles en general en el sistema probado (o la organización informante debe ponerlas a disposición en un plazo de 6 meses).

4. Solo se pueden enviar resultados de benchmark completos; no se aceptarán resultados parciales.
5. Modificación de código

Siempre que se conserve la especificación de entrada y salida, se pueden sustituir las siguientes rutinas:

- En HPL: HPL_pdgesv(), HPL_pdtrsv() (funciones de factorización y sustitución)
- Los cambios no son permitidos en el testing de DGEMM y la rutina de DGEMM (si corresponde) debe ajustarse a la definición de BLAS.
- En PTRANS: pdtrans()
- En STREAM: tuned_STREAM_Copy (), tuned_STREAM_Scale (), tuned_STREAM_Add (), tuned_STREAM_Triad ()
- En Random Access: Power2NodesMPIRandomAccessUpdate(), AnyNodesMPIRandomAccessUpdate(), y RandomAccessUpdate()

- En FFTE: fftw_malloc(), fftw_free(), fftw_create_plan(), fftw_one(), fftw_destroy_plan(), fftw_mpi_create_plan(), fftw_mpi_local_sizes(), fftw_mpi(), fftw_mpi_destroy_plan() (Todas estas funciones son compatibles con FFTW 2.1.5, por lo que el benchmark se puede vincular directamente con FFTW 2.1.5 agregando solo los indicadores del compilador y vinculador adecuados, incluidos -DUSING_FFTW)
- Se permiten cambios en partes del componente b_eff, pero es necesario preservar la portabilidad y la conformidad con el estándar MPI (MPI 1.1 o posterior). Se debe proporcionar una lista detallada de las llamadas a funciones MPI eliminadas y agregadas al momento del envío.

6. Limitaciones de Optimización

- Código con precisión de cálculo limitada

El cálculo debe realizarse con total precisión (64 bits o equivalente). Sin embargo, se permite la sustitución de algoritmos.

- Intercambio del algoritmo matemático utilizado

Cualquier cambio de algoritmos debe divulgarse en su totalidad y está sujeto a revisión por parte del Comité del HPC Challenge. Para la multiplicación de matrices en el banco de pruebas HPL, no se puede utilizar el algoritmo de Strassen, ya que cambia el recuento de operaciones del algoritmo.

Instalación de la aplicación B_eff

B_eff.- Es una aplicación de software libre que permite medir el ancho de banda efectivo de la red de comunicación de un sistema informático paralelo y / o distribuido, que utiliza diferentes tamaños de mensajes y métodos de comunicación para validar el comportamiento de la Red de Altas prestaciones (Gerrit Schulz , 2020).

A continuación, se detalla los pasos para la instalación de la aplicación:

Pasos para implementación de la aplicación:

a. Descargar el código fuente.
b. Montar la imagen en el directorio fs1/ del proyecto creado como parte de la investigación.
c. Compilar la aplicación.
d. Ejecutar la aplicación.

Para la compilación de la aplicación se utilizará el compilador xLC de IBM, que es una librería optimizada para arquitecturas “Power 8”, como se muestra en el siguiente comando:

Ecuación 2. Comando para compilar la aplicación b_eff en paralelo con la librería MPI

$$\text{mpicc } -\text{O4 } -\text{o b\_eff } -\text{DMEMORY\_PER\_PROCESSOR} = 1024\ \text{b\_eff.c (MPI} - \text{MPP)}$$

- mpicc. - comando para compilar en entorno paralelo

- -O4: variable para indicar al compilador el grado de optimización permitido por el código fuente programado.
- -o b_eff: nombre del archivo de salida compilado o ejecutable
- -DMEMORY_PER_PROCESSOR: monto de memoria asignado, basado en la cantidad disponible del hardware, este monto de memoria es la que se usará para correr el archivo compilado.
- b_eff.c.- nombre del archivo escrito en lenguaje c con el código fuente para compilar.

CAPÍTULO IV

MARCO ADMINISTRATIVO

Viabilidad

Para el desarrollo de la presente investigación se consideró lo siguiente:

Factibilidad Tecnológica: El Supercomputador Quinde I, brinda los recursos de hardware y software necesarios para el desarrollo de la presente investigación, sin necesidad de obtener un licenciamiento ya que todo la plataforma esta basada en software libre. La presente investigación utilizará los recursos existentes del Supercomputador Quinde I, lo cual permitirá obtener datos reales del Rendimiento de la Red de Altas prestaciones sobre una infraestructura de computación paralela en varios escenarios, basándose en la reglas establecidas para Benchmarks en el área de Supercomputación con la aplicación b_eff según HPC-AI Council; se estima disponer de resultados que puedan mejorar los procesos de cómputo de futuros proyectos de investigación y suplir con una guía de buenas prácticas para los usuarios finales del Servicio de Supercomputación que brinda la Empresa Pública Siembra E.P.

Factibilidad Operativa: El Supercomputador Quinde I, permite la instalación de varias aplicaciones de software libre, en función de las necesidades de cada proyecto de investigación. Además, el investigador puede instalar las aplicaciones de Software libre que requiera, siempre y cuando justifique en el formulario de acceso al Servicio de Supercomputación para obtener la aprobación de cada proyecto por parte del Comité de Acceso de la Empresa Pública Siembra E.P. En el caso de requerir

software licenciado, cada usuario proveerá en su ambiente del proyecto el licenciamiento necesario.

Valor Práctico

Beneficiario Directo, esta direccionado para los investigadores del área académica-científica del Ecuador y del mundo, que están inmersos en el campo de la Supercomputación, los cuales podrán hacer uso de los resultados de este análisis para mejorar los procesos de cómputo de varios proyectos de investigación o a su vez continuar con nuevas investigaciones o benchmarking de la Red de Altas Prestaciones InfiniBand sobre una arquitectura de computación paralela sobre aplicaciones científicas. En especial para los usuarios actuales con los que cuenta la Empresa Pública Siembra E.P. como Universidades públicas e instituciones privadas del país.

Beneficiario Indirecto, personal académico y profesional de redes de investigación en Supercomputación, estudiantes de tercero y cuarto nivel en diferentes áreas de la ciencia y tecnología. Además, esta investigación permitirá impulsar el uso del Supercomputador Quinde I y ampliar el servicio en el país.

Presupuesto

En la Tabla 5 se detalla el presupuesto que representa la presente investigación:

Tabla 5. Presupuesto para desarrollo de la investigación.

Nro.	DESCRIPCIÓN	VALOR
1	EQUIPOS, SOFTWARE Y SERVICIOS - COMPUTADOR - SERVICIOS	$ 1200
2	RECURSOS HUMANOS, TRANSPORTE - ASESORAMIENTO - SALIDAS Y VISITAS - TRANSPORTE	$ 200
3	MATERIALES Y SUMINISTROS - PAPEL RESMA - FOTOCOPIAS - IMPRESIONES	$ 200
4	MATERIAL BIBLIOGRÁFICO - LIBROS	$ 100
	SUBTOTAL	**$ 1700**
	(+) 10% IMPREVISTOS	**$ 170**
	(=) VALOR TOTAL	**$ 1870**

Nota: Elaborado por el autor.

Cronograma de actividades del Plan de Investigación

En la Tabla 6 se muestra el cronograma de actividades del Plan de Investigación:

Tabla 6. Plan de Investigación

Nro.	ACTIVIDADES	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10
1	Aprobación del uso del Supercomputador Quinde I.										
2	Levantamiento de requerimientos.										
3	Recopilación bibliográfica.										
4	Definición de variables										

5	Determinar la aplicación HPC para el análisis de Rendimiento de la Red de Altas prestaciones.	
6	Compilar y Ejecutar la aplicación HPC para análisis de Rendimiento de la Red de Altas prestaciones.	
7	Ejecución de Jobs en el Supercomputador Quinde I.	
8	Recopilar y Analizar los datos obtenidos.	
9	Presentación de los resultados	
10	Revisión y Presentación del informe de investigación	

Nota: Elaborado por el Autor.

CAPÍTULO V

SIMULACIÓN, RESULTADOS Y ANÁLISIS.

Descripción

Con base a las mejores prácticas y lineamientos establecidos para la ejecución de benchmark de Supercomputadores a nivel mundial en el HPC Challenge y en el HPC-AI Advisory Council Clúster Center, en el que principalmente apuntan hacia un propósito común de analizar el comportamiento de los componentes o de toda una infraestructura de Altas Prestaciones (HPC) frente a diferentes escenarios de prueba; esto con el fin de determinar el rendimiento de varios componentes como de la memoria, procesamiento y en el caso de la presente investigación se realizará la medición del rendimiento de la Red de Altas Prestaciones InfiniBand sobre una arquitectura de computación paralela.

Se selecciono ciertos escenarios de test, considerando la cantidad de nodos, número de procesadores, el tamaño de la memoria del procesador y las librerías para el paso de mensajes sobre MPI; esto permitirá conocer el comportamiento de la Red de Altas Prestaciones sobre una arquitectura de computación paralela.

Una vez analizado los diferentes escenarios del test de rendimiento de la Red de Altas prestaciones InfiniBand, se podrá determinar cuál de los escenarios representa un mejor comportamiento.

Se tomará uno de los mejores escenarios para aplicar a un caso práctico de un proyecto de investigación del Supercomputador Quinde I, y de esta forma validar los datos adquiridos del Análisis de Rendimiento de la Red de Altas Prestaciones, y validar el comportamiento de los procesos de cómputo en dicho proyecto con los resultados obtenidos.

Hardware Utilizado

Para ejecutar el Benchmark de Rendimiento de la Red de Altas prestaciones InfiniBand, se consideró los siguientes recursos de hardware del Supercomputador Quinde I detallados en la tabla 7, los cuales fueron solicitados mediante el formulario de acceso detallados en el Anexo A:

Tabla 7. Recursos de Hardware requeridos en el Supercomputador Quinde I.

DESCRIPCIÓN	CANTIDAD	MEDIDA
PROCESADORES	Hasta 1000	Procesadores
MEMORIA	10	GB
STORAGE	1	GB
GPU	No requerido	No requerido

Nota.- Se muestra los recursos mínimos requeridos para realizar el Análisis de Rendimiento de la Red de Altas Prestaciones InfiniBand del Supercomputador Quinde I.

Software utilizado

Para ejecutar los test de Rendimiento de la Red de Altas prestaciones InfiniBand, se requiere de los siguientes recursos de software como librerías y aplicaciones que se detallan en la Tabla 8:

Tabla 8. Recursos de Software requerido para el Análisis de Rendimiento.

SOFTWARE	DESCRIPCIÓN	TIPO DE LICENCIAMIENTO
Red Hat Enterprise Linux Server 7.2 (Maipo)	Sistema Operativo instalado en los nodos de cómputo del Supercomputador Quinde I. Operating System: Red Hat Enterprise Linux Server 7.2 (Maipo) Kernel: Linux 3.10.0-327.el7.ppc64le Architecture: ppc64-le	Software libre
Compilador C/C++ (IBM Spectrum)	El compilador de IBM tiene una gran capacidad de optimizar códigos para los CPU de IBM Power 8 Little Endian de 64 bit. Versión cc (GCC) 4.8.5	Software libre
Hpcc 1.5.0	Hpcc es el nombre del ejecutable del HPC Challenge Benchmark que combina varios test para medir el rendimiento de los componentes de una infraestructura de Altas Prestaciones (HPC Challenge, 2016)	Software libre
b_eff	Es una de las aplicaciones que son parte del hpcc Benchmark, el cual ha sido considerado para la presente investigación. Este permite medir el ancho de banda efectivo acumulado en una Infraestructura de Altas prestaciones.	Software Libre Versión 3.6 + bugfix 3.6.0.1
OpenMPI	Módulo utilizado para la interfaz de paso de mensajes, compilada en el Supercomputador Quinde I.	Software libre
Openbabel	Es una aplicación diseñada para hablar cualquier tipo de lenguajes de datos de química como manejar simulaciones de moléculas, la cual será utilizada para el Caso Práctico (Open Babel , 2016)	Software libre
MXM	La biblioteca MellanoX Messaging (MXM) proporciona mejoras a las bibliotecas de comunicaciones paralelas al utilizar completamente la infraestructura de red subyacente proporcionada por el hardware de conmutador / HCA Mellanox.	Software libre
Putty	Es una aplicación para conexión de acceso remoto al Servicio de Supercomputación mediante SSH. (putty.org, 2020)	Open source

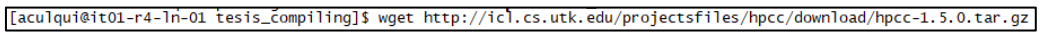
Nota: Se muestra el software a utilizar en la presente investigación. Tomado de las fuentes oficiales.

Ejecución del Análisis de Rendimiento

Para realizar el Análisis de Rendimiento de la Red de Altas prestaciones se consideró las etapas de un Benchmark que se mostró en la Figura 16 en los capítulos anteriores. A continuación se detalla los pasos que se siguieron en la presente investigación:

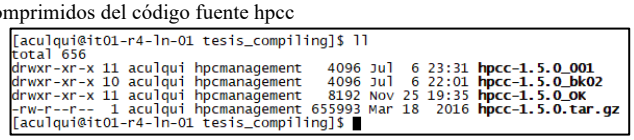
- Ingresar mediante SSH a la IP pública del login node del Supercomputador Quinde I.
- Ingresar credenciales de acceso al Supercomputador Quinde I con la cuenta asignada personal.
- Descargar el código fuente hpcc "hpcc-1.5.0.tar.gz" versión 1.5.0 , para realizar el Análisis de la Red de Altas prestaciones, de la página web https://icl.utk.edu/hpcc/software/index.html, como se muestra en la Figura 17 y 18.

Figura 17. Captura de la descarga del código fuente hpcc



Nota: Realizado por el Autor.

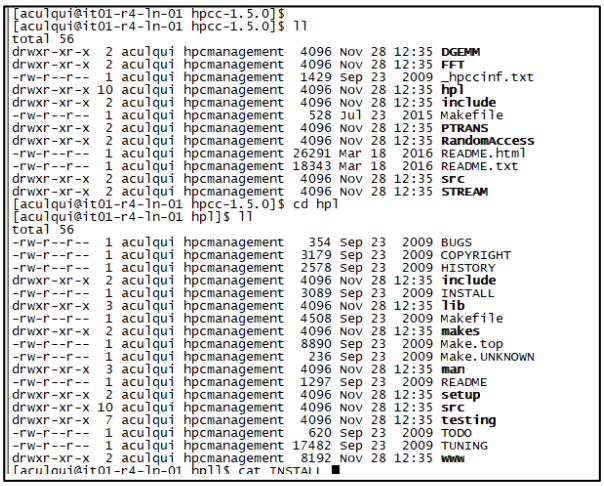
Figura 18. Archivos descomprimidos del código fuente hpcc



Nota: Realizado por el Autor.

- Ir al directorio hpl[18] desde el directorio /hpcc-1.5.0/hpl/INSTALL como se muestra en la Figura 19.

Figura 19. Ir al directorio hpl

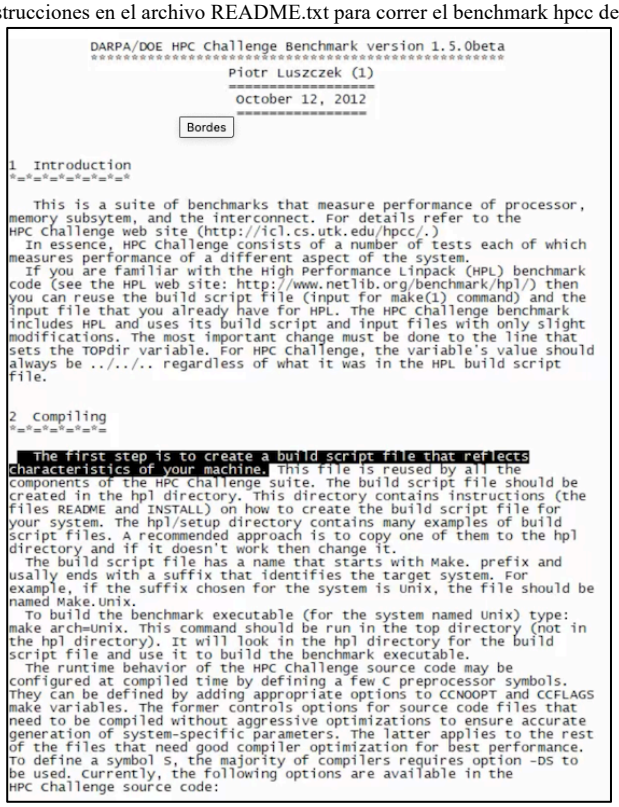


Nota: Realizado por el Autor.

- Revisar las instrucciones a seguir en el archivo README.txt para ejecutar el benchmark hpcc de acuerdo al procedimiento sugerido del HPC Challenge Benchmark como se muestra en la Figura 20.

[18] HPL.- High Performance Linpack, es parte de uno de los test de rendimiento del HPC Challenge Benchmark.

Figura 20. Verificar las instrucciones en el archivo README.txt para correr el benchmark hpcc detallado

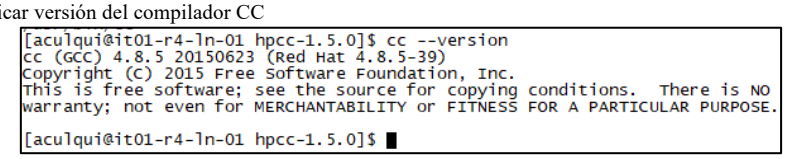


```
          DARPA/DOE HPC Challenge Benchmark version 1.5.0beta
          ***************************************************
                              Piotr Luszczek (1)
                              ==================
                               October 12, 2012
                               ================
                    Bordes

1  Introduction
*=*=*=*=*=*=*=*

   This is a suite of benchmarks that measure performance of processor,
memory subsytem, and the interconnect. For details refer to the
HPC Challenge web site (http://icl.cs.utk.edu/hpcc/.)
  In essence, HPC Challenge consists of a number of tests each of which
measures performance of a different aspect of the system.
  If you are familiar with the High Performance Linpack (HPL) benchmark
code (see the HPL web site: http://www.netlib.org/benchmark/hpl/) then
you can reuse the build script file (input for make(1) command) and the
input file that you already have for HPL. The HPC Challenge benchmark
includes HPL and uses its build script and input files with only slight
modifications. The most important change must be done to the line that
sets the TOPdir variable. For HPC Challenge, the variable's value should
always be ../../.. regardless of what it was in the HPL build script
file.


2  Compiling
*=*=*=*=*=*

   The first step is to create a build script file that reflects
characteristics of your machine. This file is reused by all the
components of the HPC Challenge suite. The build script file should be
created in the hpl directory. This directory contains instructions (the
files README and INSTALL) on how to create the build script file for
your system. The hpl/setup directory contains many examples of build
script files. A recommended approach is to copy one of them to the hpl
directory and if it doesn't work then change it.
  The build script file has a name that starts with Make. prefix and
usally ends with a suffix that identifies the target system. For
example, if the suffix chosen for the system is Unix, the file should be
named Make.Unix.
  To build the benchmark executable (for the system named Unix) type:
make arch=Unix. This command should be run in the top directory (not in
the hpl directory). It will look in the hpl directory for the build
script file and use it to build the benchmark executable.
  The runtime behavior of the HPC Challenge source code may be
configured at compiled time by defining a few C preprocessor symbols.
They can be defined by adding appropriate options to CCNOOPT and CCFLAGS
make variables. The former controls options for source code files that
need to be compiled without aggressive optimizations to ensure accurate
generation of system-specific parameters. The latter applies to the rest
of the files that need good compiler optimization for best performance.
To define a symbol S, the majority of compilers requires option -DS to
be used. Currently, the following options are available in the
HPC Challenge source code:
```

Nota: Realizado por el Autor.

- Para ejecutar el benchmark hpcc.c sobre Power 8 IBM la arquitectura correspondiente a los procesadores del Supercomputador Quinde I, se utilizará el compilador CC disponible como se muestra en la Figura 21.

Figura 21. Verificar versión del compilador CC

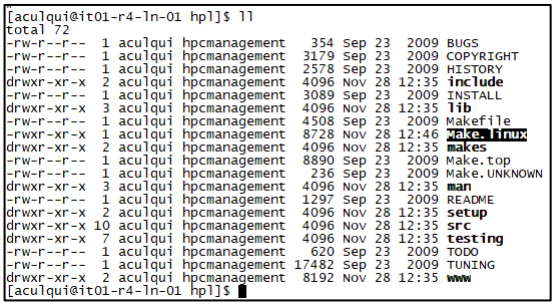


```
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ cc --version
cc (GCC) 4.8.5 20150623 (Red Hat 4.8.5-39)
Copyright (C) 2015 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
```

Nota: Realizado por el Autor.

- Ir al directorio del hpl, para validar el archivo Make, que nos permitirá ejecutar el benchmark de la red, como se muestra en la figura 22.

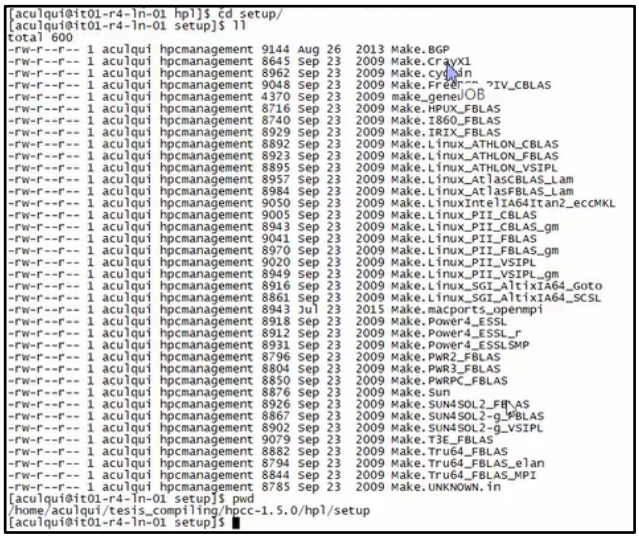
Figura 22. Directorio del hpl.



Nota: Realizado por el Autor.

- Se verifica los archivos Make para compilar hpcc sobre una arquitectura Power 8 como se observa en la figura 23. Hasta la fecha que se ejecutó los test de Rendimiento, no existía un archivo propio para esta arquitectura, por lo cual fue necesario modificar un archivo para la arquitectura Power 8 de IBM.

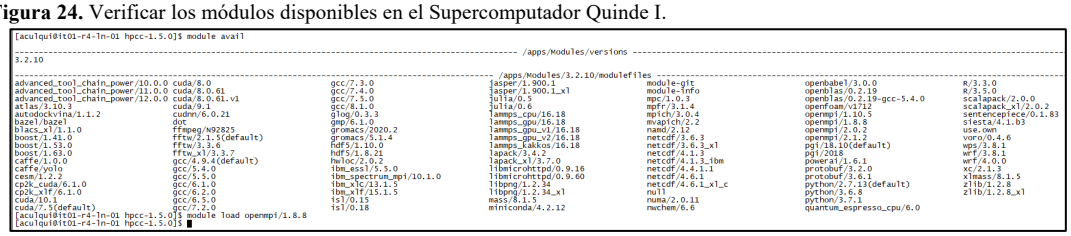
Figura 23. Verificar los archivos Make



Nota: Realizado por el Autor.

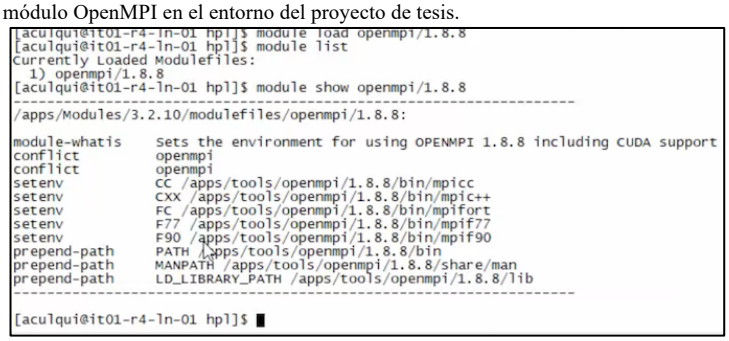
- Se instala el módulo OpenMPI en el entorno del proyecto, el cual nos permitirá enviar mensajes en paralelo sobre los nodos de cómputo, como se muestra en las figuras 24 y 25.

Figura 24. Verificar los módulos disponibles en el Supercomputador Quinde I.



Nota: Realizado por el Autor.

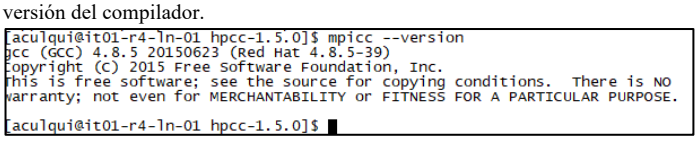
Figura 25. Cargar el módulo OpenMPI en el entorno del proyecto de tesis.



Nota: Realizado por el Autor.

- Verificar la versión del compilador mpicc como se muestra en la Figura 26, con el fin de validar si esta versión es estable y soporte la arquitectura Power 8.

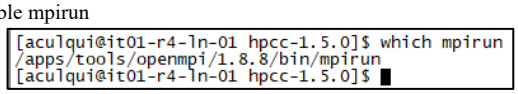
Figura 26. Verificar la versión del compilador.



Nota: Realizado por el Autor.

- Verificar la ubicación del ejecutable del mpirun como se muestra en la Figura 27, para que el path se coloque en el script del Make de la arquitectura Power 8.

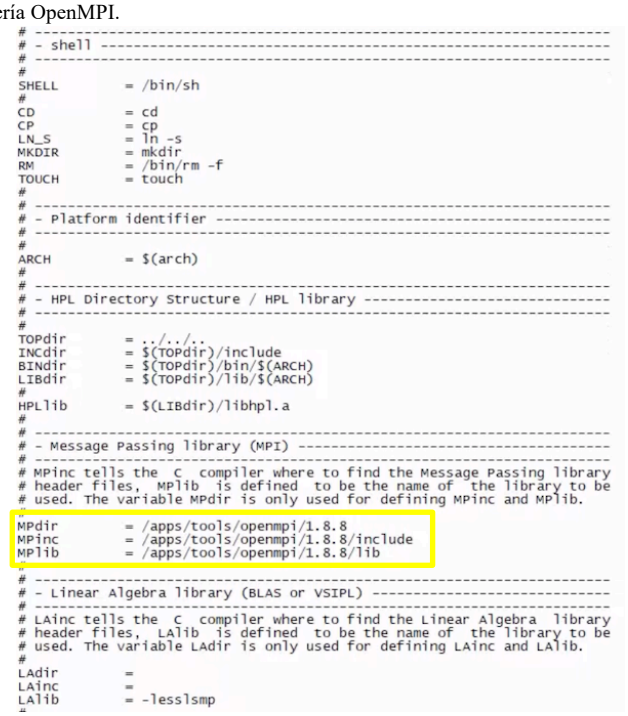
Figura 27. Ubicación del ejecutable mpirun



Nota: Realizado por el Autor.

- Se modifica el archivo "Make" donde se agrega la librería MPI, que soporta la arquitectura Power 8 del Supercomputador Quinde I; se ajusta el archivo para ejecutar los test de rendimiento de la Red de Altas Prestaciones, como se muestra en la Figura 28.

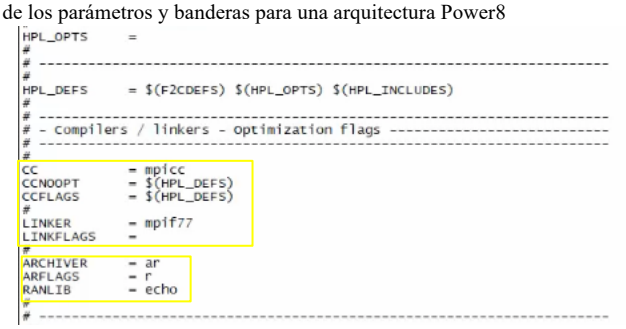
Figura 28. Agregar la librería OpenMPI.



Nota: Realizado por el Autor.

- Se modifica el archivo Make.Power8_ESSLSMP con los parámetros estándar de paralelización para ejecutar el test sobre la interfaz MPI para una arquitectura Power 8 de IBM, como se muestra en la Figura 29 y 30.

Figura 29. Configuración de los parámetros y banderas para una arquitectura Power8

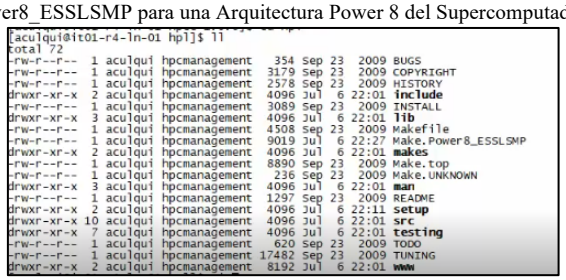


```
HPL_OPTS     =
#
# ----------------------------------------------------------------------
#
HPL_DEFS     = $(F2CDEFS) $(HPL_OPTS) $(HPL_INCLUDES)
#
# ----------------------------------------------------------------------
# - Compilers / linkers - Optimization flags ---------------------------
# ----------------------------------------------------------------------
#
CC           = mpicc
CCNOOPT      = $(HPL_DEFS)
CCFLAGS      = $(HPL_DEFS)
#
LINKER       = mpif77
LINKFLAGS    =
#
ARCHIVER     = ar
ARFLAGS      = r
RANLIB       = echo
#
# ----------------------------------------------------------------------
```

Nota: Realizado por el Autor.

- CC = mpicc
  - **Mpicc:** Comando que permite ejecutar el test utilizando el módulo OpenMPI para ambientes paralelos
- **CCFLAGS = $ (HPL_DEFS)** Se deja la configuración por defecto.
- **LINKER = mpif77**
  - **Mpif77:** comando para incorporar automáticamente las bibliotecas MPI como base el compilador fortran.
- LINKFLAGS = No require ninguna bandera para Power8
- ARCHIVER= ar, por defecto
- ARFLAGS= r, por defecto
- RANLIB= echo, por defecto

Figura 30. Archivo Make.Power8_ESSLSMP para una Arquitectura Power 8 del Supercomputador.



Nota: Realizado por el Autor.

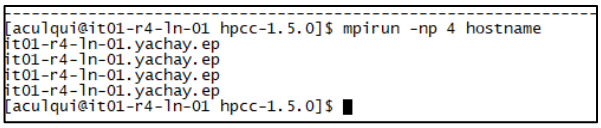
- Una vez editado el archivo Make para la arquitectura Power8, se ejecuta el archivo como se muestra en la Figura 31, con el comando make arch=Power8_ESSLSMP.
- Se valida que no exista errores al final de la compilación como se muestra en la figura 31.

Figura 31. Archivo Make.Power8_ESSLSMP

Nota: Realizado por el Autor.

- La forma exacta de ejecutar un job depende de la implementación de MPI y los detalles del sistema. En la Figura 32 se muestra un ejemplo de ejecución de un job.

Figura 32. Ejemplo de un job sobre la interfaz MPI.



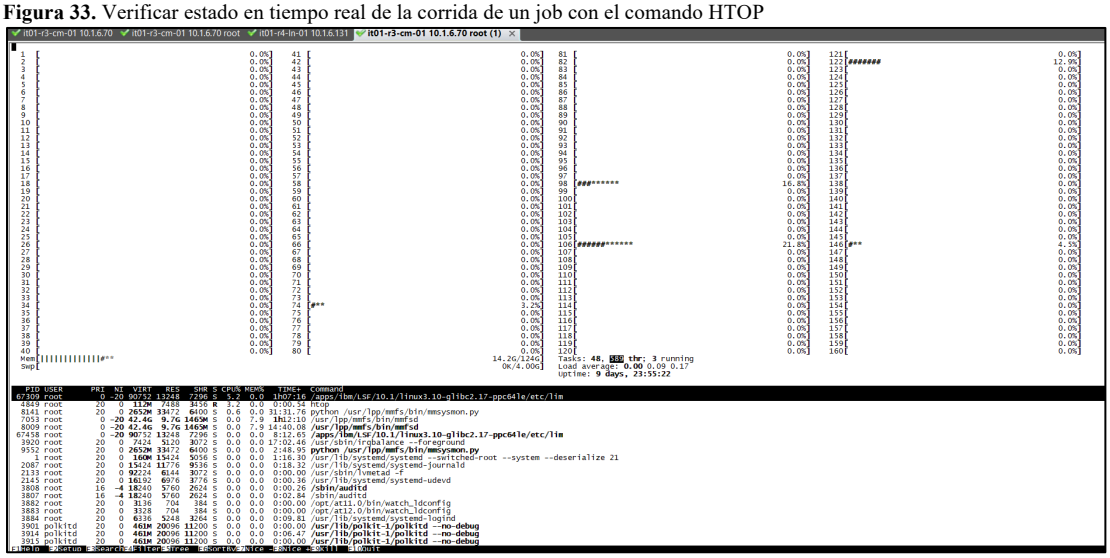
Nota: Realizado por el Autor.

Significado del comando:

- **Mpirun:** Es el comando que inicia la ejecución de un código MPI. Dependiendo del sistema, esto podría también ser aprun, mpiexec, poe, o alguno apropiado para el sistema, en el caso del Supercomputador Quinde I, se utilizará el comando mpirun para arquitecturas Power 8.
- **- Np 4 (number proccesor o número de procesadores):** Es el argumento que especifica 4 procesos MPI podrían ser iniciados.
- **"Hostname":** es el nombre del trabajo o acción a ejecutar.

- Para verificar la ejecución del job en tiempo real, se utiliza el comando htop como se muestra en la Figura 33.

Figura 33. Verificar estado en tiempo real de la corrida de un job con el comando HTOP



Nota: Realizado por el Autor.

- Se ejecuta el benchmark hpcc como se muestra en la Figura 34.

Figura 34. Ejecutar el benchmark hpcc sobre 4 procesadores.

```
Currently Loaded Modulefiles:
  1) numa/2.0.11                2) ibm_spectrum_mpi/10.1.0
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ mpirun -np 4 ./hpcc
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
```

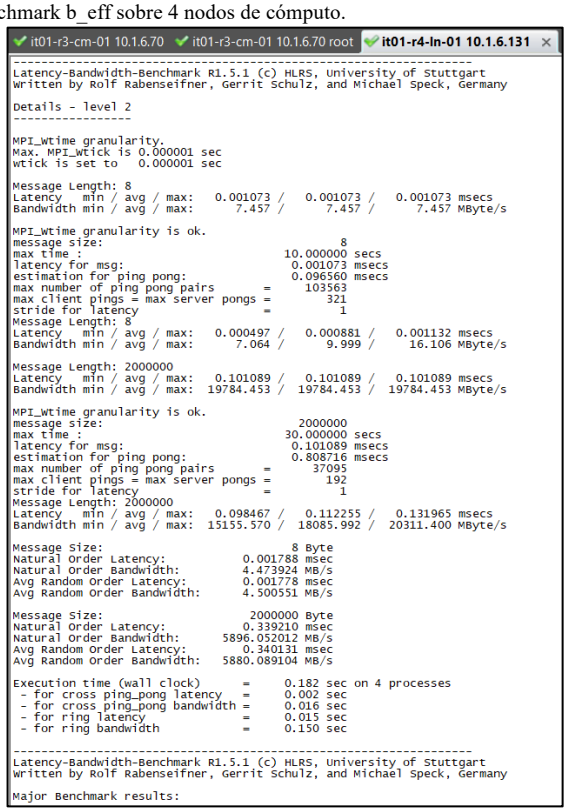
Nota: Realizado por el Autor.

Este último paso se repite con la cantidad de procesadores que se requiera hacer el benchmark hpcc.

- El número de procesos MPI debe ser lo suficientemente grande para acomodar toda la grilla de procesos especificados en el archivo hpccinf.txt. Este archivo no debe ser modificado con el fin de cumplir el HPC Challenge Benchmark.
- Una vez que se ejecuta el benchmark, se crea un archivo llamado hpccountf.txt. El cual contiene los resultados del benchmark.

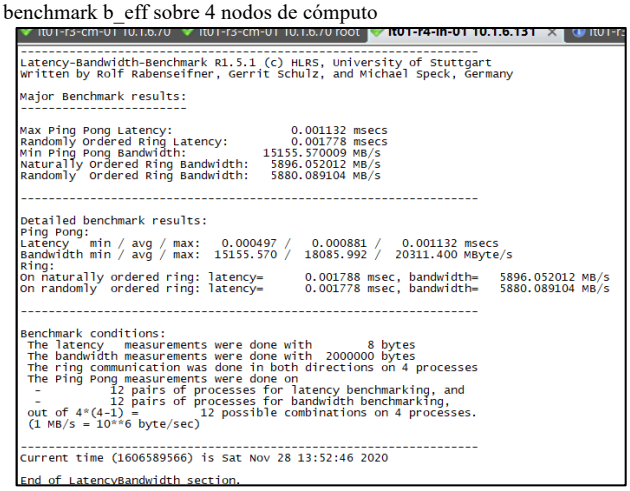
- Se verifica los resultados del benchmark b_eff como se muestra en la Figura 35 y 36. El archivo contiene los resultados del Benchmark b_eff a nivel de Latencia y Ancho de Banda.

Figura 35. Resultados del benchmark b_eff sobre 4 nodos de cómputo.



Nota: Realizado por el Autor.

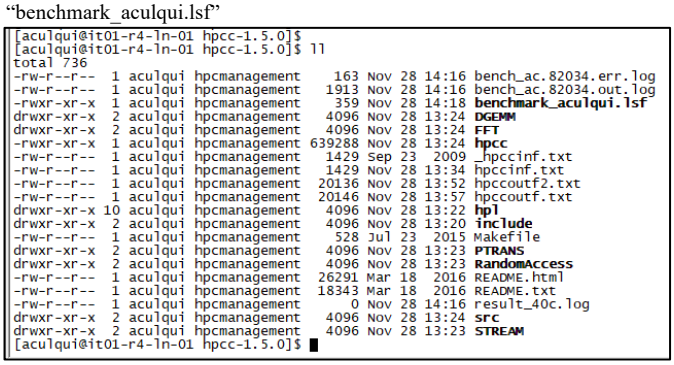
Figura 36. Resultados del benchmark b_eff sobre 4 nodos de cómputo



Nota: Realizado por el Autor.

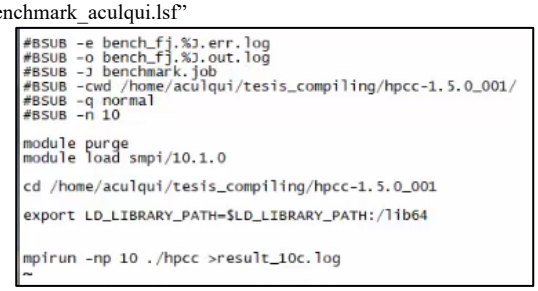
- Considerando que el Supercomputador cuenta con un balanceador de carga de Jobs de cómputo llamado LSF; por lo cual se procedió a configurar un archivo llamado “benchmark_aculqui.lsf”, con el fin de enviar sobre varios nodos de cómputo los jobs y balancear la carga de los procesos de cómputo como se muestra en las Figuras 37 y 38.

Figura 37. Archivo LSF “benchmark_aculqui.lsf”



Nota: Realizado por el Autor.

Figura 38. Editar el archivo “benchmark_aculqui.lsf”



Nota: Realizado por el Autor.

El significado de los comandos y parámetros del archivo “benchmark_aculqui.lsf”: (IBM, 2021)

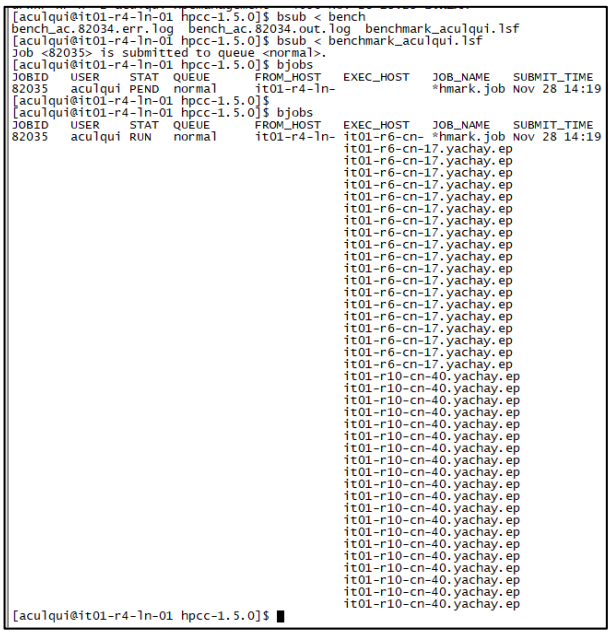
- BSUB.- es el comando que envía un job al LSF ejecutando el comando especificado y sus argumentos.

- Bsub -e bench_fj.%J.err.log: Especifica el nombre del archivo para guardar como log todos los errores que se presenten en la corrida del job.
- Busb -o bench_fj.%J.out.log: El nombre del archivo para guardar como log los resultados del job.
- Bsub -J benchmark.job: Colocar un nombre al job a ejecutar.
- Bsub -cwd /home/aculqui/tesis_compiling/hpcc-1.5.0_001/: Indica el path donde se va a correr el job.
- Bsub -q normal: Indica la cola de procesos en el LSF, se configura la cola normal ya que esta dispone de mayor cantidad de procesadores para los Jobs.
- Busb -n 10: Se reserva la cantidad de procesadores para el job a ejecutar.
- Module purge: Purgar todos los módulos cargados, como buena práctica para ejecutar un nuevo job.
- Module load smpi/10.1.0: Se recomienda cargar el módulo SpectrumMPI para aplicaciones que soporten esta versión (Para la presente investigación se utilizó la librería OpenMPI por motivos de comparación y análisis).
- Cd /home/aculqui/tesis_compiling/hpcc-1.5.0_001: Se indica el path donde se encuentra el programa HPCC 1.5.0
- Export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/lib64: Se asigna a la variable de entorno LD_LIBRARY_PATH la ruta de las librerías de 64 bits del Sistema Operativo.

- mpirun -np 10 ./hpcc: Es el comando que inicia la ejecución del job sobre 10 procesadores.
- > result_10c.log: se agrego adicionalmente para que los resultados sean almacenados en un archivo log llamado result_10c.log para mayor facilidad de identificación de los archivos del proyecto de investigación.

- Ejecutar el job sobre lsf con el comando bsub < benchmark_aculqui.lsf como se muestra en la Figura 39.

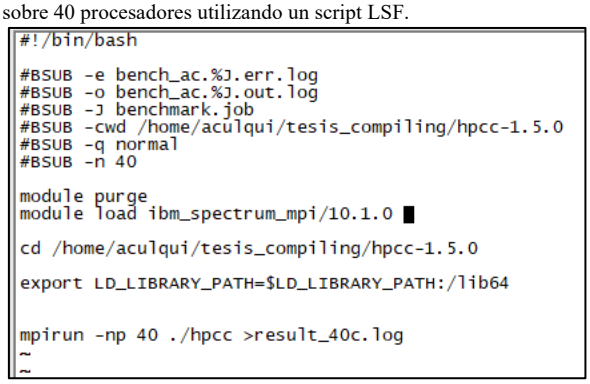
Figura 39. Ejecución del job sobre LSF.



Nota: Realizado por el Autor.

- Editar el archivo LSF para ejecutar el test sobre 40 procesadores como se muestra en la Figura 40.

Figura 40. Ejecución del job sobre 40 procesadores utilizando un script LSF.

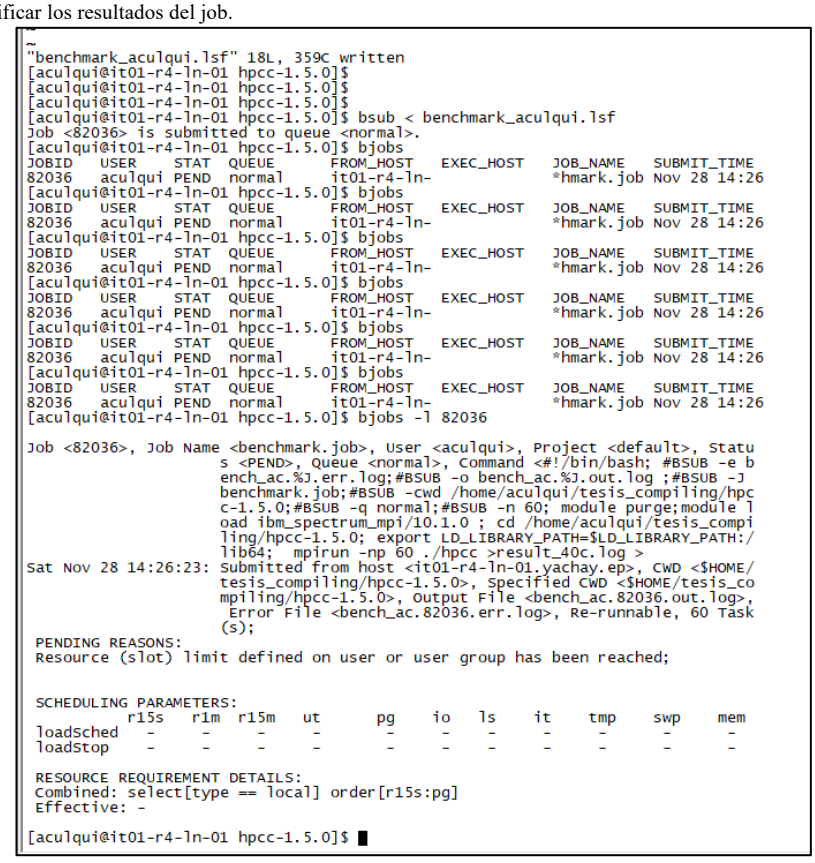


```
#!/bin/bash

#BSUB -e bench_ac.%J.err.log
#BSUB -o bench_ac.%J.out.log
#BSUB -J benchmark.job
#BSUB -cwd /home/aculqui/tesis_compiling/hpcc-1.5.0
#BSUB -q normal
#BSUB -n 40

module purge
module load ibm_spectrum_mpi/10.1.0

cd /home/aculqui/tesis_compiling/hpcc-1.5.0

export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/lib64


mpirun -np 40 ./hpcc >result_40c.log
~
~
```

Nota: Realizado por el Autor.

- Ejecutar el job con el comando “bsub” para correr los test de Rendimiento de la Red de Altas Prestaciones utilizando LSF como se muestra en la figura 41.

Figura 41. Verificar los resultados del job.



```
~
"benchmark_aculqui.lsf" 18L, 359C written
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bsub < benchmark_aculqui.lsf
Job <82036> is submitted to queue <normal>.
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
82036   aculqui PEND  normal     it01-r4-ln-             *hmark.job Nov 28 14:26
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ bjobs -l 82036

Job <82036>, Job Name <benchmark.job>, User <aculqui>, Project <default>, Statu
                     s <PEND>, Queue <normal>, Command <#!/bin/bash; #BSUB -e b
                     ench_ac.%J.err.log;#BSUB -o bench_ac.%J.out.log ;#BSUB -J
                     benchmark.job;#BSUB -cwd /home/aculqui/tesis_compiling/hpc
                     c-1.5.0;#BSUB -q normal;#BSUB -n 60; module purge;module l
                     oad ibm_spectrum_mpi/10.1.0 ; cd /home/aculqui/tesis_compi
                     ling/hpcc-1.5.0; export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/
                     lib64;  mpirun -np 60 ./hpcc >result_40c.log >
Sat Nov 28 14:26:23: Submitted from host <it01-r4-ln-01.yachay.ep>, CWD <$HOME/
                     tesis_compiling/hpcc-1.5.0>, Specified CWD <$HOME/tesis_co
                     mpiling/hpcc-1.5.0>, Output File <bench_ac.82036.out.log>,
                      Error File <bench_ac.82036.err.log>, Re-runnable, 60 Task
                     (s);
 PENDING REASONS:
 Resource (slot) limit defined on user or user group has been reached;


 SCHEDULING PARAMETERS:
           r15s   r1m  r15m   ut      pg    io   ls    it    tmp    swp    mem
 loadSched   -     -     -     -       -     -    -     -     -      -      -
 loadStop    -     -     -     -       -     -    -     -     -      -      -

 RESOURCE REQUIREMENT DETAILS:
 Combined: select[type == local] order[r15s:pg]
 Effective: -

[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
```

Nota: Realizado por el Autor.

- Verificar los resultados del test en el archivo "*result_40c.log"*como se muestra en la Figura 42, el cual contiene los datos del ancho de banda y latencia de la Red de Altas Prestaciones.
- Se recomienda realizar una copia del archivo original donde se guarda los resultados de cada test como se muestra en la Figura 42 con el comando mv hpccountf.txt hpccount_40c.txt

Figura 42. Copiar el archivo de los resultados del job a otro archivo de texto.

```
culqui@it01-r4-ln-01 hpcc-1.5.0]$ mv hpccoutf.txt hpccoutf_40c.txt
culqui@it01-r4-ln-01 hpcc-1.5.0]$ ll
tal 968
w-r--r--  1 aculqui hpcmanagement    163 Nov 28 14:16 bench_ac.82034.err.log
w-r--r--  1 aculqui hpcmanagement   1913 Nov 28 14:16 bench_ac.82034.out.log
w-r--r--  1 aculqui hpcmanagement    132 Nov 28 14:20 bench_ac.82035.err.log
w-r--r--  1 aculqui hpcmanagement   1931 Nov 28 14:20 bench_ac.82035.out.log
wxr-xr-x  1 aculqui hpcmanagement    359 Nov 28 14:18 benchmark_aculqui.lsf
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:24 DGEMM
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:24 FFT
wxr-xr-x  1 aculqui hpcmanagement 639288 Nov 28 13:24 hpcc
w-r--r--  1 aculqui hpcmanagement   1429 Sep 23  2009 _hpccinf.txt
w-r--r--  1 aculqui hpcmanagement   1429 Nov 28 13:34 hpccinf.txt
w-r--r--  1 aculqui hpcmanagement  20136 Nov 28 13:52 hpccoutf2.txt
w-r--r--  1 aculqui hpcmanagement  40304 Nov 28 14:20 hpccoutf_40c.txt
wxr-xr-x 10 aculqui hpcmanagement   4096 Nov 28 13:22 hpl
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:20 include
w-r--r--  1 aculqui hpcmanagement    528 Jul 23  2015 Makefile
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:23 PTRANS
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:23 RandomAccess
w-r--r--  1 aculqui hpcmanagement  26291 Mar 18  2016 README.html
w-r--r--  1 aculqui hpcmanagement  18343 Mar 18  2016 README.txt
w-r--r--  1 aculqui hpcmanagement      0 Nov 28 14:19 result_40c.log
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:24 src
wxr-xr-x  2 aculqui hpcmanagement   4096 Nov 28 13:23 STREAM
```

Nota: Realizado por el Autor.

- Los últimos pasos se repiten variando la cantidad de procesadores en cada benchmark en el archivo "benchmark_aculqui.lsf".
- Una vez corrido varios test en distintos escenarios, se recopila los resultados obtenidos para proceder a graficar cada escenario y analizar los datos de latencia y ancho de banda.

Resultados de los Escenarios

Una vez detallado el procedimiento del benchmark HPCC utilizado para analizar el Rendimiento de la Red de Altas Prestaciones basado el HPC Challenge Benchmark, y la interconexión de los nodos de cómputo del Supercomputador Quinde

I, se seleccionó ciertos escenarios de prueba que permitirán ver el comportamiento de la Red en función del ancho de banda y latencia.

Además, se debe considerar que cada escenario tiene ya configurado por defecto las librerías MXM de Mellanox que nos permite analizar el comportamiento de la Red InfiniBand.

A continuación, se realiza algunos test de Rendimiento de la Red de Altas prestaciones sobre varios escenarios con la aplicación Effective Bandwidth (b_eff) Benchmark indicados en la página https://fs.hlrs.de/projects/par/mpi//b_eff/:

Escenario 1

Para el escenario 1 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 2

En la Figura 43 se muestra los resultados de la latencia ping-pong con 1.905 microsegundos por mensaje, un ancho de banda ping-pong de 8774, MBytes/s y un ancho de banda efectivo de 3754.570MB/s.

Figura 43. Cuadro de resultados del test de rendimiento de la Red con la aplicación b_eff para 2 procesadores.

Thu Mar 12 23:52:57 2020 on Linux it01-r6-cn-13.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

$\mathbf{b_{eff}} = \mathbf{3754.570\ MB/s}$ = 1877.285 * 2 PEs with 128 MB/PE

	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	2	3755	1 MB	16107	16108	2.340	2.283	1.905	8774
per process		1877	8053	8054					

Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 1.905 microsec per message Bandwidth: 8774.219 MB/s (with MB/s = 10^6 byte/s)

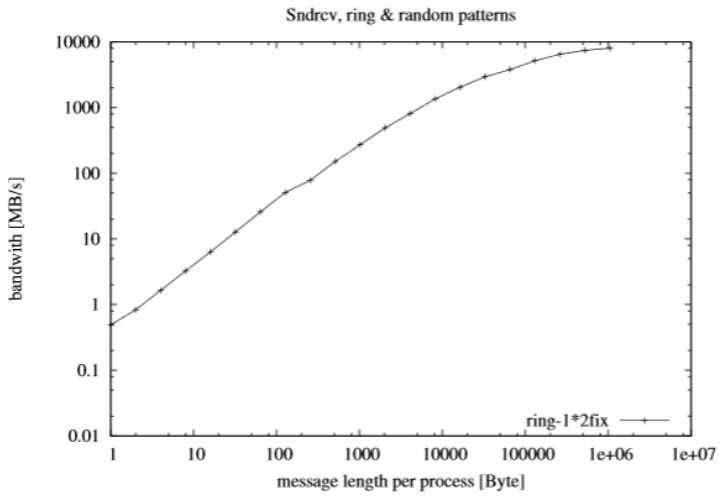
Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

Descripción de la tabla de resultados:

- Lmax: Configuración propia del benchmark b_eff para realizar el test de cada escenario; el cual establece un valor de cache mínimo en 1MB con el fin que pueda compararse con otras infraestructuras de computación paralela en ambientes similares.
- B_eff: Benchmark de Ancho de Banda Efectivo de la Red de Altas prestaciones, este dato es parte del análisis de la presente investigación.
- Latencias (rings&random y rings only): son datos de la latencia de la Red, que arroja el benchmark b_eff en función del tipo de patrón de comunicación utilizado para cada test, estos datos no son parte del análisis principal de la presente investigación.
- Latencia ping-pong: es el tiempo requerido para enviar mensajes de 8 bytes desde un nodo a otro, el cual es motivo de análisis en la presente investigación.
- Ancho de Banda ping-pong: es aquel que se obtiene a través del tamaño del mensaje divido para el tiempo que toma en transmitir los mensajes sobre MPI, el cual es motivo de análisis en la presente investigación.

En la Figura 44 se presenta la gráfica de los resultados del Escenario 1, en función del tamaño del mensaje por proceso y el Ancho de Banda.

Figura 44. Gráfica de los resultados del test de rendimiento con la aplicación b_eff para 2 procesadores.



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En el escenario 1 se puede observar que a mayor flujo de información o aumento de información a ser procesada, el ancho de banda se incrementa para su mejor desenvolvimiento asegurando así que los procesos se puedan ejecutar de la mejor forma posible, pero el detalle aquí es que en este escenario no se hace uso de la red InfiniBand esto debido a que todo el procesamiento se está realizando sobre el mismo nodo de cómputo, eso quiere decir que está usando el ancho de banda del bus de información entre los procesadores, memoria, y Caché.

Escenario 2

En el escenario 2 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 4

En la Figura 45 se muestra los resultados del benchmark con 4 procesadores.

Figura 45. Resultados del test de rendimiento utilizando el benchmark b_eff con 4 procesadores.

Effective Bandwith Benchmark (b_{eff}) Version 3.5

High-Performance Computing-Center Stuttgart, HLRS

Thu Mar 12 23:47:53 2020 on Linux it01-r6-cn-12.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

b_{eff} = 6136.625 MB/s = 1534.156 * 4 PEs with 128 MB/PE

	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	4	6137	1 MB	22932	20523	1.896	1.920	1.593	8821
per process		1534	5733	5131					

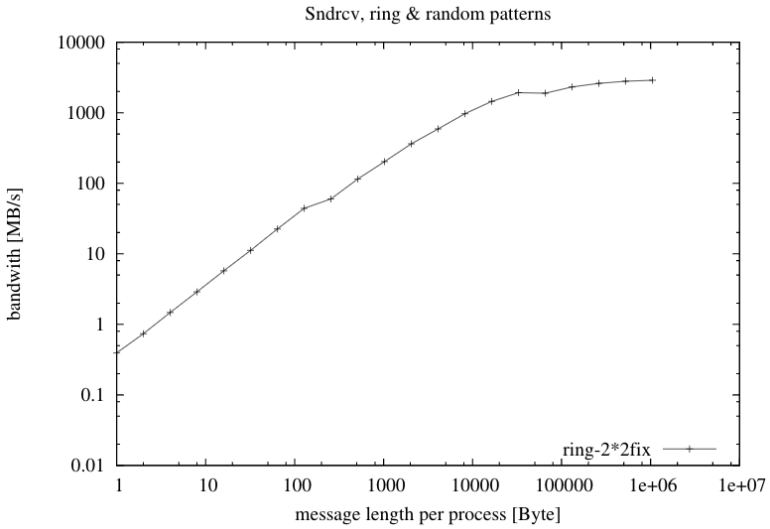
Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 1.593 microsec per message Bandwidth: 8821.407 MB/s (with MB/s = 10^6 byte/s)

Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En la figura 45, podemos observar que tenemos los resultados del benchmark b_eff con 4 procesadores sobre MPI para ambientes en paralelo. La red InfiniBand presenta un Ancho de banda efectivo de 6136.625MB/s, ancho de banda ping-pong de 8821 Mbyte/s y una latencia ping-pong de 1.593 microsegundos por mensaje.

En la Figura 46 se presenta la gráfica de los resultados del Escenario 2, en función del tamaño del mensaje por proceso y el Ancho de Banda.

Figura 46. Gráfica de los resultados del test de rendimiento con la aplicación b_eff



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En el escenario 2 se puede observar que a mayor flujo de información o tamaño de los mensajes a ser procesada, el ancho de banda se incrementa para mejor desenvolvimiento asegurando así que los procesos se puedan ejecutar de la mejor forma posible, pero el detalle aquí es que en este escenario no se hace uso de la red InfiniBand esto debido a que todo el procesamiento se está realizando sobre el mismo nodo de cómputo, eso quiere decir que está usando el ancho de banda del bus de información entre los procesadores, memoria, y Caché.

Escenario 3

En el escenario 3 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 16

En la Figura 47 se muestra los resultados del benchmark con 16 procesadores.

Figura 47. Resultados del test de rendimiento utilizando la aplicación benchmark b_eff con 16 procesadores.

Effective Bandwith Benchmark (b_{eff}) Version 3.5

High-Performance Computing-Center Stuttgart, HLRS

Mon Mar 9 11:12:45 2020 on Linux it01-r14-cn-63.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

b_{eff} = 16263.114 MB/s = 1016.445 * 16 PEs with 128 MB/PE

	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	16	16263	1 MB	45370	40448	1.771	1.824	0.796	15802
per process		1016	2836	2528					

Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 0.796 microsec per message Bandwidth: 15802.041 MB/s (with MB/s = 10^6 byte/s)

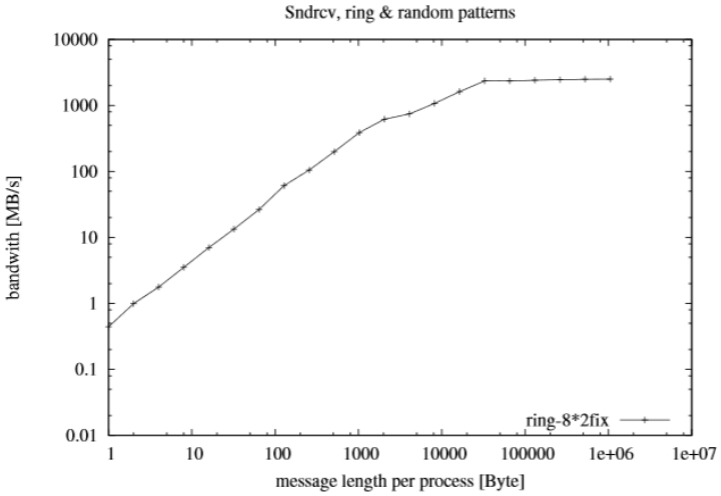
Nota: Realizado por el Autor utilizando el editor Overleaf[19] en formato Latex.

En la figura 47, podemos observar que tenemos los resultados del benchmark b_eff con 16 procesadores sobre MPI para ambientes en paralelo. La red InfiniBand presenta un Ancho de banda efectivo de 16263.114MB, ancho de banda ping-pong de 15802Mbyte/s y una latencia de 0,796 microsegundos por mensaje.

En la Figura 48 se presenta la gráfica de los resultados del Escenario 3, en función del tamaño del mensaje por proceso y el Ancho de Banda.

[19] Overleaf es un editor colaborativo de LaTeX basado en la nube que se utiliza para escribir, editar y publicar documentos científicos. Se asocia con una amplia gama de editores científicos para proporcionar plantillas oficiales de LaTeX para revistas y enlaces de envío directo.

Figura 48. Gráfica de los resultados del test de rendimiento con la aplicación b_eff



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En el escenario 3 se puede observar que a mayor flujo de información o aumento de información a ser procesada, el ancho de banda se incrementa para su mejor desenvolvimiento asegurando así que los procesos se puedan ejecutar de la mejor forma posible, pero el detalle aquí es que en este escenario no se hace uso de la red InfiniBand esto debido a que todo el procesamiento se está realizando sobre el mismo nodo, eso quiere decir que está usando el ancho de banda del bus de información entre los procesadores, memoria, y Caché.

Escenario 4

En el escenario 4 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 32

En la Figura 49 se muestra los datos del resultado del benchmark con 32 procesadores. La red InfiniBand presenta un Ancho de banda efectivo de 22739.318 MB, ancho de banda ping-pong de 16437 Mbyte/s y una latencia ping-pong de 0,790 microsegundos por mensaje.

Figura 49. Resultados del test de rendimiento utilizando el benchmark b_eff con 32 procesadores

Effective Bandwith Benchmark (b_{eff}) Version 3.5

High-Performance Computing-Center Stuttgart, HLRS

Mon Mar 9 23:44:02 2020 on Linux it01-r14-cn-63.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

b_{eff} = 22739.318 MB/s = 710.604 * 32 PEs with 128 MB/PE

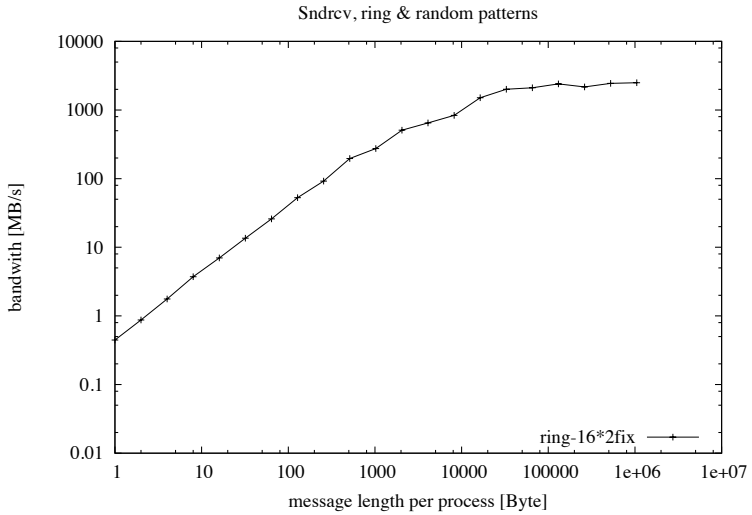
	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	32	22739	1 MB	61864	80581	2.107	1.946	0.790	16437
per process		711	1933	2518					

Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 0.790 microsec per message Bandwidth: 16436.906 MB/s (with MB/s = 10^6 byte/s)

Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En la Figura 50 se presenta la gráfica de los resultados del Escenario 2, en función del Ancho de Banda por el tamaño del mensaje por proceso.

Figura 50. Gráfica de los resultados del test de rendimiento con la aplicación b_eff para 32 procesadores.



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

El escenario 4 es más complejo que los anteriores y a la vez permite observar el uso de la red InfiniBand, esto debido a que se está procesando entre dos nodos de cómputo diferentes, por lo tanto, se puede deducir lo siguiente, en el escenario 1, la latencia es de 1.905 microsegundos por mensaje, mientras que al realizar el procesamiento entre dos nodos y estos interconectados a través de la red InfiniBand, la latencia es de 0.790 microsegundos, esto nos da a entender que a pesar de estar en nodos separados el performance de la red InfiniBand es mucho mejor que el escenario 1, presentando un escenario más óptimo en cuestión de latencia.

Escenario 5

En el escenario 5 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 64

En la Figura 51 se muestra los datos del resultado del benchmark con 64 procesadores. La red InfiniBand presenta un Ancho de banda efectivo de 62367.350 MB/s, ancho de banda ping-pong de 20682 Mbyte/s y una latencia ping-pong de 0,787 microsegundos por mensaje.

Figura 51. Gráfica de los resultados del test de rendimiento con la aplicación b_eff para 96 procesadores

Effective Bandwith Benchmark (b_{eff}) Version 3.5

High-Performance Computing-Center Stuttgart, HLRS

Sun Mar 15 23:34:36 2020 on Linux it01-r10-cn-39.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

b_{eff} = 62367.350 MB/s = 974.490 * 64 PEs with 128 MB/PE

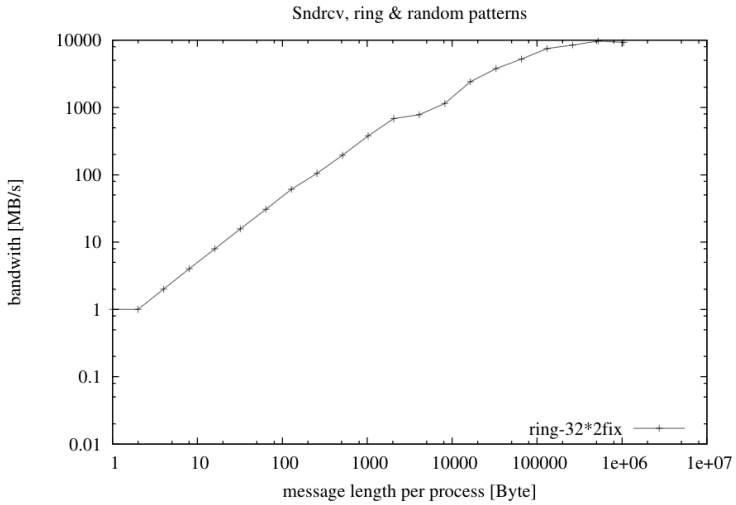
	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	64	62367	1 MB	180979	250604	1.918	1.699	0.787	20682
per process		974	2828	3916					

Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 0.787 microsec per message Bandwidth: 20682.090 MB/s (with MB/s = 10^6 byte/s)

Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En la Figura 52 se presenta la gráfica de los resultados del Escenario 5, en función del Ancho de Banda por el tamaño del mensaje por proceso.

Figura 52. Gráfica estadística de los resultados del test de rendimiento con la aplicación b_eff para 64 procesadores.



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

El escenario 5 permite observar el uso de la red InfiniBand, esto debido a que se está procesando entre varios nodos diferentes, por lo tanto, se puede deducir lo siguiente, en el escenario 1, la latencia es de 1.905 microsegundos, mientras que al realizar el procesamiento entre varios nodos y estos interconectados a través de la red InfiniBand, la latencia es de 0.787 microsegundos, esto nos da a entender que a pesar

de estar en nodos separados el performance de la red InfiniBand mejora siendo más óptimo por uno pocos microsegundos.

Escenario 6

En el escenario 6 se consideró los siguientes parámetros:

- Modulo OpenMPI versión 1.8.8
- Benchmark hpcc versión 1.5.0
- Cantidad de procesadores: 96

En la Figura 53 se muestra los datos del resultado del benchmark con 96 procesadores. La red InfiniBand presenta un Ancho de banda efectivo de 90757.993 MB/s, ancho de banda ping-pong de 19547 Mbyte/s y una latencia ping-pong de 0.788 microsegundos por mensaje.

Figura 53. Gráfica de los resultados del test de rendimiento con la aplicación b_eff para 96 procesadores

Sun Mar 15 23:46:44 2020 on Linux it01-r10-cn-36.yachay.ep 3.10.0-327.el7.ppc64le 1 SMP Thu Oct 29 17:31:13 EDT 2015 ppc64le

$\mathbf{b_{eff}} = \mathbf{90757.993\ MB/s} = 945.396$ * 96 PEs with 128 MB/PE

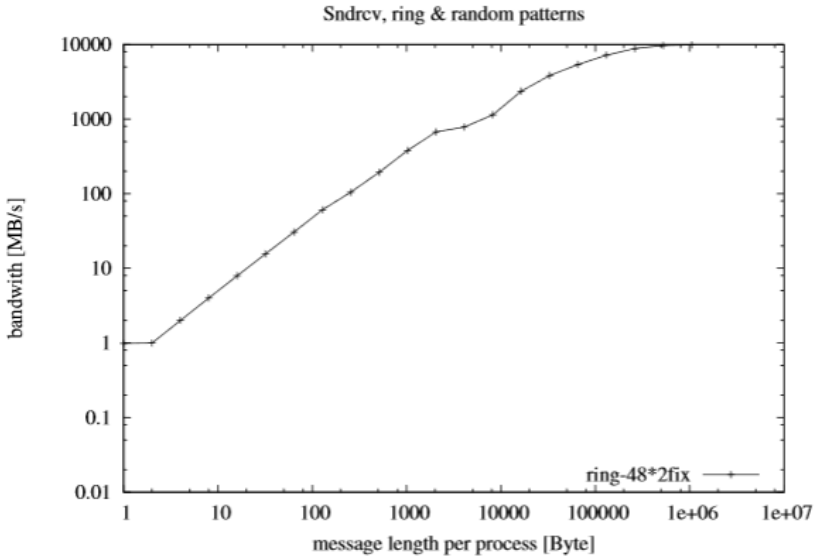
	number of processors	b_{eff} MByte/s	Lmax	b_{eff} at Lmax rings& random MByte/s	b_{eff} at Lmax rings only MByte/s	Latency rings& random mircosec	Latency rings only microsec	Latency ping-pong microsec	ping-pong bandwidth MByte/s
accumulated	96	90758	1 MB	256298	324727	1.961	1.744	0.788	19547
per process		945	2670	3383					

Ping-Pong result (only the processes with rank 0 and 1 in MPI_COMM_WORLD were used):
Latency: 0.788 microsec per message Bandwidth: 19546.873 MB/s (with MB/s = 10^6 byte/s)

Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

En la Figura 54 se presenta la gráfica de los resultados del Escenario 6, en función del Ancho de Banda por el tamaño del mensaje por proceso.

Figura 54. Gráfica estadística de los resultados del test de rendimiento con la aplicación b_eff para 96 procesadores.



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

El escenario 6 permite observar el uso de la red InfiniBand, esto debido a que se está procesando entre 4 nodos de cómputo diferentes, por lo tanto, se puede deducir lo siguiente, en el escenario 1, la latencia es de 1.905 microsegundos, mientras que al realizar el procesamiento entre varios nodos de cómputo y estos interconectados a través de la red InfiniBand, la latencia es de 0.788 microsegundos, esto nos da a entender que a pesar de estar en nodos separados el performance de la red InfiniBand mejora siendo más óptimo por uno pocos microsegundos. Se nota una diferencia mínima en comparación al escenario 4 donde utiliza 32 procesadores en dos nodos de cómputo.

Además, se observa que la latencia de la Red de Altas prestaciones se mantiene dentro de un margen de 0.787 a 0.79 microsegundos por mensaje al utilizar la red Infiniband entre varios nodos de cómputo, mientras que al ejecutar los Jobs dentro de un mismo nodo de cómputo la latencia es casi el doble como se presentó en el escenario 1 de 1.905 microsegundos por mensaje.

Análisis Comparativo

Resultados de los escenarios del Benchmark de la Red InfiniBand

En la tabla 9, se puede observar un resumen de los datos de cada escenario considerado en el análisis de rendimiento de la red de Altas prestaciones InfiniBand, mediante la aplicación b_eff:

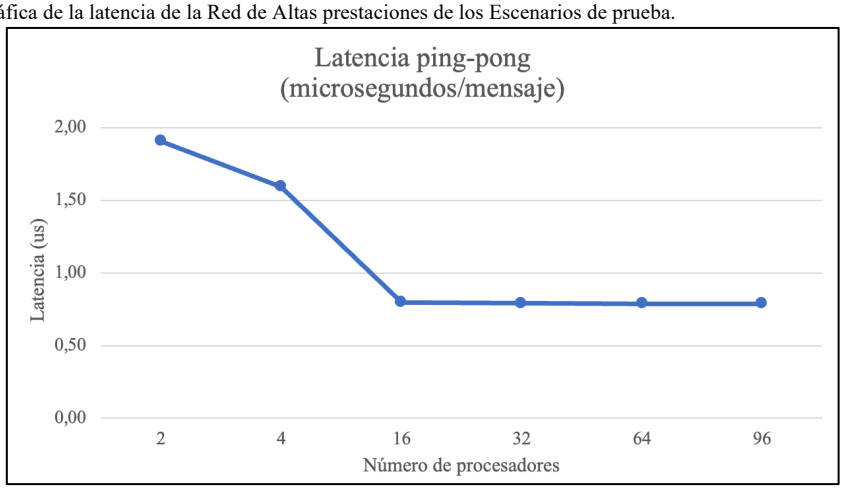
Tabla 9. Cuadro comparativo de los escenarios de prueba realizados a la Red de Altas Prestaciones InfiniBand.

Escenario	Cantidad de procesadores	Cantidad de nodos de cómputo	Bandwidth b_eff (Mbyte/s)	Latencia (microsegundos)	Bandwidth Ping/pong (MByte/s)
1	2	1	3754.570	1.905	8774
2	4	1	6136.625	1.593	8821
3	16	1	16263.114	0.796	15802
4	32	2	22739.318	0.79	16437
5	64	4	62367.350	0.787	20682
6	96	5	90757.993	0.788	19547

Nota: Realizado por el Autor.

En la Figura 55 se presenta la gráfica de los resultados de los seis escenarios ejecutados del Análisis de Rendimiento de la Red de Altas prestaciones, en función de la latencia ping-pong con la cantidad de procesadores utilizados.

Figura 55. Gráfica de la latencia de la Red de Altas prestaciones de los Escenarios de prueba.



Nota: Realizado por el Autor.

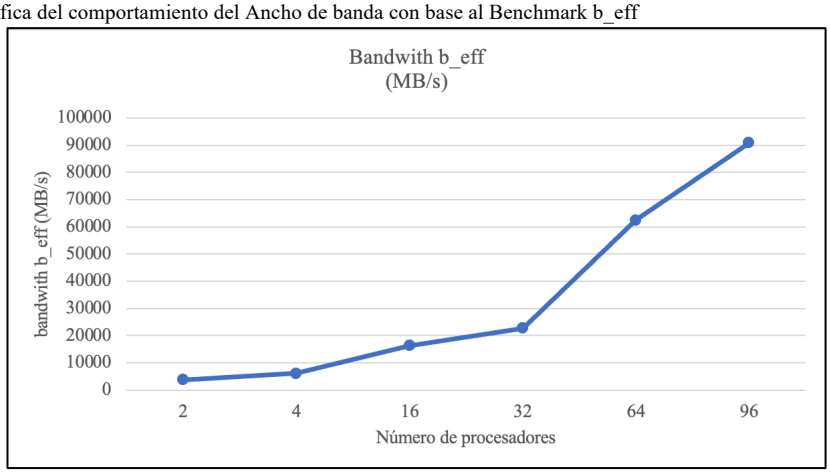
La latencia de la comunicación es una métrica de rendimiento que representa el tiempo que tarda la información en fluir de un nodo de cómputo a otro o de un lugar a otro. Por lo tanto, se vuelve muy importante no solo comprender cómo medir la latencia de una Red de Altas prestaciones sino también comprender cómo esta latencia afecta el rendimiento de los Jobs de cómputo o de las aplicaciones HPC. La latencia puede afectar la sincronización de los Jobs y su tiempo total de ejecución, lo cual es muy importante que cada usuario investigador considere la latencia de la red para cada uno de sus trabajos de investigación.

Como se muestra en la figura 55 la latencia de la red por mensaje es menor mientras mayor cantidad de procesadores se utilice, resaltando que los escenarios que tienen 2, 4 y 16 procesadores no utilizan la red InfiniBand ya que están procesando en dentro del mismo nodo de cómputo, lo cual demuestra que la latencia es mucho mayor procesando solo en un nodo de cómputo en comparación que realizar el procesamiento entre varios nodos de cómputo utilizando la red InfiniBand.

A partir de los 16 procesadores la curva se estabiliza, lo cual implica que la latencia no tiene una fluctuación tan pronunciada mientras aumenta la cantidad de procesadores, manteniendo un promedio de latencia de alrededor de 0.79 microsegundos por mensaje.

Análisis del comportamiento del Ancho de Banda efectivo.

Figura 56. Gráfica del comportamiento del Ancho de banda con base al Benchmark b_eff

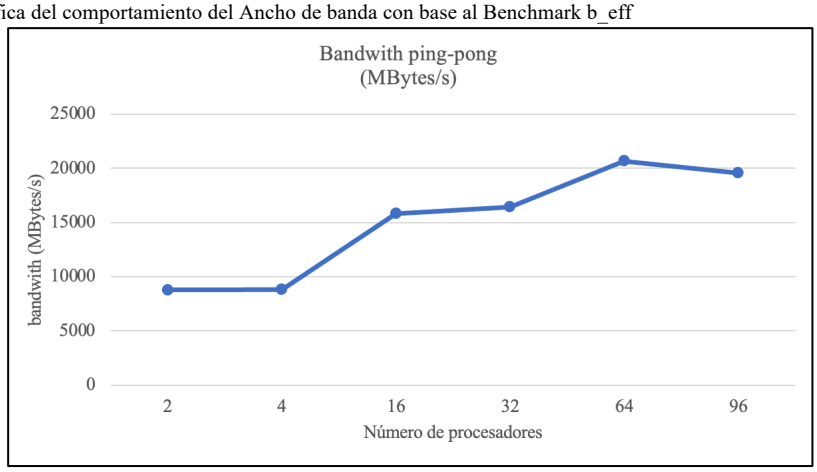


Nota: Realizado por el Autor.

En la figura 56 se puede observar que la Red de Altas prestaciones InfiniBand tiene mayor consumo del Ancho de Banda efectivo mientras mayor cantidad de procesadores se utilice, se observa una curva ascendente.

Análisis del comportamiento del Ancho de Banda ping-pong

Figura 57. Gráfica del comportamiento del Ancho de banda con base al Benchmark b_eff



Nota: Realizado por el Autor.

En la figura 57 se puede observar que la Red de Altas prestaciones InfiniBand tiene mayor consumo del Ancho de Banda ping-pong mientras mayor cantidad de procesadores se utilice. A partir de los 16 procesadores se ve un consumo promedio de alrededor de 18888 MBytes/s, teniendo una tendencia a estabilizarse el consumo del Ancho de banda entre nodos de cómputo.

Comparación de resultados con otra Infraestructura

En la presente investigación, se utilizó como referencia los datos de uno de los Benchmark del HPC Advisory Council, que se ejecutó sobre parámetros similares para fines de comparación con otras infraestructuras. (HPC-AI Advisory Council, 2020)

Se utilizó como referencia del benchmark del sistema de computación Cray T3E-900 con 512+32 procesadores y 128 MByte/procesador con el fin de comparar

con el Supercomputador Quinde I en función de los resultados obtenidos de los escenarios de prueba anteriores que se muestran en la Tabla 10. (Gerrit Schulz , 2020)

Cray T3E-900 cuenta con las siguientes características: (Bhakhra, 2019)

- Frecuencia 450Mhz.
- Tiene implementado MPI mpt.1.3.0.2.
- Es un Multiprocesador de memoria compartida escalable.
- Arquitectura RISC.
- Arquitectura diseñada para tolerar latencia y mejorar la escalabilidad.
- Topología 3D Torus en la que los nodos están conectados a sus nodos vecinos más cercanos en una malla 3D. (Arquitecturas Paralelas Introducción William Stallings, 2014)
- Cray tiene su propia memoria local.

Tabla 10. Cuadro comparativo entre Cray T3E-900 y el Supercomputador Quinde I.

Cray T3E-900

Número de procesadores	Memoria	b_eff (Mbyte/s)
2	128 MB	182,989
4	128 MB	355,045
16	128 MB	1063,217
32	128 MB	1893,872
64	128 MB	3158,554
96	128 MB	4180,723

Supercomputador Quinde I

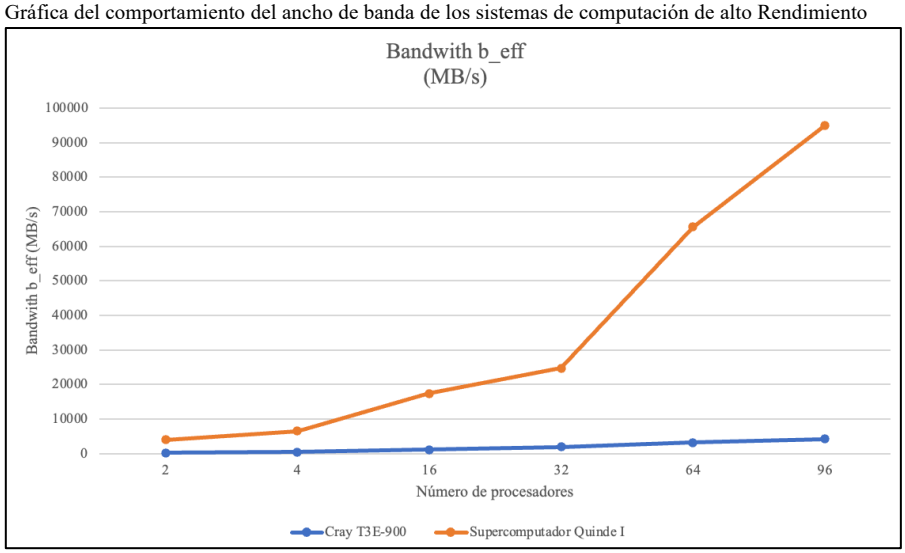
Número de procesadores	Memoria	b_eff (Mbyte/s)
2	128 MB	3754,570

4	128MB	6136,625
16	128MB	16263,114
32	128MB	22739,318
64	128MB	62367,35
96	128MB	90757,993

Nota: Los datos del Supercomputador Quinde I fueron resultado de los test realizados en los ítems anteriores y los datos de Cray TE3-900 fueron tomados de (Gerrit Schulz , 2020)

En la figura 58 se puede observar el comportamiento de la Red de Altas prestaciones InfiniBand del Supercomputador Quinde I en comparación con la otra Infraestructura Cray T3E-900 en función del Ancho de Banda Efectivo (b_eff) y la cantidad de procesadores.

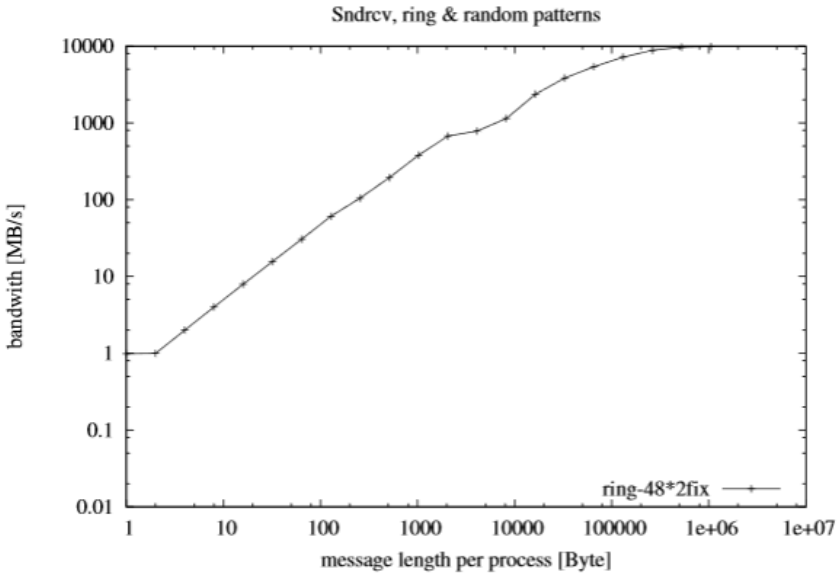
Figura 58. Gráfica del comportamiento del ancho de banda de los sistemas de computación de alto Rendimiento



Nota: Elaborado por el Autor.

Para motivos de comparación con la otra infraestructura, se utilizará la figura 59 del Escenario 6 del test de Rendimiento de la Red de Altas prestaciones InfiniBand del Supercomputador Quinde I, el cual muestra una gráfica ascendente con tendencia en cierto punto a estabilizarse.

Figura 59. Gráfica del Escenario 6 del test del Ancho de Banda del Supercomputador Quinde I con 96 procesadores.



Nota: Realizado por el Autor utilizando el editor Overleaf en formato Latex.

Como se muestra en la Figura 58, el Supercomputador Quinde I tiene una subida exponencial del ancho de banda mientras CRAY mantiene un estado altamente lineal a mayor cantidad de procesadores, esto nos puede llevar a una interpretación errónea de la comparativa del Ancho de banda; por lo cual es importante resaltar que CRAY en general son equipos diseñados para propósitos definidos, en este caso el CRAY T3E-900 fue diseñado para tolerar latencia y mejorar la escalabilidad sin usar redes InfiniBand, que tolera el tipo de carga en una sola infraestructura de hardware, mientras el Quinde I esta compuesto por varios componentes interconectados a través de la red InfiniBand, esto implica que la carga se distribuye de manera uniforme.

Al comparar la Figura 58 con la Figura 59 de uno de los resultados del Escenario 6 del test realizado en los ítems anteriores, se puede denotar que mientras más grande es el tamaño del mensaje en el Quinde I, se visibiliza una curva exponencial hasta llegar a los 10000 MB/s y luego la curva tiene un punto donde se mantiene cierta estabilidad a mayor longitud del mensaje por proceso.

Debemos tener claro que existen sistemas distribuidos y sistemas compartidos a nivel de hardware, y que esto puede ayudar en algunos caso como no en otros, sin embargo en el tema de equipos de Supercomputación donde abarca la clusterización de equipamiento a nivel bajo, para luego mostrarlo como una sola infraestructura de altas prestaciones a nivel de usuario de múltiples propósitos de investigación científica ante los equipos clúster en su mayoría diseñados para propósitos definidos.

Actualmente los supercomputadores en el mundo se encuentran diseñados para múltiples propósitos lo que ha permitido que centenares de investigaciones se puedan lograr en tiempos extremadamente cortos, esto gracias a su capacidad de manejar múltiples propósitos de investigación en un solo equipo.

Cabe señalar que la grafica de la Figura 58 muestra un escalamiento exponencial diferenciado al equipo CRAY; esto nos da una clara seguridad de como los equipos son diseñados para sus objetivos, esto debido a que a pesar de que en la grafica el ancho de banda sube de manera exponencial contra CRAY, la red InfiniBand inteligentemente estabiliza la escalabilidad en un cierto punto como se muestra en la figura 59, esto nos da un comportamiento a mantenerse en el manejo de los paquetes a través de redes InfiniBand ya que esta permite usar su gran ancho de banda y al mismo tiempo controlar el uso del Ancho de Banda hasta un punto de llegar a un equilibrio en orden a mantener la estabilidad del paso de paquetes.

Comprendiendo un poco mas el Sistema Cray, tomando como ejemplo un trabaja de dinámica molecular los dos computadores de Altas Prestaciones lo puede

hacer; sin embargo, en el caso de ejecutar trabajos de minería de bitcoins, Cray no esta diseñado específicamente para este tipos de cargas de trabajo mientras el Quinde I si puede hacerlo.

Cray tiene una infraestructura de hardware mas compleja que el Supercomputador Quinde I, mientras el Quinde I es mucho más Escalable a nivel de Hardware y Software.

Análisis Referencial con el Test de Linpack

Adicional se realizó un análisis de los datos obtenidos respecto al Ancho de banda con referencia al Test de linpack del fabricante realizado el 28 marzo de 2016 en el Supercomputador Quinde I.

El Test de Linpack es una medida de la tasa de ejecución de punto flotante de una computadora. Se determina ejecutando un programa de computadora que resuelve un sistema denso de ecuaciones lineales. Durante años las características del Test de Linpack han ido cambiando un poco a medida que la tecnología avanza en nuestro medio. (netlib, 2007)

El Test de Linpack surgió del proyecto de software Linpack el cual originalmente, estaba destinado a dar una idea a los usuarios de cuánto tiempo llevaría resolver ciertos problemas de matriz en ciertos programas. (netlib, 2007)

Además, Linpack al realizar cálculos con matrices es un test fácilmente paralelizable, por lo que es muy utilizado para medir la eficiencia de sistemas multiprocesador; los Supercomputadores más robustos a nivel mundial se basan en el Linpack para mostrar la eficiencia de los mismos en la página en Internet del "Top 500" (http://www.top500.org/).

Una vez descrito a que se refiere un Test de Linpack, con base al objeto de la presente investigación se tomó solamente los datos referente al Ancho de banda de uno de los archivos del Test de Linpack, el cual muestra el comportamiento de la Red de Altas prestaciones durante el Benchmark realizado al Supercomputador Quinde I por parte de la Empresa proveedora IBM.

En la tabla 11 se muestra un resumen del comportamiento del Ancho de Banda en varios escenarios aleatorios en los que se ejecutó el Test de Linpack con parámetros estándar predefinidos en la herramiento run_jlink[20] utilizada para verificar el estado de salud de la Red de altas prestaciones o detectar errores.

Valor mínimo: 1000

Valor máximo possible: 1000000

Total de nódos de cómputo=84

[20] La herramienta jlink se usa para medir el ancho de banda entre nodos de cómputo y para descubrir errores o enlaces de bajo rendimiento.

Tabla 11. Datos del Ancho de Banda del Test de Linpack del Supercomputador Quinde I.

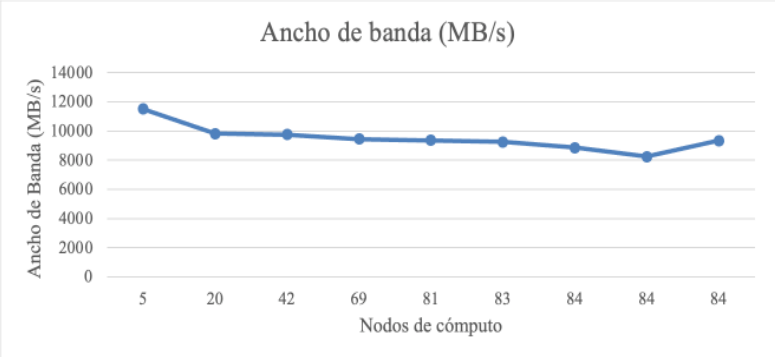
Escenario	Nodos de cómputo	Avg_bw MB/s	Agg_bw MB/s
1	5	11502	69010
2	20	9825	216139
3	42	9762	673611
4	69	9447	1662746
5	81	9369	4000620
6	83	9257	9192196
7	84	8848	61690.319
8	84	8247	30638625
9	84	9346	21916419

Nota: Realizado por el Autor con datos obtenidos del Test de Linpack del año 2016.

Avg_bw (MB/s). – Ancho de banda promedio.

Agg_bw (MB/s). – Ancho de banda agregado.

Figura 60. Gráfica del comportamiento del Ancho de Banda en el Test de Linpack del Supercomputador Quinde I.



Nota: Realizado por el Autor con datos obtenidos del Test de Linpack del año 2016.

En la Figura 60 se puede observar el comportamiento del Ancho de banda en función de la cantidad de nodos de cómputo durante el Test de Linpack. La gráfica muestra un comportamiento lineal estable del ancho de banda, variando dentro del rango de 8000 MB a 12000MB/s con n nodos de cómputo. Es importante resaltar los

valores de la Tabla 11 superan los valores mínimos y máximos establecidos en el Test de linpack para una Red InfiniBand para esta infraestructura de computación Paralela, lo cual refleja en ese entonces un buen rendimiento de la red de Altas prestaciones.

En función del tema de investigación se realizará un análisis referencial con los datos del escenario 6 expuestos en la Figura 53 donde se utilizó 96 procesadores sobre 5 nodos de cómputo para ejecutar el test de rendimiento de la Red de Altas Prestaciones, en comparación con los datos del Test del linpack con una cantidad de 5 nodos de cómputo. En la Tabla 12 se muestran los datos del escenario 6 y del Test 1 del Test de Linpack.

Tabla 12. Datos del Ancho de Banda del Escenario 1 del Test de Linpack

Escenario	Nodos de cómputo	Avg_bw MB/s	Agg_bw MB/s
1	5	11502	69010

Nota: Realizado por el Autor.

Tabla 13. Datos del Ancho de Banda Escenario 6 del Test de Rendimiento b_eff

Escenario	Nodos de cómputo	Ping-pong bandwidth MB/s	b_eff MB/s
6	5	19547	90758

Nota: Realizado por el Autor.

Como se muestra en la tabla 12 y 13, el Test de Linpack Nro. 1 realizado sobre 5 nodos de cómputo, presentó un Ancho de banda (Avg_bw) de 11502 MB/s y un Ancho de Banda agregado (Agg_bw) de 69010 MB/s, mientras que en el Escenario 6 utilizando la aplicación b_eff muestra un Ancho de Banda (ping-pong) de 19547MB/s y un Ancho de Banda efectivo de 90758MB/s; analizando estos valores se puede determinar que el Ancho de Banda del Escenario 5 es mayor que los datos del Test de Linpack, lo cual implica que la Red de Altas prestaciones tiene un buen rendimiento considerando que esta dentro de los valores mínimos y máximos establecidos para un buen estado de salud de la Red de Altas prestaciones Infiniband.

Adicional es importante resaltar el concepto del Ancho de Banda efectivo el cual representa la mayor cifra de velocidad de transmisión fiable de una Red. Este valor por lo general es menor que el máximo teórico, que muchas veces se considera el mejor ancho de banda utilizable, y resulta fundamental para comprender la cantidad de tráfico que puede admitir una conexión. En este caso se puede observar que el Ancho de Banda Efectivo es mayor que el Ancho de Banda del Test de Linpack.

Con base al análisis realizado se puede concluir que la Red de Altas prestaciones InfiniBand sobre una Arquitectura paralela sobre mensajes MPI presenta un buen rendimiento.

Caso Práctico

Gracias a la colaboración del equipo de uno de los proyectos de investigación desarrollados en el Supercomputador Quinde I liderado por la Empresa Pública Siembra E.P., se pudo aplicar los parámetros analizados en el presente documento

respecto al rendimiento de la red InfiniBand; con el fin de analizar el comportamiento real de los trabajos de cómputo en una aplicación científica.

La investigación *"Natural Products as Potential Inhibitors for SARS-CoV-2 Papain-Like Protease: An in Silico Study"* actualmente ya se encuentra publicada en la página de Springer Link en el siguiente enlace https://link.springer.com/chapter/10.1007%2F978-3-030-65775-8_25.

A continuación, se describe de manera general una parte de la investigación científica utilizada para aplicar los parámetros estudiados anteriormente a modo de caso práctico: (Alvarado Huayhuaz, Jimenez, Cordova Serrano, Camps, & Puma Zamora, s.f.)

Actualmente en nuestro país se está viviendo una crisis Sanitaria de alto impacto y se hace imprescindible el uso de Supercomputadores en el "TIEMPO MAS CORTO" a través de simulaciones computacionales con el fin de ayudar a desarrollar productos que tienen el potencial de salvar millones de vidas en todo el mundo.

Es así que el aporte de esta investigación brindará información relevante real acerca de una aplicación que se usó en una simulación computacional de búsqueda de inhibidores que ayuden a desactivar el "SARS-COV2" o también denominado COVID19, Coronavirus, uno de los proyectos que se desarrollaron en el Supercomputador Quinde I, liderado por la Empresa Pública Siembra E.P.

La investigación usó la aplicación (OPENBABEL), esta aplicación es un proyecto abierto y colaborativo que permite a cualquier persona buscar, convertir, analizar o almacenar datos de modelado molecular, química, materiales de estado sólido, bioquímica o áreas relacionadas.

A través de esta aplicación se convirtió las moléculas del formato *.smiles al formato *.pdb; este paso si se lo hubiera realizado sobre computadores normales hubiera tomado meses de procesamiento, es así que la aplicación OPENBABEL fue paralelizada usando (OPENMPI), esto debido a que su código fuente se encuentra con las subrutinas necesarias para trabajar en paralelo.

La versión de OPENMPI compilada, usó la bandera o característica de compilación siguiente:

$ ./configure --prefix=/home/fjimenez/prueba_curly/openmpi-4.0.3/compiled --with-mxm=/opt/mellanox/mxm --enable-mpi-cxx

- $ ./configure: se llama a la aplicacion para ejecutarla.
- --prefix=/home/fjimenez/prueba_curly/openmpi-4.0.3/compiled: ubicación del archivo o ejecutable.
- --with-mxm=/opt/mellanox/mxm: Comando para llamar a las librerías de Mellanox.
- --enable-mpi-cxx: Habilitar la opción de paso de mensajes en paralelo para C++.

Se puede denotar en las letras de color azul la bandera de configuración para decirle al compilador que se construya con mxm **"--with-mxm=/opt/mellanox/mxm"**, esto quiere decir que se construya o compile usando la herramientas propias del fabricante de la redes InfiniBand, con el objetivo de obtener todo el poder de la red de altas prestaciones y por consiguiente mejor rendimiento al momento de enviar a procesar grandes cantidades de información, esto nos asegura que el paso de mensajes sobre las aplicaciones de las diferentes ramas científicas sea la más óptima, sin embargo hay que denotar que también dependerá del grado de paralelismo de cada una de las aplicaciones.

Se ilustra en la Ecuación 3 los tiempos tomados para el trabajo de conversión de moléculas *.smiles a *.pdb, usando el siguiente archivo LSF:

Ecuación 3. Comando para compilar la aplicación b_eff en paralelo con la librería MPI

$$\text{mpirun} - \text{np } 20 \text{ obabel} - \text{ismiles} * . \text{smiles} - \text{opdb} - \text{O} * . \text{pdb} - \text{p } 7.4 \\ - - \text{gen3D} > \& \text{ result.log}$$

Nota: Realizado por el Autor.

Aquí no se agrega sobre el comando la opción siguiente "-tcp -x PAMI_IBV_ADAPTER_AFFINITY=0 -x PAMI_IBV_DEVICE_NAME=mlx5_0:0 -x PAMI_IBV_DEVICE_NAME=mlx5_1:0 --bind-to core" esto debido a que ya previamente se construyó OPENMPI con la opción MXM para que de manera predeterminada para aprovechar las bondades de tener este tipo de redes de altas prestaciones, esto nos ayuda a garantizar de mejor forma que nuestra aplicación "de

cualquier rama de la ciencia" si se encuentra paralelizada esta puede escalar de mejor forma.

En la Tabla 14 se muestra los resultados obtenidos producto del uso de la aplicación OBABEL con 7463 moléculas:

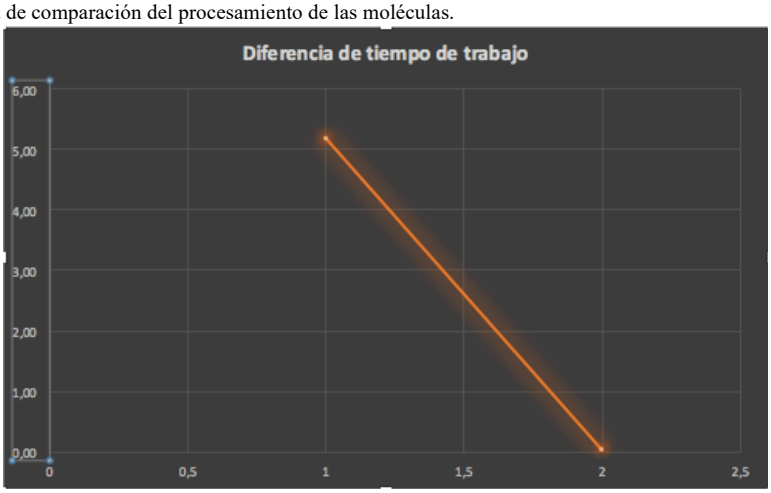
Tabla 14. Cuadro comparativo del procesamiento de moléculas en una infraestructura en paralelo y serial.

Número de moléculas	Tiempo de conversión promedio x molécula en segundos	Tipo	Total (segundos)	Total (Días)	Observaciones
7463	60	Serial	447780	5,18	Realizar el procesamiento de estas moléculas en serial tardaría alrededor de 5 días y medio.
7463	0,33	Paralelo	2462,79	0,03	Realizar el procesamiento de estas moléculas en paralelo tardaría alrededor de 0,72 horas (43 minutos)

Nota: Tomado de la investigación desarrollada en el Supercomputador Quinde I.

Sobre un computador serial convertir una molécula tiene un tiempo de procesamiento de alrededor de 60 segundos mientras que en paralelo usando la opción **"mxm"** y usando la **librería OpenMPI** antes denotada toma alrededor de 0,33 segundos por molécula como se ilustra en la Figura 61.

Figura 61. Gráfica de comparación del procesamiento de las moléculas.



Nota: Tomado de la investigación desarrollada en el Supercomputador Quinde I.

Los resultados obtenidos nos indican que procesar las 7463 moléculas sobre una infraestructura de "casa" nos tomaría alrededor de cinco días y una horas más, mientras haciendo el mismo proceso sobre un supercomputador utilizando la opción "mxm" sobre OPENMPI antes descrita nos tomaría alrededor de cuarenta y tres minutos y treinta segundos, claramente se puede denotar la diferencia notable de tiempo, y esta velocidad es lo que se necesita para la obtención rápida de resultados en un trabajo de investigación científica usando la red InfiniBand de Altas prestaciones que dispone el Supercomputador Quinde I.

Resumen de datos obtenidos en el caso práctico:

- CPU time: 2580.00 sec.
- Max Memory: 797 MB
- Average Memory: 597.41 MB
- Max Swap: 17592186044415 MB
- Max Processes: 24

- Max Threads: 29
- Run time: 480 sec.
- Turnaround time: 480 sec.

Guía de Buenas prácticas

Como parte del presente trabajo de investigación en el Anexo B se presenta la guía de buenas prácticas, recopilando la experiencia de los diferentes parámetros y aspectos aplicados en cada Escenario del Análisis de Rendimiento de la Red de Altas prestaciones Infiniband sobre el paso de mensajes MPI; el cual será un apoyo complementario para la ejecución de los procesos de cómputo de futuros trabajos de investigación en el Supercomputador Quinde I.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

A través del análisis de Rendimiento de la Red de Altas prestaciones Infiniband del Supercomputador Quinde I, se pudo determinar con la aplicación b_eff un buen rendimiento en cuanto a ancho de banda y latencia bajo varios escenarios de prueba, y sobre todo obtener una guía de buenas prácticas para la ejecución de los jobs de cómputo.

Con base a los resultados obtenidos en cada escenario de prueba se determina que la red de Altas prestaciones Infiniband presenta un buen rendimiento en función del ancho de banda y latencia, considerando que se debe ejecutar los Jobs de cómputo sobre procesadores de diferentes nodos de cómputo; la latencia no tiene una fluctuación tan pronunciada mientras aumenta la cantidad de procesadores, manteniendo un promedio de latencia muy bajo de 0.79 us por mensaje, siendo un valor aceptable para redes InfiniBand de 100Gb/s.

A nivel del Ancho de banda se ve un buen rendimiento de la Red de Altas Prestaciones InfiniBand, a mayor cantidad de procesadores el Ancho de banda aumenta, en cierto punto el Ancho de banda se mantiene dentro de un rango establecido para evitar la sobrecarga de todo el canal de la Red, teniendo una respuesta eficiente en cada escenario donde se usa la red Infiniband.

La ejecución de Jobs o trabajos de cómputo pueden tener un rendimiento variado de un supercomputador a otro, dependiendo mucho del grado de paralelismo de los proyectos de investigación y el hardware utilizado. Además, un código eficiente para un problema en una máquina paralela, puede ser muy diferente al código eficiente para el mismo problema en otra máquina, ya que esto depende mucho de la arquitectura de los procesadores, el tipo de red y las librerías a utilizar u otros factores que pueden influir en el rendimiento de la red de Altas prestaciones.

La Red de Altas prestaciones presenta una infraestructura escalable, ya que tenemos un ambiente estable en función a la latencia y ancho de banda en los escenarios de prueba. Además, que esta infraestructura cuenta con una arquitectura de memoria distribuida los cuales son baratos y fáciles de escalar a un gran número de procesadores. Sin embargo, es importante resaltar que el rendimiento de la Red y la escalabilidad también depende del buen uso de las librerías y el grado de paralelismo en cada proyecto de investigación.

Mellanox EDR 100Gb/s ahora le pertenece a la marca Nvidia, esta tecnología es una solución que responde a las necesidades actuales y futuras de los proyectos de investigación del Supercomputador Quinde I. Además, esta tecnología esta diseñada para utilizar el procesamiento propio de las tarjetas de red que aumentan el rendimiento de los procesos de cómputo.

El paso de mensajes utilizando Remote Direct Memory Access (RDMA) a través de la interconexión InfiniBand permite que las páginas de memoria se anclen

automáticamente y las transferencias del buffer se manejan directamente mediante el adaptador InfiniBand sin la participación del procesador del nodo de cómputo.

La Red InfiniBand ahora es una de las tecnologías que está en creciente desarrollo, presentando al mercado soluciones hasta 400Gb/s de velocidad y latencias menores a 0,6us, la cual se está convirtiendo en una de las tecnologías mas utilizadas en infraestructuras de Altas Prestaciones como Supercomputadores, Cloud Computing, Big Data y entre otros. Además, que InfiniBand es una de las redes mas utilizadas por el TOP 500 de Supercomputador a nivel mundial.

Al comparar con otra infraestructura de Altas prestaciones, podemos denotar un consumo creciente del ancho de banda efectivo; sin embargo, es importante diferenciar que el Supercomputador Quinde I es una infraestructura para multipropósitos mientras la otra infraestructura tiene un propósito específico por lo cual el comportamiento es variado.

Con referencia al Test de Linpack, el comportamiento del ancho de banda de la Red de Altas Prestaciones representa un buen rendimiento ante los escenarios de prueba realizados con la aplicación b_eff, que demuestra que la Red cumple con los parámetros de fábrica iniciales desde la fecha que fue instalada la infraestructura.

Los resultados obtenidos en el Caso Práctico nos muestran que procesar 7463 moléculas sobre una infraestructura de “casa” nos tomaría alrededor de cinco días, mientras haciendo el mismo proceso sobre un supercomputador utilizando la opción “mxm” de InfiniBand sobre OPENMPI tomó alrededor de cuarenta y tres minutos y

treinta segundos, esto implica un buen rendimiento y alta velocidad de respuesta en un trabajo de investigación científica usando la red de Altas prestaciones Infiniband que dispone el Supercomputador Quinde I.

Recomendaciones

Es recomendable utilizar versiones estables de la librería OpenMPI para futuros trabajos de investigación que cumplan con el paso de mensajes en paralelo y sobre todo sea compatible con la aplicación científica o para benchmarks.

Considerando los resultados del Análisis de Rendimiento de la Red de Altas prestaciones, es importante que al momento de ejecutar un job, este se ejecute sobre procesadores de diferentes nodos de cómputo con el fin de tener un mejor rendimiento de la Red, en función del consumo del ancho de banda y baja latencia.

Es recomendable utilizar la librería MXM propia de la tarjeta HCA InfiniBand con el fin de utilizar toda la capacidad de la Red de Altas prestaciones, y garantizar la ejecución eficiente de los Jobs de cómputo.

El Supercomputador Quinde I cuenta con un equipo LSF el cual permite balancear la carga de trabajo de los Jobs de cómputo; con el fin de evitar problemas de saturación de los nodos de cómputo o baja disponibilidad de los recursos de los mismos. Este es un factor importante que puede influir en la eficiencia de la ejecución de los Jobs de cómputo.

Es importante considerar el grado de paralelismo de los trabajos de investigación previo a ejecutar cualquier job de cómputo, ya que esto puede ser un factor muy importante en el rendimiento de la Red de Altas prestaciones.

REFERENCIAS BIBLIOGRÁFICAS

Ocampo Yahuarcani, I., & Campos Baca, L. E. (2017). *Introducción a la Supercomputación en el Perú.* Perú: Depósito Legal en la Biblioteca Nacional del Perú N° 2017-08981.

Aguilar Castro, J. L., & Leiss, E. (2004). *Introducción a la Computación Paralela.* Venezuela.

internetpasoapaso. (s.f.). *internetpasoapaso*. Obtenido de internetpasoapaso: https://internetpasoapaso.com/flops/

Secretaría Nacional de Planificación y Desarrollo - Senplades. (2017). Plan Nacional para el Buen Vivir 2017-2021. Quito, Pichincha, Ecuador.

Empresa Pública Siembra E.P. (2020). *Servicio de Supercomputación.* Obtenido de Servicio de Supercomputación.: https://hpc.yachay.gob.ec

Aguilar Castro, J. L., & Leiss, E. (2004). *Introducción a la Computación Paralela.* Mérida, Venezuela: Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Los Andes.

INFINIBAND. (s.f.). *WIKIPEDIA.* Obtenido de https://es.wikipedia.org/wiki/InfiniBand: https://es.wikipedia.org/wiki/InfiniBand

LATENCIA. (s.f.). *WIKIPEDIA.* Obtenido de https://es.wikipedia.org/wiki/Latencia: https://es.wikipedia.org/wiki/Latencia

OPENBABEL. (s.f.). *Openbabel.* Obtenido de http://openbabel.org/wiki/Main_Page: http://openbabel.org/wiki/Main_Page

OPENMPI. (s.f.). *https://www.open-mpi.org/.* Obtenido de https://www.open-mpi.org/: https://www.open-mpi.org/

insideHPC. (s.f.). *InsideHPC*. Obtenido de InsideHPC: https://insidehpc.com/hpc-basic-training/what-is-hpc/

García Nocetti, F. (junio de 2014). *www.inegi.org.mx.* Obtenido de www.inegi.org.mx: https://www.inegi.org.mx/eventos/2014/big-data/doc/P-DemetrioGarcia.pdf

IBM. (2020). *www.ibm.com*. Obtenido de www.ibm.com: https://www.ibm.com/support/knowledgecenter/linuxonibm/liaag/wcm2/l0wcm200_nw_shared_disk_nsd.htm

Red Hat. (s.f.). *www.redhat.com*. Obtenido de www.redhat.com: https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces

Jorba Esteve, J. (2006). *Análisis automático de prestaciones de aplicaciones paralelas basadas en paso de mensajes.* Barcelona.

open-mpi. (11 de junio de 2020). *www.open-mpi.org*. Obtenido de www.open-mpi.org: https://www.open-mpi.org/

HPC-AI Advisory Council. (2020). *www.hpcadvisorycouncil.com*. Obtenido de www.hpcadvisorycouncil.com: https://www.hpcadvisorycouncil.com/best_practices.php

Rolf Rabenseifner, G. S. (2020). *fs.hlrs.de*. Obtenido de fs.hlrs.de: https://fs.hlrs.de/projects/par/mpi//b_eff/

ICL UT. (s.f.). *icl.cs.utk.edu*. Obtenido de icl.cs.utk.edu: http://icl.cs.utk.edu/hpcc/

Empresa Pública Siembra E.P. (2020). *Informes técnicos* . Urcuquí.

Charte, F. (s.f.). *https://fcharte.com/.* Obtenido de https://fcharte.com/: https://fcharte.com/tutoriales/0000_programacionyparalelismo/

Gerrit Schulz , R. R. (2020). *https://fs.hlrs.de/projects/par/mpi//b_eff/.* Obtenido de https://fs.hlrs.de/projects/par/mpi//b_eff/: https://fs.hlrs.de/projects/par/mpi//b_eff/

Senplades. (2017). *Plan Nacional para el Buen Vivir 2017-2021.* Quito.

Constitucional., P. (10 de diciembre de 2019). Decreto Ejecutivo Nro. 945. *Decreto Ejecutivo Nro. 945*. Quito, Pichincha, Ecuador.

Wikipedia. (2 de septiembre de 2019). *wikipedia.org.* Obtenido de wikipedia.org: https://es.wikipedia.org/wiki/General_Parallel_File_System

https://icl.utk.edu. (s.f.). Obtenido de https://icl.utk.edu: https://icl.utk.edu/hpcc/hpcc_record.cgi?id=492

Acosta Berlinghieri, C. A. (2009). *Implementación de Computación de Alto Rendimiento y Programación Paralela en Códigos Computacionales.* Universidad EAFIT, Medellin.

iperf. (2020). *iperf*. Obtenido de iperf: https://iperf.fr/

Universidad de Oregon. (2020). *Performance Research Lab*. Obtenido de Performance Research Lab: http://www.cs.uoregon.edu/research/tau/home.php

nvidia. (2020). *nvidia High Performance Computing* . Obtenido de nvidia High Performance Computing : https://developer.nvidia.com/cuda-zone

Open Babel . (2016). *Open Babel* . Obtenido de Open Babel : http://openbabel.org/wiki/Main_Page

National Library of Medicine. (2013). Obtenido de National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/23813626/

gromacs. (2018). *gromacs*. Obtenido de gromacs: http://www.gromacs.org/About_Gromacs

putty.org. (2020). *putty.org*. Obtenido de https://www.putty.org/

Universidad de Alicante. (2011). *web.ua.es*. Obtenido de web.ua.es: https://web.ua.es/es/cluster-iuii/compiladores/compiladores-intel.html

top500. (noviembre de 2020). *top500*. Obtenido de top500: https://www.top500.org/project/top500_description/

Mellanox Technologies. (s.f.). *https://www.mellanox.com*. Obtenido de https://www.mellanox.com: https://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf

Villar Ortiz, J. A. (julio de 2004). Estudio y puesta en marcha de una subred. *Proyecto de Fin de Carrera*. UNIVERSIDAD DE CASTILLA-LA MANCHA.

EcuRed. (s.f.). *www.ecured.cu*. Obtenido de www.ecured.cu: https://www.ecured.cu/InfiniBand#Arquitectura

Infiniband Trade Association. (22 de diciembre de 2014). *www.infinibandta.org.* Obtenido de www.infinibandta.org: https://www.infinibandta.org/

Grun, P. (2010). *Introduction to InfiniBand for End Users.* Infiniband Trade Association.

HPC Challenge. (2016). *https://icl.utk.edu/.* Obtenido de https://icl.utk.edu/: https://icl.utk.edu/hpcc/index.html

Announcing the Mellanox ConnectX-5 100G Infiniband Adapter. (15 de junio de 2016). *https://es.slideshare.net/.* Obtenido de slideshare: https://es.slideshare.net/insideHPC/announcing-the-mellanox-connectx5-100g-infiniband-adapter

IBM. (2021). *IBM Spectrum LSF*. Obtenido de www.ibm.com: https://www.ibm.com/docs/en/spectrum-lsf/10.1.0?topic=reference-bsub

Alvarado Huayhuaz, J., Jimenez, F., Cordova Serrano, G., Camps, I., & Puma Zamora, W. (s.f.). *https://link.springer.com/*. Obtenido de Springer Link: https://link.springer.com/chapter/10.1007%2F978-3-030-65775-8_25

Bhakhra, S. (14 de junio de 2019). *GeeksforGeeks*. Obtenido de www.geeksforgeeks.org: https://www.geeksforgeeks.org/cray-t3e-architecture/

Arquitecturas Paralelas Introducción William Stallings, O. (2014). *www.electro.fisica.unpl.edu.ar.* Obtenido de www.electro.fisica.unpl.edu.ar: http://electro.fisica.unlp.edu.ar/arq/transparencias/ARQII_07-Paralelo-Redes-2014.pdf

netlib. (8 de mayo de 2007). *www.netlib.org*. Obtenido de netlib: http://www.netlib.org/utk/people/JackDongarra/faq-linpack.html#_Toc27885709

Paessler. (s.f.). *Paessler*. Obtenido de www.es.paessler.com: https://www.es.paessler.com/it-explained/bandwidth

icl. (s.f.). *HPC Challenge FAQ*. Obtenido de https://icl.utk.edu/: https://icl.utk.edu/hpcc/faq/index_print.html

Yachay E.P. (2019). *Servicio Nacional de Supercomputación.* Obtenido de Servicio Nacional de Supercomputación.: https://hpc.yachay.gob.ec

Secretaría Nacional de Planificación y Desarrollo, Senplades. (2017). *Plan Nacional para el Buen Vivir 2017-2021.* Quito.

ANEXOS

ANEXO A.- FORMULARIO DE SOLICITUD DE ACCESO A LOS RECURSOS DE SUPERCOMPUTADOR "QUINDE 1"

ANEXO B.- GUIA DE BUENAS PRÁCTICAS PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO.

ANEXO A.- FORMULARIO DE SOLICITUD DE ACCESO A LOS RECURSOS DEL SUPERCOMPUTADOR "QUINDE I"

1. **Lugar y Fecha:**

Urcuquí, 20 de noviembre de 2019.

2. **Institución auspiciante:**

Universidad Técnica del Norte

3. **Nombre del Proyecto:**

"ANÁLISIS DE RENDIMIENTO DE LA RED DE ALTAS PRESTACIONES EN UNA INFRAESTRUCTURA DE COMPUTACIÓN PARALELA, A TRAVÉS DE UNA APLICACIÓN HPC, COMO GUÍA PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO"

4. **Línea de Investigación:**

Innovación tecnológica y productos de Telecomunicación.

5. **Descripción del proyecto:**

El presente trabajo de investigación comprende realizar un Análisis de Rendimiento de la Red de Altas prestaciones InfiniBand sobre una arquitectura de computación paralela del Supercomputador Quinde I, mediante el paso de mensajes sobre la interfaz MPI, utilizando una aplicación HPC; donde se define los parámetros base para analizar el rendimiento de la red sobre ambientes paralelos. Una vez obtenidos los datos a través de la aplicación HPC, se realizará un análisis de los escenarios de prueba y la comparación de los datos con otra infraestructura de Altas Prestaciones. Parte de la presente investigación es aplicar los parámetros utilizados en el Análisis de Rendimiento de la Red de Altas prestaciones sobre un caso práctico con el fin de analizar el comportamiento de los procesos de cómputo en un ambiente real dentro de un proyecto de investigación científica. Al finalizar el análisis se presenta una guía de buenas prácticas para la ejecución de procesos de cómputo en arquitecturas paralelas con el fin de obtener el mejor Rendimiento de la Red de Altas prestaciones InfiniBand en cualquier proyecto de investigación que se desarrolle en el Supercomputador Quinde I.

6. **Resultados esperados:**

- Obtener datos del ancho de banda y latencia de los Jobs o procesos de cómputo ejecutados.
- Comparar los resultados obtenidos con otros trabajos de investigación en otro supercomputador.
- Presentar en gráficas estadísticas los resultados del análisis de los datos obtenidos.

- Presentar un caso práctico en una de las investigaciones ya realizadas en el Supercomputador "Quinde I" con el fin de verificar los resultados obtenidos del análisis.
- Presentar la guía de buenas prácticas para obtener un mejor Rendimiento de la Red de Altas Prestaciones en los jobs de cómputo.

7. Justificación de la Investigación respecto al uso del Supercomputador:

Con base a lo publicado en el sitio web de la Empresa Pública Yachay E.P. *"La supercomputación, High Performance Computing, HPC, computación de altas prestaciones y en los últimos años denominada informática de alto rendimiento o informática de gama alta, es una herramienta estratégica fundamental para el Ecuador, que ha sido implementada por la Empresa Pública Yachay E.P, y que potenciará el progreso científico, el desarrollo y la innovación industrial, la seguridad nacional y permitirá afrontar los retos sociales del país de mejor manera a través del modelamiento y la simulación computacional."* (Yachay E.P., 2019)

"El Servicio Nacional de Supercomputación de Yachay EP, cuenta con un Modelo de Gestión; el mismo que ha sido estructurado en tres fases, que comprende al final la triple hélice: ACADEMIA, INDUSTRIA Y ESTADO, para fomentar la Gestión del conocimiento técnico, científico y coadyuvar al desarrollo industrial y productivo del país. El Modelo de Gestión cuenta con el aval del SENESCYT y del Comité de Acceso Quinde 1." (Yachay E.P., 2019)

En el campo de la investigación, la supercomputación tiene una gran relevancia en diferentes áreas de la ciencia, tecnología, física, matemática, salud, etc.; el Supercomputador Quinde I, actualmente cuenta con varios proyectos en diferentes líneas de investigación que están transformando la academia, industria y estado, por lo tanto los investigadores requieren contar con los recursos y guías necesarias para la ejecución eficiente de sus procesos de cómputo sobre una red de altas prestaciones Infiniband en una infraestructura de computación paralela.

Esta temática se encuentra dentro de la línea de investigación de Innovación tecnológica y productos de telecomunicación de la Universidad Técnica del Norte, que permitirá coadyuvar al campo de la investigación, academia e industria de la zona norte del país.

Con base a todos los antecedentes expuestos, el desarrollo de este trabajo es de vital importancia, ya que permitirá presentar a la comunidad académica – científica datos reales del rendimiento de la red de altas prestaciones Infiniband y una guía de buenas prácticas para la ejecución de sus procesos de cómputo en una infraestructura de computación paralela del Supercomputador Quinde I. Además, este trabajo permitirá impulsar proyectos de investigación futuros que aporten al desarrollo científico y

social del país, convirtiendo a Ecuador un referente en el campo de la investigación a nivel regional.

8. Listado de productos a obtener:

- Comportamiento del Ancho de banda de la Red de Altas Prestaciones
- Latencia de la Red de Altas Prestaciones

9. Participantes del proyecto que requieren acceso:

NOMBRES Y APELLIDOS	DOCUMENTO DE IDENTIFICACIÓN	INSTITUCIÓN	CORREO	TELÉFONO DE CONTACTO
Alexandra Nataly Culqui Medina	100292537-6	UTN	anculquim@utn.edu.ec	0980438045

10. Recursos de altas prestaciones a utilizar:

DESCRIPCIÓN	CANTIDAD	MEDIDA
PROCESADORES	Hasta 1000	Procesadores
MEMORIA	10	GB
STORAGE	1	GB
GPU	No requerido	No requerido

11. Tiempo requerido para ejecución del proyecto (días): 360 días.

ANEXO B.- GUIA DE BUENAS PRÁCTICAS PARA LA EJECUCIÓN DE PROCESOS DE CÓMPUTO.

Índice

1. Introducción

Con base a los resultados del Análisis de Rendimiento de la Red de Altas prestaciones y la experiencia adquirida durante el desarrollo del presente trabajo de Investigación, se detalla a continuación la siguiente guía de buenas prácticas para la ejecución de los procesos de cómputo o para el desarrollo de otros tipos de benchmarks en el Supercomputador Quinde I.

2. Conceptos Clave

InfiniBand: Es una nueva y potente tecnología diseñada para soportar la conectividad E/S para infraestructuras de computación. InfiniBand es compatible con los principales proveedores de servidores OEM[21] como un medio para expandirse más allá y crear el estándar de interconexión E/S de próxima generación de servidores. InfiniBand es el único en proporcionar tanto una solución de backplane "in the box", una interconexión externa y "Ancho de banda out the box", por lo que proporciona conectividad de una manera previamente reservada solo para interconexiones de redes tradicionales.

RDMA: Remote direct memory access (RDMA) o su traducción al castellano acceso remoto directo a memoria se refiere al acceso directo desde la memoria principal de un ordenador en la de otro sin intervención del sistema operativo, permitiendo alta rendimiento en sistemas de cómputo y de baja latencia. (HPC-AI Advisory Council, 2020)

Job: en español trabajo es una o varias tareas lógicas de cómputo que se ejecutan en una infraestructura de computación para obtener un resultado específico del algoritmo o programa ejecutado.

Benchmark: Es una prueba de rendimiento o de evaluación comparativa de un equipo computacional, con el fin de medir la capacidad real o el rendimiento de uno de sus componentes o de toda la infraestructura de Altas prestaciones. (EcuRed, s.f.)

Login node: son equipos que permiten el acceso al usuario científico a la consola de línea de comandos del Supercomputador Quinde I. Cuenta con dos login node, con el fin de balancear la cantidad de accesos de los usuarios. (Empresa Pública Siembra E.P., 2020)

[21] OEM (del inglés, Original Equipment Manufacturer) o fabricante de equipos originales que confecciona piezas, un subsistema o software que se utilizan en los productos de otras empresas. Algunos ejemplos son los sistemas operativos y los microprocesadores en equipos.

LSF: Plataforma LSF (Load Sharing Facility) es un conjunto de productos de administración de recursos distribuidos que permite: (Empresa Pública Siembra E.P., 2020)

MPI: de sus siglas en inglés Message Passing Interface o Interfaz de paso de mensajes, es un estándar que permite la implementación de librerías de paso de mensajes, siendo el principal objetivo el permitir la interacción entre la eficiencia y portabilidad, que proporciona una librería de funciones para C, C++ o Fortran que son empleadas para comunicar datos entre procesos. MPI es portable, fue diseñada y optimizada principalmente para trabajar sobre arquitecturas de memoria distribuida. A partir de esta especificación se desarrollan librerías de software libre como lo es OpenMPI (Jorba Esteve, 2006)

OpenMPI (Message Passing Interface): Es una API de código abierto utilizada para programación paralela y/o distribuida, que se basa en el estándar MPI. (open-mpi, 2020):

B_eff: Es una aplicación de software libre que permite medir el ancho de banda efectivo de la red de comunicación de un sistema informático paralelo y/o distribuido, que utiliza diferentes tamaños de mensajes y métodos de comunicación para validar el comportamiento de una Red de Altas prestaciones (Gerrit Schulz , 2020).

Librería Mellanox: La biblioteca MellanoX Messaging (MXM) proporciona mejoras a las bibliotecas de comunicaciones paralelas utilizando completamente la infraestructura de red proporcionada por el hardware de Mellanox HCA / switch. (Mellanox Technologies, s.f.)

3. Consideraciones iniciales

Formulario de Acceso

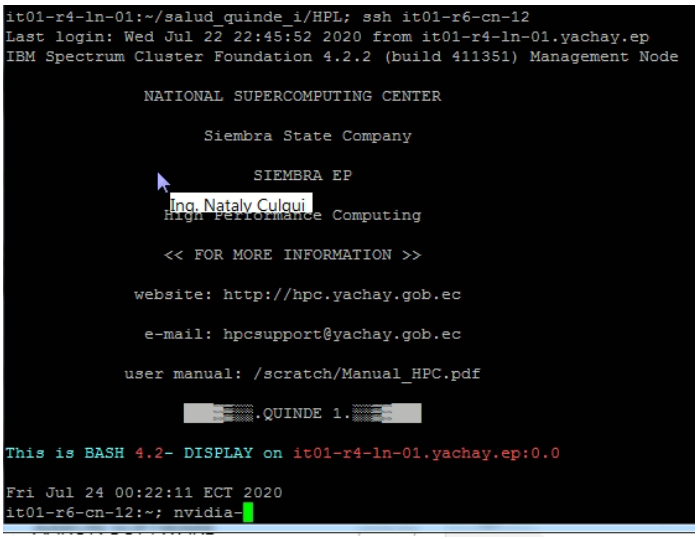
Previo al desarrollo de cualquier trabajo de investigación, es importante llenar el formulario de solicitud de acceso al Supercomputador Quinde I para la aprobación respectiva del comité de acceso de la ex Empresa Pública Siembra E.P. En dicho formulario se debe explicar la justificación del proyecto y los recursos necesarios para ejecutar el proyecto de investigación que se detalló en el Anexo A.

Manual de Uso del Supercomputador Quinde I

Para la conexión remota e ingreso al directorio del proyecto asignado en el sistema de Almacenamiento del Supercomputador Quinde I, es necesario seguir las instrucciones del “Manual de Uso del Supercomputador Quinde I” que se encuentra

en la pagína web *hpc.yachay.gob.ec* o en el sistema de archivos /scratch como se muestra en la figura 1.

Figura 62. ***Ingreso al servicio de Supercomputación***



Nota: Elaborado por el autor.

Conexión remota al Supercomputador Quinde I

Es importante considerar que para ingresar por el agente SSH de forma remota al login node, el entorno debe estar configurado con un usuario y contraseña proporcionado por el equipo de Soporte Técnico del Supercomputador Quinde I.

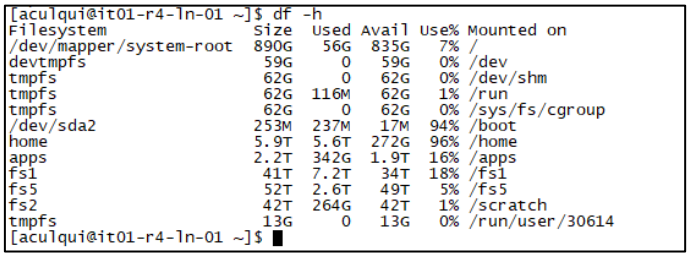
Se debe cumplir con las medidas de seguridad indicadas por el personal de Soporte Técnico para el manejo de las credenciales de acceso al Supecomputador Quinde I, ya que es de uso exclusivo por cada proyecto de investigación.

El login node es un componente del Supercomputador Quinde I que es directamente accesible para el usuario final. Solo el personal de administración tiene acceso directo a otros components de la Infraestrcutura de Altas Prestaciones.

Filesystem

Previo a ejecutar los jobs de cómputo del proyecto de investigación es necesario revisar las particiones del sistema de Almacenamiento del login node, con el fin de tener el espacio disponible para las nuevas aplicaciones y logs de los resultados de todo el proyecto de investigación con el comando “df -h” como se muestra en la figura 2.

Figura 63. Revisión de espacio de almacenamiento



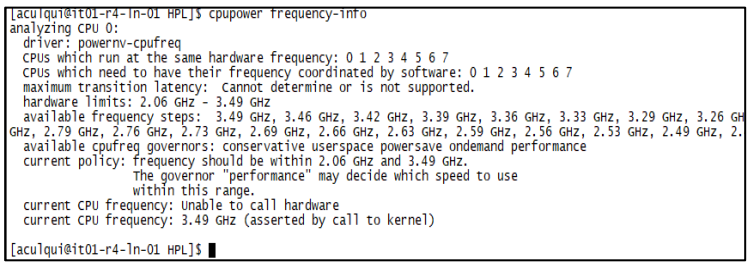
Nota: Elaborado por el autor.

Frecuencia de reloj

La frecuencia de reloj de la CPU es un buen indicador del desempeño del procesador por lo cual es muy importante revisar la frecuencia de reloj de los nodos de cómputo que deben estar configurados dentro del rango 2.92GHz a 2.926GHz, siendo un parámetro recomendado por el proveedor de la infraesrtructura, el cual permite trabajar adecuadamente los jobs de cómputo.

Como se muestra en la figura 3 se verifica la frecuencia configurada

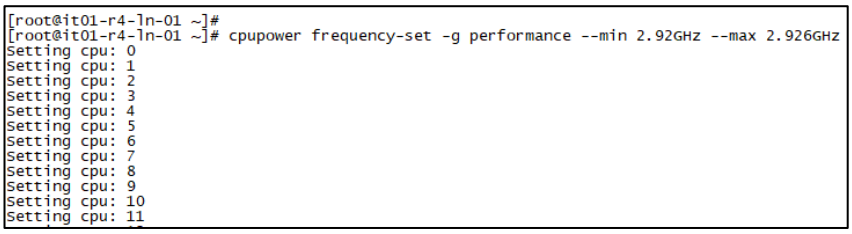
Figura 64. Verificar la frecuencia de reloj de los nodos de cómputo.



Nota: Elaborado por el autor.

Para configurar la frecuencua de reloj, en el caso de ser necesario se utiliza el comando indicado en la figura 4.

Figura 65. Verificar la frecuencia de reloj de los nodos de cómputo.



Nota: Elaborado por el autor.

4. Consideraciones generales para entornos paralelos

Al momento de ejecutar los jobs de cómputo sobre una infraestructura paralela, es importante considerar los siguiente aspectos:

Uso de la cantidad de nodos de cómputo

Usar más de dos nodos de cómputo para ejecutar los jobs, ya que esto permitirá utilizar la Red Infiniband para el paso de los mensajes, con base a los escenarios de prueba realizados en la presente investigación, se obtuvo una latencia muy baja al ejecutar los jobs de cómputo sobre varios nodos de cómputo.

Arquitectura Power 8 de los procesadores

Definir el enfoque/paradigma apropiado para la arquitectura Power 8 de IBM de los procesadores que tiene el Supercomputador Quinde I, ya que esto permitirá optimizar de mejor manera los jobs de cómputo.

Verificar que el compilador a utilizar sea una versión estable y soporte arquitecturas Power 8. Para el caso de la presente investigación se utilizó el compilador mpicc.

Grado de paralelismo

El grado de paralelismo es el número de acciones (tareas) que se pueden ejecutar en paralelo durante la ejecución de una aplicación. El grado de paralelismo puede ser diferente a través de las diferentes partes del programa o aplicación HPC a utilizar en el proyecto de investigación. Por lo tanto, se recomienda definir el grado de paralelismo de los jobs, ya que esto puede implicar en el uso ineficiente de los recursos del Supercomputador Quinde I.

La ejecución de Jobs o trabajos de cómputo pueden tener un rendimiento variado en supercomputador como en otro, dependiendo mucho del grado de paralelismo de los proyectos de investigación y el hardware utilizado. Además, un código eficiente para un problema en una máquina paralela, puede ser muy diferente al código eficiente para el mismo problema en otra máquina, ya que esto depende mucho de la arquitectura de los procesadores, el tipo de red y las librerías a utilizar u otros factores que pueden influir en el rendimiento de la red de Altas prestaciones.

Es importante resaltar que el rendimiento de la Red y la escalabilidad también depende del buen uso de las librerías y el grado de paralelismo en cada proyecto de investigación.

Tiempos de comunicación

Se recomienda separar los tiempos de comunicación, de Entradas/Salidas y de cálculo en la aplicación, si es posible (esto permitirá identificar qué partes de la aplicación se deben optimizar).

Escenarios de prueba

Realizar pruebas en varios escenarios y situaciones de la infraestructura paralela.

Uso de LSF

Considerar que el Supercomputador tiene una arquitectura de memoria distribuida, por lo tanto es importante ejecutar los jobs sobre el equipo LSF con el fin de evitar que el uso de recursos desbalanceado de los nodos de cómputo, y se tenga una mala percepción del rendimiento del Supercomputador Quinde I. El no uso del LSF puede incurrir en un monitoreo complejo de los jobs ejecutados en toda la infraestructura.

Además, el equipo LSF garantiza una cola de trabajo específico para cada proyecto de investigación, dando prioridad a los jobs para un uso balanceado de los recursos de los nodos de cómputo con base al requerimiento de los usuarios.

Se recomienda una configuración básica del archivo LSF como se muestra en la figura 5, que permitirá ma

Figura 66. Editar el archivo "benchmark_aculqui.lsf"

Nota: Realizado por el Autor.

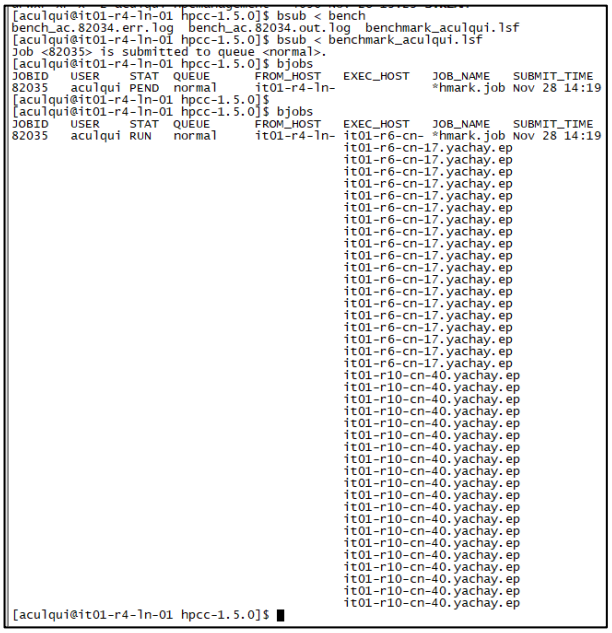
El significado de los comandos y parámetros del archivo "benchmark_aculqui.lsf": (IBM, 2021)

- BSUB.- es el comando que envía un job al LSF ejecutando el comando especificado y sus argumentos.
- Bsub -e bench_fj.%J.err.log: Especifica el nombre del archivo para guardar como log todos los errores que se presenten en la corrida del job.
- Busb -o bench_fj.%J.out.log: El nombre del archivo para guardar como log los resultados del job.
- Bsub -J benchmark.job: Colocar un nombre al job a ejecutar.
- Bsub -cwd /home/aculqui/tesis_compiling/hpcc-1.5.0_001/: Indica el path donde se va a correr el job.

- Bsub -q normal: Indica la cola de procesos en el LSF, se configura la cola normal ya que esta dispone de mayor cantidad de procesadores para los Jobs.
- Busb -n 10: Se reserva la cantidad de procesadores para el job a ejecutar.
- Module purge: Purgar todos los módulos cargados, como buena práctica para ejecutar un nuevo job.
- Module load smpi/10.1.0: Se recomienda cargar el módulo SpectrumMPI para aplicaciones que soporten esta versión (Para la presente investigación se utilizó la librería OpenMPI por motivos de comparación y análisis).
- Cd /home/aculqui/tesis_compiling/hpcc-1.5.0_001: Se indica el path donde se encuentra el programa HPCC 1.5.0
- Export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/lib64: Se asigna a la variable de entorno LD_LIBRARY_PATH la ruta de las librerías de 64 bits del Sistema Operativo.
- mpirun -np 10 ./hpcc: Es el comando que inicia la ejecución del job sobre 10 procesadores.
- > result_10c.log: se agrego adicionalmente para que los resultados sean almacenados en un archivo log llamado result_10c.log para mayor facilidad de identificación de los archivos del proyecto de investigación.

Para ejecutar el job sobre LSF se utiliza el comando bsub < benchmark_aculqui.lsf como se muestra en la Figura 6.

Figura 67. Ejecución del job sobre LSF.



Nota: Realizado por el Autor

Fuentes confiables

Descargar las aplicaciones o librerias necesarias de fuentes confiables, por seguridad de los ambientes de cada proyecto de investigación.

5. Interfaz OpenMPI

La interfaz MPI es utilizado específicamente en ambientes de computación paralela, ya que permiten el paso de mensajes de los diferentes procesos de cómputo sobre varios procesadores al mismo tiempo, reduciendo tiempos de ejecución de los Jobs de cómputo de los trabajos de investigación del Supercomputador Quinde I.

Por el uso de pase de mensajes en el que está basado la presente investigación, el rendimiento puede verse afectado sino se utiliza de manera adecuada el nivel de paralelismo en los scripts de programación de cada proyecto de investigación.

Para proyectos futuros es importante considerar el uso de la librería SprectrumMPI si esta librería es compatible con las aplicaciones científicas

En la presente investigación se consideró el módulo OpenMPI versión 1.8.8 siendo la versión más estable en el período que se ejecutó el benchmark de la Red de Altas Prestaciones, por lo que es recomendable para futuras investigaciones instalar versiones estables de las librerías o software que se requiera en cualquier proceso de computación paralela.

Instalar un versión estable del módulo OpenMPI que soporte la arquitectura Power8 de los procesadores de IBM de los nodos de cómputo.

Se recomienda compilar el módulo OpenMPI en CC, la cual no presentó ningun problema en la ejecución.

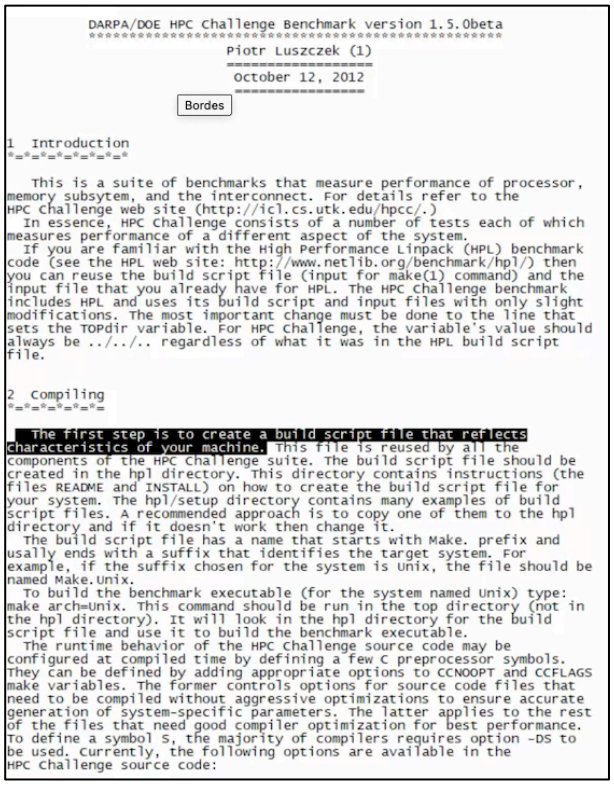
Por motivos de análisis y comparación con otras infraestructuras de computación paralela, se utilizó la librería estándar OpenMPI, la cual es compatible con varias aplicaciones científicas. Sin embargo, la infraestructura del Supercomputador Quinde I tiene su propia librería MPI llamada SpectrumMPI propia de la marca IBM, la cual está optimizada para ciertas aplicaciones científicas, por lo tanto se recomienda utilizar está librería siempre y cuando se valide la compatibilidad y funcionalidad con la aplicación a utilizar en el proyecto de investigación.

6. Benchmark b_eff

Previo a ejecutar b_eff, es importante entender que este pertenece al ejecutable del hpcc según el HPC Challenge Benchmark, el cual incluye 7 tipos de pruebas de rendimiento de una infraestructura de Altas prestaciones, con base al tipo de análisis y datos que se requiera en la investigación. Para la presente investigación se utilizó el benchmark b_eff con el fin de analizar el rendimiento de la Red de Altas Prestaciones Infiniband, para lo cual es importante tener en cuenta las siguientes consideraciones:

- Revisar las instrucciones a seguir en el archivo README.txt para ejecutar el benchmark hpcc de acuerdo al procedimiento sugerido del HPC Challenge Benchmark como se muestra en la Figura 7.

Figura 68. Verificar las instrucciones en el archivo README.txt para correr el benchmark hpcc



```
DARPA/DOE HPC Challenge Benchmark version 1.5.0beta
*****************************************************
                    Piotr Luszczek (1)
                    ==================
                    October 12, 2012
                    ================
```

Bordes

```
1  Introduction
*=*=*=*=*=*=*=*

   This is a suite of benchmarks that measure performance of processor,
memory subsytem, and the interconnect. For details refer to the
HPC Challenge web site (http://icl.cs.utk.edu/hpcc/.)
  In essence, HPC Challenge consists of a number of tests each of which
measures performance of a different aspect of the system.
  If you are familiar with the High Performance Linpack (HPL) benchmark
code (see the HPL web site: http://www.netlib.org/benchmark/hpl/) then
you can reuse the build script file (input for make(1) command) and the
input file that you already have for HPL. The HPC Challenge benchmark
includes HPL and uses its build script and input files with only slight
modifications. The most important change must be done to the line that
sets the TOPdir variable. For HPC Challenge, the variable's value should
always be ../../.. regardless of what it was in the HPL build script
file.


2  Compiling
*=*=*=*=*=*=

  The first step is to create a build script file that reflects
characteristics of your machine. This file is reused by all the
components of the HPC Challenge suite. The build script file should be
created in the hpl directory. This directory contains instructions (the
files README and INSTALL) on how to create the build script file for
your system. The hpl/setup directory contains many examples of build
script files. A recommended approach is to copy one of them to the hpl
directory and if it doesn't work then change it.
  The build script file has a name that starts with Make. prefix and
usally ends with a suffix that identifies the target system. For
example, if the suffix chosen for the system is Unix, the file should be
named Make.Unix.
  To build the benchmark executable (for the system named Unix) type:
make arch=Unix. This command should be run in the top directory (not in
the hpl directory). It will look in the hpl directory for the build
script file and use it to build the benchmark executable.
  The runtime behavior of the HPC Challenge source code may be
configured at compiled time by defining a few C preprocessor symbols.
They can be defined by adding appropriate options to CCNOOPT and CCFLAGS
make variables. The former controls options for source code files that
need to be compiled without aggressive optimizations to ensure accurate
generation of system-specific parameters. The latter applies to the rest
of the files that need good compiler optimization for best performance.
To define a symbol S, the majority of compilers requires option -DS to
be used. Currently, the following options are available in the
HPC Challenge source code:
```

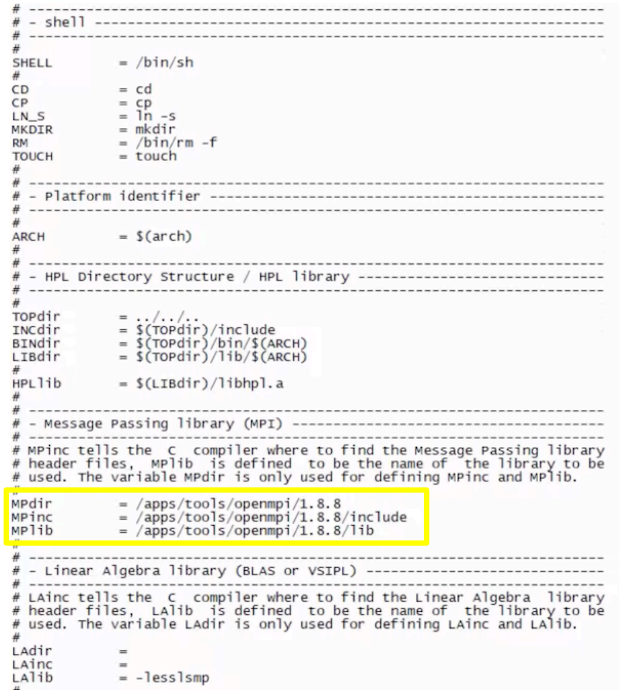
Nota: Realizado por el autor.

- El test B_eff (Efective bandwidth y latencia) también se puede ajustar o realizar tuning, lo cual significa ajustar los parámetros y variables a la arquitectura de la infraestructura de Altas Prestaciones en este caso a la arquitectura Power 8.

- Es importante recordar que el benchmark b_eff debe usar solo llamadas MPI estándar para realizar comparaciones con otras infraestructuras.

- Al compilar la aplicación b_eff se recomienda definir adecuadamente el path de las librerias a utilizar dentro del script modificado para la arquitectura Power 8, en este caso de la librería OpenMPI como se muestra en la figura 8.

Figura 69. Agregar la librería OpenMPI.

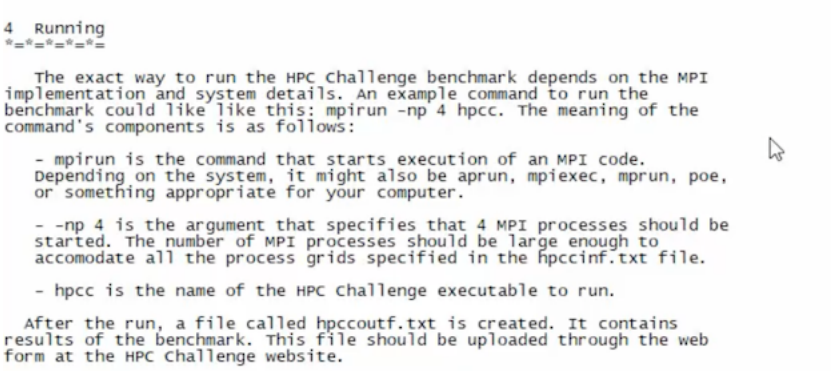


```
# ----------------------------------------------------------------------
# - shell --------------------------------------------------------------
# ----------------------------------------------------------------------
#
SHELL        = /bin/sh
#
CD           = cd
CP           = cp
LN_S         = ln -s
MKDIR        = mkdir
RM           = /bin/rm -f
TOUCH        = touch
#
# ----------------------------------------------------------------------
# - Platform identifier ------------------------------------------------
# ----------------------------------------------------------------------
#
ARCH         = $(arch)
#
# ----------------------------------------------------------------------
# - HPL Directory Structure / HPL library ------------------------------
# ----------------------------------------------------------------------
#
TOPdir       = ../../..
INCdir       = $(TOPdir)/include
BINdir       = $(TOPdir)/bin/$(ARCH)
LIBdir       = $(TOPdir)/lib/$(ARCH)
#
HPLlib       = $(LIBdir)/libhpl.a
#
# ----------------------------------------------------------------------
# - Message Passing library (MPI) --------------------------------------
# ----------------------------------------------------------------------
# MPinc tells the  C  compiler where to find the Message Passing library
# header files,  MPlib  is defined  to be the name of  the library to be
# used. The variable MPdir is only used for defining MPinc and MPlib.
#
MPdir        = /apps/tools/openmpi/1.8.8
MPinc        = /apps/tools/openmpi/1.8.8/include
MPlib        = /apps/tools/openmpi/1.8.8/lib
#
# ----------------------------------------------------------------------
# - Linear Algebra library (BLAS or VSIPL) -----------------------------
# ----------------------------------------------------------------------
# LAinc tells the  C  compiler where to find the Linear Algebra  library
# header files,  LAlib  is defined  to be the name of  the library to be
# used. The variable LAdir is only used for defining LAinc and LAlib.
#
LAdir        =
LAinc        =
LAlib        = -lesslsmp
```

Nota: Realizado por el Autor.

- Para ejecutar correctamente los jobs se recomienda revisar el punto 4 del Readme.txt del hpcc como se muestra en las figuras 9 y en la figura 10 se muestra un ejemplo de como correr un jon en paralelo sobre 4 procesadores.

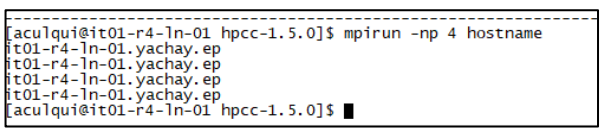
Figura 70. Guía para correr los jobs



```
4  Running
*=*=*=*=*=

   The exact way to run the HPC Challenge benchmark depends on the MPI
implementation and system details. An example command to run the
benchmark could like like this: mpirun -np 4 hpcc. The meaning of the
command's components is as follows:

   - mpirun is the command that starts execution of an MPI code.
   Depending on the system, it might also be aprun, mpiexec, mprun, poe,
   or something appropriate for your computer.

   - -np 4 is the argument that specifies that 4 MPI processes should be
   started. The number of MPI processes should be large enough to
   accomodate all the process grids specified in the hpccinf.txt file.

   - hpcc is the name of the HPC Challenge executable to run.

  After the run, a file called hpccoutf.txt is created. It contains
results of the benchmark. This file should be uploaded through the web
form at the HPC Challenge website.
```

Nota: Realizado por el Autor.

Figura 71. Ejemplo de un job sobre la interfaz MPI.

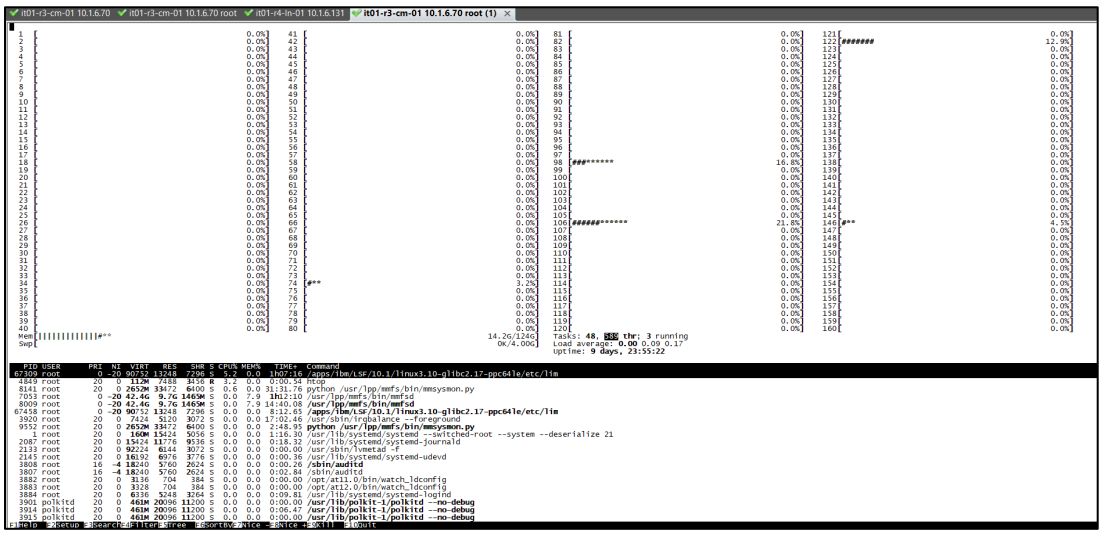


```
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$ mpirun -np 4 hostname
it01-r4-ln-01.yachay.ep
it01-r4-ln-01.yachay.ep
it01-r4-ln-01.yachay.ep
it01-r4-ln-01.yachay.ep
[aculqui@it01-r4-ln-01 hpcc-1.5.0]$
```

Nota: Realizado por el Autor.

- Se recomienda monitorear el job ejecutado en tiempo real con el comando htop como se muestra en la figura 11.

Figura 72. Verificar estado en tiempo real de la corrida de un job con el comando HTOP



Nota: Realizado por el Autor.

7. RDMA

IBM implementa el paso de mensajes utilizando Remote Direct Memory Access (RDMA) a través de la interconexión InfiniBand. Este mecanismo permite que las páginas de memoria se anclen automáticamente y las transferencias del buffer se manejan directamente mediante el adaptador InfiniBand sin la participación de la CPU del nodo de cómputo, como resultado un mejor rendimiento y respuesta de los procesos de cómputo.

8. Librería MXM

Es recomendable utilizar la librería MXM de la tarjeta HCA Infiniband propia de la Infraestructura de Altas Prestaciones, con el fin de utilizar toda la capacidad de la Red Infiniband y mejorar la ejecución de los Jobs.

Se recomienda utilizar Mellanox OFED 1.5.3 o superior, o una versión estable.

Usar OpenMPI 1.8 o superior o una versión estable que soporte la arquitectura Power 8 del Supercomputador Quinde I.

MLNX_OFED v1.8 o superior viene pre instalado OpenMPI 1.6 la cual esta listo para ser configurado con MXM.

Habilitando MXM en OpenMPI

MXM v1.1 es seleccionado automáticamente por OpenMPI cuando el número de procesos (NP) es mayor o igual a 128. Para habilitar MXM para cualquier NP debe usar el siguiente parámetro para OpenMPI "-mca mtl_mxm_np <number>".

La versión de OPENMPI compilada para la presente investigación, usó la bandera o característica de compilación siguiente:

$ ./configure --prefix=/home/fjimenez/prueba_curly/openmpi-4.0.3/compiled --with-mxm=/opt/mellanox/mxm --enable-mpi-cxx

- $ ./configure: se llama a la aplicacion para ejecutarla.
- --prefix=/home/fjimenez/prueba_curly/openmpi-4.0.3/compiled: ubicación del archivo o ejecutable.
- --with-mxm=/opt/mellanox/mxm: Comando para llamar a las librerías de Mellanox.
- --enable-mpi-cxx: Habilitar la opción de paso de mensajes en paralelo para C++.

Se puede denotar en las letras de color azul la bandera de configuración para decirle al compilador que se construya con mxm **"--with-mxm=/opt/mellanox/mxm"**, esto quiere decir que se construya o compile usando la herramientas propias del fabricante de la redes InfiniBand, con el objetivo de obtener todo el poder de la red de altas prestaciones y por consiguiente mejor rendimiento al momento de enviar a procesar grandes cantidades de información, esto nos asegura que el paso de mensajes sobre las aplicaciones de las diferentes ramas científicas sea la más óptima, sin embargo hay que denotar que también dependerá del grado de paralelismo de cada una de las aplicaciones.