National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name: Ankitha Nelaturi……..……

Student ID: x23152346……

Programme: Masters in Cloud Computing **Year:** 2023……

Module: Cloud Architecture……

Lecturer: ……

Submission Due Date: 27.11.2023……

Project Title: Upscaling On premise HPC of Institute Pasteur

Word Count: ……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Ankitha Nelaturi ……

Date: 27/11/2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Cloud Architecture]

[Upscaling On premise HPC of Institute Pasteur]

Your Name/Student Number	Course	Date
Ankitha Nelaturi/x23152346	Masters in Cloud Computing	27.11.2023

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
N/A	N/A	N/A

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[N/A]	
[N/A]	[N/A]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[N/A]

Additional Evidence:

[N/A]

Upscaling on premise HPC of Institute Pasteur

Ankitha Nelaturi
23152346
Department of Computing
National College of Ireland, Dublin, Ireland
x23152346@student.ncirl.ie

***Abstract*—In view of the present-day growth in biological data dimensions, scaling of computational capabilities is necessary. Analyzing these expanding repositories to drive research transformation, progressively stresses existing infrastructure as high-throughput devices continue to exponentially increase. In order to maximize significance, targeted migration beyond standard designs is required. This paper illustrates how organizations like Institute Pasteur will benefit from the adoption of Petascale HPC architecture with the goal to eliminate computational difficulties and enhance application performance. It also provides some insight into the underlying research being conducted in this area. Furthermore, the research analyzes the cost implications and performance improvements that occurred in the petascale solution.**

***Index Terms*—High performance computing(HPC), Total cost of ownership(TCO), Amdhal's Law, Literature based discovery(LBD)**

I. INTRODUCTION

The emergence of the COVID-19 pandemic in 2019 highlighted the crucial importance of biomedical research, emphasizing how staying up to date on scientific advances is critical in averting global calamities. The World Health Organization (WHO) saw firsthand the consequences of not being fully informed and prepared. This wake-up call highlighted the critical importance of biomedical research in understanding, combating, and minimizing the effects of new infectious illnesses. Biomedical research has made extraordinary contributions to understanding biological functioning and has provided important methodologies for disease prevention, such as hypertension control, cancer screening, and vaccines[1]. Swanson's methodology is one of the several approaches available for creating new findings in this discipline. It is a literature-based discovery(LBD) method that focuses on implicit links between documents, opening up a broader research field for information retrieval and knowledge discovery[2]. It is predicted on a well-known concept of transitivity: in the event that entities A and C have no known direct relationship but A and B and C have published relationships, one may conjecture the existence of a conceivable, novel, but unpublished indirect relationship between A and C[3]. Furthermore, this is ethically advantageous, as this form of research assures that no individuals will be subjected to harm throughout the course of the study.

To establish connections among different biological entities, the initial step involves creating a knowledge graph based on the existing database. PubMed and Spoke are primary repositories that experts are leveraging for their data sourcing. Pubmed is a free online database of biomedical literature, maintained by the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in the United States. It contains over 30 million citations and abstracts of biomedical literature, including articles from MEDLINE, life science journals, and online books. SPOKE is a biomedical knowledge network that integrates over 40 data sources into a graph with more than 50 million vertices (of 20 types) and more than 100 million edges (of 55 types). The data in SPOKE includes genomic associations with disease, chemical compounds and their binding targets, and metabolic reactions from select bacterial organisms of relevance to human health. SPOKE draws from seven standard biomedical ontologies in the National Center for Biomedical Ontology's BioPortal repository, and there are knowledge graphs for both. It is note worthy that SPOKE and PubMed represent two different perspectives of relations among medical concepts. While PubMed is a free online database of biomedical literature, SPOKE is a biomedical knowledge network that integrates over 40 data sources into a graph. The main aim here is to first calculate all pair shortest path in each graph so that we can understand how different entites are related together then compare the paths in the both graphs to check for swanson linking.

To compute the connections in the knowledge graphs one may use the well known APSP (All pair shortest path) algorithms. But implementing this on a dataset with 50 million vertices will have intensive calculations. Therefore to handle this task we have to use a supercomputer which is capable of handling this much of data. There are many different types of supercomputers that can be used and for our research we will be using a Petascale supercomputer. Petascale supercomputers can process one quadrillion flops. A supercomputer can compute much larger portion of the graph in a small amount of time. A thorough understanding of the complex interactions between genes, diseases, medications, and other biological entities is fostered by the seamless integration of disparate datasets, spanning from genetic studies to clinical trials. Using petascale computers allows researches to do intricate network analysis revealing relationships and patterns with enormous datasets. Its utility goes beyond expedited drug development, it enables the effective simulation of drug-bio entity interactions at a never-before-seen magnitude. The computational power of predictive modeling is extremely beneficial. Moreover, the possibility of real-time analysis of streaming biological data materializes, enabling scientists to react quickly to new trends and abnormalities.

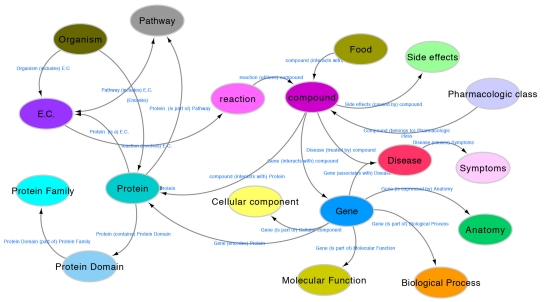

Fig. 1. Spoke representation.[4]

II. BACKGROUND

Founded in 1887 by the illustrious Louis Pasteur, the Institute Pasteur is a renowned biomedical research centre that epitomises scientific brilliance with its unwavering dedication to unravelling the mysteries of microbiology, immunology, and infectious illnesses. With a focus on virology, bacteriology, parasitology, mycology, immunology, neuroscience, genetics, and public health, the institute's broad research portfolio contributes to a thorough understanding of the biomedical sciences. Institute Pasteur has a long history of developing vaccines to prevent diseases like pertussis, diphtheria, and rabies. The organisation is still at the forefront of the search for novel treatments and preventative measures against infectious diseases. Its worldwide influence can be observed in the affiliated institutes and cooperative projects across the globe that prioritise information exchange and group efforts in the worldwide battle against infectious diseases. Beyond research, the center is important in teaching, with a thriving scientific community that hosts seminars and conferences that attract experts from all over the world. The museum, archive resources, and active participation in public health programs attest to Institute Pasteur's numerous accomplishments, making it a lasting leader in biomedical research with a major impact on world health.

Their research department's objective is to make it possible for everyone, not just them, to explore new fields. The post-Covid-19 era has witnessed a significant intensification of research efforts in the biological field, driven by heightened curiosity and a proactive approach. During the epidemic, the serious effects of bio-hazards on human health were made abundantly clear. The African research team sets a noteworthy example by quickly identifying the third strain of Covid-19, dubbed Omicron. This incident serves as a reminder of how crucial it is to increase our knowledge of bio-entities and promote teamwork in order to effectively mitigate and address new biological threats. Drawing inspiration from instances like these, the European High Performance Computing Joint Undertaking (EuroHPC JU) a joint initiative between the EU, European countries and private partners to develop a World Class Supercomputing Ecosystem in Europe[5]. This initiative taken in July 2021 allocates funds equivalent to 7 billion through 2021-2027.

At the Institute Pasteur, a focal point of current research revolves around literature-based discovery methodologies.The current state of research on literature-based discovery and knowledge graphs is impressive, as evidenced by the dense web of connections shown in Figure 2. The green node represents research undertaken at Institute Pasteur. In 2013, Michael J Cairelli and his team performed semantic predictions and literature-based discovery to elucidate a mechanism behind the obesity paradox [6]. Closely connected is a 2014 effort by Andrej Kastrin and colleagues, who explored link prediction using various similarity measures like common neighbors and Jaccard index [7]. More distantly related is the 2020 Wang node, referring to work by Lucy Lu Wang's team on the COVID-19 Open Research Dataset CORD-19. These examples demonstrate substantial efforts toward literature-based discovery. The cluster in the bottom right comprises research to improve knowledge graph algorithms, like optimizing multi-GPU computation of all-pairs shortest paths [8]. The upper cluster contains work related to Swanson's concepts of literature-based discovery and connecting findings from disjoint literature sets. As research continues, teams at Institute Pasteur are seeking to combine findings from PubMed and SPOKE. However, as established earlier, SPOKE is an enormous dataset. Biomedical repositories grow in size daily. Current supercomputing capabilities are only able to process 1/10,000th of SPOKE, creating a computational bottleneck.

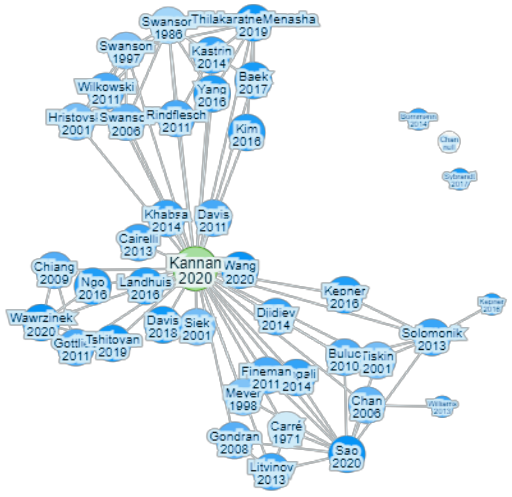

Fig. 2. correlation of research domains

Inspired by the EuroHPC JU initiative, the Pasteur Institute is committed to elevating its current High-Performance Computing (HPC) capabilities. The goal is to push this capacity into the petaflop range, which would represent a major advancement in biomedical knowledge extraction, with a theoretical peak of 86.1 TFlops. While our current infrastructure allows us to extract graphs from vast databases, the aspiration is to delve deeper. As a reflection of our commitment to advancing biomedical research and knowledge extraction, an improved computational framework is necessary to enable the comparative analysis of these complex graphs.

III. CURRENT STATE OF ART

Back in 2004, as PubMed was approaching towards its 16 million citations milestone, the Institute Pasteur tackled the challenge head-on by developing a supercomputer boasting an impressive 86.1 TFlops per clock cycle. This achievement gained particular significance when compared to the Blue Gene/L, which had clinched the title of the fastest supercomputer in the 2004 TOP500 list with a theoretical peak performance of 70.72 TFlops[9]. What stood out was the Institute Pasteur's adept resource utilization, surpassing the capabilities of contemporaneous supercomputers. Their success not only showcased technological prowess but also underscored a commitment to advancing research through effective resource management in the scientific community.

Algorithm 2 A Baselines 2D Distributed Block FW

1: **Input**: Distributed sparse matrix A;
2: **On each MPI process** pid **do in parallel:**
3: **for** k = 1, 2, 3 . . . nb **do**
4: Synchronize all processes
Diagonal Update
5: **if** pid owns A(k, k) **then**
6:A(k, k) $\leftarrow$ FW (A(k, k))
7: Send A(k, k) to Px(k) and Py(k)
Panel Update
8: **if** $p_{id} \in (k)$**then**
9: Wait for A(k, k)
10: A(k, :) $\leftarrow$ A(k, :)$\oplus A(k,k)(k,:)$
11: Send A(k, :) blocks to needed processes in Px(:)
12: **else**
13: Receive A(k, :) blocks if needed
14: **if** $p_{id} \in Px(k)$**then**
15: Wait for A(k, k)
16: A(:, k) $\leftarrow$ A(:, k)$\oplus A(:,k)(k,k)$
17: Send A(:, k) blocks to required processes in Py(:)
18: **else**
19: Receive A(:, k) blocks if required
MinPlus Outer Product
20: **for** i = 1, 2 . . . , nb, i $\oplus =$ k**do:**
21: **for** j = 1, 2 . . . , nb, j $\oplus =$ k**do:**
22: A(i, j) $\leftarrow$ A(i, j)$\oplus A(i,k)(k,j)$

The above algorithm is used to find the links between different biomedical concepts. This is specifically constructed to improve the parallel distribution of the tasks.

As we can see in the figure, the compute nodes are linked like a large matrix or grid. To improve the algorithm's efficiency, researchers have implemented a blocked version of the Floyd-Warshall algorithm. In this approach, all the nodes are mapped to coordinates on a grid (px,py), and the input matrix A is divided among these nodes to apply the Floyd-Warshall algorithm individually. After computing the results locally, there is a need to collect and combine them to get the final output. This is achieved efficiently by having each node broadcast its results only to its immediate horizontal and vertical neighbors in the grid, rather than all nodes. Specifically, a server at grid coordinates (pi,pj) shares its results with NODES(,j) and NODES(i,). By only transmitting results to neighboring nodes, the communication overhead is reduced compared to a full broadcast.

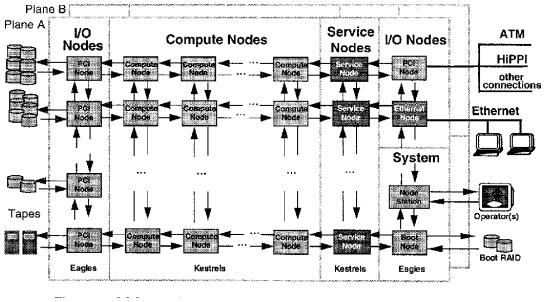

Fig. 3. Terascale supercomputer[10]

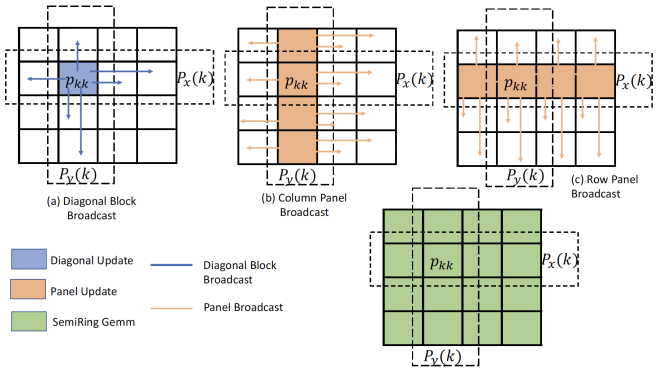

Fig. 4. Blocked Floyd-Warshall [11]

TABLE I
CURRENT SUPERCOMPUTER CONFIGURATIONS

Aspect	Specifications
Processors	IBM Power3-II Processors
Nodes	3584
Processors/Node	16
Total Processors	57,344
Peak Performance	86.1 teraflops
Memory	2 terabytes chnage this
Interconnect	IBM SP Switch2
Year of Deployment	2005

IV. PROPOSED HPC PETASCALE ARCHITECTURE

PubMed has about 36 million citations and abstracts of biomedical literature as of August 15, 2023[12], and this archive is still growing. We can now efficiently process the full PubMed graph utilizing a petascale high-performance computing (HPC) device by leveraging cutting-edge technologies. The need to keep up with technology is especially important for an institution that employs 2500 people and welcomes 900 students each year.

A. User Requirements

The computational workload driving the requirements for the HPC system primarily consists of optimized implementa-

tions of the Floyd-Warshall algorithm for solving the All Pairs Shortest Path (APSP) problem on large graphs. Asymptotic complexity analysis reveals that the standard Floyd-Warshall algorithm has a time complexity given by O(n^3), where n is the number of vertices in the graph.Based on the size of the graphs being analyzed, which scale up to 50 million vertices for "Spoke" graph constructions, and adjusting this vertex count to account for near-term growth projections to 100 million, the corresponding value for n used in design studies is 50 x 10^5.

Thus, the estimated computational load is quantified by the n^3 term, equating to 125 x 10^{15} floating point operations. Mapping this to high performance computing capabilities, it amounts to a system requirement of 1.25 petaflops (10^{15}) for real-time or interactive processing. The planned HPC installation at Institute Pasture will enable such cutting-edge computations using state-of-the-art equipment. Moreover, the extensive in-house expertise with over 2,500 employees and a continuous influx of 900 graduate student researchers per year will facilitate maximizing the scientific output and benefits of the new HPC infrastructure.

TABLE II
APSP ALGORITHM COMPARISON

Algorithm	Time Complexity
Floyd-Warshall	$O(n^3)$
Johnson's algorithm	$O(n^2 \log n + n^3)$
Bellman-Ford algorithm	$O(n^3)$
Dijkstra's algorithm (run for each node)	$O(n^3 \log n)$

$$\text{Computation required} = n^3 \tag{1}$$

$$\text{Total nodes in Spoke} = 50 \text{ Million} \tag{2}$$

$$\text{Adjusting size to } \frac{1}{100}th \tag{3}$$

$$\text{thus } n = 5 \times 10^5 \tag{4}$$

$$n^3 = 125 \times 10^{15} \tag{5}$$

B. Architecture

With the computational demands and usage scenarios established through close collaboration with the intended scientific user base, the process of architecting a petascale high performance computing (HPC) system next shifts to careful design tradeoffs and component selection considerations. Guided by both usage requirements and current technological capabilities, we utilize a co-design philosophy targeting an optimal balance of capabilities in various aspects of the holistic system.

I. Processing Subsystem and GPU's

The processing backbone leverages AMD's latest 4th Gen EPYC CPUs, specifically the 9654 model, as it demonstrates a 123% performance gain over 3rd gen EPYC processors like the 7763 [13]. As one of AMD's leading Zen 4 server processors, the EPYC 9654 provides 96 high-performance CPU cores and simultaneously supports 192 threads with SMT enabled. The CPU cores operate at a base frequency of 2.4 GHz, which at this frequency enables up to 3.686 double-precision teraflops of theoretical compute performance per chip[13]. The design also allows flexible frequency boosting up to 3.7 GHz, enabled by the TSMC 5nm manufacturing process. Each EPYC 9654 processor also packs 384 MB of L3 cache and 12 DDR5 memory channels. We paired the cutting-edge EPYC CPUs on the dual-socket nodes with NVIDIA's brand new H100 GPUs as accelerators. The flagship 80 GB H100 delivers up to 3 TB/s bandwidth between its HBM3 memory and the GPU cores. It achieves up to 60 TFLOPS of FP32 compute for HPC workloads via its fourth-generation Tensor Cores, which also reach 30 TFLOPS on reduced precision FP8 math for AI applications.By combining AMD's latest EPYC 9654 processors and NVIDIA's top-of-the-line H100 GPUs, we enable an exceptional heterogenous compute environment.

II. Server Nodes and Compute Nodes

The modular server node design is built on the ASUS RS700A-E12-RS4U platform, packing dual AMD EPYC 9654 series processors, 1TB of DDR5 memory, 80GB NVIDIA H100 GPUs, and high-speed local NVME storage in a dense 2U form factor.Instead of traditional HDDs, NVMe SSD storage provides massively higher throughput and IOPS for accessing large genomic and biomedical datasets. Specifically, each node integrates two high-performance 3.2TB NVMe M.2 SSDs delivering over 15 GB/s sequential reads. Additionally, two 7.6 TB U.2 NVMe SSDs add ultra-fast data storage.This optimized server configuration extracts maximum performance from the EPYC CPUs and H100 GPUs while avoiding storage bottlenecks for data-intensive access patterns. Furthermore, dual-port HDR InfiniBand adapters allow for redundancy and increased throughput. The modular node building blocks scale efficiently thanks to the high-performance interconnect fabric.Dedicated service nodes handle scheduling, management, monitoring and storage services. These 4 independent RS700A-E12-RS4U servers are equipped with two 16-core EPYC™ 9174F processors, for 32 cores per service node, paired with 768 GB of memory for infrastructure support duties across the cluster.

$$\text{No. Flops per CPU: } 3.686 \text{ TFlops} \tag{6}$$

$$\text{No. of FLOPS per GPU of FP32 instruction set: } 51 \text{ TFlops} \tag{7}$$

$$\text{Total FLOPS per server: } 58.372 \text{ TFlops} \tag{8}$$

$$\text{No. of compute nodes required: } \frac{125,000 \times 10^{12}}{59 \times 10^{12}} \approx 2141.437 \tag{9}$$

IV. Racks and Interconnect Network

To house the high performance computing cluster nodes, we selected rack enclosures optimized for density, thermal management, and ease of access during servicing. Specifically, the APC NetShelter SX 42U model provides ample vertical space at 1991mm height for integrating dual-socket server nodes and GPU accelerators, along with 600mm width ensuring sufficient horizontal clearance. The integrated cooling fans and 1070mm

TABLE III
ASUS RS700A-E12-RS4U PRICING

Aspect	Specification	Elaboration	Price (EUR)
PROCESSORS	2x AMD EPYC™ 9654	96 cores, 192 threads, 2.4GHz, 384MB Cache, 360W TDP	€16,354.00
GPU	1x Nvidia H100	80GB GDDR6, PCIe Gen5, Passive Cooling	€27,493.51
MEMORY	1.5 TB	24x 64GB DDR5 4800 MHz	€4,211.38
M.2 STORAGE	2x 800GB	M.2 NVMe, 5000 MBps Read / 1400 MBps Write, 3DWPD	€250.84
NVME STORAGE	4x 1.92TB	2.5" NVMe 4x4, 6800 MBps Read / 2700 MBps Write, 1DWPD	€693.94
PCIE EXPANSION	NVIDIA HDR100	4x Single Port, PCIe 4.0 x16	€877.37
TPM MODULE	Added	TPM 2.0 Module	€41.66
WARRANTY	5 Year	Aspen 5-yr System Advanced Parts Replacement Warranty	€1,077.38
OPERATING SYSTEM	Ubuntu 22.04 LTS Server	Latest stable release	€55.25

depth accommodate efficient front-to-rear airflow across all components.For node interconnect, we chose the Mellanox MCP1650 series cables and InfiniBand adapters to obtain 200 Gb/s HDR link speed meeting our application performance targets. The MCP1650 provides a passive copper twinax direct attach cable (DAC) rated for 3 meters length. Combined with Mellanox ConnectX-6 adapters inside each server, this provides 200Gb/s per port bi-directional bandwidth, conforming to the latest HDR200 InfiniBand standard.

$$\text{Number of Nodes:} \quad 2146 \tag{10}$$

$$\text{Nodes per Rack:} \quad 42 \tag{11}$$

$$\text{Total Racks:} \quad \frac{2146}{42} \approx 52 \tag{12}$$

VI. Cooling and Energy considerations

One of the best cooling systems for High-Performance Computing (HPC) is the 2-phase immersion cooling system offered by LiquidStack. LiquidStack's 2-phase immersion cooling system provides cooling capacities for unmatched server density, removing thermal bottlenecks at the device and facility levels. But using water becomes troublesome when it comes to managing the hot water, if we try to release it direcly into water bodies it will kill the organism and storing the water till it cools off will be another overhead. Rather than that we should use Cold Plate Cooling: This method involves attaching cooling plates directly to the heat-generating components, such as CPUs or GPUs, to efficiently remove heat. Cold plate cooling can be an effective alternative to liquid cooling for specific components within a server. Thus also helping with sustainability where no unecessary equipemts are used. Along the same terms of sustainability, we can harness 15% of energy requirements by installing solar panels.

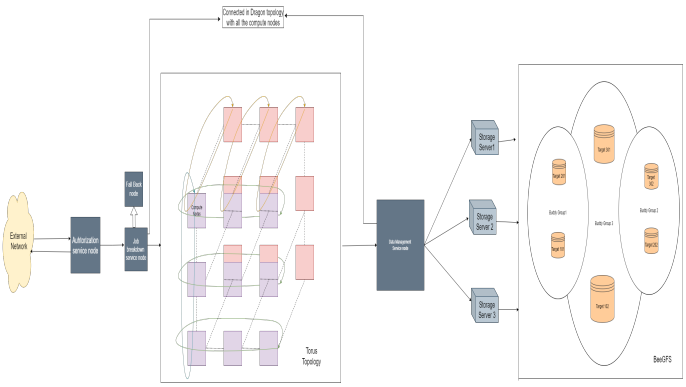

Fig. 5. Proposed High performance computing solution

C. Scalability and Redundancy Considerations

Scalability is crucial for any solution, enabling it to efficiently accommodate growth and changing demands. This attribute optimizes resource usage, ensuring cost efficiency by preventing unnecessary provisioning. Scalable systems can adapt to peak workloads, maintaining reliability and user satisfaction. Beyond immediate benefits, scalability future-proofs solutions, providing a competitive edge by facilitating innovation and rapid responses to market dynamics. In essence, scalability is the key to maintaining optimal performance and adaptability in dynamic environments.

At Institute Pasteur, the computational biomedical research demands are expected to grow exponentially, requiring infrastructure that flexibly scales. In order to provide such flexibility the proposed HPC involves generating intricate knowledge graphs from vast repositories covering interactions between innumerable biological entities. For this system to be scalable, each modular server building block grants storage, memory and acceleration headroom for future expansion. Moreover, the integrated BeeGFS parallel file system transparently grows filesystem capacity and access concurrency linearly with server nodes added. Overall for Institute Pasteur, the planned HPC design delivers excellent scalability on two fronts — adaptive capacity to handle data and job increases from the researchers in a pay as you grow cost effective model. Essentially the infrastructure can sustainably scale in-step with our trailblazing work.

To assess future scalibility andhals and guftan

1) Amdahl's Law: Amdahl's Law establishes the theoretical speedup limit achievable through parallelization, considering the proportion of serial and parallel code within an algorithm.This understanding is critical for recognizing the dimensions of scalability. In the context of the Floyd-Warshall algorithm, which can be represented algebraically as the matrix closure of a weight matrix, its primary operations involve sequential addition followed by a minimum operation. This inherent sequence results in approximately 50 percent of the algorithm remaining inherently serial, even when scaled to the level of distributed computing. However, this limitation can be mitigated through the adoption of optimized versions like MAGMA (Matrix Algebra on GPU and Multi-core Architectures)[14]. Researches can further apply vectorization, pipe-

lining, and auto-tuning to produce a faster implementation. This can lead to approximately 80 percent parallelization.

Assumptions:

- 4,284 processors
- 80% of code is parallelized
- 20% remains serial

Applying the values:

$$\text{Serial Percentage} = 0.2$$
$$\text{Parallel Percentage} = 0.8$$
$$\text{NumProcessors} = 4,284$$
$$\text{Speedup} = \frac{1}{(0.2 + \frac{0.8}{4,284})}$$
$$\approx \frac{1}{(0.2 + 0.000198)}$$
$$=4.995 \approx 5$$

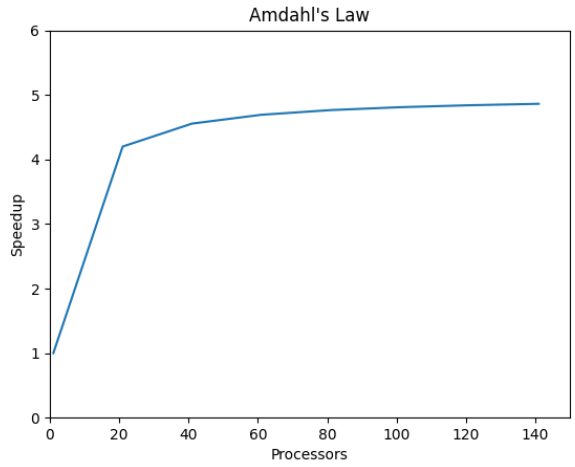

Fig. 6.

2) Gustafson's Law: Gustafson's Law serves as a counter part of Amdhals law, emphasizing the correlation between the number of processors and input size.

Gustafson's Equation:

$$\text{Scaled Speedup} = \text{Serial Percentage} + (\text{NumProcessors} - 1) \times \text{Parallel Percentage}$$

Plugging in the values: Serial Percentage = 0.2

$$\text{NumProcessors} = 4,284$$
$$\text{Parallel Percentage} = 0.8$$
$$\text{Scaled Speedup} = 0.2 + (4,284 - 1) \times 0.8$$
$$= 3426.6$$

In contrast to Amdahl's fixed 5x limit, Gustafson shows a scaled speedup of 2,600x using 4,050 processors. This better reflects parallel scaling for modern HPC systems. The large graph analysis workloads stand to gain tremendously via scaled acceleration.

D. Availability calculations of the proposed HPC

Understanding reliability of a system is just as crucial for organizations operating large-scale information systems due to the substantial data loss and outage risks failures introduce. MTBF (Mean Time Between Failures) and MTTF (Mean Time To Failure) constitute crucial reliability measures that determine the viability of deploying any computing infrastructure, whether cloud platforms or on-premise data centers. We determine MTBF after finding MTTF and MTTR. IT managers may reduce the risk of failure, guarantee system lifespan, and determine the overall cost of large computing systems by using precise MTBF and MTTF analytics. There are several reasons for for system failure, node level redundancy, modularity, power outages, circuit breakers and component RAS Features. Considering these points we can reasonably predict that the petascale HPC system will fail thrice a year.

- **MTBF (Mean Time Between Failures)** This is the average time a system operates between failures. Usually it is expressed in hours. It can be calculated using following formula:

$$\text{MTBF} = \frac{\text{Total Operating Time}}{\text{Number of Failures}}$$
$$= \frac{5 * 365 * 24}{3}$$
$$= 14600 \text{ hours}$$

- **MTTR (Mean Time To Repair)** This is the average time it takes to repair a system after a failure. We will computer MTTR and assume that the total downtime is 36 hours in a year. It can be calulated using following formula:

$$\text{MTTR} = \frac{\text{Total Downtime}}{\text{Number of Failures}}$$
$$= \frac{36 * 5}{3}$$
$$= 60 \text{ hours}$$

- **MTTF (Mean Time To Failure)** This is the average time a system operates before it fails. It can be calulated using following formula:

$$\text{MTTF} = \text{MTBF - MTTR}$$
$$= 14600 \text{ hours - } 60 \text{ hours}$$
$$= 14540 \text{ hours}$$

- **MTBR (Mean Time Between Repairs** This is the average time a system operates between repairs. It can be calulated using following formula:

$$\text{MTBR} = \text{MTTF + MTTR}$$
$$= 14540 + 60$$
$$= 14600 \text{ hours}$$

Now to calculate the availability of the proposed petascale HPC system:

$$\text{Availability} = \frac{\text{MTTF}}{\text{MTBF}} * 100$$

$$= \frac{14540}{14600} * 100$$
$$= 99.59\%$$

E. Ethical and Data governance

In order to improve healthcare delivery, research, and policy-making throughout the EU, the European Health Data Space (EHDS) intends to protect patient privacy and security while facilitating improved sharing and access to various forms of health data. The main legal framework regulating the EHDS is the EU Data Governance Act which sets baseline rules for voluntary data sharing [15]. Furthermore, guidelines for the lawful processing of personal data and the unrestricted flow of non-personal data are established by the EU General Data Protection Regulation (GDPR) [16]. To specifically address health data sharing for research, the EU adopted a Recommendation in 2021 which advises members states to facilitate cross-border access and establish agreements while adhering to ethical and legal requirements regarding consent, purpose limitation and data minimization [15].

Conducting research through the European Health Data Space (EHDS) necessitates strict compliance with the General Data Protection Regulation's (GDPR) data minimization, purpose limitation, storage, and access control requirements[15]. De-identified datasets must first be certified by the Data Protection Officer before loading for secondary analysis use cases like personalized medicine research. As per GDPR Article 5, health data shared in the EHDS will be subject to limited retention periods - deleted every 5 years and updated as necessary by data holders. Usage scope limitations will get defined in data processing agreements, allowing only analyses related to approved clinical trials. Only authenticated researchers associated with approved projects and institutions can gain time-bound access to pseudonymized datasets after formal agreement. The researchers must provide the necessary documentations confirming their credentials, research objectives, data needs and confirm to data protection obligations before any access. Overall, state-of-the-art data safeguards in conjunction with uncompromising transparency establish necessary public confidence in the realization of data-driven medicine through HPC innovation.

F. Estimated Linpack benchmark

The LINPACK Benchmarks are a measure of a system's floating-point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering[17]. As our supercomputer is not yet deployed Linpack algorithm can not be run on it and hence we benchmarking can not be done. But if we take boost clock speed of processor we can estimate Linpack benchbark. The top-bin Epyc 9654 part comes in a 320-400 watt TDP and provides 5.376 teraflops of peak double-precision performance running at max boost frequency of 3.5 GHz — over 10 teraflops in a dual-socket server[13].

No. of processors = 2 x 2142 = 4284 number of GPU = 2142 performance = 4284 x 10 TFLOPS + 2142 x 51 TFLOPS = 152.08 petaflops.

TABLE IV
PETA-SUPERCOMPUTER CONFIGURATIONS

Aspect	Specifications
Processors	AMD EPYC™ 9684X
Nodes	2025
GPU	NVIDIA H100
Total Processors	4,050
Peak Performance	125 petaflops
Memory	32 GB (per node)
Interconnect	Mellanox quantum
Year of Deployment	2024

V. MULTI-VARIABLE COST

A. Location Assumptions

- **Location:** Paris, France
- **Data Center Facility:** 800 square meters
- **Power Cost:** €0.18 per kWh The average electricity cost in France is roughly around €0.15 to €0.18 per kilowatt-hour (kWh), including all taxes[18]
- **Cooling Requirements:** 250 kW facility equipment output
- **User Support Team:** 4 full-time IT staff

B. Server Hardware Costs

- **Number of Server Nodes:** 2142
- **Per Node Cost:** €57158.88
- **Number of Service Nodes:** 4
- **Per Node Cost:** €45064.26
- **Total Servers Cost:** €122479385.22

C. Network Infrastructure Cost

We will use HPE P36051-001 InfiniBand HDR/Ethernet 200Gb 2-port QSFP56 PCIe4 x16 MCX653106A-HDAT Adapters.

- **Required Switches:** 52 switches
- **NVIDIA Quantum-2 MQM9700 switch Switch:** €44,263.24 [19]
- **Total Switches Cost:** $52\,\text{switches} \times 44,263.24 = 2,301,688$
- **Cabling Required:** 2142 meters
- **Cost per 3 meters of Cabling:** €96.81
- **Total Cabling Cost:** $2142\,\text{meters} \times \frac{96.81}{3} = 69,122$
- **Cost per Rack:** €2,117.50[20]
- **Total Racks Cost:** $52\,\text{racks} \times 2,117.50 = 110,110$
- **HPE HDR Adapters:** 4,284 adapters
- **Cost per Adapter:** €2959.09
- **Total Adapters Cost:** $4,284\,\text{adapters} \times 2959.09 = 12,676,741$

Total Network and Infrastructure Cost: $2,301,688 + 69,122 + 110,110 + 12,676,741 = 15,157,661$

D. Cooling System

- **Total processors needing cooling:** 4,292 CPUs
- **Total GPUs needing cooling:** 2146 GPUs
- **Total cooling plates needed:** 6,438

- **Motivair's Dynamic® Cold Plate cost estimation:** €300 per plate
- **Estimated Cooling System Cost:** €1,931,400

E. Power Consumption

- **Total Facility Load Estimate:** 1.1 MW
- **Power Cost Calculation:** $1.1\,\text{MW} \times 24\,\text{hrs/day} \times 365\,\text{days/yr} \times 0.18\,\text{per kWh}$
- **Annual Power Cost:** €1734.48

F. User Support Team

- **Team:** 4 IT support staff
- **Loaded Cost per Staff:** €72,000 per year (salary, benefits, taxes)
- **Total User Support Cost:** $4 \times 72,000 = 288000$ per year

G. Total Cost of Ownership Per Year

- **Server Hardware Cost:** €122479385.22
- **Network Infrastructure Cost :** €15157661
- **Cooling System Cost :** €1931400 per year
- **Power Consumption:** €1734.48 per year
- **User Support Staff:** €288000 per year
- **Total OPex for 5 years :** €4480244.4
- **Total CAPex :** €137637046.22
- **Total TCO for 5 years:** €142117290.62

H. Utility Computing

The presented supercomputer infrastructure has an annual total cost of ownership (TCO) of €28,423,458. This equates to an estimated hourly TCO rate of around €3,245. Institute Pasteur may promote the ideas of utility supercomputing to fellow researchers and collaborators by providing this world-class facility as a platform for collaborative research across the European Union. Allowing transparent access to external groups and institutions for worthy projects is consistent with EuroHPC JU's aim of cultivating a thriving environment that advances scientific breakthroughs.

VI. CONCLUSION AND FUTURE WORK

As we embarked on this expedition to architect a customized petascale supercomputer in response to the Continuous growth of biomedical datasets, this current achievement marks but a fleeting success. Our sights must turn with prescience towards the next milestone - crossing into the exascale computing regime in anticipation of ever-expanding data generation.

The Frontier system's debut heralds this forthcoming era, achieving 1.5 exaflops on mixed-precision workloads. Though immense in its present scale, our proposed system's modular architecture enables incremental expansion by deploying additional compute nodes. Incorporating next-generation processors and GPUs, networking fabrics supporting greater bandwidth, and storage media advancing in density can propel our 125 petaflop platform to surpass the exaFLOPS threshold within a couple years.

More broadly, continuity of scientific discovery and computational versatility has distinguished Institute Pasteur since its inception. Dedication towards matching infrastructural capabilities to researchers' trajectory has been integral throughout our organization's rich history. We continue standing at the vanguard today by proactively designing our flexible supercomputer for impending waves of clinical and genomic data. Seamless exascale upgrades will empower investigators with revolutionary modeling resolution for human cell simulations, multidimensional visualizations for real-time molecular interactions, and intricately layered omics integration. Pasteur's systems thinking approach synergizing foresight, modular designs and user collaboration persists catapulting possibilities at frontiers of life sciences.

REFERENCES

[1] Shine, K., 1998. The health sciences, health services research, and the role of the health professions.. Health services research, 33 3Pt 1, pp.439-45.

[2] L. Wang, W. Lin, and M. Zhao, "Literature-based discovery: a classical approach of information science," in ITM Web of Conferences, vol. 47, 2022, paper 02053, https://doi.org/10.1051/itmconf/20224702053.

[3] D. R. Swanson, N. R. Smalheiser, and V. I. Torvik, "Ranking indirect connections in literature-based discovery: The role of medical subject headings," Journal of the American Society for Information Science and Technology, vol. 57, no. 11, pp. 1427–1439, 2006.

[4] "Home," Scalable Precision Medicine Open Knowledge Engine. https://spoke.ucsf.edu/.

[5] European Commission. "High-Performance Computing Joint Undertaking." Digital Strategy, European Commission, https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing-joint-undertaking.

[6] M. J. Cairelli, C. M. Miller, M. Fiszman, T. E. Workman, and T. C. Rindflesch, "Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox."

[7] A. Kastrin, "Link Prediction on the Semantic MEDLINE Network," presented at the International Conference on Discovery Science, Oct. 2014.

[8] H. Djidjev, S. Thulasidasan, G. Chapuis, R. Andonov and D. Lavenier, "Efficient Multi-GPU Computation of All-Pairs Shortest Paths," 2014 IEEE 28th International Parallel and Distributed Processing Symposium, Phoenix, AZ, USA, 2014, pp. 360-369, doi: 10.1109/IPDPS.2014.46.

[9] "November 2004 — TOP500," www.top500.org. https://www.top500.org/lists/top500/2004/11/

[10] "ASCI White — Semantic Scholar," www.semanticscholar.org. https://www.semanticscholar.org/topic/ASCI-White/1749631 (accessed Nov. 23, 2023).

[11] R. Kannan et al., "Scalable Knowledge Graph Analytics at 136 Petaflop/s," SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2020, pp. 1-13, doi: 10.1109/SC41405.2020.00010.

[12] PubMed, "PubMed Overview," PubMed, 2019. https://pubmed.ncbi.nlm.nih.gov/about/

[13] HPCwire. "AMD's 4th Gen EPYC 'Genoa': 96 5nm Cores Across 12 Compute Chiplets." HPCwire, November 10, 2022. [Online]. Available: https://www.hpcwire.com/2022/11/10/amds-4th-gen-epyc-genoa-96-5nm-cores-across-12-compute-chiplets/

[14] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, and I. Yamazaki, "Accelerating numerical dense linear algebra calculations with gpus," Numerical Computations with GPUs, pp. 1–26, 2014.

[15] "Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance (Data Governance Act)," Regulation (EU) 2022/868, May 2022.

[16] "Data protection and the general data protection regulation (gdpr)," https://enterprise.gov.ie/en/Data-Protection/.

[17] Wikipedia contributors. "LINPACK benchmarks." Wikipedia, The Free Encyclopedia, [Year last updated, e.g., 2022]. [Online]. Available: https://en.wikipedia.org/wiki/LINPACK-benchmarks

[18] E. Team, "Electricity in Paris: Costs and How to Get it in Your Apartment," Culture Travel, Aug. 09, 2023. https://culturetravel.com/electricity-in-paris/

[19] “NVIDIA Quantum-2 MQM9700 — www.shi.com,” www.shi.com. https://www.shi.com/product/45990799/NVIDIA-Quantum-2-MQM9700 (accessed Nov. 26, 2023).
[20] “APC NetShelter SX, Server Rack Enclosure, 42U, Black, 1991H x 600W x 1070D mm - APC Netherlands,” www.apc.com. https://www.apc.com/shop/nl/en/products/APC-NetShelter-SX-Server-Rack-Enclosure-42U-Black-1991H-x-600W-x-1070D-mm/P-AR3100 (accessed Nov. 26, 2023).