

**PRAIRIE VIEW A&M UNIVERSITY
ROY G. PERRY COLLEGE OF ENGINEERING
DEPARTMENT OF MECHANICAL ENGINEERING
ELEG 4363- Introduction to High Performance Computing
FALL 2020**

I. COURSE DESCRIPTION IN ABET FORMAT

II ASSESSMENT RESULTS FOR CURRENT AND PAST THREE PERIODS

Program Outcomes Assessed in Course

Assessment Summary of Class Semester Averages for Outcomes

End of Semester Course Outcomes Assessment Reports

II. TESTS

Summary of Topics covered by Test

Original Test Questions & Master Solution/grading scheme used

(Summary of topics can be a separate page or include on the assignment itself)

Above Average (Graded original work)

Average (Graded original work)

Below Average (Graded original work)

III. HOMEWORK ASSIGNMENTS

Summary of Topics covered by Homework

Original Homework Questions & Master Solution/grading scheme used

Above Average (Graded original work)

Average (Graded original work)

Below Average (Graded original work)

IV. PROJECT

Summary of Topics Covered by Project

Project Problem Statement and Grading Scheme

Above Average (Graded original work)

Average (Graded original work)

Below Average (Graded original work)

[Include any other evaluation used and omit any method above method (Test, Quizzes, Homework, Laboratory and Project reports) that was not used]

Course Description in ABET Format

ELEG 4363 Introduction to High Performance Computing

1.	Course number and name:	ELEG 4363 Introduction to High Performance Computing
2.	Credits and contact hours:	3 credit hours; 3 contact hours
3.	Instructor's or course coordinator's name:	Dr. Suxia Cui
4.	Text book: Title:	N/A
	Author:	
	Publisher, Year & Edition:	
	a. Other Supplemental material	Instructor's handout
5.	Specific course information	
	a. Brief description of the content of the course (catalog description)	The course will inspect high performance computing hardware architecture, software tools, and applications.
	b. Prerequisites or Co-requisites	ELEG 3073.
	c. *Indicate whether a required, elective, or selected elective (as per definition provided) course in the program	Elective.
6.	Specific goals for the course	Students will be able to:
	a. Specific outcomes of instruction (e.g. The student will be able to explain the significance of current research about a particular topic.)	understand fundamentals of parallel programming; run jobs on high performance computing cluster; and know how to evaluate the performance of an HPC system.
	b. Explicitly indicate which of the student outcomes listed in Criterion 3 or any other outcomes are addressed by the course.	This course addresses: N/A
7.	Brief list of topics to be covered	Introduction (4) Unix/Linux Review (8) HPC Hardware (8) Parallel Computing (8) Performance Evaluation and Benchmark– STREAM (4) HPC Resource Allocation and Cybersecurity Awareness (4) Data Cloud and Internet of Things (IoTs) – ThingSpeak (4) Data Transfer and Science DM (4) HPC Applications (4)

II ASSESSMENT RESULTS FOR CURRENT AND PAST THREE PERIODS
Program Outcomes Assessed in this course

This course does not access any ABET outcomes.

Assessment Summary of Class Semester Averages for Outcomes
(Should be provided for Fall 2019 through Spring 2022 Assessment Period)

Outcomes	
Semester	Semester Class Average (%)
Fall 2020	
Comments on Trend of Data (Improving or deteriorating)	

*Expected Semester Class Average = 70% (or as defined by the program)
Ideal Semester Class Average = 75% (or as defined by the program)

End of Semester Course Assessment Reports Fall 2020 and Last Three Assessment Periods.

Each summary contains the following Information

1. Performance Statistics

- Total number of students in class
- The class average performance (**stated as %**) in each course outcome assessed in this class and the **two previous times course was assessed**.
- The acceptable class average for the class in each outcome area as set by your Department.
- Percentage of students who scored below the expected average in each outcome.

2. Implementation Summary

- Brief summary of plans implemented during the semester based on last assessment report

3. Perceived Problems

- Instructor's **critical evaluation** of perceived problems that affected students' performance. **Realistic problems should be identified throughout the semester and summarized in this section of the report.**

4. New Plans for Addressing Problems

- Instructor's plans for addressing the perceived problems the next time the class is taught. **These should be specific enough for another instructor to implement. Most of these should be things that, you, the instructor can implement to help students to improve. Be creative.**

5. Overall Trend over last three assessment periods

- Instructors summary of overall performance trend for each outcome measured over the last three periods.

5. Were Expectations Met?

- Answer **YES** if the **class average was above the expected average** and the **stated percentage of students** scored at or above the expected average, else answer **NO**.

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: MIDTERM EXAM

TOPICS COVERED IN MIDTERM EXAM

Midterm exam covered the following topics:

1. Concepts of HPC
2. Architecture
3. Hardware
4. Parallel computing

ELEG 4363 Advanced Topic in Computer Engineering

Midterm Exam Solution

Name:

1. (2 points)

(1) The `ls -l` command gives all the information except:

- a. The current date
- b. The last modified date
- c. File size
- d. File owner

Answer: A

(2) You want to move `halitosis.txt` to the file named `disease`. Which of the following command lines would you use to compute this?

- a. `mv disease halitosis.txt`
- b. `mv disease halitosis.txt/`
- c. `mv halitosis.txt disease/`
- d. `mv halitosis.txt disease`

Answer: D

(3) Which of the following is the correct command for removing a file?

- a. `rmv`
- b. `rm`
- c. `rm -f`
- d. `rm -p`

Answer: B

(4) The `whoami` command returns:

- a. Your username
- b. Your IP address
- c. Your name
- d. Your current date, time and location

Answer: A

2. (2 points) What is Top 500 List? Briefly introduce the top 5 on the most recent list.

The TOP500 project was launched in 1993 to improve and renew the Mannheim supercomputer statistics, which had been in use for many years.

Top 500 is a benchmark to rank systems and to decide on whether or not they qualify for the TOP500 list. The benchmark was Linpack, which means that systems are ranked only by their ability to solve a set of linear equations, $Ax = b$, using a dense random matrix A . Therefore, any supercomputer – no matter what its architecture is – can make it into the TOP500 list, as long as it is able to solve a set of linear equations using floating point arithmetic.

The top 500 list is a project that ranks with details of the 500 most powerful supercomputer in the world.

See top500.org for current ranking

3. (2 points) Explain Von Neumann architecture and its bottleneck.

Von Neumann architecture consists of a central processing unit that carries out the computations, and main memory that stores data and instructions. The separation of CPU and memory is often called the Von Neumann bottleneck since it limits the rate at which instructions can be executed.

The von Neumann architecture comprises of

- a processing unit that contains ALU and processor registers
- a control unit that contains an instruction register and program counter
- Memory that stores data and instruction
- External mass storage
- Input and output mechanism

The shared bus between the program memory and data memory has led to the von Neumann bottleneck.

(page 71)

4. (2 points) Explain the differences between shared-memory system vs. distributed-memory system.

Shared memory systems offers a single memory space used by all autonomous processors connected to the memory systems via an interconnect whereas in distributed memory systems all processors has its own private memory.

Shared memory and distributed memory are low-level programming abstractions that are used with certain types of parallel programming. Shared memory allows multiple processing elements to share the same location in memory (that is to see each other's reads and writes) without any other special directives, while distributed memory requires explicit commands to transfer data from one processing element to another.

Shared memory cache-coherent multiprocessor systems. The systems communicated with each other and with shared main memory over a shared bus. This meant that any access from any processor to main memory would have equal latency. Today these types of systems are not manufactured. Instead there are various point-to-point links between processing elements and memory elements. However, the idea of communicating directly through memory remains a useful programming abstraction. So in many systems this is handled by the hardware and the programmer does not need to insert any special directives. Some common programming techniques that use these abstractions are Open MP and Thread's.

Distributed memory has traditionally been associated with processors performing computation on local memory and then once it using explicit messages to transfer data with remote processors. This adds complexity for the programmer, but simplifies the hardware implementation because the system no longer has to maintain the illusion that all memory is actually shared. This type of programming has traditionally been used with supercomputers that have hundreds or thousands of processing elements. A commonly used technique is MPI.

5. (3 points) Please explain structure of Linux.

The Linux structure is made of the following components

The Kernel: The kernel handles tasks such as memory management, input output requests and program scheduling. It provides software connection to the hardware

The Shell or Graphical User Interface (GUI): This provides command line interface (CLI) that allows user to input command and translates this command into a form that the kernel will understand

The built-in system utilities: They are programs that allow users to perform tasks that involves complex actions but are basic to the operating systems. These actions are too complex to be built into the shell.

The application programs: These are additional program that come with the operating system or separately. This can be a basic utility or a full scale commercial application

6. (3 points) What is Amdahl's law? If 20% of the execution time may be speedup, and the improvement makes the affected part twice as fast, Use Amdahl's law to calculate the overall speedup of applying the improvement.

Amdahl's law: is a formula which gives the theoretical speedup in latency of the execution of a task at fixed workload that can be expected of a system whose resources are improved.

Amdahl's law can be formulated in the following way:

$$\begin{aligned} S(\text{latency}) &= 1/(1-p) + p/s \\ &= 1/(1-0.2) + 0.2/2 \\ &= 1.11 \end{aligned}$$

7. (3 points) Briefly introduce MPI and OpenMP. Compare their differences.

To use Intel MPI, the binary must be compiled and linked against the Intel MPI library with the Intel compiler. Include the module load command in your script as shown below to enable impi at runtime.

Module load impi

Mpirun -n [number of processes] ./executable_name.exe

To use Open MPI, the binary must be compiled and linked against the OpenMPI library with either the Intel compiler or GNU compiler.

Include the module load command in your script as shown below to enable openmpi at runtime. Also load the compiler that you used to compile your executable.

Module load openmpi

Module load compiler_modulefile_name/n.n.n

Mpirun -n [number of processes] ./executable_name.exe

8. (3 points) **What are processes, multitasking, and threads?**

A process is an instance of a program that is being executed. A process has the following components: an executable machine language program, a block of memory, descriptor of resources that the OS has allocated to the process, security information and information about the state of the process.

Multitasking gives the illusion that a single processor system is running multiple programs simultaneously. Each process takes turns running. (Time slice). After its time is up, it waits until it has a turn again. (Blocks)

A thread is a sequence of instructions or entity within a process that can be executed independently of others code.

ABOVE AVERAGE

Score for this quiz: 18.5 out of 20

Submitted Oct 6 at 2:59pm

This attempt took 59 minutes.

Question 1

/ 0.5 pts

The `ls -l` command gives all the information except:

File size

Correct!

The current date

File owner

The last modified date

Additional Comments:

Question 2

/ 0.5 pts

You want to move `halitosis.txt` to the file named `disease`. Which of the following command lines would you use?

Correct Answer

`mv halitosis.txt disease/`

You Answered

`mv halitosist.txt disease`

`mv disease halitosis.txt/`

`mv disease halitosis.txt`

Additional Comments:

Question 3

/ 0.5 pts

Which of the following is the correct command for removing file?

`rm -f`

Correct!

`rm`

`rm -p`

`rmv`

Additional Comments:

Question 4

/ 0.5 pts

The whoami command returns:

Your current date, time and location

Your IP address

Correct!

Your username

Your name

Additional Comments:

Question 5

/ 2 pts

What is Top 500 list? Briefly introduce the top 5 on the most recent list.

Your Answer:

The Top 500 list ranks the world's 500 most powerful supercomputers based on their performance as measured by the Linpack benchmark.

The top 5 on the most recent list are:

1) SUPERCOMPUTER FUGAKU - SUPERCOMPUTER FUGAKU, A64FX 48C 2.2GHZ, TOFU INTERCONNECT D By Fujitsu

2)SUMMIT - IBM POWER SYSTEM AC922, IBM POWER9 22C 3.07GHZ, NVIDIA VOLTA GV100, DUAL-RAIL MELLANOX EDR INFINIBAND BY IBM

3)SIERRA - IBM POWER SYSTEM AC922, IBM POWER9 22C 3.1GHZ, NVIDIA VOLTA GV100, DUAL-RAIL MELLANOX EDR INFINIBAND BY IBM

4)SUNWAY TAIHULIGHT - SUNWAY MPP, SUNWAY SW26010 260C 1.45GHZ, SUNWAY BY NRCPC

5)TIANHE-2A - TH-IVB-FEP CLUSTER, INTEL XEON E5-2692V2 12C 2.2GHZ, TH EXPRESS-2, MATRIX-2000 BY NUDT

Additional Comments:

Question 6

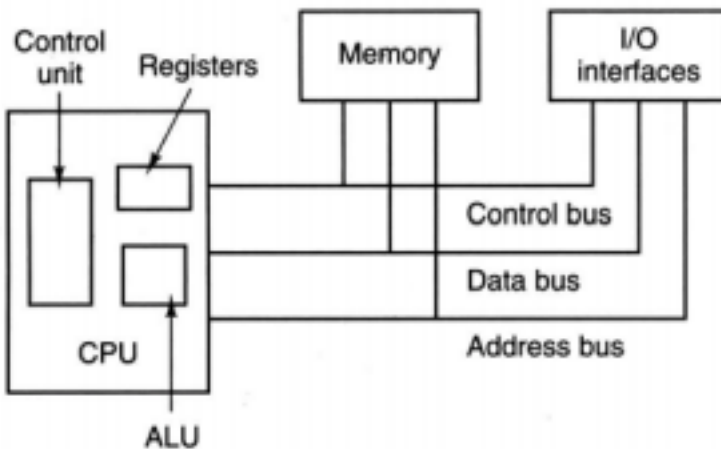
/ 2 pts

Explain Von Neumann architecture and its bottleneck.

Your Answer:

Von Neumann architecture

Von Neumann architecture is composed of three distinct components (or sub-systems): a central processing unit (CPU), memory, and input/output (I/O) interfaces.



bottleneck

The von Neumann bottleneck is a limitation on throughput caused by the standard personal computer architecture

Additional Comments: Program & data share the same bus causes the bottleneck.

Question 7

/ 2 pts

Explain the differences between shared-memory system vs. distributed-memory system.

Your Answer:

Distributed memory is a multiprocessor computer system in which each processor has its own private memory. Computational tasks can only operate on local data, and if remote data is required, the computational task must communicate with one or more remote processors. In contrast, a shared memory multiprocessor offers a single memory space used by all processors. Processors do not have to be aware where data resides, except that there may be performance penalties, and that race conditions are to be avoided.

Additional Comments:

Question 8

/ 3 pts

Please explain structure of Linux.

Your Answer:

The Linux Operating System's architecture primarily has these components: the Kernel, Hardware layer, System library, Shell and System utility.

The kernel is the core part of the operating system, which is responsible for all the major activities of the LINUX operating system.

System libraries are special functions, that are used to implement the functionality of the operating system and do not require code access rights of kernel modules

System Utility programs are liable to do individual, and specialized-level tasks.

Hardware layer of the LINUX operating system consists of peripheral devices such as RAM, HDD, CPU.

The shell is an interface between the user and the kernel, and it affords services of the kernel. It takes commands from the user and executes kernel's functions

Additional Comments:

Question 9

/ 3 pts

What is Amdahl's law? If 20% of the execution time may be speedup, and the improvement makes the affected part twice as fast, use Amdahl's law to calculate the overall speedup of applying the improvement.

Your Answer:

Amdahl's law is a formula which gives the theoretical speedup in latency of the execution of a task at fixed workload that can be expected of a system whose resources are improved.

Formula: $1/(F+(1-F)/P)$

Additional Comments: Please put 20% in the equation and calculate the result.

Question 10

/ 3 pts

Briefly introduce MPI and OpenMP. Compare their differences.

Your Answer:

MPI stands for Message Passing Interface . These are available as API(Application programming interface) or in library form for C,C++ and FORTRAN.

OpenMP stand for Open Multiprocessing . OpenMP is basically an add on in compiler. It is available in gcc (gnu compiler) , Intel compiler and with other compilers.

Their difference is that OpenMP is Only thread based parallelism, it has No such overheads, as thread can share variables. its threads have both private as well shared variable.

Process in MPI has private variable only, no shared variable. There are overheads associated with transferring message from one process to another.MPI target both distributed as well shared memory system

Additional Comments:

Question 11

/ 3 pts

What are processes, multitasking, and threads?

Your Answer:

A process is the execution of a program that allows you to perform the appropriate actions specified in a program.

Multithreading refers to multiple threads of execution within an operating system. In simple terms, two or more threads of a same process are executing simultaneously.

Thread is an execution unit that is part of a process. A process can have multiple threads, all executing at the same time. It is a unit of execution in concurrent programming. A thread is lightweight and can be managed independently by a scheduler. It helps you to improve the application performance using parallelism

AVERAGE

Score for this quiz: 17.5 out of 20
Submitted Oct 6 at 2:55pm
This attempt took 55 minutes.

Question 1

/ 0.5 pts

The `ls -l` command gives all the information except:

You Answered

The last modified date

Correct Answer

The current date

File owner

File size

Additional Comments:

Question 2

/ 0.5 pts

You want to move `halitosis.txt` to the file named `disease`. Which of the following command lines would you use?

Correct!

`mv halitosis.txt disease/`

`mv disease halitosis.txt/`

`mv disease halitosis.txt`

`mv halitosis.txt disease`

Additional Comments:

Question 3

/ 0.5 pts

Which of the following is the correct command for removing file?

Correct!

`rm`

`rm -p`

`rmv`

`rm -f`

Additional Comments:

Question 4

/ 0.5 pts

The whoami command returns:

Correct!

Your username

Your IP address

Your name

Your current date, time and location

Additional Comments:

Question 5

/ 2 pts

What is Top 500 list? Briefly introduce the top 5 on the most recent list.

Your Answer:

The Top 500 list is a list with the 500 most powerful computers in the world.

1. The Tianjin 1A is at the National Supercomputer Center in Tianjin is manufactured by NUDT, it is located in China, has 186,388 cores and uses 4.04 Mega Watts of power.
2. The Jaguar is at the Oak Ridge National Laboratory is manufactured by Cray, it is located in USA, has 224,162 cores and uses 6.95Mega Watts of power.
3. The Nebulae is at the National Supercomputing Centre in Shenzhen is manufactured by Dawning, it is located in China, has 120,640cores and uses 2.58 Mega Watts of power.
4. The Tsubame-2 is at the GSIC, Tokyo Institute of Technology is manufactured by NEC/HP, it is located in Japan, has 73,278 cores and uses 1.40 Mega Watts of power.
5. The Hopper is at the DOE/SC/LBNL/NERSC is manufactured by Cray, it is located in USA, has 153,408 cores and uses 2.91 Mega Watts of power.

Additional Comments: Tianjin 1A is not the top one.

Question 6

/ 2 pts

Explain Von Neumann architecture and its bottleneck.

Your Answer:

The Von Neuman architecture is relatively inexpensive and simple. Instructions and data share the memory device. It has only one set of data bus and address bus shared by the instruction and the data memory. Allows only one memory access at any time instant (either read instructions or access data in the data memory). Its Bottleneck is that its instructions and data must be continuously fed to the control arithmetic units, so that the speed of the memory interface poses a limitation on compute performance.

Additional Comments:

Question 7

/ 2 pts

Explain the differences between shared-memory system vs. distributed-memory system.

Your Answer:

A shared memory system the data is distribution is completely transparent to the user. In a distributed memory system the user has to distribute the data over the processors and also the data exchange between processor has to be performed explicitly when using message passing parallelization model.

Additional Comments: Shared memory?

Question 8

/ 3 pts

Please explain structure of Linux.

Your Answer:

The structure of Linux is

- A. Kernel – handles memory management, input and output requests, and program scheduling. (OS)
- B. Shell and GUI – basic Linux shells provides a “command line” interface which allows the user to type in commands
- C. Built in System Utilities – programs that allow a user to perform tasks which involve complex actions.
- D. Applications Software and Utilities – these are not part of the operating system.
- E. The SHELL communicates to the KERNEL and vice versa
- F. The APPLICATION communicates with both.

Additional Comments:

Question 9

/ 3 pts

What is Amdahl's law? If 20% of the execution time may be speedup, and the improvement makes the affected part twice as fast, use Amdahl's law to calculate the overall speedup of applying the improvement.

Your Answer:

Amdahl's law roughly says that unless virtually all of a serial program is parallelized, the possible speedup is going to be very limited regardless of the number of cores available.

Overall speedup = 1.23

F = .2

$$S = 2$$

$$1 / [(1-F)+(F/S)] = \text{overall speed up}$$

$$1 / [(1-0.2)+(0.2/20)] = 1.23$$

Additional Comments:

Question 10

/ 3 pts

Briefly introduce MPI and OpenMP. Compare their differences.

Your Answer:

Message Passing Interface (MPI) – focus is both on distributed and shared memory system

1. Available as API (Application programming interface)
2. Library form for C, C++ and FORTRAN
3. different MPI follow similar functionality, but with different performance efficiency.

OpenMP is basically an add on in compiler

1. target shared memory systems
2. it is available in gcc (gnu compiler), Intel compiler and with other companies.

Additional Comments:

Question 11

/ 3 pts

What are processes, multitasking, and threads?

Your Answer:

Processes are an instance of a computer program that is being executed.

Multitasking gives the illusion that a single processor system is running multiple programs simultaneously. Each process takes turns running and after its time is up, it waits until it has a turn again.

Threading are threads contained within programs processes. They allow programmers to divide their programs into (more or less) independent tasks.

BELOW AVERAGE

Score for this quiz: 15 out of 20
Submitted Oct 6 at 3pm
This attempt took 60 minutes.

Question 1

/ 0.5 pts

The `ls -l` command gives all the information except:

File size

Correct!

The current date

File owner

The last modified date

Additional Comments:

Question 2

/ 0.5 pts

You want to move `halitosis.txt` to the file named `disease`. Which of the following command lines would you use?

You Answered

`mv halitosis.txt disease`

Correct Answer

`mv halitosis.txt disease/`

`mv disease halitosis.txt/`

`mv disease halitosis.txt`

Additional Comments:

Question 3

/ 0.5 pts

Which of the following is the correct command for removing file?

`rm -p`

`rm -f`

`rmv`

Correct!

`rm`

Additional Comments:

Question 4

/ 0.5 pts

The whoami command returns:

Your name

Your IP address

Correct!

Your username

Your current date, time and location

Additional Comments:

Question 5

/ 2 pts

What is Top 500 list? Briefly introduce the top 5 on the most recent list.

Your Answer:

The Top 500 list is a list of the top project ranks & 500 most powerful computer systems in the world.

1. Supercomputer Fugaku- Supercomputer made in Japan it has 7,299,072 cores, Rmax= 415,530.0 TFlop/s , Rpeak= 513,854.7 TFlop/s, Power= 28,335 kW
2. Summit- IBM Power system made in the U.S. it has 2,414,592 cores, Rmax= 148,600.0 TFlop/s , Rpeak= 200,794.9 TFlop/s, Power= 10,096 kW
3. Sierra- IBM Power System made in the U.S. it has 1,572,480 cores, Rmax= 94,640.0 TFlop/s , Rpeak= 125,712.0 TFlop/s, Power= 7,438 kW
4. Sunway TaihuLight- Sunway MMP made in China it has 10,649,600 cores, Rmax= 93,014.6 TFlop/s , Rpeak= 125,435.9 TFlop/s, Power= 15,371 kW
5. Tianhe-2A - TH-IVB-FEP Cluster made in China it has 4,981,760 cores, Rmax= 61,4r4.5TFlop/s , Rpeak= 100,678.7 TFlop/s, Power= 18,482 kW

Additional Comments:

Question 6

/ 2 pts

Explain Von Neumann architecture and its bottleneck.

Your Answer:

The Von Neumann Architecture is the standard model of computer hardware. It consists of a CPU that carries out computations, main memory that stores data & instructions. It's bottleneck is the separation of CPU & Memory which limits the rate in which instruction is executed

Additional Comments:

Question 7

/ 2 pts

Explain the differences between shared-memory system vs. distributed-memory system.

Your Answer:

Distributed Memory System

- processors can see limited memory

Shared Memory System

-

Additional Comments:

Question 8

/ 3 pts

Please explain structure of Linux.

Your Answer:

the Kernel

Additional Comments:

Question 9

/ 3 pts

What is Amdahl's law? If 20% of the execution time may be speedup, and the improvement makes the affected part twice as fast, use Amdahl's law to calculate the overall speedup of applying the improvement.

Your Answer:

Amdahls Law states unless virtually all of a serial program is parallelized, the possible speed up is going to be very limited.(regardless of the number of cores available.

$$T_{\text{parallel}} = 0.2 \times T_{\text{serial}} / P$$

Additional Comments:

Question 10

/ 3 pts

Briefly introduce MPI and OpenMP. Compare their differences.

Your Answer:

Message Passing Interface, MPI

- originally for distributed memory systems but now used for both distributed memory & shared memory

- Available as API

- good for problems involving large memory

Open Multiprocessing OpenMP

- launches a single process which it can create a number of threads as desired
- based on thread approach
- targets shared memory systems
- based on fork and join method
- programming in OpenMP is relatively easy
- user determines number of threads used.

Additional Comments:

Question 11

/ 3 pts

What are processes, multitasking, and threads?

Your Answer:

Processes- an executing instance of program or running program.

Thread- a subset of process.

Multitasking- rapidly switching among executing programs

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: HOMEWORK 1

TOPICS COVERED IN HOMEWORK 1

1. Background of HPC

ELEG 4363 Intro to HPC
Homework 1

Based on introduction slides, find one topic or keyword of your interest to develop one-page (single space) informative memo. You are allowed for one page of further explanation on a topic or a key word, and please be sure to cite and reference external resources in your memo. (See attached a guidance)

Suggested heading:

To: Intro to HPC Class

From: Your Name

Date:

SUBJECT: The topic or keyword of your choice

Memo Begins.

ABOVE AVERAGE

To: Intro to HPC Class

From: Brychan Johnson-Jackson

10

Date: 09/01/2020

SUBJECT: HPC Cloud Computing

HPC is an acronym for High Performance Computing and is also known as Supercomputing. It uses parallel processing for running advanced program applications effectively, efficiently, quickly, and reliably. HPC is used for computationally intensive tasks such as weather forecasting, galaxy formation, gene sequencing, quantum mechanics, signal processing, and much more. Super computing is all about increasing the speed of computing and decreasing size of the technology.

There are different forms of HPC. One form is The Commodity of HPC Cluster. A HPC Cluster is a collection of hundreds to tens of thousands of servers connected, working together on a single task. Another form of HPC is a Dedicated Supercomputer. A Supercomputer is an extremely powerful computer with the highest level of performance.

An additional form of HPC is Grid Computing. Grid computing is when multiple HPC Clusters are connected together. To create a single system that has the power of multiple systems focusing on a single assignment. An alternative form of HPC is HPC Cloud Computing. HPC Cloud Computing is similar to Grid Computing, but with the end user having less control. HPC Cloud Computing is composed of virtual HPC Clusters that can be accessed remotely. HPC Cloud Computings has the capability to offer both infrastructure and software as services.

There are many benefits of HPC Cloud Computing. It increases scalability and flexibility, increases productivity, accuracy, can be accessed remotely, and is cost effective. Risk management is another benefit of cloud computing by allowing businesses to transfer risk to their vendors. Types of risk that can be avoided include capacity utilization, development, regulatory, adoption, and risk of operational failure.

Flexibility and scalability are accomplished through combining internal resources into a private cloud and public cloud resources. Productivity is increased by running on large clusters, using parallel computing, parametric sweeps, performance optimizations to analyze many options for design. The accuracy is increased by removing CPU and memory limitations. The HPC Cluster being on a Cloud gives users remote access to the system. The Cloud places layers between the hardware and the user.

AVERAGE

9.5

To: Intro to HPC Class
From: Racine McLean
Date: September 1st, 2020
SUBJECT: Parallel Computing

The Computer Science discipline, Parallel Computing have been an area of interest for years now, it dates to the late 50's with advancements surfacing in the form of supercomputers throughout the 60's and 70's. According to GeeksforGeeks Parallel computing "...is the use of multiple processing elements simultaneously for solving any problem. Problems are broken down into instructions and are solved concurrently as each resource which has been applied to work is working at the same time." Due to this process a computer can take on much larger tasks.

The stages in which Parallel computing developed over the years are; (1950's, 1960's, 1970's) shared memory multiprocessors- multiple processors working side by side on shared data; (1980's) massively parallel processors(MPPs)- dominate the top end of computing, with the ASCI Red supercomputer computer in 1997 breaking the barrier of one trillion floating point operations per second; (late 80's) Clusters- A cluster is a type of parallel computer built from large numbers of off-the-shelf computers connected by an off-the-shelf network.

Currently Parallel computing is becoming popular due to Multi-core processors, most computers sold today have dual-core microprocessors and quad-core readily available. According to Intel "...Chip manufacturers have begun to increase overall processing performance by adding additional CPU cores. The reason is that increasing performance through parallel processing can be far more energy-efficient than increasing microprocessor clock frequencies. In a world which is increasingly mobile and energy conscious, this has become essential."

The future of parallel computing, over the years we have seen a progression of CPUs getting better and better, and with the introduction of GPUs computer processing power are remarkable. Chip maker companies are always looking to make their technology better than it was yesterday, according to Glenn Reid, "These days, AI and Big Data (in addition to games) are areas where parallel processing play big time. GPU's are in use today, and Google's TPU's have shown that newer chip designs can run faster and consume less power - very valuable in the AI market, primarily for performance reasons. Software has already been created in the Big Data environment to manage decomposed processes that run on 1000's of machines simultaneously to give the needed performance boost." the main takeaway here is that there are application that need parallel computing, without it these applications are unusable.

Bibliography

universedecoderCheck out this Author's contributed articles., et al. "Introduction to Parallel Computing." GeeksforGeeks, 14 Dec. 2018, www.geeksforgeeks.org/introduction-to-parallel-computing/.
https://www.intel.com/pressroom/kits/upcrc/ParallelComputing_background.pdf

BELOW AVERAGE

8.5

ELEG 4363 Intro to HPC Homework 1

Suggested heading: Cloud Computing

To: Intro to HPC Class

From: Virgie Leyva

Date: 09/01/2020 SUBJECT: Cloud Computing Good evening,

Cloud computing is used in our everyday lives. One form of cloud computing that is used daily is Google's Gmail. Emails are checked multiple times a day, everyday. The way it works is Gmail users check their email by utilizing applications and files hosted by Google through the internet. This allows the user to check their email through various methods such as tablet, laptop, phone, and watches. This makes cloud computing efficient because applications and data are made readily available from any device connected to the internet.

There are three types of cloud computing. The first type is IaaS which stands for infrastructure as a service. IaaS is the lowest level cloud service because it has instant computing infrastructure and is solely managed over the internet. This service can provide extra storage or enable access to a HPC. Some examples are Microsoft Azure, Amazon Web Service, and Google Compute Engine. There are some disadvantages to IaaS such as security, flexibility and technical problems. Another type of cloud computing is PaaS which stands for platform as a service. This service delivers a platform to users so that they may develop, initialize and manage applications. Some examples of PaaS are Google's App Engine, IBM BlueMix and Apache Stratos. Just like any platform there are some disadvantages such as data security issues, and outages. Lastly, SaaS which stands for software as a service. This is an on demand type of product that is delivered to the user over the web. Most SaaS come with a subscription that the user pays for monthly. Dropbox, Google Apps, and Concur are some types of SaaS. Some disadvantages of SaaS are low performance, integration can be troublesome, and data security.

Cloud computing has help make many advances throughout the years. Through Cloud Computing, technical departments can fulfill any of their needs.

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: HOMEWORK 2

TOPICS COVERED IN HOMEWORK 2

- 1. Processors for HPC cluster**
- 2. CPU vs. GPU**
- 3. IBM HPC cluster**

ELEG4363 Intro to HPC

Homework 2

Based on various HPC clusters from different manufacturers, the selection of microprocessors (including CPU and GPU) is also different. Please list at least four popular CPU and/or GPU in HPC cluster. Briefly introduce and compare them in one article. (2 pages)

Pick one of the microprocessor from the above article and develop PowerPoint slides to further introduce the product in details. Suggest to have 10 slides or more with manufacture, architecture, application, and etc. You can also compare one group of microprocessors. For example, my IBM iDataPlex cluster at PVAMU has Intel Xeon processor E5-2600 series, find the introduction of it, and also compare with the top on the list in the E5 series to see the development for the past years. You can use class handouts as a starting point. Search for other resources to enhance your writing. Cite all the references at the bottom of your paper.

ABOVE AVERAGE

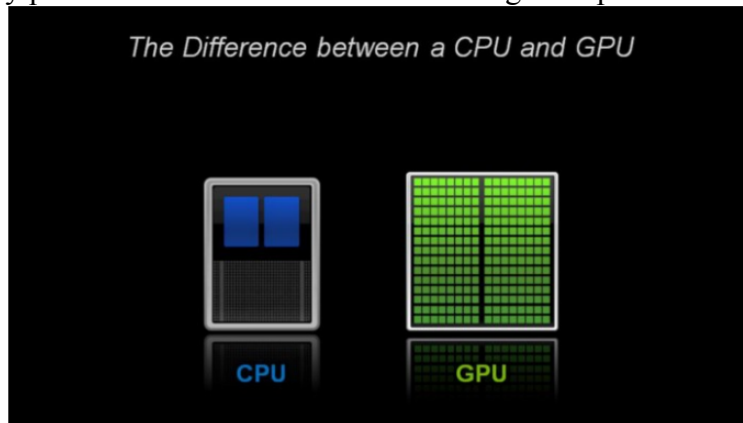
9.5

To: Intro to HPC Class

From: Racine McLean

Date: September 8th, 2020

CPUs and GPUs process tasks in different ways. Regarding interrelations, they are often compared with brain and brawn. A CPU (the brain) can work on a variety of different calculations, while a GPU (the brawn) is best at focusing all the computing abilities on a specific task. That is because a CPU consists of a few cores (up to 24) optimized for sequential serial processing. It is designed to maximize the performance of a single task within a job; however, the range of tasks is wide. On the other hand, a GPU uses thousands of smaller and more efficient cores for a massively parallel architecture aimed at handling multiple functions at the same time.



Source: blogs.nvidia.com

NVIDIA TESLA V100 - GPU

DATA CENTER GPU_s



NVIDIA V100 FOR NVLINK

Ultimate performance for deep learning.

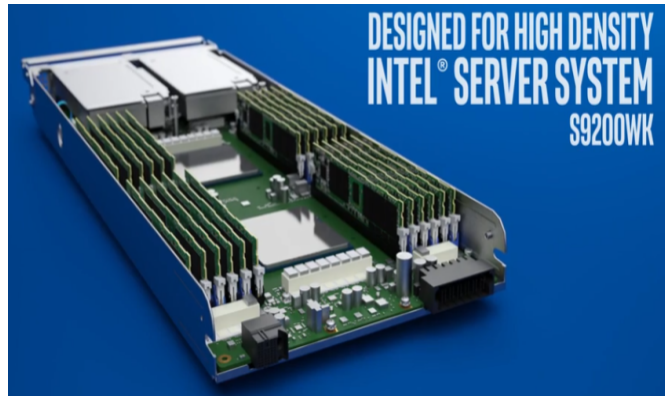


NVIDIA V100 FOR PCIe

Highest versatility for all workloads.

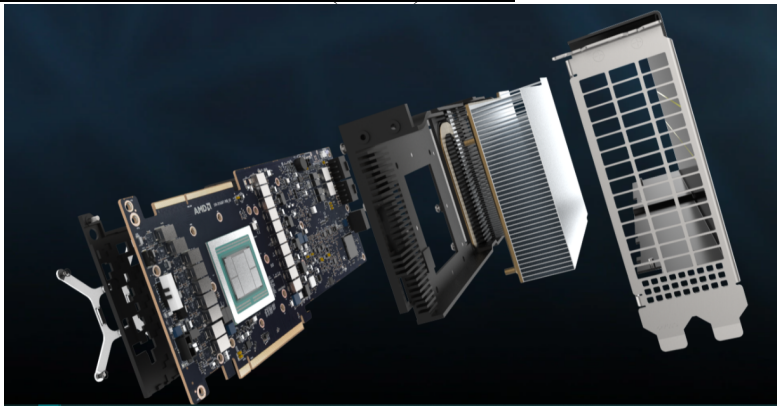
NVIDIA® Tesla® V100 is built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta™, the NVIDIA Tesla models offer the performance of 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once impossible. NVIDIA® V100 Tensor Core is the most advanced data center GPU ever built to accelerate AI, high performance computing (HPC), data science and graphics. It is powered by NVIDIA Volta architecture, comes in 16 and 32GB configurations, and offers the performance of up to 32 CPUs in a single GPU. Data scientists, researchers, and engineers can now spend less time optimizing memory usage and more time designing the next AI breakthrough

Intel® Xeon® Platinum 9200 Processors - CPU



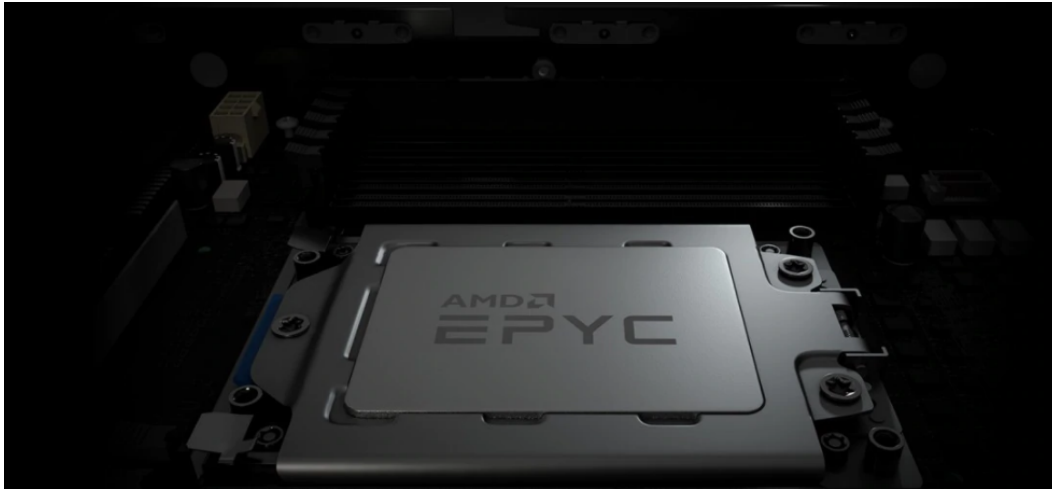
A new class of Intel® Xeon® Scalable processors is here! Designed to optimize performance for high performance computing (HPC) and AI (Artificial Intelligence) workloads: Simulation, Modeling, and AI Training and Inference. Introducing the advanced performance of Intel® Xeon® Platinum 9200 processors. Leadership compute performance with 32, 48 and 56 high-performance cores. Architected for the most demanding workloads. LINPACK up to 5.8x1 (Estimated vs. AMD EPYC 7601). Stream Triad up to 1.4x2 (Estimated vs. AMD EPYC 7601). Unprecedented memory bandwidth with 12 DDR4 memory channels to fuel real-world applications.

AMD Radeon Instinct™ MI50 Accelerator (32GB) - GPU



The Radeon Instinct™ MI50 compute card is designed to deliver high levels of performance for deep learning, high performance computing (HPC), cloud computing, and rendering systems. This new accelerator is designed with optimized deep learning operations, exceptional double precision performance, and hyper-fast HBM2 memory delivering 1 TB/s memory bandwidth speeds. The Radeon Instinct™ MI50 server accelerator designed on the world's first 7nm FinFET technology process brings customers a full-feature set based on the industry newest technologies. the Radeon Instinct™ MI50 brings customers the compute and memory performance needed for enterprise-class, mid-range compute capable of training complex neural networks for a variety of demanding machine deep learning applications in a cost-effective design.

AMD EPYC™ 7002 Series Processors - CPU



Whether you run enterprise applications, virtualized and cloud computing environments, software-defined infrastructure, high-performance computing, or data analytic applications. According to AMD “EPYC™ processor-based systems are #1 on industry benchmarks, including those measuring integer, floating-point, virtualization, database, and HPC performance¹. AMD EPYC™ 7742 processor has set new world records that establish AMD as THE performance leader.” A balanced and optimized IT infrastructure, workload acceleration, and data protection is a must for enterprises to be competitive. AMD has been on the forefront of solving these challenges by delivering leadership innovations. The new AMD EPYC™ 7Fx2 Series Processors provide businesses advanced security features and ground-breaking performance with Hyper Converged Infrastructure (HCI), Databases (DB) and Commercial High-Performance Computing (HPC) applications – enabling faster business insights and time to volume while helping optimize TCO.

AMD EPYC™ has been engineered for data centers that rely on CPU performance. From oil and gas exploration, to in-memory databases, to big data analytics to production rendering to standard data center applications, highly parallel workloads have more cores to work with. AMD EPYC™ 7002 generation processors scale from 8 to 64 cores (16 to 128 threads per socket). No other x86 vendor today enables such a core density in the market⁵.

In an article Olena states “GPUs are not replacements for CPU architecture. Rather, they are powerful accelerators for existing infrastructure. GPU-accelerated computing offloads compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. From a user’s perspective, applications just run much faster.” The industry of HPC still has room to grow and many of these manufacturers are pushing the bar each year, after all consumers always need faster computing power.

BIBLIOGRAPHY

<https://www.amd.com/en/processors/epyc-7002-series>

<https://www.intel.com/content/www/us/en/products/servers/server-chassis-systems/server-board-s9200wk-systems.html>

<https://www.nvidia.com/en-us/data-center/v100/>

<https://medium.com/altumea/gpu-vs-cpu-computing-what-to-choose>

AVERAGE

To: Intro to HPC Class

From: Kobi Tioro

Date: 09/01/2020

SUBJECT: Based on various HPC clusters from different manufacturers, the selection of microprocessors (including CPU and GPU) is also different. Please list at least four popular CPU and/or GPU in HPC cluster. Briefly introduce and compare them in one article. (2 pages)

A Central Processing Unit, also called CPU or processor is an electronic circuitry within a computer that executes instructions that make up a computer program. It is also called the brain of the computer because it performs basic arithmetic, logic, controlling, and input/output (I/O) operations specified by the instructions in the program.

According to guru99 website, Core i5 is a mid-range processor developed and manufactured by Intel. It is dual-core or quad-core processors. You can use it on both desktop and laptop computers. Core i7 is a family of high-end performance x86 64-bit processors developed by Intel. It is designed for high-end desktops and laptops. It is the first processor generation using Nehalem Microarchitecture.

The advantage of the intel core i5 is that it has a high-speed performing rate so that system are able to perform at the maximum CPU rate of 3.6 GHz.. The Turbo technology is present in the device which helps you to boost up the working speed Also, It offers 64-bit architecture to get reliable working. For the intel core i7, the benefits are multiple. It has a very fast processing speed. It also offers highly reliable cooling system. Four cores allow for handling software that requires lots of computations. It also provides high data visualization to users that help them to get high-quality images and video graphics.

One disadvantage of intel core i5 is that it does not support high data visualization technology for users to view high-quality images and video graphics. Also the power consumption of core-i7 is not better compared to the core-2 duo processor type. It demands newer motherboards. The i5 Processor is sensitive to higher voltages. For the disadvantages of intel core i7, It is Relatively costly Processor, the Power consumption is high compared to other processors. The i7 processors work only with DDR3 Memory, which means that users upgrading from DDR2 will require to have a new motherboard. The is not many software needed for multithreading, which means the average users do not get much performance gain.

Intel 5	Intel 7
Core i5 CPU score is 13,179	Core i7 CPU score of 13,995
Released on January 2010 to April 2018.	Released on September 2009 to April 2018.
Video transcoding time of Core i5 CPU took 18 minutes and 28 seconds.	The time of video transcoding is 17 minutes and 19 seconds.

The clock speed of Core i5 is 1.2-3.6 GHz	The clock speed of corei7 is 1.3-3.5 GHz.
The battery of Core i5 processor last for 14 hours and 45 minutes.	The Core i7 battery survived for 10 hours and 49 minutes, which is slightly better than i5 Processor.
Core i5 Processor is dual-core or quad-core.	Core i7 processor is dual-core, quad-core and hex-core processor architectures.
Hyper-Threading is not possible in Core i5.	Hyper-Threading is possible in Core i7.
Offer mid-range graphics.	It offers best in class graphics
The cache size of Core i5 is 3MB-6MB.	The cache size of Core i7 is 4MB-8MB.
Numbers of 4 cores for desktop and 2 Core for Mobile.	4 cores/6 core for desktop and 2 Core and 4 Cores for Mobile.
It uses the DMI bus type.	It uses the QPI bus type.
Offers embedded options.	Not offers an embedded option.
Video transcode a 6.27GB, 12-minute-and-30-second 4K video to 1080p.	The Core i7, did this in 17 minutes and 19 seconds, putting a 1:09 (6.2 percent) gap between them.

A graphics processing unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.

AMD Radeon 5700 vs. Nvidia GeForce RTX 2060 Super

	AMD Radeon RX 5700	GeForce RTX 2060 Super
Architecture	RDNA	Turing
Date of Introduction	6/10/2019	7/2/2019
Manufacturing Process / Die Size	7nm / 251mm ²	12nm / 445mm ²
Transistor Count	10.3 billion	10.8 billion
Processing Cores	2,304	2,176
Base/Boost GPU Clocks	1,465MHz / 1,725MHz	1,470MHz / 1,650MHz
Video Memory	8GB GDDR6	8GB GDDR6
Memory Interface / Bandwidth	256-bit / 448GBps	256-bit / 448GBps
Board Power	180 watts	175 watts
Power Connectors	One six-pin, one eight-pin	One eight-pin
Recommended Launch Price	\$349	\$399

Table from [pcmag.com](https://www.pcmag.com)

According to the table from, we see both GPUs are using different architectures. The AMD Radeon RX 5700 is built using RDNA architecture and the GeForce RTX 2060 is using Turing architecture. Both graphic cards has the same video memory of 8GB GDDR6 and the same memory interface of 256 bit/448GBps. The AMD Radeon RX 5700 costs \$50 more than the GeForce RTX 2060. The AMD Radeon RX 5700 has more transistors than the GeForce RTX 2060. The AMD Radeon RX 5700 has higher Boost GPU clocks, more processing cores, more power connectors and higher board power than the GeForce RTX 2060.

Work Cited

“Intel i5 vs i7: Ultimate Showdown.” *Guru99*, www.guru99.com/i5-vs-i7.html.

“Nvidia GeForce RTX 2060 Super vs. AMD Radeon RX 5700: Which Midrange Gaming Card Is Better?” *PCMag*, www.pcmag.com/comparisons/nvidia-geforce-rtx-2060-super-vs-amd-radeon-rx-5700-which-midrange-gaming.

BELOW AVERAGE

8.8

Virgie Leyva Dr. Cui September 8, 2020

ELEG 4363 Intro to HPC Homework 2

As much as HPC clusters vary so do their components. Different manufacturers select different CPUs or GPUs used within their clusters. There are various HPC clusters made by different manufacturers such as IBM, AMD, Lenovo, Cluster Vision and Dell EMC to name a few. Each providing a specific need to their customers.

The IBM HPC cluster at Prairie View A&M utilizes the E5 2630 processor. This cluster was with specific accommodations in mind. Some constraints are the amount of floor space available, how much power needed and cooling conditions. One of the key features is that a user can use a highly efficient low voltage 60W processor to high-performance 130W processor. The data width is 64 bits with 6 cores, and 12 threads. The floating-point unit for this processor is integrated. The E5 2630 has low power features of the Enhanced SpeedStep Technology. Some security features are NX/XD/execute disable bit and Trusted Execution Technology (TXT). The thermal design power is 95 watts.

Another HPC cluster manufacturer is AMD. AMD utilizes the AMD EPYC processor. This processor provides a many key specifications such as the data width is 64 bits, it contains 64 cores, and 128 threads. This processor also contains an integrated floating-point unit. The security features for this processor are EVP, secure encrypted virtualization, SMAP, secure memory encryption. The AMD EPYC processor has the low power feature of PowerNow. The thermal design power is 225 watts.

Dell HPC system for manufacturing is designed specifically for digital manufacturing workloads. It bridges the gap between workstations and full scale HPC data centers. The company has used the Intel Xeon E5 2967A processor for some of its manufacturing needs. The Xeon E5 2967A has a data width 64 bit, 16 cores and 32 threads. This processor has an integrated floating point unit. Some security features are it utilizes the Execute Disable Bit and Trusted Execution Technology. It contains low power features of Enhanced SpeedStep Technology and a thermal design power of 145 watts.

One microprocessor used is the Intel Core i9. The Intel Core i9 has a data width of 64 bits with 10 central processing unit cores and 20 threads. It has an integrated floating point unit. Some security features it contains are the Executable Disable Bit the Memory Protection extensions. It also has the Supervisor Mode Access Prevention and the Secure Mode Execution Protection. The low power features are Enhanced SpeedStep Technology and has a thermal design power of 165 watts.

As you can see there are many specifications and features that vary between processors used by different manufacturers. There are many factors that go into a manufacturer's decision when choosing a processor. Each one chosen on that can be used to full fill specific customer needs and budgets.

References

EMC, Dell. "i.dell.com." 2016.

https://i.dell.com/sites/csdocuments/Business_solutions_whitepapers_Documents/en/Dell EMC_HPC_System_for_Manufacturing_Sol_fact_sheet.pdf. 05 09 2020.

"IBM Support." 27 08 2020. *<https://www.ibm.com>*. 05 09 2020.

World, CPU. "http://www.cpu-world.com/CPUs/Core_i9/Intel-Core%20i9%20i9-10900X.html." 28 08 2020. *<http://www.cpu-world.com>*. 05 09 2020.

—. "<http://www.cpu-world.com/CPUs/Xeon/Intel-Xeon%20E5-2630.html>." 05 06 2020. *<http://www.cup-world.com>*. 05 09 2020.

—. "<http://www.cpu-world.com/CPUs/Xeon/Intel-Xeon%20E5-2697A%20v4.html>." 28 08 2020. *<http://www.cpu-world.com>*. 05 09 2020.

—. "<http://www.cpu-world.com/CPUs/Zen/AMD-EPYC%207742.html>." 14 04 2020. *<http://www.cup-world.com>*. 05 09 2020.

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: HOMEWORK 3

TOPICS COVERED IN HOMEWORK 3

- 1. UNIX/LINUX**
- 2. Opensource**

ELEG4363 Intro to HPC

Homework 3

Part 1: Most servers and supercomputers use Unix/Linux operating system. After review Unix/Linux instructions, give examples of 10 instructions with short description and screenshot to illustrate the result.

Part 2: Find a Github link of your interest, such as <https://github.com/vt-vl-lab/3d-photo-inpainting> Follow the guidance to download the code, install the environment, and run the code for results. Record the procedures for a report.

ABOVE AVERAGE

ELEG4363 Intro to HPC

Homework 3

Pat 1: Most servers and supercomputers use Unix/Linux operating system. After review Unix/Linux instructions, give examples of 10 instructions with short description and screenshot to illustrate the result.

Instruction 1: Change Directory to my desktop

```
Last login: Tue Sep 15 19:52:20 on ttys000
[kobi@Kobis-MacBook-Pro-3 ~ % cd ~/Desktop
```

Instruction 2: Create a new folder named INTRO_TO_HPC

```
Last login: Tue Sep 15 19:52:20 on ttys000
[kobi@Kobis-MacBook-Pro-3 ~ % cd ~/Desktop ]
[kobi@Kobis-MacBook-Pro-3 Desktop % mkdir INTRO_TO_HPC ]
```

Instruction 3: Change Directory to INTRO_TO_HPC

```
Last login: Tue Sep 15 19:52:20 on ttys000
[kobi@Kobis-MacBook-Pro-3 ~ % cd ~/Desktop ]
[kobi@Kobis-MacBook-Pro-3 Desktop % mkdir INTRO_TO_HPC ]
[kobi@Kobis-MacBook-Pro-3 Desktop % cd ~/Desktop/INTRO_TO_HPC ]
```

Instruction 4: create a folder Week 1 into the INTRO_TO_HPC folder

```
[kobi@Kobis-MacBook-Pro-3 INTRO_TO_HPC % mkdir WEEK_1
Last login: Tue Sep 15 19:52:20 on ttys000
[kobi@Kobis-MacBook-Pro-3 ~ % cd ~/Desktop ]
[kobi@Kobis-MacBook-Pro-3 Desktop % mkdir INTRO_TO_HPC ]
[kobi@Kobis-MacBook-Pro-3 Desktop % cd ~/Desktop/INTRO_TO_HPC ]
[kobi@Kobis-MacBook-Pro-3 INTRO_TO_HPC % mkdir WEEK_1 ]
```

Instruction 5: create a folder Week 2 into the INTRO_TO_HPC folder

```
Last login: Tue Sep 15 19:52:20 on ttys000
[kobi@Kobis-MacBook-Pro-3 ~ % cd ~/Desktop ]
[kobi@Kobis-MacBook-Pro-3 Desktop % mkdir INTRO_TO_HPC ]
[kobi@Kobis-MacBook-Pro-3 Desktop % cd ~/Desktop/INTRO_TO_HPC ]
[kobi@Kobis-MacBook-Pro-3 INTRO_TO_HPC % mkdir WEEK_1 ]
[kobi@Kobis-MacBook-Pro-3 INTRO_TO_HPC % mkdir WEEK_2 ]
```

Instruction 6: Install install -r requirements

```
((kobi) (base) kobi@Kobis-MBP-3 INTRO_TO_HPC % cd bird-bot ]
((kobi) (base) kobi@Kobis-MBP-3 bird-bot % sudo pip3 install -r requirements.txt ]
[Password: ]
WARNING: The directory '/Users/kobi/Library/Caches/pip' or its parent directory
is not owned or is not writable by the current user. The cache has been disabled
. Check the permissions and owner of that directory. If executing pip with sudo,
you may want sudo's -H flag.
Collecting PyQt5==5.14.2
  Downloading PyQt5-5.14.2-5.14.2-cp35.cp36.cp37.cp38-abi3-macosx_10_6_intel.whl
(43.1 MB)
|████████████████████████████████████████| 43.1 MB 649 kB/s
Collecting js2py==0.68
  Downloading Js2Py-0.68-py3-none-any.whl (605 kB)
|████████████████████████████████████████| 605 kB 761 kB/s
```

Instruction 7: List of folder in my desktop

```
kobi@Kobis-MacBook-Pro-3 Desktop % ls
$RECYCLE.BIN
ARTLIGHT
CODE FOR GOOD
Credera- Answers.pdf
DOCS
FALL_20
FL Studio 10.lnk
GOOGLE
GS
HW3.pdf
INTRO_TO_HPC
McKinsey
NEW SAUCEZ
PRESETS PREMIERE PRO
PROD MP3
PVAMU
PYTHON
Personality-of-color-infographic-by-website-builder-expert.jpg
SACEM -DOCS
SENIOR DESIGN I
SPIKE
kobi@Kobis-MacBook-Pro-3 Desktop % touch Kobi_file
```

Instruction 8: sort the List of folder in my desktop by date

```
kobi@Kobis-MacBook-Pro-3 Desktop % ls -t
Kobi_file
INTRO_TO_HPC
Tioro, Kobi - Resume TMCf.pdf
Tioro, Kobi - Resume M20.pdf
Tioro, Kobi - Resume Fiber.pdf
HW3.pdf
FALL_20
McKinsey
NEW SAUCEZ
PVAMU
ARTLIGHT
SUBSTITUTION
DOCS
PROD MP3
CODE FOR GOOD
```

Instruction 9: show the current date

```
38804 tlys000 0.00.00 -Zsn
kobi@Kobis-MacBook-Pro-3 Desktop % date
Tue Sep 15 20:29:01 CDT 2020
kobi@Kobis-MacBook-Pro-3 Desktop %
```

Instruction 10: show the disk usage

```

kobi@Kobis-MacBook-Pro-3 Desktop % df
Filesystem                                512-blocks      Used Available Capa
city iused      ifree %iused  Mounted on
/dev/disk1s5                                685574104    21975216 162050816    1
2%  488606 3427381914    0%  /
devfs                                      385          385          0    10
0%    666          0 100%  /dev
/dev/disk1s1                                685574104    491678544 162050816    7
6% 1809437 3426061083    0%  /System/Volumes/Data
/dev/disk1s4                                685574104     8390832 162050816
5%    5 3427870515    0%  /private/var/vm
/dev/disk0s3                                269455360    237384176    32071184    8
9%   8142 4294959137    0%  /Volumes/BOOTCAMP
/dev/disk0s4                                20217856     16023424     4194432    8
0%    112 4294967167    0%  /Volumes/VOLUME
map auto_home                                0            0            0    10
0%    0          0 100%  /System/Volumes/Data/home
/dev/disk1s3                                685574104    1033016 162050816

```

Part 2: Find a Github link of your interest, such as <https://github.com/natewong1313/bird-bot>. Follow the guidance to download the code, install the environment, and run the code for results. Record the procedures for a report.

```

[(kobi) (base) kobi@Kobis-MBP-3 INTRO_TO_HPC % cd bird-bot
[(kobi) (base) kobi@Kobis-MBP-3 bird-bot % sudo pip3 install -r requirements.txt
Password:
WARNING: The directory '/Users/kobi/Library/Caches/pip' or its parent directory
is not owned or is not writable by the current user. The cache has been disabled
. Check the permissions and owner of that directory. If executing pip with sudo,
you may want sudo's -H flag.
Collecting PyQt5==5.14.2
  Downloading PyQt5-5.14.2-5.14.2-cp35.cp36.cp37.cp38-abi3-macosx_10_6_intel.whl
(43.1 MB)
|████████████████████████████████████████| 43.1 MB 649 kB/s
Collecting js2py==0.68
  Downloading Js2Py-0.68-py3-none-any.whl (605 kB)
|████████████████████████████████████████| 605 kB 761 kB/s

```

```

Last login: Sat Sep 19 18:48:32 on tty000
(base) kobi@Kobis-MBP-3 ~ % pipenv shell
Launching subshell in virtual environment.
. /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/bin/activate
(base) kobi@Kobis-MBP-3 ~ % . /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/bin/activate
(kobi) (base) kobi@Kobis-MBP-3 ~ % cd desktop
(kobi) (base) kobi@Kobis-MBP-3 desktop % cd INTRO_TO_HPC
(kobi) (base) kobi@Kobis-MBP-3 INTRO_TO_HPC % git clone https://github.com/natewong1313/bird-bot
Cloning into 'bird-bot'...
remote: Enumerating objects: 63, done.
remote: Counting objects: 100% (63/63), done.
remote: Compressing objects: 100% (36/36), done.
remote: Total 335 (delta 30), reused 49 (delta 22), pack-reused 272
Receiving objects: 100% (335/335), 7.96 MiB | 812.00 KiB/s, done.
Resolving deltas: 100% (198/198), done.
(kobi) (base) kobi@Kobis-MBP-3 INTRO_TO_HPC % cd bird-bot
(kobi) (base) kobi@Kobis-MBP-3 bird-bot % sudo pip3 install -r requirements.txt
Password:
(WARNING: The directory '/Users/kobi/Library/Caches/pip' or its parent directory is not owned or is not writable by the current user. The cache has been disabled. Check the permissions and owner of that directory. If executing pip with sudo, you may want sudo's -H flag.)
Collecting pyqt5==5.14.2
  Downloading PyQt5-5.14.2-5.14.2-cp36-cp37-cp38-abi3-macosx_10_6_intel.whl (43.1 MB)
    | 43.1 MB 649 kB/s
Collecting js2py==0.68
  Downloading Js2Py-0.68-py3-none-any.whl (605 kB)
    | 605 kB 761 kB/s
Collecting requests==2.23.0
  Downloading requests-2.23.0-py2.py3-none-any.whl (58 kB)
    | 58 kB 899 kB/s
Collecting lxml==4.5.0
  Downloading lxml-4.5.0-cp38-cp38-macosx_10_9_x86_64.whl (4.6 MB)
    | 4.6 MB 754 kB/s
Collecting darkdetect==0.1.1
  Downloading darkdetect-0.1.1-py3-none-any.whl (4.5 kB)
Collecting colorama==0.4.3
  Downloading colorama-0.4.3-py2.py3-none-any.whl (15 kB)
Collecting selenium==3.141.0
  Downloading selenium-3.141.0-py2.py3-none-any.whl (904 kB)
    | 904 kB 738 kB/s
Collecting chromedriver-py==81.0.4044.69
  Downloading chromedriver_py-81.0.4044.69-py3-none-any.whl (16.4 MB)
    | 16.4 MB 926 kB/s
Collecting pycryptodome==3.4.3
  Downloading pycryptodome-3.4.3.tar.gz (6.5 MB)
    | 6.5 MB 743 kB/s
Collecting PyQt5-sip<13,>=12.7
  Downloading PyQt5_sip-12.8.1-cp38-cp38-macosx_10_9_x86_64.whl (63 kB)
    | 63 kB 1.3 MB/s
Collecting pyjsparser==2.5.1
  Downloading pyjsparser-2.7.1.tar.gz (24 kB)
Requirement already satisfied: six>=1.10 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from js2py==0.68->-r requirements.txt (line 2)) (1.15.0)
Collecting tzlocal>=1.2
  Downloading tzlocal-2.1-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: idna<3,>=2.5 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from requests==2.23.0->-r requirements.txt (line 3)) (2.10)
Collecting tzlocal>=1.2
  Downloading tzlocal-2.1-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: idna<3,>=2.5 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from requests==2.23.0->-r requirements.txt (line 3)) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from requests==2.23.0->-r requirements.txt (line 3)) (2020.6.20)
Requirement already satisfied: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from requests==2.23.0->-r requirements.txt (line 3)) (1.25.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/lib/python3.8/site-packages (from requests==2.23.0->-r requirements.txt (line 3)) (3.0.4)
Collecting pytz
  Downloading pytz-2020.1-py2.py3-none-any.whl (510 kB)
    | 510 kB 747 kB/s
Building wheels for collected packages: pycryptodome, pyjsparser
  Building wheel for pycryptodome (setup.py) ... done
  Created wheel for pycryptodome: filename=pycryptodome-3.4.3-cp38-cp38-macosx_10_9_x86_64.whl size=6672452 sha256=86fda6cc98364c114f037a29ef4702fab7cdf15c22d7b5da31f4e9d21d46e0a
  Stored in directory: /private/tmp/pip-ephem-wheel-cache-0r5kls9s/wheels/1a/72/8a/c7c67a5683b57bc9da3a7eb20cd7d9c98b4a119e50a7b011f1
  Building wheel for pyjsparser (setup.py) ... done
  Created wheel for pyjsparser: filename=pyjsparser-2.7.1-py3-none-any.whl size=25998 sha256=20426306dc34c461a0b0c70d605f4cea2bc97277ad0e732b3f0f102e6dfdc92
  Stored in directory: /private/tmp/pip-ephem-wheel-cache-0r5kls9s/wheels/d5/88/34/cc5bb40eb3178a134eb2936c363928c5bcfb40b91031db76
Successfully built pycryptodome pyjsparser
Installing collected packages: PyQt5-sip, pyqt5, pyjsparser, pytz, tzlocal, js2py, requests, lxml, darkdetect, colorama, selenium, chromedriver-py, pycryptodome
  Attempting uninstall: requests
    Found existing installation: requests 2.24.0
    Uninstalling requests-2.24.0:
      Successfully uninstalled requests-2.24.0
Successfully installed PyQt5-sip-12.8.1 chromedriver-py-81.0.4044.69 colorama-0.4.3 darkdetect-0.1.1 js2py-0.68 lxml-4.5.0 pycryptodome-3.4.3 pyjsparser-2.7.1 pyqt5-5.14.2 pytz-2020.1 requests-2.23.0 selenium-3.141.0 tzlocal-2.1
WARNING: You are using pip version 20.2.2; however, version 20.2.3 is available.
You should consider upgrading via the '/Users/kobi/.local/share/virtualenvs/kobi-qdJ5-6s6/bin/python -m pip install --upgrade pip' command.
(kobi) (base) kobi@Kobis-MBP-3 bird-bot % python3 app.py
Welcome To Bird Bot
/Users/kobi/Desktop/INTRO_TO_HPC/bird-bot/pages/createdialog.py:113: SyntaxWarning: "is not" with a literal. Did you mean "!="?
  self.maxprice_checkbox.setChecked(task_tab.max_price is not '')

```

Terminal Shell Edit View Window Help

Bird Bot

Tasks

0 Total Tasks 0 Total Carts 0 Checkouts

New Task Start All Stop All Delete All

Image ID Site Product Profile Status Actions

```
Download pytz-2020.1-py2.py3-none-any.whl (510 kB)
Building wheels for collected packages: pycryptodome, pyjsparser
Building wheel for pycryptodome (setup.py) ... done
Created wheel for pycryptodome: filename=pycryptodome-3.4.3-cp38-cp38-macosx_10_9_x86_64.whl size=6672452 sha256=86fde6cc698364c114f037a29ef4702fab7cdf15c22d7b5da31f4e9d21d44e0a
Stored in directory: /private/tmp/pip-ephem-wheel-cache-0r5kls9s/wheels/1a/72/8a/c7c67a5683b57bc9da3a7eb20cd7d9c98b4a119e50a7b011f1
Building wheel for pyjsparser (setup.py) ... done
Created wheel for pyjsparser: filename=pyjsparser-2.7.1-py3-none-any.whl size=25998 sha256=20426306dc34c461a0b0c70d605f4cea2bc9727ad0e732b3f0f102e6dfdc92
Stored in directory: /private/tmp/pip-ephem-wheel-cache-0r5kls9s/wheels/d5/88/34/cdb5b40eb3178a134eb293e6c363928c5b0fba0b91031db76
Successfully built pycryptodome pyjsparser
Installing collected packages: PyQt5-sip, pyqt5, pyjsparser, pytz, tzlocal, js2py, requests, lxml, darkdetect, colorama, selenium, chromedriver-py, pycryptodome
Attempting uninstall: requests
Found existing installation: requests 2.24.0
Uninstalling requests-2.24.0:
Successfully uninstalled requests-2.24.0
Successfully installed PyQt5-sip-12.8.1 chromedriver-py-81.0.4044.69 colorama-0.4.3 darkdetect-0.1.1 js2py-0.68 lxml-4.5.0 pycryptodome-3.4.3 pyjsparser-2.7.1 pyqt5-5.14.2 pytz-2020.1 requests-2.23.0 selenium-3.141.0 tzlocal-2.1
```

Screen Shot 2020-0...20.20 PM

HW3 HPC.docx

Screen Shot 2020-0...16.24 PM

Screen Shot 2020-0...15.54 PM

Screen Shot 2020-0...42.24 PM

Screen Shot 2020-0...56.44 PM

Screen Shot 2020-0...15.45 PM

AVERAGE

9

CHMOD

On Unix-like operating systems, a set of flags associated with each file determines who can access that file, and how they can access it. These flags are called file *permissions* or *modes*, as in "mode of access." The command name **chmod** stands for "change mode." It restricts the way a file can be accessed.

```
(3DP) racine@racine-VirtualBox:~/Desktop/3d-photo-inpainting-master$ chmod +x download.sh
(3DP) racine@racine-VirtualBox:~/Desktop/3d-photo-inpainting-master$ ./download.sh
mkdir: cannot create directory 'checkpoints': File exists
downloading from filebox ...
--2020-09-15 20:20:09-- https://filebox.ece.vt.edu/~jbhuang/project/3DPhoto/model/color-model.pt
h
Resolving filebox.ece.vt.edu (filebox.ece.vt.edu)... 128.173.88.43
Connecting to filebox.ece.vt.edu (filebox.ece.vt.edu)|128.173.88.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 206331633 (197M)
Saving to: 'color-model.pth'

color-model.pth      100%[=====] 196.77M  1.34MB/s   in 2m 56s
```

cd command

To navigate through the Linux files and directories, use the **cd** command. It requires either the full path or the name of the directory, depending on the current working directory that you're in.

```
(base) racine@racine-VirtualBox:~$ cd Desktop/
(base) racine@racine-VirtualBox:~/Desktop$
```

mkdir command

Use **mkdir** command to make a new directory — if you type **mkdir Music** it will create a directory called **Music**.

```
(base) racine@racine-VirtualBox:~/Desktop$ mkdir example
(base) racine@racine-VirtualBox:~/Desktop$ cd example/
(base) racine@racine-VirtualBox:~/Desktop/example$
```

pwd command

Use the **pwd** command to find out the path of the current working directory (folder) you're in.

```
(base) racine@racine-VirtualBox:~$ pwd
/home/racine
```

ls command

The **ls** command is used to view the contents of a directory. By default, this command will display the contents of your current working directory.

```
(base) racine@racine-VirtualBox:~$ ls
anaconda3  Documents  GNUstep   Pictures  PycharmProjects  snap      Videos
Desktop    Downloads  Music     Public    Python-eg         Templates
```

cat command

cat (short for concatenate) is one of the most frequently used commands in Linux. It is used to list the contents of a file on the standard output (stdout). To run this command, type **cat** followed by the file's name and its extension.

```
(base) racine@racine-VirtualBox:~/Desktop/3d-photo-inpainting-master$ cat requirements.txt
opencv-python==4.2.0.32
vispy==0.6.4
moviepy==1.0.2
transforms3d==0.3.1
networkx==2.3
cynetworkx
scikit-image
```

rm command

The **rm** command is used to delete directories and the contents within them.

```
(base) racine@racine-VirtualBox:~/Desktop/example$ rm example.txt
(base) racine@racine-VirtualBox:~/Desktop/example$ ls
(base) racine@racine-VirtualBox:~/Desktop/example$
```

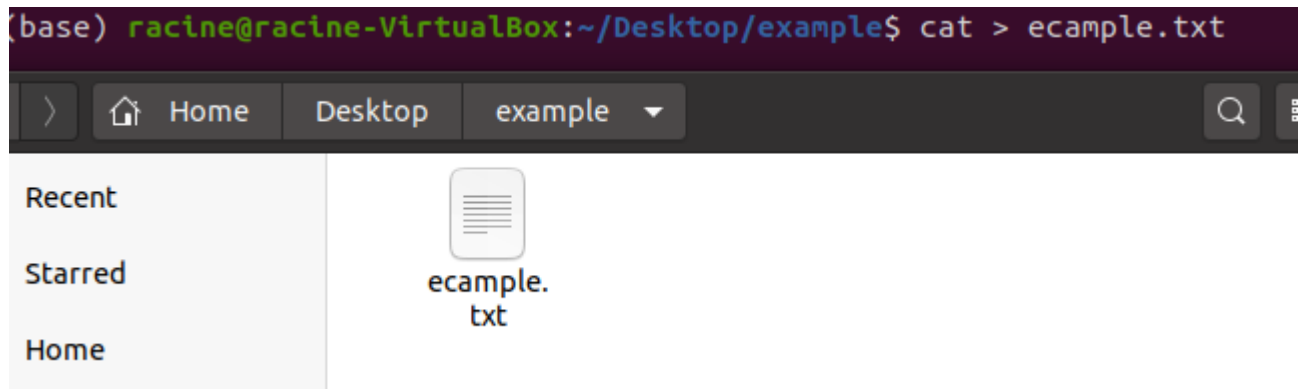
touch command

The **touch** command allows you to create a blank new file through the Linux command line. As an example, enter **touch touch example.py** to create a python file entitled **example.py** under the **example** directory.

```
(base) racine@racine-VirtualBox:~/Desktop/example$ touch example.py
(base) racine@racine-VirtualBox:~/Desktop/example$ ls
example.py
(base) racine@racine-VirtualBox:~/Desktop/example$
```

cat command

cat (short for concatenate) is one of the most frequently used commands in Linux. It is used to list the contents of a file on the standard output (stdout). To run this command, type **cat** followed by the file's name and its extension.



sudo command

Short for “**SuperUser Do**”, this command enables you to perform tasks that require administrative or root permissions. However, it is not advisable to use this command for daily use because it might be easy for an error to occur if you did something wrong.

```
(base) racine@racine-VirtualBox:~/Desktop/example$ sudo apt install mlocate
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libfprint-2-tod1 libllvm9 linux-headers-5.4.0-26 linux-headers-5.4.0-26-generic
  linux-image-5.4.0-26-generic linux-modules-5.4.0-26-generic linux-modules-extra-5.4.0-26-generic
  python3-click python3-colorama
```

When you have been using Linux for a certain period, you'll quickly notice that you can run hundreds of commands every day. As such, running **history** command is particularly useful if you want to review the commands you have entered before.

```
(base) racine@racine-VirtualBox:~/Desktop/example$ locate -i Python.exe
(base) racine@racine-VirtualBox:~/Desktop/example$ ls
example.py
(base) racine@racine-VirtualBox:~/Desktop/example$ grep linux example.py
(base) racine@racine-VirtualBox:~/Desktop/example$ python --version
Python 3.8.3
(base) racine@racine-VirtualBox:~/Desktop/example$ history
 1 python3
 2 cd Python-eg
 3 cd Home
 4 racine
 5 cd racine
 6 cat > newfile.txt
 7 history
 8 clear
 9 ~cd
```

BELOW AVERAGE

8.8

ELEG 4363 Intro to HPC Homework 3

Part 1: Most servers and supercomputers use Unix/Linux operating systems. After reviewing Unix/Linux instructions, give examples of 10 instructions with short description and screenshot to illustrate the result.

1. Listing files and directories a. % ls

i. The ls command lists the contents of your current working directory

2. Making Directories a. % mkdir

i. The command mkdir is used to make a subdirectory in your home directory to hold the files you will be creating and using

3. Changing to a different directory a. % cd

i. The command cd means to change the current working directory to 'directory'

4. Pathnames

a. % pwd

Virgie Leyva Dr. Cui September 14, 2020

```
[virgieleyva@webminal.org ~]$ls
dir1 dir2
[virgieleyva@webminal.org ~]$

[virgieleyva@webminal.org ~]$mkdir -v dir1
mkdir: created directory 'dir1'
[virgieleyva@webminal.org ~]$

./dir2/dir3/dir4:
[virgieleyva@webminal.org ~]$cd dir2
[virgieleyva@webminal.org dir2]$
```

i. The command pwd means to print working directory. The full pathname will be displayed.

5. Copying Files a. % cp

i. The command cp means to copy a file in the current working directory and call it 'file2'

6. Moving between directories a. % cd -

i. The command cd - means to move the previous directory

7. Removing files and directories a. % rm

i. The command rm means to delete or remove a file.

8. Clearing display screen a. % clear

- The command clear means to clear the terminal window of the previous commands so the output of the following commands can be clearly understood.

```
#####
# webminal.org - your linux ~ #
#####

- Share files with others, See /common_pool/README.txt
- See 'Root' menu for Webminal Desktop Root and Webminal Root

features
- For Students, partial sudo (plus C programming) platform available - mail us with your college/school id.
[virgieleyva@webminal.org ~]$pwd
/home/virgieleyva
[virgieleyva@webminal.org ~]$

-----
[virgieleyva@webminal.org dir2]$cp -v hello.txt dir3
'hello.txt' -> 'dir3/hello.txt'
[virgieleyva@webminal.org dir2]$

./dir2/dir3/dir4:
[virgieleyva@webminal.org ~]$cd dir2
[virgieleyva@webminal.org dir2]$
m: remove regular file 'dir2/hello.txt'? yes
virgieleyva@webminal.org ~]$rm dir2
[virgieleyva@webminal.org dir2]$
```

9. Displaying Contents a. % cat

- The command cat can be used to display the contents of a file on the screen.

10. Disk usage a. % du

- The command du displays the disk usage of current directory

Part 2: Find a Github link of your interest, such as <https://github.com/vt-vl-lab/3d-photo-inpainting> Follow the guidance to download the code, install the environment, and run the code for results. Record the procedures for a report.

I am unsure what you mean by downloading the code, install what environment and run code for results. I do not see an environment to install or where it states which environment to install.

MICROPROCESSOR SYSTEMS DESIGN

FALL 2020

Title of Assignment: HOMEWORK 4

TOPICS COVERED IN HOMEWORK 4

- **Data transmission interface**
- **PCIe vs. infiniband**

ELEG4363 Intro to HPC

Homework 4

Data transmission within an HPC cluster is very crucial for the performance. There are so many choices. Please search for related articles from IEEE or ACM, compare the pros and cons of PCIe and infiniband. Make a conclusion of if you are going to build an HPC cluster, which one are you going to choose for better performance?

You can use class handouts as a starting point. Search for other resources to enhance your writing. Below three links can be your resources:

https://www.theregister.co.uk/2012/06/20/mellanox_infiniband_ethernet_isc/

<http://www.hpctoday.com/news/will-pci-express-replace-ethernet-and-infiniband/>

<https://www.networkworld.com/article/2167120/tech-primers/pci-express-based-fabrics--a-low-cost-alternative-to-infiniband.html>

Cite all the references at the bottom of your paper. You can work by yourself or have a partner. If two students work in a group, the paper should be at least four pages long.

ABOVE AVERAGE

9.8

Racine McLean

ELEG -HPC

September 24, 2020

Dr. Cui

Homework #4

Prompt: Data transmission within an HPC cluster is very crucial for the performance. There are so many choices. Please search for related articles from IEEE or ACM, compare the pros and cons of PCIe and InfiniBand. Make a conclusion of if you are going to build an HPC cluster, which one are you going to choose for better performance?

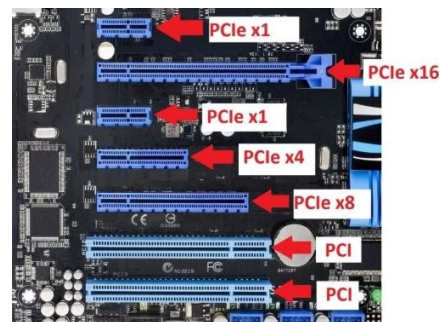
High-performance computing (HPC) is the ability to process data and perform complex calculations at high speeds. From simulations and modeling to machine learning, high-performance computing is now needed in almost every industry and it is not done with giant computers anymore. High-performance computing today high-performance is achieved through the use of clusters of small servers these clusters can be small, medium, or large but in all cases, the servers need to communicate with each other very quickly to operate in Unison and achieve high performance. This is where InfiniBand started but it did not stop there today many different industries share the need for a fast interconnect.

InfiniBand is the fastest networking technology available today in a world where we have so much data moving and needing to be analyzed the last thing the application needs is to waste valuable CPU time either transporting or waiting for the data, instead with native our RDMA (Remote Direct Memory Access) capabilities meaning there is little overhead on the CPU data can move between servers without distracting the CPU from the application. InfiniBand delivers business advantages and competitive leadership for any data center size as it enables maximum performance from the CPU's and storage. The in-network computing capabilities of InfiniBand allow data to be analyzed while in transit offering the best application performance and return on investment for cluster computing infrastructure as a public standard InfiniBand can be processor agnostic supporting many vendors processors and co-processor technologies. InfiniBand also has the lowest latency of any interconnect allowing data to make round trip journeys in a fraction of the time enabled by other technologies and InfiniBand can be deployed in many fabric configurations from fat-tree, Torus, hypercube, dragonfly, and many other topologies allowing users to match the network design to their applications needs.

What's more InfiniBand products are backward and forward compatible meaning that prior and current generations of devices will work with the devices up tomorrow this extends the life of past investments and realize a much better return on investment than proprietary technologies which require rip and replace forklift upgrades.



A. InfiniBand



B. PCIe

The competition PCI Express, technically Peripheral Component Interconnect Express but often seen abbreviated as PCIe or PCI-E, is a standard type of connection for internal devices in a computer. Generally, PCI Express refers to the actual expansion slots on the motherboard that accept PCIe-based expansion cards and to the types of expansion cards themselves. Computers do contain a mix of various expansion slots, but PCI Express is considered the standard internal interface. Many computer motherboards today are manufactured using PCIe slots only.

PCIe can also allow the sharing of input and output devices using standard multifunction networking and telecommunications hardware and software, something InfiniBand cannot easily do. Tim Fisher states “The most powerful advantage for PCIe is that it is already a near-universal interconnection technology. Based on its technical elegance and simplicity, and its historical widespread adoption, almost every device in a cloud or data center system has at least one -- and often many -- PCIe connections.” This includes CPUs, storage devices, input and output devices, and special-purpose bridges. In fact, even InfiniBand host bus adaptors connect to systems through a PCIe port. It is this fact that makes PCIe so much less expensive and lower power to use in fabric. Tim states “With InfiniBand, a Host Channel Adapter (HCA) is required to translate from PCIe -- where it starts at the host -- to the InfiniBand protocol, then translated back on the other end and while that makes sense for connections traveling over relatively long distances, it is significant overkill for a rack, or a small cluster of racks, in a data center.” While power and cost can be significant motivators, one of the most powerful advantages of PCIe for data centers is its ability to act as a truly converged rack-level fabric. Regarding power and cost, it is assumed that an InfiniBand backplane would simply be replaced with a PCIe version.

In summary, PCIe seems like the better option so if I were going to make a cluster, I would go with that. According to Larry and Krishna “It can offer performance comparable to a QDR InfiniBand solution, with a much lower cost and power envelope. InfiniBand technology has its place in those applications that require high data transfer rates without regard to power or cost, and those willing to pay a premium for this requirement will continue to use the technology. But for cloud and data center applications that need and value the three P's -- performance, power, and price -- PCIe is the superior option.” With the unpredictable future of technology, one may speculate that InfiniBand may replace PCIe but for now, PCIe controls the 3P's and that makes a big difference in any industry.

Works Cited

Mallampati, Larry Chisvin and Krishna. “PCI Express-Based Fabrics: A Low-Cost Alternative to InfiniBand.” *Network World*, Network World, 10 June 2013, www.networkworld.com/article/2167120/pci-express-based-fabrics--a-low-cost-alternative-to-infiniband.html.

Fisher, Tim. “What Is PCI Express (PCIe)?” Lifewire, 8 May 2020, www.lifewire.com/pci-express-pcie-2625962.

AVERAGE

Maritza Garcia Homework 4

InfiniBand Architecture is an industry standard which offers low latency and high bandwidth as well as many advanced features such as Remote Direct Memory Access (RDMA), atomic operations, multicast and QoS. Currently, InfiniBand products in the market can achieve a latency of several microseconds for small messages and a bandwidth of up to 700-900 MB/s. As a result, it is becoming increasingly popular as a high speed interconnect technology for building high performance clusters. PCI has been the standard local I/O bus technology for the last ten years. However, currently more and more applications require lower latency and higher bandwidth than what a PCI bus can provide. As an extension to PCI, PCI-X offers higher performance and efficiency. However, it can still become a bottleneck for today's demanding applications, especially for those running over InfiniBand. For example, a 64 bit/133 MHz PCI-X bus can sustain around 1 GB/s aggregate bandwidth at most. However, current 4x InfiniBand HCAs have a peak bandwidth of 1 GB/s in each link direction, resulting in an aggregate bandwidth of 2 GB/s for each port. To make matters worse, some of these InfiniBand HCAs have two ports which can deliver up to 4 GB/s combined theoretical bandwidth. Even PCI-X with Double Data Rate (DDR) cannot fully take advantage of their performance potential. Another issue with PCI and PCI-X buses is that a bus can be shared with other I/O devices. Therefore, communication performance can be adversely impacted by I/O operations of other devices on the same bus. Recently, PCI Express has been introduced as the next generation local I/O interconnect. Unlike PCI, PCI Express uses a serial, point-to-point interface. Compared with PCI, PCI Express can achieve lower latency by allowing I/O devices to be connected directly to the memory controller. More importantly, it can deliver scalable bandwidth by using multiple lanes in each point-to-point link. For example, an 8x PCI Express link can achieve 2 GB/s bandwidth in each direction (4 GB/s total), which matches perfectly with the requirement of current InfiniBand HCAs.

The InfiniBand Architecture is substantially different from local bus architectures such as PCI and PCI-Express. Because both PCI-Express and PCI are strictly local interconnect technologies, it is much more natural to consider PCI-Express as a “replacement” for the PCI bus. In reality even this is an over-simplification as PCI-Express offers an upgrade path rather than replacement. A separate PCI-Express connector has been defined to augment the PCI connector. Overtime PCIeExpress may effectively replace the PCI bus in PCs however history has proven that this may take a decade or more. In some sense PCI-Express may finally force the legacy ISA slots out of PCs (ten years after the addition of PCI slots) and thus it can be viewed as the replacement for ISA slots. The following table highlights the key features that distinguish InfiniBand from simple local bus architectures such as PCI and PCI-Express:

Feature	InfiniBand	PCI/PCI-Express
I/O Sharing	Yes	No
System Hierarchy	Channel Based Fabric	Memory Mapped Tree
Kernel Bypass	Memory Protection & Address Translation	No
System Software Interface	Transport level connections	Low Level load/store

As can be seen from the table above, the InfiniBand architecture supports features that make it substantially more powerful than a simple local bus point of attachment. Thus it does not make

sense to classify InfiniBand as a local bus like these other technologies. InfiniBand delivers reliable, transport level connections in hardware and supports both message passing and memory semantics on the wire. In addition InfiniBand's I/O sharing capabilities make it possible to package computers in fundamentally new ways. Utilizing InfiniBand as an I/O connection to external RAID arrays or JBODs (just a bunch of disks) enables storage to be shared by many servers connected via the InfiniBand fabric. The PCI model requires that each server include a PCI card to connect to clustering, communications, and storage networks constraining the system configuration to be a hierarchy of many I/O nodes with a single host. InfiniBand supports a system configuration of many-to-many where all I/O resources are shared by all the servers.

If I am going to build an HPC cluster, I will probably choose the PCIe. Even though the InfiniBand does have a lot of good pros compared to the PCIe, the PCIe builds on the legacy architecture of PCI, so that it has been quick and easy to migrate to the newer, faster connections. While InfiniBand has achieved very low latency with a relatively complex protocol through special-purpose hardware and software drivers that have been tuned over many years, PCIe starts out with low latency and simplicity based on its construction, giving it an advantage in adaptability. The most powerful advantage for PCIe is that it's already a near-universal interconnection technology. With PCIe, all those costly (in the hundreds of dollars) HCAs are replaced with a simple re-timer device to ensure a clean signal, which could cost as low as a few dollars in high volume. For a 36-slot rack, or a mini-cluster that can span up to hundreds of nodes, the math easily shows the cost differential. The power advantage of PCIe over InfiniBand is similar, and also flows directly from the ability to use a simple re-timer rather than an HCA. A single re-timer device will dissipate less than 1W, while an InfiniBand HCA is in the 5-7W range. And, as with the cost aspect, this power difference adds up to a substantial advantage for PCIe in a data-center rack application. But while power and cost can be significant motivators, one of the most powerful advantages of PCIe for data centers is its ability to act as a truly converged rack-level fabric. With regard to power and cost, it's assumed that an InfiniBand backplane would simply be replaced with a PCIe version. PCIe can offer performance comparable to a QDR InfiniBand solution, with a much lower cost and power envelope. InfiniBand technology has its place in those applications that require high data transfer rates without regard to power or cost, and those willing to pay a premium for this requirement will continue to use the technology.

Citations

Liu, Jiuxing, A. Mamidala, A. Vishnu, and D.k. Panda. "Performance Evaluation of InfiniBand with PCI Express." *Proceedings. 12th Annual IEEE Symposium on High Performance Interconnects*, n.d. <https://doi.org/10.1109/conect.2004.1375193>.

Mallampati, Larry Chisvin and Krishna. "PCI Express-Based Fabrics: A Low-Cost Alternative to InfiniBand." *Network World*. Network World, June 10, 2013. <https://www.networkworld.com/article/2167120/pci-express-based-fabrics--a-low-cost-alternative-to-infiniband.html>.

Mallampati, Larry Chisvin and Krishna. "PCI Express-Based Fabrics: A Low-Cost Alternative to InfiniBand." *Network World*. Network World, June 10, 2013. <https://www.networkworld.com/article/2167120/pci-express-based-fabrics--a-low-cost-alternative-to-infiniband.html>.

BELOW AVERAGE

8.5

ELEG 4363 Intro to HPC Homework 4

In the world of High-Performance Computers internet reliability and speed are what is talked about. There are many different components to choose from one being the PCIe versus the InfiniBand. The PCI-express is a high-speed serial connection that operates more like a network than a bus. Its current competition is the InfiniBand. The InfiniBand is a computer network communications standard used in high performance computing that features very high throughput and very low latency. It is used for data interconnect both among and within the computer. There are many advantages and disadvantages to both.

Regarding the PCI-express there are many advantages. For example, the speed of the PCIe supports up to 958 Mega Bytes per second per lane. PCIe devices can support 1x, 2x, 4x, 8x, or 16x lanes. That means a PCIe device has the capability of a transfer speed up to 15.76 Giga Bytes per second. The PCIe is bidirectional which means it can read and write simultaneously. This protocol eliminates the need for translations of SAS, SATA, or FC storage. This capability makes it faster than external devices. Another key point to the PCIe is the installation. The process to install is simple, which allows the PCIe SSD only needing to be plugged to the PCIe slot and that is it. The PCIe card puts the SSD closest to CPU memory.

With advantages come disadvantages and one for the PCIe is the price. The per Giga Bytes cost for the PCIe's SSD is much more expensive than the normal SATA SSD. Another con is that you're giving up a slot within your system for the sake of speed. Power consumption is another disadvantage. Normal tasks that are CPU and RAM intensive won't require any difference in consumption. Reading and transferring data will require the PCIe SSDs to use more energy. Maintenance could be another deal breaker for many. With its internal being a bus connected device the PCIe SSD will cause you to reboot every time you install or replace. This can make your system unavailable for a certain amount of time. The PCIe cannot be shared with other servers. This can be an advantage for some, but it can also be a disadvantage for others. Some may see as an advantage providing their system an additional layer of security.

Virgie Leyva Dr. Cui September 23, 2020

Just like the PCIe there are advantages of the InfiniBand. The InfiniBand's switches provide performance and bandwidth to support up to 40 Giga Bytes per second of host connectivity. The inter-switch connectivity can reach 120 Giga Bytes per second of connectivity. This is the highest supported speed for interconnectivity that is available in the market today. The InfiniBand provides low latency enabling for a better performance in High Performance Computing. The cost for InfiniBand should have one of the lowest with competing technologies. The InfiniBand provides a roadmap that allows the capability to consolidate networks, clustering, and storage data. This reduces the power, cost and property required for managing multiple devices and technologies. The InfiniBand is highly reliable. It performs data transfer correctly. Infiniband gives direct support for advanced transport protocols such as Remote Direct Memory Access (RDMA).

The InfiniBand has a long list of advantages to provide the consumer with, but there are some disadvantages as well. One important disadvantage is the InfiniBand Architecture. With lossless traffic, congestions that is initially concentrated at a single hot spot can spread throughout the

entire network. It has a complex design which does not allow for many platforms to support it. Its is bleeding edge technology and will require users to perform extensive testing.

As with all technologies there are many varieties that are available to the consumer. Each one consisting of advantages and disadvantages. The PCIe and the InfiniBand each have varying pros and cons. In conclusion if I were going to build an HPC cluster I would use the InfiniBand technology for better performance.

References

Bihan, Stephane. "<http://www.hpctoday.com/news/will-pci-express-replace-ethernet-and-infiniband/>." 24 March 2014. <http://www.hpctoday.com>. 20 September 2020.

Chatterjee, Puja. "<https://ecomputertips.com/what-is-pcie/>." 27 June 2020. <https://ecomputertips.com>. 20 September 2020.

Chisvin, Larry and Krishna Mallampati. "<https://www.networkworld.com/article/2167120/pci-express-based-fabrics--a-low-cost-alternative-to-infiniband.html>." 10 June 2013. <https://www.networkworld.com>. 20 September 2020.

Morgan, Timonth Prickett. "https://www.theregister.com/2012/06/20/mellanox_infiniband_ethernet_isc/." 20 June 2012. <https://www.theregister.com>. 20 September 2020.

Williams, Philip. "Congestion in InfiniBand Networks." *Stanford University EE382C* (2020): 11.

Wilson, Tracy V V. "<https://computer.howstuffworks.com/pci-express.htm>." 17 August 2005. <https://howstuffworks.com>. 20 September 2020.

**INTRODUCTION TO HIGH PERFORMANCE COMPUTING
FALL 2020**

Title of Assignment: HOMEWORK 5

TOPICS COVERED IN HOMEWORK 5

- 1. Parallel programming, OpenMP**
- 2. Shared memory**

ELEG4363 Intro to HPC
Homework 5

Option1: Explore parallel programming using OpenMP. And also practice how to control the threads. Run the three “Hello World” programs to analyze the results. Include the output in the report and with conclusions on different ways to control threads.

Option 2. Download at least 4 articles from ACM or IEEE to write your paper on shared memory (3 pages).

ABOVE AVERAGE

9.5

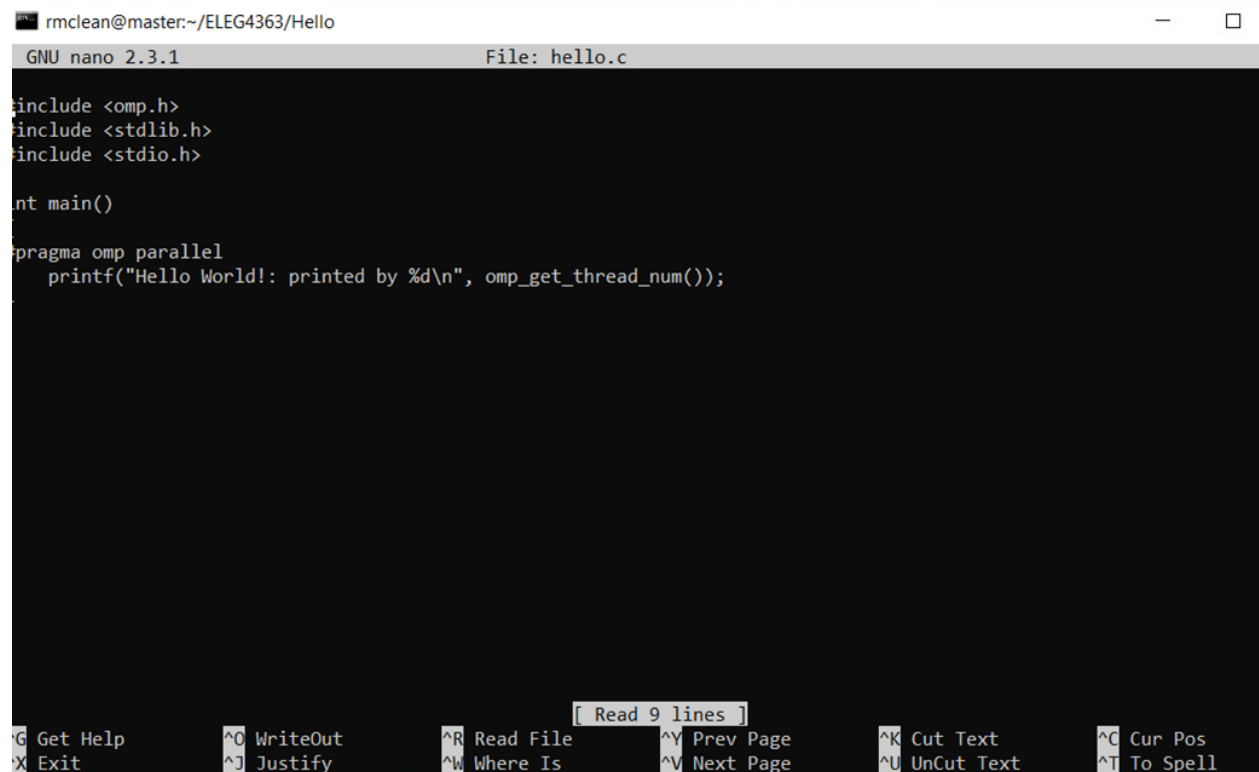
Racine McLean

ELEG4363 Intro to HPC

Homework 5

Option1: Explore parallel programming using OpenMP. And also practice how to control the threads. Run the three “Hello World” programs to analyze the results. Include the output in the report and with conclusions on different ways to control threads.

OpenMP is a library for parallel programming in the SMP (symmetric multi-processors, or shared-memory processors) model. When programming with OpenMP, all threads share memory and data. OpenMP supports C, C++ and Fortran. The OpenMP functions are included in a header file called `omp.h`



```
rmclean@master:~/ELEG4363/Hello
GNU nano 2.3.1 File: hello.c

#include <omp.h>
#include <stdlib.h>
#include <stdio.h>

int main()

#pragma omp parallel
    printf("Hello World!: printed by %d\\n", omp_get_thread_num());

[ Read 9 lines ]
^G Get Help  ^O WriteOut  ^R Read File  ^Y Prev Page  ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is   ^V Next Page  ^U UnCut Text ^T To Spell
```

```
Command Prompt
[rmclean@master ~]$ ls
ELEG4363
[rmclean@master ~]$ cd ELEG4363/
[rmclean@master ELEG4363]$ ls
Hello
[rmclean@master ELEG4363]$ cd Hello/
[rmclean@master Hello]$ ls
a.out hello.c
[rmclean@master Hello]$ nano hello.c
[rmclean@master Hello]$ gcc -fopenmp hello.c
[rmclean@master Hello]$ ls
a.out hello.c
[rmclean@master Hello]$ ./a.out
Hello World!: printed by 1
Hello World!: printed by 9
Hello World!: printed by 3
Hello World!: printed by 6
Hello World!: printed by 2
Hello World!: printed by 5
Hello World!: printed by 11
Hello World!: printed by 4
Hello World!: printed by 8
Hello World!: printed by 0
Hello World!: printed by 7
Hello World!: printed by 10
[rmclean@master Hello]$ exit
logout
Connection to 129.207.46.224 closed.
H:\>
```

AVERAGE

ELEG 4363 Intro to HPC Homework 5

OpenMP is a library for parallel programming in the SMP (symmetric multi-processors, or shared-memory processors) model. OpenMP has directives that allow the programmer to:

- • specify the parallel region
- • specify whether the variables in the parallel section are private or shared
- • specify how/if the threads are synchronized
- • specify how to parallelize loops
- • specify how the work is divided between threads (scheduling)

When programming with OpenMP, all threads share memory and data. Different ways to control a thread through the parallel region (`#pragma omp parallel num_threads()`), the run-time function (`omp_set_num_threads()`), and the environment variable (`export OMP_NUM_THREADS=n`).

Virgie Leyva Dr. Cui September 29, 2020

```
[vleyva@master ~]$ ls
ELEG4363HW5
[vleyva@master ~]$ cd ELEG4363HW5
[vleyva@master ELEG4363HW5]$ ls
Hello Hello2
[vleyva@master ELEG4363HW5]$ cd Hello
[vleyva@master Hello]$ ls
a.out hello.c
[vleyva@master Hello]$ gcc -fopenmp hello.c
[vleyva@master Hello]$ ./a.out
Hello, World!!
Hello, World!!
Hello, World!!
Hello, World!!
Hello, World!!
Hello, World!!
Hello, World!!
Hello, World!!
[vleyva@master Hello]$
```

```
GNU nano 2.3.1 File: hello.c
#include <omp.h>
#include <stdlib.h>
#include <stdio.h>

int main()
{
    omp_set_num_threads(9);
    #pragma omp parallel
    {
        printf("Hello, World!!\n");
    }
    return 0;
}
```

```
GNU nano 2.3.1 File: hello2.c
#include <omp.h>
#include <stdlib.h>
#include <stdio.h>

int main()
{
    #pragma omp parallel
    {
        printf("Hello World!: printed by %d\n", omp_get_thread_num());
    }
    return 0;
}
```

BELOW AVERAGE

Maritza Garcia

HPC Homework 5

Shared Memory

A shared memory is an extra piece of memory that is attached to some address spaces for

their owners to use. As a result, all of these processes share the same memory segment and have access to it. Consequently, race conditions may occur if memory accesses are not handled properly. Shared memory is a feature supported by UNIX System V, including Linux, SunOS and Solaris. One process must explicitly ask for an area, using a key, to be shared by other processes. This process will be called the server. All other processes, the clients, that know the shared area can access it. However, there is no protection to a shared memory and any process that knows it can access it freely. To protect a shared memory from being accessed at the same time by several processes, a synchronization protocol must be set up. A shared memory segment is identified by a unique integer, the shared memory ID.

One parallel computing architecture uses a single address space. Systems based on this consist of shared memory multiprocessors which allows processor communication through variables stored in a shared address space. We consider a configuration with multiple processors and shared memory modules connected together by a backplane bus. Each processor has its own local memory and a data cache. The local memory in a processor contains the code and local data for all tasks bound to that processor, and the cache is used only to cache globally shared data. The local memory is essential to avoid bus contention for normal task execution which can not only slow down tasks but also lead to unpredictability. As a result, the shared memory is used only for accessing globally shared resources like data structures and devices. We also assume that a hardware mechanism such as bus snooping is used to maintain data coherence. Static priority preemptive scheduling is used, and any given processor always runs the highest priority ready task eligible to run on that processor. Read-modify-write instructions such as test-and-set are available to implement binary semaphores residing in shared memory. A semaphore that is accessed by tasks allocated to different processors is referred to as a global semaphore, and a critical section guarded by a global semaphore is referred to as a global critical section (or gcs in short). And, by assumption, all global semaphores shall reside in the shared memory space. Similarly, a semaphore that is accessed by tasks allocated to a single processor is called a local semaphore, and a critical section guarded by a local semaphore is referred to as a local critical section (lcs). To avoid generating additional bus traffic, all local semaphores shall reside in the local memory space of each processor. Note that if there are no global semaphores in the system, the multiprocessor synchronization problem decomposes into multiple uniprocessor problems and the uniprocessor priority ceiling protocol, for example, can be used very effectively on each processor. Since nested global critical sections can potentially lead to large increases in blocking durations, we shall assume in this section that global critical sections cannot nest other critical sections or be nested inside other critical sections. Globally shared resources can be a significant source of blocking since multiple tasks allocated to different processors can attempt to access them simultaneously. A primary goal of the shared memory synchronization protocol is that the blocking duration of a task should only be a function of critical section execution times. As a result, a gcs may need to execute at a priority higher than any non-critical section code. Such a requirement arises from the following theorem.

Theorem 2: The remote blocking time of a job blocked on a gcs will be a function of critical sections if and only if the gcs cannot be preempted by jobs that are executing outside critical sections.

Proof: First, if a gcs cannot be preempted by non-critical section code, then the waiting time for a task to enter a gcs can only be a function of critical sections. Second, if a gcs on a processor can be preempted by the non-critical section code of J, then a job on a remote processor blocked by this gcs will wait for the execution time of J.

The shared memory synchronization protocol is an extension of the uniprocessor priority ceiling protocol in that if the number of processors in the system is 1, the protocol reduces to the priority ceiling protocol. The protocol is used on each processor in the system. Recall that we assume that there are no nested or overlapping global critical sections. Under this protocol, resources guarded by local semaphores are accessed using the uniprocessor priority ceiling protocol locally on each processor, and globally shared resources are accessed using read-modify-write instructions to obtain global semaphores. In addition, all global critical sections are executed at their predefined priority levels. The protocol is as follows.

1. A job uses its assigned priority unless it is within a critical section.
2. The uniprocessor priority ceiling protocol is used for all requests to local semaphores. This means the following: (a) When a job J requests the local semaphore S on processor p, let S' be the semaphore with the highest priority ceiling of all local semaphores locked by jobs other than J on processor p. (b) A job J on a processor p can obtain the local semaphore S if its priority is higher than the priority ceiling of S' . Else, J is said to be blocked by S' . In this case, J inherits S' 's priority until it releases S' . (c) A job J can preempt another job JL if its priority is higher than the priority, assigned or inherited, at which job JL is executing.
3. A job J within a global critical section guarded by the global semaphore S, has the priority assigned to its gcs.
4. A job J within a gcs can preempt another job J' within a gcs if the assigned priority of J's gcs is greater than that of J's gcs.
5. When a job J requests a global semaphore S, S, can be granted to J by means of an atomic transaction on shared memory, if S, is not currently held by another job.
6. If a request for a global semaphore S, cannot be granted, the job J is added to a prioritized queue on S, without being preempted (i.e J holds the processor until it is inserted into the queue). The priority used as the key for queue insertion is the normal priority assigned to J.
7. When a job J attempts to release a global semaphore S, the highest priority job JH waiting for S, is signaled and becomes eligible for execution at JH's host processor at its gcs priority. If there is no job suspended on S, the semaphore is released.

Distributed shared memory is an attractive abstraction because it allows processes uniform access to local and remote information. This uniformity of access simplifies

programming] eliminating the need for separate mechanisms to access local state and remote state. However, consistent distributed shared memory (DSM) can be difficult to implement efficiently. Most DSM implementations to date use variants of multiprocessor cache consistency algorithms that perform poorly in high latency distributed systems. Weakly consistent memories allow implementations better suited to the high latencies encountered in distributed systems. Traditionally] a shared memory is correct if reads return the value of the "most recent write" to the location being read. Atomic memory satisfies this "register property" by regarding reads and writes as operation intervals on a global timeline and requiring that operations "take effect" at some point within the operation interval [17]. Under this model, each operation corresponds to a

distinct point (operations may not “take effect” simultaneously) on the global timeline and, for any read operation] the most recent write is well-defined. While the order of overlapping writes may not be determined until a subsequent read operation “chooses” which write is the most recent, the resulting execution must still obey the register property. Sequential consistency [12] is a weakening of atomic memory that relaxes the requirement that operations take effect during their operation intervals. Several researchers [2012, 61] have sought to exploit the considerable flexibility provided by sequential consistency over atomic memory yet the requirement that sequentially consistent executions appear “as if” they obey the register property is severely restrictive.

Citations

R. Rajkumar, "Real-time synchronization protocols for shared memory multiprocessors,"

in *Proceedings., 10th International Conference on Distributed Computing Systems*, url:
<https://doi.ieeecomputersociety.org/10.1109/ICDCS.1990.89257>

M. Ahamad, P. Hutto and R. John, "Implementing and programming causal distributed shared memory," in *Proceedings 11th International Conference on Distributed Computing Systems*, url:
<https://doi.ieeecomputersociety.org/10.1109/ICDCS.1991.148677>

B. Bergsten, M. Couprie and P. Valduriez, "Prototyping DBS3, a shared-memory parallel database system," in *Proceedings of the First International Conference on Parallel and Distributed Information Systems*, url:
<https://doi.ieeecomputersociety.org/10.1109/PDIS.1991.183107>

R. Ahmed, R. Frazier and P. Marinos, "Cache-aided rollback error recovery (CARER) algorithm for shared-memory multiprocessor systems," in *Digest of Papers. Fault-Tolerant Computing: 20th International Symposium*, url:
<https://doi.ieeecomputersociety.org/10.1109/FTCS.1990.89338>

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: HOMEWORK 6

TOPICS COVERED IN HOMEWORK 6

- 1. Current development of HPC**
- 2. Supercomputing centers**

ELEG4363 Intro to HPC
Homework 6

Write 2-page report to summarize what you've learned from Oklahoma Supercomputing Symposium 2020.

ABOVE AVERAGE

9.7

ELEG 4363 Intro to HPC Homework 6

Write a 2-page report to summarize what you've learned from the Oklahoma Supercomputing Symposium 2020.

Henry Neeman from the University of Oklahoma was one of the first speakers at the Oklahoma Supercomputing Symposium. His topic was over OSCER State of the Center Address. The Oklahoma Supercomputing Symposium is a large event. It has over 740 pre-registered attendees from over 48 U.S. states and 4 U.S. territories. It also has attendees from 14 other countries. Attendees from 229 institutions and 43 private companies were also pre-registered to attend. The symposium has been going on for the last 18 years.

Mr. Neeman spoke about the University of Oklahoma's resources. One of its resources is the Dell Intel Haswell HPC Cluster. This cluster has a peak base speed of 537 TFLOPs and max turbo of 730 TFLOPs. A Teraflop rating measures the performance of the GPU. Unlike a gigahertz (GHz), which measures a processor's clock speed, TFLOP is a direct mathematical measurement of a computer's performance. A teraflop refers to the capability of a processor to calculate one trillion floating-point operations per second. For example, if something has five TFLOPs this means that the processor setup is capable of handling on average five trillion floating-point calculations every second.

The University of Oklahoma's has their own PetaStore. This PetaStore is a large-scale tape library that is available to researchers at OU as well as statewide. The University of Oklahoma is also getting some new supercomputer features coming soon. Some of their networks will be updated with InfiniBand and Ethernet. These new features will be better equipped to handle Artificial Intelligence (AI) and Machine Learning. This will be capable by containing more hardware capacity, more variety of hardware components. The emphasis for this transition is so that the features are better adapted for AI and Machine Learning instead of floating point. Another new feature is an auto-archive directory for each user in each filesystem. A daemon script will constantly walk the filesystems looking for new and modified files to archive.

Virgie Leyva Dr. Cui October 13, 2020

OURdisk which is the University of Oklahoma's Research disk will be the lowest cost disk offering in Oklahoma University's information technology history. This will be available at the OURRstore. The OURRstore tape Archive, available at Oklahoma University, is a Regional Research Store. This tape archive will be available to researchers in all EPSCoR jurisdictions and all Great Plains Network full member states. Institutions in non-EPSCoR and non-GPN jurisdictions can buy a tape cartridge expansion cabinet. The OURRstore has some new features. One feature is auto archiving, this allows users to place files in a specific directory and a daemon process archives them automatically. With a few clicks, a file owner can designate a file to be downloadable by a specific user, a specific group, or the whole world. Also, the files live on the disk front end until they are the least recently used and need to be cleared out to make room for incoming files. This allows for popular files to be place on disk and tape. Unpopular files are on tape only. Researcher cost for Linear Tape Open 7 (LTO-7), which its native raw data capacity is 6.0 Terabytes is 3.6 times bigger and half the cost per Terabyte of LTO-6 that is available on the PetaStore. The LTO-7 has a compressed capacity of 15 Terabytes and a max uncompressed speed of 300 Megabytes per second. The time to write a full tape at max

uncompressed speed is 5.33 hours. With the LTO-7 it has both WORM and encryption capabilities. The max number of partitions is four.

Another speaker was John Shalf of the Lawrence Berkeley National Laboratory. Mr. Shalf is the Department Head of Computer Science and spoke on Pathfinding for Post-Exascale HPC.

Exascale computing refers to computing systems capable of at least one exaflop. An exaflop capable system can calculate one billion billion floating point calculations per second. That is 50 times faster than the most powerful supercomputers being used today. The scaling trend of rising transistors, thread performance, clock frequency and power may determine that exascale will happen sometime soon. It can be estimated that exascale computations will occur around 2021 to 2023. Dennard scaling, also known as MOSFET scaling, is a scaling law based on a 1974 paper co-authored by Robert H. Dennard. It states that as a transistor gets smaller, their power density stays constant. So, the power used stays in proportion with the area. This is nature's way of extracting more performance in resource limited environments. Some powerful general-purpose systems are Xeon and power. There are many lightweight chips such as KNL AMD, Cavium/Marvell, and GPU. As well as many different specialized data centers like Apple, Google, Amazon, and Samba Nova.

There are many things I learned from the Oklahoma Supercomputing Symposium 2020. One of which is technology is constantly advancing. I learned more about High Performance Computer and their hardware. I will definitely be attending again next year.

AVERAGE

9.2

Maritza Garcia HPC HW 6

The context and history of today's digital transformation that permeates everyday life through everyday products, has a parallel origin story that goes back to the early 1800's. The transitions from Water to Rail to Roads to Wires has spawned innovations and winners and also left losers who failed to adapt. Many Scientists and Engineers still see innovation as a cycle of Build, Fly, Crash, Fix, confined to the physical world. High Performance Computing has been with us in some form for 40+ years, and for many industries it has become norm driving innovation. However, with 'low tech' goods, adoption of the latest technologies that Innovate how to innovate has continued to be a slog. We will cover many examples where there are some surprising successes from the bathroom, to the laundry room and beyond.

Machine learning and 'big' data has now emerged as a source of data and a way to simplify the adoption of results from what were very expensive, technical and tedious calculations. Machine learning does not 'replace' HPC; it leverages it and can enhance its output. The digital age and HPC in particular has created whole new careers dedicated to our latest challenges, which are amazingly similar, but clearly different, to the ones we faced 40 years ago.

With the inevitable demise of Moore's Law, it will be challenging to obtain the time to solve goals for future IC computing requirements with conventional off the shelf CPU designs. The last decade of HPC has seen the growth of GPUs and accelerators, and the underlying philosophy is that in order to achieve the required speeds, we must develop hardware accelerators tailored for science in addition to the general purpose CPUs. The mega datacenter (Facebook, Microsoft Azure, Google, and Amazon) have deployed multiple strategies to introduce specialized hardware to continue performance growth for their data centers in the absence of transistor scaling. The three strategies are application specific acceleration (such as the Google TPU), heterogeneous acceleration

within System On Chips (SOCs) (Amazon Graviton and Graviton2 and Apple's iPhone chips), and resource disaggregation (Google, Facebook). The central question for us to consider is which of those strategies (or combination thereof) could be applied to serve the needs of scientific computing, and what would we need to change or augment to make these specialization strategies work for science.

Very Large Scenario list with simple Simulation, an HPC Opportunity: Big Data challenges are across all sciences Simulations are simple but cases of interest are very large. Lack of understanding for how HPC can help. Software connective tissue to run as embarrassingly parallel problems in HPC is still scarce. The central problem in technology driven large scale infrastructure systems is providing services that are predictable, reliable, and broadly accessible, without allowing them to become obsolete, stale, and under

resourced. For Cyberinfrastructure (CI), this problem is particularly acute, as the technologies themselves advance very rapidly, user demand and expectations escalate, and the cyberinfrastructure is itself dynamic. This talk provides an overview of programs in the Office of Advanced Cyberinfrastructure, (OAC) at the National Science Foundation, and explores ways OAC manages this tension between provisioning the national research CI and enabling innovation in the CI systems and services that nurture development of new capabilities to transform the next generation of scientific research.

BELOW AVERAGE

8.5

Racine McLean
ELEG -HPC
October 13th, 2020
Dr. Cui
Homework #6

Prompt: Write 2-page report to summarize what you have learned from Oklahoma Supercomputing Symposium 2020.

Presenter(s): David Shannahoff-Khalsa (UCSD) Mona Wong (SDSC) Jeff Sale (SDSC)

For the portion of the presentation I attended I listened to David Shannahoff-Khalsa, Mona Wong and Jeff Sale presented about High Resolution Spatial Temporal Analysis of Whole-Head 306-Channel Magnetoencephalography & 66-channel Electroencephalography Brain Imaging in Humans During Sleep. They discuss the various methods they used to collect the data about the human brain during the sleep cycle.

Some of the resources they used were MATLAB, Fieldtrip Toolbox, Open Science Framework, Seed Me/SDSC HPC Share and they process most of their data on XSEDE Comet in interactive mode. They wrote custom MATLAB code to do most of their processing steps, first the neuro mag files were stored in separate in what they called trial files that end in a dot fif(.fif) and so they converted the proprietary data files into a MATLAB format using that field trip toolbox and each subject had over a dozen data files which were stitched together gave them millions of recorded points over several hours. They went on to discuss the processes they took to collect and analyze their data so they can get results. they had hundreds of files as they saved the process results of MATLAB in ASCII format, so file management strategies were very important.

Some of the most important thing I learned from this event about HPC is that it requires large data sets be beneficial, another thing that stands out to me is that the individuals that did this research were not necessarily Computer Scientist, Some were medical professionals and researchers. The individuals had to learn to use MATLAB software and python coding language to

complete their research.

An important remark one of the presenters made was that the XSEDE functions were not able to support OpenGL hardware acceleration in MATLAB which significantly slowed their production of their visualizations and animations, but they work around it. This goes to show that there are a lot of obstacles to overcome when doing a research, but you should always try to find a way to work around it to achieve your goal.

INTRODUCTION TO HIGH PERFORMANCE COMPUTING FALL 2020

Title of Assignment: HOMEWORK 7

TOPICS COVERED IN HOMEWORK 7

- 1. HPC performance evaluation**
- 2. Benchmarks**
- 3. Programming**

ELEG4363 Intro to HPC

Homework 7

Objectives

- Gain and understanding of sustained memory bandwidth
- Become familiar with the effect of data size on sustained memory bandwidth
- Become familiar with the STREAM benchmark
- Use STREAM to predict performance for Sparse Matrix-Vector product

Tasks

Part 1:

Write a program that measures the performance of this operation (in C):

```
double c[n], a[n];
for (k=0; k<nk; k++) {
    for (i=0; i<n; i++)
        c[i] = a[i];
    //See note below
}
```

This is a simple copy from an array a to an array c. Fill in the following table:

- Time for loop (s)
- Rate (MB/s)

(The note below) Depending on your compiler and the optimization settings, you may need to keep the compiler from eliminating code that it believes unnecessary. Add a call to a routine

```
void dummy(double *a, double *c) {}
```

after the copy loop but within the loop over k. Place this routine in a separate file (that keeps the compiler from finding and eliminating it as well).

Submit

Plot the results, using a semilog plot. Turn in both the table and the plot. Also report on which system you ran your test. Submit your code with your table.

n	Time for loop (s)	Rate (MB/s)
250		
1000		
5000		
10000		
50000		
100000		
500000		
1000000		

Note

A MB is a Megabyte = 10^6 bytes (note: 2^{20} bytes is a Mebibyte, written MiB, not a Megabyte). To time the loop, you may use any accurate timer. You may also look at the code in Part 2 and use the same timer used for that code.

When timing the loop, pick nk such that the time measured for both loops is about one second. Report the total time divided by nk as the “Time for loop”. When computing the rate, count each byte written or read as a separate byte. I.e., for $n=10$ and for 8 byte doubles, this copy loop moves $n*8*2=10*8*2=160$ bytes.

Part 2:

Download the STREAM benchmark from <http://www.cs.virginia.edu/stream/> and run it. You may use either the C or Fortran version. Make sure that the array size is chosen so that it is at least 4 times the size of the cache on the processor that you are using. Report the results and indicate which version you ran (C or Fortran) in submission.

Part 3:

Using the COPY bandwidth from STREAM (Part 2) to compute a performance expectation for the sparse matrix-vector multiply operation. Assume that the read and write bandwidths are the same. Turn in the formula that you used, the values of the parameters, and the performance expectation in GigaFLOPS.

Think about the following (but you don’t need to turn in):

1. How could you estimate a value of nk given an assumption that memory bandwidth is between 1 and 20 Gigabytes/sec?
2. How could you choose N to efficiently plot the different regions of memory performance?
3. How does the performance expectation in Part 3 compare to (a) the peak performance of your processor and (b) the performance of a sparse matrix-vector multiply (if you want to do this, either write a simple test program or use a library routine that performs a sparse matrix-vector multiply).

ABOVE AVERAGE

9.8

Kobi Tioro ELEG4363 Intro to HPC Homework 7

```

HPC > HPC > stream.c > No Selection
22 /* 3a. In order to be referred to as "STREAM benchmark results", */
23 /* published results must be in conformance to the STREAM */
24 /* Run Rules, (briefly reviewed below) published at */
25 /* http://www.cs.virginia.edu/stream/ref.html */
26 /* and incorporated herein by reference. */
27 /* As the copyright holder, John McCalpin retains the */
28 /* right to determine conformity with the Run Rules. */
29 /* 3b. Results based on modified source code or on runs not in */
30 /* accordance with the STREAM Run Rules must be clearly */
31 /* labelled whenever they are published. Examples of */

```

```

-----
This system uses 8 bytes per array element.
-----
Array size = 10000000 (elements), Offset = 0 (elements)
Memory per array = 76.3 MiB (= 0.1 GiB).
Total memory required = 228.9 MiB (= 0.2 GiB).
Each kernel will be executed 10 times.
The *best* time for each kernel (excluding the first
iteration)
will be used to compute the reported bandwidth.
-----
Your clock granularity/precision appears to be 1
microseconds.
Each test below will take on the order of 33493
microseconds.
(= 33493 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time
Max time
Copy:         5999.5    0.028998     0.026669
0.030920
Scale:        6178.8    0.028311     0.025895
0.031604
Add:          8003.8    0.031597     0.029986
0.033626
Triad:        7995.2    0.032505     0.030018
0.035386
-----
Solution Validates: avg error less than 1.000000e-13 on
all three arrays
-----
Program ended with exit code: 0

```

Figure 1: Results of the STREAM benchmark running on my local machine.

```
ktioro@master:~/STREAM
[ktioro@master STREAM]$ ls
stream.c
[ktioro@master STREAM]$ nano stream.c
[ktioro@master STREAM]$ nano stream.c
[ktioro@master STREAM]$ gcc -O stream.c -o stream
[ktioro@master STREAM]$ ls
stream stream.c
[ktioro@master STREAM]$ ./stream
-----
STREAM version $Revision: 5.10 $
-----
This system uses 8 bytes per array element.
-----
Array size = 10000000 (elements), Offset = 0 (elements)
Memory per array = 76.3 MiB (= 0.1 GiB).
Total memory required = 228.9 MiB (= 0.2 GiB).
Each kernel will be executed 10 times.
The *best* time for each kernel (excluding the first iteration)
will be used to compute the reported bandwidth.
-----
Your clock granularity/precision appears to be 1 microseconds.
Each test below will take on the order of 12831 microseconds.
(= 12831 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time     Max time
Copy:         12808.0    0.013467    0.012492    0.013964
Scale:        12554.0    0.013715    0.012745    0.014252
Add:          14182.7    0.018387    0.016922    0.019327
Triad:        14207.9    0.018168    0.016892    0.019101
-----
Solution Validates: avg error less than 1.000000e-13 on all three arrays
-----
[ktioro@master STREAM]$
```

Figure 2: shows the results of the STREAM benchmark running on the clusters

AVERAGE

9.5

Racine McLean

ELEG4363 Intro to HPC Homework 7

Comments: the benchmark was run on both my local machine and the HPC cluster on campus.
My device:



Intel® Core™ i7-9750H Processor (12M Cache,
up to 4.50 GHz)

- 12 MB Intel® Smart Cache
- 6 Cores
- 12 Threads
- 4.50 GHz Max Turbo Frequency
- H - High performance graphics
- 9th Generation

```
[rmclean@master Homework_7]$ gcc -fopenmp benchmark.c
[rmclean@master Homework_7]$ ls
a.out benchmark.c stream.c
[rmclean@master Homework_7]$ ./a.out

-----
STREAM version $Revision: 5.10 $
-----
This system uses 8 bytes per array element.
-----
Array size = 10000000 (elements), Offset = 0 (elements)
Memory per array = 76.3 MiB (= 0.1 GiB).
Total memory required = 228.9 MiB (= 0.2 GiB).
Each kernel will be executed 10 times.
The *best* time for each kernel (excluding the first iteration)
will be used to compute the reported bandwidth.
-----
Number of Threads requested = 12
Number of Threads counted = 12
-----
Your clock granularity/precision appears to be 1 microseconds.
Each test below will take on the order of 5801 microseconds.
    (= 5801 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time     Max time
Copy:         20027.7    0.008677    0.007989    0.010379
Scale:        19029.3    0.009119    0.008408    0.012267
Add:          20949.7    0.012032    0.011456    0.013563
Triad:        21639.2    0.011598    0.011091    0.013529
-----
Solution Validates: avg error less than 1.000000e-13 on all three arrays
-----
[rmclean@master Homework_7]$
```

Figure 1: shows the results of the STREAM benchmark running on the Cluster

```

-----
STREAM version $Revision: 5.10 $
-----
This system uses 8 bytes per array element.
-----
Array size = 10000000 (elements), Offset = 0 (elements)
Memory per array = 76.3 MiB (= 0.1 GiB).
Total memory required = 228.9 MiB (= 0.2 GiB).
Each kernel will be executed 10 times.
The *best* time for each kernel (excluding the first iteration)
will be used to compute the reported bandwidth.
-----
Your clock granularity/precision appears to be 972 microseconds.
Each test below will take on the order of 20944 microseconds.
(= 21 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time     Max time
Copy:         8457.6    0.019385    0.018918    0.019947
Scale:        10027.0    0.016856    0.015957    0.018981
Add:          12666.3    0.021957    0.018948    0.026960
Triad:        12033.1    0.022337    0.019945    0.024932
-----
Solution Validates: avg error less than 1.000000e-013 on all three arrays
-----
PS C:\Users\kultu\Desktop\FALL_2020\High_Performance_Computing\week_9>

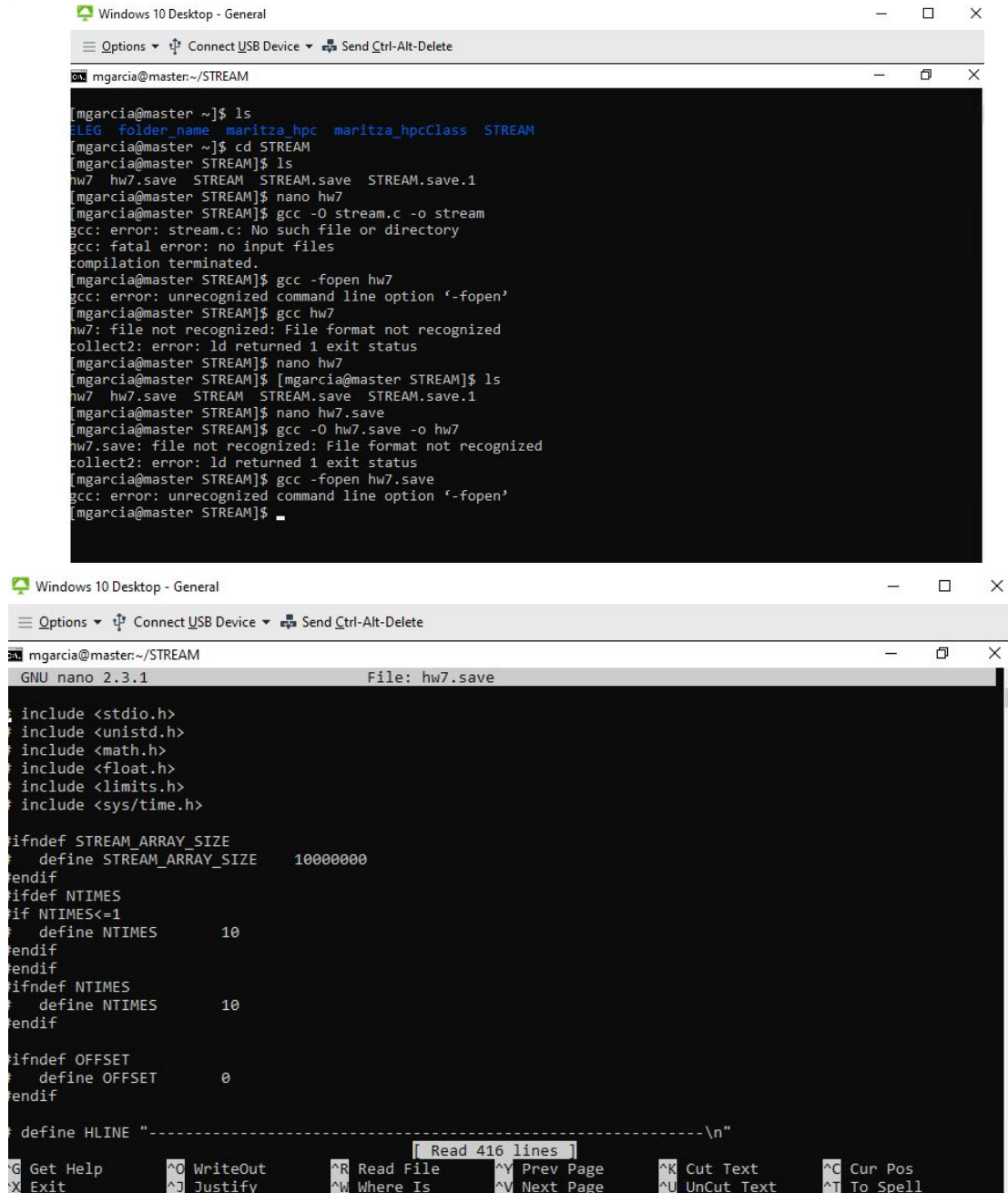
```

Figure 2: shows the results of the STREAM benchmark running on my local machine

BELOW AVERAGE

8.5

I keep getting these errors and I cannot seem to open nor run the file. Below are partial screenshots of the code I copied from the benchmark c files. I watched the recordings from class but they were still not helpful. I do not know how to download the file directly to the cluster. I wasn't able to find videos on how to do it either.



The first screenshot shows a terminal window with the following commands and output:

```
[mgarcia@master ~]$ ls
LEG folder_name maritza_hpc maritza_hpcClass STREAM
[mgarcia@master ~]$ cd STREAM
[mgarcia@master STREAM]$ ls
hw7 hw7.save STREAM STREAM.save STREAM.save.1
[mgarcia@master STREAM]$ gcc -O stream.c -o stream
gcc: error: stream.c: No such file or directory
gcc: fatal error: no input files
compilation terminated.
[mgarcia@master STREAM]$ gcc -fopen hw7
gcc: error: unrecognized command line option '-fopen'
[mgarcia@master STREAM]$ gcc hw7
hw7: file not recognized: File format not recognized
collect2: error: ld returned 1 exit status
[mgarcia@master STREAM]$ nano hw7
[mgarcia@master STREAM]$ [mgarcia@master STREAM]$ ls
hw7 hw7.save STREAM STREAM.save STREAM.save.1
[mgarcia@master STREAM]$ nano hw7.save
[mgarcia@master STREAM]$ gcc -O hw7.save -o hw7
hw7.save: file not recognized: File format not recognized
collect2: error: ld returned 1 exit status
[mgarcia@master STREAM]$ gcc -fopen hw7.save
gcc: error: unrecognized command line option '-fopen'
[mgarcia@master STREAM]$
```

The second screenshot shows a code editor window displaying the contents of `hw7.save`:

```
GNU nano 2.3.1 File: hw7.save
#include <stdio.h>
#include <unistd.h>
#include <math.h>
#include <float.h>
#include <limits.h>
#include <sys/time.h>

#ifndef STREAM_ARRAY_SIZE
#define STREAM_ARRAY_SIZE 10000000
#endif

#ifndef NTIMES
#define NTIMES 10
#endif

#ifndef NTIMES
#define NTIMES 10
#endif

#ifndef OFFSET
#define OFFSET 0
#endif

#define HLINE "-----\n"
```

The editor status bar at the bottom shows "Read 416 lines" and various keyboard shortcuts.

Windows 10 Desktop - General

Options Connect USB Device Send Ctrl-Alt-Delete

mgarcia@master:~/STREAM

GNU nano 2.3.1 File: hw7.save

```
# define MIN(x,y) ((x)<(y)?(x):(y))
# endif
# ifndef MAX
# define MAX(x,y) ((x)>(y)?(x):(y))
# endif

# ifndef STREAM_TYPE
# define STREAM_TYPE double
# endif

static STREAM_TYPE      a[STREAM_ARRAY_SIZE+OFFSET],
                        b[STREAM_ARRAY_SIZE+OFFSET],
                        c[STREAM_ARRAY_SIZE+OFFSET];

static double    avgtime[4] = {0}, maxtime[4] = {0},
                  mintime[4] = {FLT_MAX,FLT_MAX,FLT_MAX,FLT_MAX};

static char      *label[4] = {"Copy:      ", "Scale:      ",
                              "Add:        ", "Triad:      "};

static double    bytes[4] = {
    2 * sizeof(STREAM_TYPE) * STREAM_ARRAY_SIZE,
    2 * sizeof(STREAM_TYPE) * STREAM_ARRAY_SIZE,
    3 * sizeof(STREAM_TYPE) * STREAM_ARRAY_SIZE,
};

G Get Help      ^O WriteOut      ^R Read File
X Exit          ^J Justify        ^W Where Is
```

mgarcia@master:~/STREAM

Windows 10 Desktop - General

Options Connect USB Device Send Ctrl-Alt-Delete

mgarcia@master:~/STREAM

GNU nano 2.3.1 File: hw7.save

```
ain()
{
    int          quantum, checktick();
    int          BytesPerWord;
    int          k;
    ssize_t      j;
    STREAM_TYPE  scalar;
    double       t, times[4][NTIMES];

    printf(HLINE);
    printf("STREAM version $Revision: 5.10 $\n");
    printf(HLINE);
    BytesPerWord = sizeof(STREAM_TYPE);
    printf("This system uses %d bytes per array element.\n",
        BytesPerWord);

    printf(HLINE);
# ifndef N
    printf("***** WARNING: *****\n");
    printf("    It appears that you set the preprocessor variable N when compiling this code.\n");
    printf("    This version of the code uses the preprocessor variable STREAM_ARRAY_SIZE to control the ar$
    printf("    Reverting to default value of STREAM_ARRAY_SIZE=%llu\n", (unsigned long long) STREAM_ARRAY_$
    printf("***** WARNING: *****\n");
# endif
endif

G Get Help      ^O WriteOut      ^R Read File      ^Y Prev Page
X Exit          ^J Justify        ^W Where Microsoft Edge Next Page      ^K Cut Text      ^C Cur Pos
                                  ^U UnCut Text      ^T To Spell
```

**INTRODUCTION TO HIGH PERFORMANCE COMPUTING
FALL 2020**

FINAL PROJECT

TOPICS COVERED IN FINAL PROJECT

FINAL PROJECT

DEVELOP AN ARTICLE ON HPC KNOWLEDGE LEARNED

PRAIRIE VIEW A&M UNIVERSITY
COLLEGE OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
ELEG 4363 Introduction to HPC
Final Project

In this mandatory assignment, students will work on using the HPC cluster environment to develop a research project. The deliverables include:

1. A 4-page double column IEEE formatted report
2. A PowerPoint Slides with presentation

There are recommended topics provided on course website.

Your project will be evaluated as a group with the rubric:

Quality of Visuals and or written materials _____ (6 points)
(i.e readable text, no typos, figures)

Demonstrated knowledge of research and quality of the
system _____ (8 points)

Delivery of Information
(quality of presentation, organization, clarity) _____ (4 points)

Ability to address questions, anticipated issues posed during
presentation. _____ (2 points)

TOTAL _____ (20 points)

ABOVE AVERAGE

Virtual Reality in Big Data Applications

Maritza Mateos

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Suxia Cui

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Abstract—The recent appearance of low cost virtual reality (VR) technologies is attracting the attention of users and researchers suggesting it may be the next largest stepping stone in technological innovation. However, the history of VR technology is longer than it may seem. The concept of VR was formulated in the 1960s and the first commercial VR tools appeared in the late 1980s. For this reason, during the last 20 years, 100s of researchers explored the processes, effects, and applications of this technology producing 1000s of scientific papers.

Many virtual and augmented reality systems have been proposed to support renal interventions. This paper reviews such systems employed in the treatment of renal cell carcinoma and renal stones.

To explore the effects of virtual reality (VR) and augmented reality (AR) in the treatment of claustrophobia, the potential effects of VR and AR on induced anxiety were investigated in this paper. During the experiment, 34 subjects were randomly selected and distributed in AR and VR scenes in a sequence. The skin conductance and heart rates of the subjects were measured throughout the entire process, and the anxiety scale was used to assess the subjective anxiety when the task in each scene was completed. The results showed the following: (1) AR and VR scenes led to feelings of discomfort, but the subjective anxiety scores obtained in the two scenes were not significantly different; (2) the skin conductance level of the subjects significantly increased from the baseline when the subjects entered the experimental scene but remained active in the two scenes without showing significant difference between the scenes; and (3) the heart rate index significantly increased from the base line after the subject centered the scene and then gradually decreased.

Eye tracking data collection in the virtual reality context is typically carried out in laboratory settings, which usually limits the number of participants or consumes at least several months of research time. In addition, under laboratory settings, subjects may not behave naturally due to being recorded in an uncomfortable environment. In this work, we propose a proof-of-concept eye tracking data collection protocol and its implementation to collect eye tracking data from remotely located subjects, particularly for virtual reality using Ethereum blockchain and smart contracts. With the proposed protocol, data collectors can collect high quality eye tracking data from a large number of human subjects with heterogeneous socio-demographic characteristics. The quality and the amount of data can be helpful for various tasks in data driven human-computer interaction and artificial intelligence.

Many studies have been conducted on the use of virtual reality in education and training. This article lists examples of such research. Reasons to use virtual reality are discussed. Advantages and disadvantages of using virtual reality are presented, as well as suggestions on when to use and when not to use virtual reality.

A model that can be used to determine when to use virtual reality in an education or training course is presented. Virtual technologies can improve students' academic performance and motivation, e.g. students social and collaborative skills, and students' psychomotor and cognitive skills. Among them Virtual Reality (VR) and Augmented Reality (AR) have long been a popular design space for educational technology, and recently, Mixed Reality (MR) also increasingly applied for educational use. These immersive technologies have the potential to increase learner motivation and engagement, promote a full student-centered learning experience, support collaborative and situated learning and enable learners to more concretely and tangibly access previously physically inaccessible/invisible content

I. INTRODUCTION

Virtual reality (VR), augment reality (AR), and extended reality (xR) are recently developed tools for assisting people understand the real word around us.

A. Virtual Reality

Virtual Reality (VR) is an emerging technique which places a more and more important role in our daily lives. Recent development of high performance computing (HPC) platform makes it accessible to many researchers in the field of big data analysis and visualization [1]. Previous virtual-reality simulations of curved space, which typically present honeycombs or other periodic structures, have proven effective in letting mathematicians experience curved space directly. By contrast, for students and other non-mathematicians, a game like Non-Euclidean Billiards is more effective because it gives students not just something to see, but also something to do in the curved space. However, such simulations encounter a geometrical problem: they must track the player's hands as well as her head, and in curved space the effects of holonomy would quickly lead to violations of body coherence. This is, what the player sees with her eyes would disagree with what she feels with her hands. The article presents a solution to the body coherence problem, as well as several other questions that arise in interactive VR simulations in curved space. [2]

As the mainstream big data distributed computing framework cannot meet the requirements of virtual reality applications for high real-time business and data-intensive services, this paper designs a distributed server cluster system, which is deployed on Alibaba Cloud and Tencent Cloud. In the case of high concurrency, it provides users with a variety of high real-time data services to meet the requirements of real-time,

high availability, and high scalability, and solves the problems of high-concurrency user interaction and massive data storage and transmission in virtual reality systems. The distributed server cluster system can customize different policies according to specific business and cooperate with cloud computer resources to ensure data security and transmission efficiency. The experimental results show that the modules and algorithms of the system designed in this paper can run correctly, and the distributed server cluster system designed in this paper can meet the needs and goals of large-scale virtual reality scenarios. [1]

B. Augment Reality

Many virtual and augmented reality systems have been proposed to support renal interventions. Many computer assistance systems rely on virtual reality (VR) or augmented reality (AR) approaches. This review aims to provide a comparative and critical overview of those systems that are used for the renal interventions. The focus on these interventions results from the predominance of these areas for VR/AR applications. VR is described as “the use of computer modeling and simulation that enables a person to interact with an artificial three dimensional (3D) visual or other sensory environment” Whereas VR therefore relies on purely virtual environments, AR “allows the user to see the real world, with virtual objects superimposed upon or composited with the real world” Both aspects are used for clinical purposes by enabling the presentation of additional information pre- and intraoperatively. Preim and Botha provide a comprehensive overview of medical VR/AR displays. The scope of this review covers VR and AR systems for support in renal interventions with focus on PN, RN, PCNL and lithotripsy. VR is in this context considered as a virtual scene (e.g. virtual 3D model) that can be manipulated by human-computer interaction. [3]

C. Extended Reality

Recently, the HCI community has seen an increased interest in applying Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR) into educational settings. Despite many literature reviews, there still lacks a clear framework that reveals the different design dimensions in educational Extended Reality (XR) systems. Addressing this gap, they synthesize a broad range of educational XR to propose the XR-Ed framework, which reveals design space in six dimensions (Physical Accessibility, Scenario, Social Interactivity, Agency, Virtuality Degree, Assessment). Within each dimension, they contextualize the framework using existing design cases. Based on the XR-Ed Design framework, they incorporated instructional design approaches to propose XR-Ins, an instruction-oriented, step-by-step guideline in educational XR instruction design. Jointly, they aim to support practitioners by revealing implicit design choices, offering design inspirations as well as guide them to design instructional activities for XR technologies in a more instruction-oriented and learner-centered way. [4]

II. APPLICATIONS OF VR, AR, AND XR

AR and VR have many applications including data analysis, medical, and education.

A. Applications in Data Analysis

Eye tracking data collection in the virtual reality context is typically carried out in laboratory settings, which usually limits the number of participants or consumes at least several months of research time. In addition, under laboratory settings, subjects may not behave naturally due to being recorded in an uncomfortable environment. In this work, they propose a proof-of-concept eye tracking data collection protocol and its implementation to collect eye tracking data from remotely located subjects, particularly using Ethereum blockchain and smart contracts. With the proposed protocol, data collectors can collect high quality eye tracking data from a large number of human subjects with heterogeneous socio-demographic characteristics. The quality and the amount of data can be helpful for various tasks in data-driven human-computer interaction and artificial intelligence. [5]

We have developed a library for simulation of DC traction system. The modeling, installation and reliability aspects of Traction system in purview of current developments in HPC and Virtual reality are exemplified. We have proposed a model to predict and analyze traction system more comprehensively and have proposed a modeling structure regarding the same. Distribution of various computations an HPC setup has been shown and implemented. The advantages of Virtual reality in installation in city and intercity, driving has been implemented on C. VR framework for installation of DC traction thus has been implemented on Windows Compute Cluster Server. The proposal is then applied to model the system for installation at a crowded city using images taken from the satellites. The proposal is then explained through a set of codes which are derived from the library. The proposal is implemented using state-of-art tools for modeling like Extreme optimization library, XNA and WCCS. Recent research in Traction system taking in account the proposed techniques and computations required are also covered. The information can be used to enhance the functionality of the library developed. [6]

Spatial Augmented Reality (SAR) systems can be suitably combined with other existing Extended Reality (XR) technologies to support collaboration. In existing strategies, users unencumbered by a viewing technology, such as a tablet interface or a headmounted display, must rely on the transmission of their collaborators' positioning through interpreting a first-person camera view. This design creates a seam between a user's experience of the augmented physical environment in SAR, and their collaborators' experience inside the virtual environment. To assist in development and evaluation of spatial cues to support spatial awareness in SAR environments, an egocentric spatial-communication taxonomy is presented given two determining dimensions, a cue's attachment (physical/virtual) and animation (local/world). We developed four egocentric cues which characterize the four independent dimensions of the matrix: arrow, path, glow, and radial, and a single exocentric

world in miniature visualization. Our study shows that virtual attachment cues are preferred, providing the highest accuracy, highest performance when collaborators are occluded, and produce the least mental effort when used with a single virtual collaborator. For multiple collaborators however, the virtual attached, world animated radial cue produces significant increases in mental load and reductions in preference, demonstrating the impact of visual augmentation clutter. The single exocentric visualization produced higher levels of head movement, and poorer accuracy, however the novelty of the visualization produced positive qualitative results [7]

B. Applications in Mental Health

[8] VR is a type of advanced human–computer interface that allows users to interact with computers in real time and immerse themselves in the environment created by computers. In psychological treatment, VR technology can present audio-visual experiences more realistically than traditional therapy technology and provides a safe and controllable environment where patients can be treated. The scenario is customized according to the patient’s condition, such that the patient’s nervousness, fear, joy, and other emotions can be fully expressed. Various sensing devices are used to collect data and obtain feedback. Botella exposed a participant to eight kinds of claustrophobic space for 35–45 min, and the therapeutic outcome was evaluated. The anxiety level and avoidance reaction decreased. Bruce and Regenbrecht constructed a VR system for treating claustrophobia and included 18 subjects for the within-subject design experiment. Their result showed that these scenarios enabled the subjects to immerse in the VR environment, and the Subjective Unit of Discomfort scale (SUDS) score showed that the subjects felt discomfort. The two published study cases showed positive results. Some studies have proven that the construction of a closed room leads to subject discomfort. However, randomized controlled trials are required to assess the effectiveness of VR exposure treatment in the future.

Another EX was gradually developed, namely, augmented reality exposure treatment (ARET). Azuma reported that AR must satisfy the following conditions: (1) combined real and virtual objects in real environments, (2) interactive and real-time implementation, and (3) associated real and virtual objects. One of the purposes of virtual elements is to deepen user experience and knowledge. ARET is only suitable for small animal phobia, because the direction and position of the user relative to the screen are tracked through visual symbols in ARET. The virtual stimulus disappears completely when the user loses sight of the visual symbols. This detail hinders ARET from treating some phobias, such as large animal, public, and lightning phobias. The advantage of ARET is that the patient can be exposed to actual environments wherein he/she feels scared and anxious. The virtual object that the patient fears would be presented based on this environment. [3]

C. Applications in Education

The use of virtual reality (VR) in education can be considered as one of the natural evolutions of computer-assisted instruction (CAI) or computer-based training (CBT). Use of computers as instructional aids has a long history going back to the early 1950s. Serious studies began in the early 1960s. Since the advent of the microcomputer in 1977, computers, particularly microcomputers or personal computers (PCs), have become a growing and recognized delivery system for many forms of education. Virtual reality, which can be used on all types of computers, has followed that trend. In her extensive bibliography on virtual reality in education and training, Pantelidis (1991-2009) lists over 800 printed resources, such as articles and reports, on this application of virtual reality, going back to 1989. The list is by no means complete and comprehensive. The advantages of using VR to teach educational objectives are similar in many ways to the advantages of using a computer or interactive simulation, particularly a three-dimensional computer simulation. Computer-based simulations have been used for many years in computer-assisted instruction (CAI). In fact, advantages of computer based simulations are well known. Zacharia (2003), referring to Chou (1998) asserts that “researchers attribute success of simulations to the empowerment of students, the unique instructional capabilities, the support for new instructional approaches, the development of cognitive skills, and the development of attitudes”. Ferry et al. (2004) state that “Whilst we acknowledge that a simulation is only a representation of real-life, there are features that can enhance real-life experience. For example, a simulation can provide authentic and relevant scenarios, make use of pressure situation that tap users’ emotions and force them to act, they provide a sense of unrestricted options and they can be replayed”, referencing Aldrich (2004). Steinberg (2000) contends that “students should know that simulations make it possible to explore new domains, make predictions, design experiments, and interpret results”. [9]

III. EXPECTED FUTURE IMPROVEMENTS

By 2030, IDTechEx predicts the augmented, virtual and mixed reality market to be over 30Bn. With COVID limiting physical interaction, virtual communication and interaction will be the normal for many years to come. Although seen as futuristic by some, augmented, virtual and mixed reality devices have shown that they have an important part to play in many different industries. They are truly the technology of the future, today. Over the past two decades, there have been big strides in the technological development of XR products. In the early 90’s a number of headsets were released which allowed the user to view a video on a headset from an external device. Then in the next decade, devices such as the google glass explorer were released. These captured the imagination of many people, and it was one of the first “augmented reality” products available to consumers. Although this did not take off, it paved the way for other products to be developed. Later in the decade, more VR and AR products were released, for example in 2019, Magic Leap released its first product after

years of funding. Oculus also released a VR headset which does not need to be connected to a computer – the Quest – proving that VR does not need to be tethered to a computer.

IV. CONCLUSIONS

This review shows the extent to which VR/AR support is already used in clinical practice. Furthermore, there is a desire for an increased use of AR systems in urology. VR/AR has the potential to improve the safety and outcome of renal interventions in the future. Although results from large clinical studies are not yet reported in the literature, many advances in the last ten years have led to sophisticated systems. Further interdisciplinary research is required to cope with current limitations of VR/AR assistance in clinical environments.

The study indicated that the proposed AR and VR systems effectively induced anxiety in subjects. Moreover, by measuring the subjective scale scores, skin conductance, and heart rate, no significant difference was found between the proposed AR and VR systems in terms of inducing anxiety. Future research will further quantify the contribution of each type of stimulus from human perceptions (visual, auditory, haptic, or smell) and investigate whether AR or VR system is more effective in inducing anxiety.

Virtual reality has a place in education and training. Research on educational applications of VR, as well as research on the educational use of simulations has shown its value. There are many reasons to use VR and advantages to using VR. The educator or trainer has only to determine when to use it. The use of a model can help make that determination. Such a model can play a part in the continuing search for ways to use virtual reality in education and training courses.

A. Figures and Tables

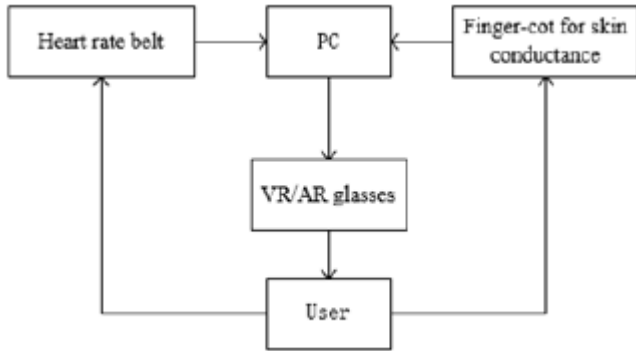


Fig. 1. System Architecture of VR and AR to Induce Anxiety.

ACKNOWLEDGMENT

The work of this paper is a research of different aspects that virtual reality is used. From education to health and to infrastructure. This article is solely for learning purposes and research.

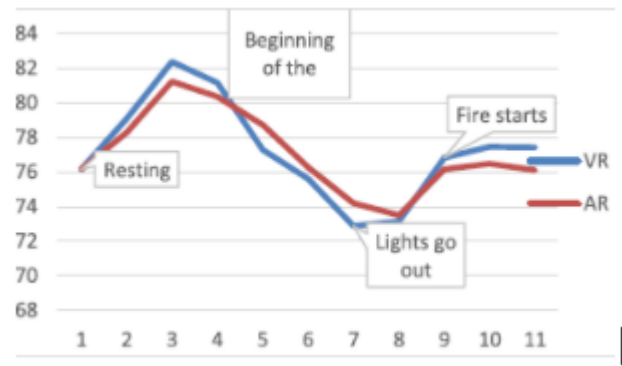


Fig. 2. Changes in the mean value of heart rates in the AR and VR scenarios to Induce Anxiety.

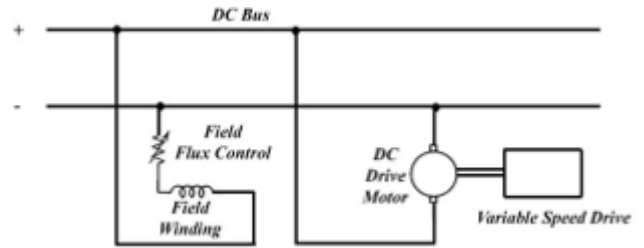


Fig. 3. Basic schematic diagram of the classical variable-speed dc drive for Modeling of Traction Systems.

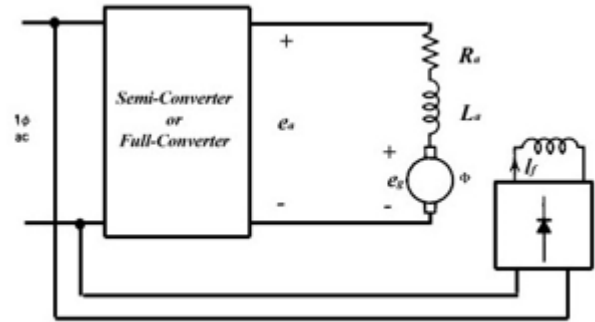


Fig. 4. Basic single-phase circuit for the speed control of a separately excited dc motor for Modeling of Traction Systems.



Fig. 5. Virtual Reality in Education.

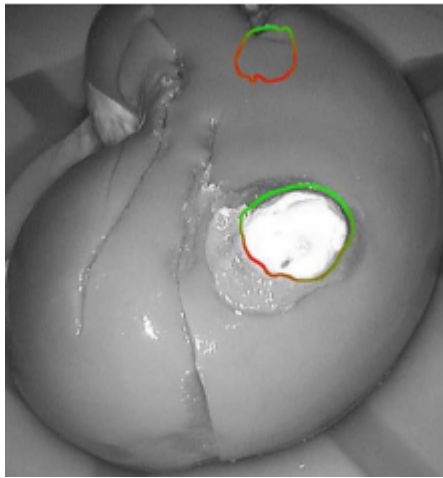


Fig. 6. AR view of tumor margins on ex vivo lamb kidney for Systems for Renal Interventions.

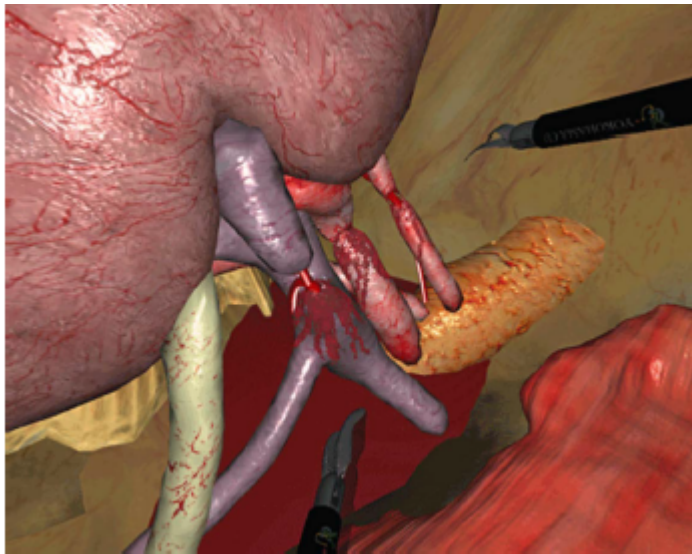


Fig. 7. View of VR environment of in simulator for "rehearsal" laparoscopic renal surgery for Systems for Renal Interventions.

REFERENCES

- [1] B. Gan and C. Zhang, "Research on application of distributed server architecture for virtual reality scenarios in big data environment," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, 2020, pp. 52–55.
- [2] J. Weeks, "Virtual reality simulations of curved spaces," 2020.
- [3] F. J. Detmer, J. Hettig, D. Schindele, M. Schostak, and C. Hansen, "Virtual and augmented reality systems for renal interventions: A systematic review," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 78–94, 2017.
- [4] K. Yang, X. Zhou, and I. Radu, "Xr-ed framework: Designing instruction-driven and learner-centered extended reality systems for education," 2020.
- [5] E. Bozkir, S. Eivazi, M. Akgün, and E. Kasneci, "Eye tracking data collection protocol for vr for remotely located subjects using blockchain and smart contracts," 2020.
- [6] S. Ahmed, R. Pathak, A. Jain, S. Katiyal, A. Ludhiyani, and S. Joshi, "Applying hpc and virtual reality for modeling of traction systems," *International Journal of Academic Research*, vol. 2, pp. 107–116, 05 2010.

- [7] A. Irlitti, T. Piumsomboon, D. Jackson, and B. H. Thomas, "Conveying spatial awareness cues in xr collaborations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 11, pp. 3178–3189, 2019.
- [8] S. C. Yeh, Y. Y. Li, C. Zhou, P. H. Chiu, and J. W. Chen, "Effects of virtual reality and augmented reality on induced anxiety," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1345–1352, 2018.
- [9] V. S. Pantelidis, "Reasons to use virtual reality in education and training courses and a model to determine when to use virtual reality," pp. 59–70.

AVERAGE

Deep Learning for Face Recognition

Racine Mclean

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Suxia Cui

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Abstract—This document is a model and instructions for L^AT_EX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

Face recognition is a challenging task in various environments for different applications [1]. There are both traditional methods and machine learning methods. Recently, deep learning provides a new paradigm utilizing with training data-sets to improve the performance [2], [3]. Face recognition has been around for years now, the dawn of facial recognition was early 1960s but due to the lack in computing power and popularity the technology was unpopular.

The earliest pioneer of facial recognition was Woodrow Wilson Bledsoe. In 1964 Bledsoe began work using computers to recognise the human face. Due to the funding of the project originating from an unnamed intelligence agency, much of his work was never published. However it was later revealed that their initial work involved the manual marking of various “landmarks” on the face such as eye centres, mouth etc. These were then mathematically rotated by a computer to compensate for pose variation. The distances between landmarks were also automatically computed and compared between images to determine identity [4].

In today’s modern age digital technology developments and methods like artificial intelligence and Internet of Things, Artificial Intelligence facial recognition is now a highly developing technology. The category of machine learning, deep learning plays a vital role in modern day face recognition. Deep learning uses ML algorithms and lots of data to educate deep neural networks for enhanced accuracy. In this paper we will focus on deep learning method of face recognition. To utilize Deep learning algorithms to process images for face recognition, large data-sets are fed into the algorithm for training so that the network can learn light, expression, angle and other invariant characteristics from the data.

In this paper, Section II will discuss the traditional approach for face recognition, in Section III we will go over the Deep Learning approach of face recognition and the basic structure of deep learning convolutional neural network. Section IV

This research is funded by NSF grant OAC 1827243 and HRD 1332566.



Fig. 1. An image of Woodrow Bledsoe from a 1965 study. PHOTOGRAPH: DAN WINTERS

provides an explanation of the the approach used in our face recognition research project, then Section V will provide a breakdown of our results and constraints and finally in Section VI will convey the conclusions of this paper.

II. TRADITIONAL APPROACH

Before the age of deep learning there were numerous traditional mathematical method of face recognition. The main function of the face detection step are (1) To determine the image is a face image, and (2) The location of the face in the picture. The expected output of this step is the group of images that have been sorted from the input image.

While the feature extraction step include dimension reduction, salience extraction, and noise cleaning. After that, the image is transformed into a certain dimension vector according to the number of extracted features. Some literature reveals that feature extraction goes into the steps of face recognition [11]. The various classifier is used in classification step to identification and verification. To achieve recognition automatically, the face data set needs to be created, for each person consists of several images with different poses and then its features are executed and stored in the data set [11].

In face recognition, the performance of an algorithm depends on the conditions of training data and circumstances changes. High accuracy is not only factor in determining the performance of an algorithm. Eigenface and Fisherman methods are the methods used in face recognition. The methods have advantages and disadvantages which are further developed with a combination of Principal Component Analysis (PCA) algorithms, Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA) and so on. Various existing studies will help in determining the traditional method of optimal for face recognition.

A. Eigenface Method

The aim of face recognition is to distinguish the input signal in the form of image data to several classes (person). The input signal has a high noise caused by differences in lighting, pose and so forth. Each face does have different characteristics but has a similar pattern that can be detected, for example with the eye, mouth, nose, and range of distance between the object. Facial recognition based on characteristics of this feature is known as Eigenface. The objects can be extracted by mathematically using Principal Component Analysis (PCA). The PCA is tasked with transforming each original image into its corresponding Eigenface [12].

Eigenface is a method that is included in appearance-based approach. The basic principle of this face recognition is to cite the unique information from face then encoded and compared with the previous decoded results. In the eigenface method, decoding is done by calculating the eigenvector and then represented in a large matrix [13]. Here are the stages of this method: [14]

- 1) The M vector of size N (image line $\times$ image column) represents the set of the sample image.
- 2) The center average is reduced by the average image of each image vector to know the eigenvector and eigenvalue of the covariance matrix.
- 3) Set back the eigenvectors and eigenvalue and calculate the cumulative energy content for each eigenvector.
- 4) Select a subset of the eigenvector as the base vector.
Project the data z scores into the new base.

B. Fisherface Method

Fisherface is particularly useful when facial images have great variations in their illumination and expression. The Fisherface method is an improvement of the Eigenface method using Fisher's Linear Discriminant Analysis which is abbreviated as FLDA or LDA for dimensional reduction [15]. FLDA is one example of class-specific method, because this method

forms the distance (scatter) between classes and intra classes so as to produce a better classification [16].

A fisherfaces method is a holistic approach used in the facial recognition stage. The holistic stage is an identification based on the local features of the face (eyes, nose, and mouth). Fisherface approach is a widely used method for feature extraction in face images. This approach tries to find the direction of projection in which images belonging to different classes are separated maximally. Mathematically, it tries to find the projection matrix (weight) in such a way that the ratio between the scatter matrix class and the image in it can be maximally projected.

C. Support Vector Machine (SVM) Support Vector Machine is a learning algorithm that analyzes data for classification and regression analysis purpose. SVM is a separation or discrimination between two classes. SVM classifies by finding a hyper plane that maximizes the margin between the two classes. Stages of the SVM algorithm can be seen in Figure 3 [17].

The larger amounts of data from the training set, the rate of recognition of the SVM method will be smaller because it affects the length of computation and the error rate that makes this method does not produce satisfactory output for large data [18]. So SVM is suitable to be applied to small and irregular sets for faster computing and better accuracy [19].

III. DEEP LEARNING APPROACH

Deep learning is a type of machine learning that trains a computer to perform human-like tasks, such as recognizing speech, identifying images or making predictions. Instead of organizing data to run through predefined equations, deep learning sets up basic parameters about the data and trains the computer to learn on its own by recognizing patterns using many layers of processing. An important deep learning algorithm that is applied to face recognition is convolutional neural networks (CNNs).

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. [5].

Convolution Neural Network (CNN) has four layer architectures to check the degrees of the shift, scale, and distortion. The four layers are [6]:

A. Convolutional Layer

This layer is the main layer that underlies the CNN process. The convolution process is the process of applying a function into the output of another function repeatedly. In image processing techniques, the purpose of convolution is to extract features from the inserted image. The result of the extraction process is linear transformation data.

B. Subsampling Layer

Subsampling is the process of reducing the size of an image data. Most of CNN, the widely used subsampling method is max pooling. This max pooling technique works by dividing the output from the convolution layer into several small grids. Then the highest value of each grid is arranged in a matrix.

C. Full Connection Layer

Full connection layer is a layer that serves to perform the transformation on the data dimension so that data can be classified linearly. Each neuron in the convolution layer needs to be transformed first so that the information contained is not lost.

D. Output Layer

The output layer is the last layer as a result of the Convolution Neural Network (CNN) process.

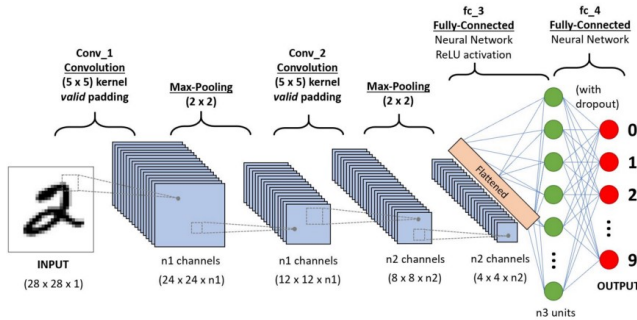


Fig. 2. Structure of CNN

Most face recognition system that follow a similar algorithm structure like early ones with, with modifications to improve the accuracy and performance of the algorithm. there are both advantages and disadvantages of using deep learning for face recognition.

Advantages

As one method, the Convolution Neural Network (CNN) also has several advantages and disadvantages. Here are the advantages and disadvantages of the Convolution Neural Network (CNN)[8] [7]:

- Can be implemented in various image resolutions.
- Computing is so detailed that the rate of error is likely small.
- Convolution Neural Network (CNN) is able to solve problems that have a high complexity that has many parameters to be computed.
- Can classify the face shape of known and unknown data.

Disadvantages

- Not suitable to be simple
- Process long enough
- Computing is very complex, directly proportional to the complexity of the problems encountered.
- Can't describe on the face with a certain position.

IV. FACE RECOGNITION USING DEEP LEARNING AND COMPUTER VISION

In this project we utilize deep learning and computer vision to develop a face recognition attendance system. There are

a few python libraries that are required to perform our face recognition. These libraries are computer vision(CV2), numpy, and face recognition. One important non-python library used is Dlib. Dlib is a modern C++ toolkit containing machine learning algorithms and tools for creating complex software in C++ to solve real world problems. It is used in both industry and academia in a wide range of domains including robotics, embedded devices, mobile phones, and large high performance computing environments. Dlib's open source licensing allows you to use it in any application, free of charge.

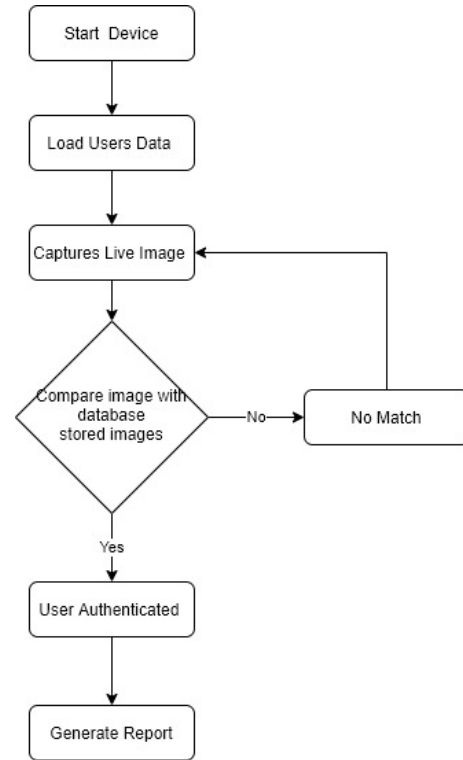


Fig. 3. Flow chart showing the software process.

The software was develop to run on the Jetson Nano microcomputer which has a 128-core Maxwell GPU. NVIDIA® Jetson Nano™ Developer Kit lets you run multiple neural networks in parallel for applications like image classification, object detection, segmentation, and speech processing. All in an easy-to-use platform that runs in as little as 5 watts.

Adam Geitgey who is one of the main contributors to the face-recognition python API stated "... the Jetson Nano supports the exact same CUDA libraries for acceleration that almost every Python-based deep learning framework already uses. This means that you can take an existing Python-based deep learning app and often get it running on the Jetson Nano with minimal modifications and still get decent performance." [8] The Jetson Nano utilize and external camera for image processing applications. Various cameras were considered for this application and as you know for face recognition high quality images are required, we decided to go with the 8 megapixel raspberry pi camera. Though there are much more powerful cameras that are compatible with the Jetson Nano,

our decision was not only based on performance but also cost. In selecting the raspberry pi camera our system not only meets performance standards but also cost effective. With these two hardware components our system is now ready to process faces.

V. RESULTS

The system was successful in recognizing my face and recording my attendance. The application identified numerous pictures of individuals and recorded their attendance as well. Due to limited access to individuals to test the system on, we had to use images on a phone and hold it up to the camera. The system was successful in identifying the pictures and recording the attendance. With the system recognizing an individual face from a picture one might see this as a failed test case. This is not the case at all, the system is only required to recognise a face and record the person name and time seen, whether it be a live feed or an image held up to the camera. Though I was unable to test the application with numerous faces from live camera feed, I still manage to find some faults with the face recognition API, which I believe stems from the data-set fed into the face-recognition algorithm. The system was miss identifying people with darker complexion, in troubleshooting the program I realize that that this can stem from the data set that was loaded into the neural network

VI. CONCLUSION & FUTURE WORK

In the future I plan on editing the data-set that is going to be used to train the neural network. Now that we have discuss face-recognition traditional and deep learning recognition. One can see the reason for shifting from the traditional hand calculation method to the deep learning approach. From the evidence in this research paper we can conclude that deep learning method for face recognition is the best solution thus far Daniel Saez stated "...face recognition systems based on CNNs have become the standard due to the significant accuracy improvement achieved over other types of methods." [9]. Everyday advancements are being made in the field of deep learning. I will advice everyone to get into AI and Machine learning since this is the future.

REFERENCES

- [1] Y. Zhou, D. Liu, and T. Huang, "Survey of face detection on low-quality images," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 769–773.
- [2] C. Peng, N. Wang, J. Li, and X. Gao, "Soft semantic representation for cross-domain face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 346–360, 2021.
- [3] M. Wang and W. Deng, "Deep face recognition: A survey," *CoRR*, vol. abs/1804.06655, 2018. [Online]. Available: <http://arxiv.org/abs/1804.06655>
- [4] J. Bar-Ilan, "The history of information security: A comprehensive handbook 2008" edited by karl de leeuw and jan bergstra. the history of information security: A comprehensive handbook . oxford: Elsevier 2007. 887 pp. (hard cover), isbn: 9780444516084," *Library Hi Tech*, vol. 26, pp. 682–683, 11 2008.
- [5] X. Qu, T. Wei, C. Peng, and P. Du, "A fast face recognition system based on deep learning," in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 01, 2018, pp. 289–292.
- [6] S. Setiowati, Zulfanahri, E. Franita, and I. Ardiyanto, "A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods," 10 2017, pp. 1–6.
- [7] S. Setiowati, Zulfanahri, E. L. Franita, and I. Ardiyanto, "A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods," in *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2017, pp. 1–6.
- [8] A. Geitgey, "Build a hardware-based face recognition system for \$150 with the nvidia jetson nano and python," 2019.
- [9] L. M. Daniel Saez Trigueros, "Face recognition: From traditional to deep learning methods."

BELOW AVERAGE

Machine Learning Based Job Scheduling Strategies on Cloud Computing

Kobi Tioro

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Suxia Cui

*Electrical and Computer Engineering
Prairie View A&M University
Prairie View, USA
email address or ORCID*

Abstract—This document is a model and instructions for L^AT_EX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

Cloud Computing provides delivery of infrastructure, storage, databases, and networking to its users. With the help of cloud computing, users can manipulate and store large datasets for real-time processing at a local resource-limited environment. Job scheduling methods are very important to utilize cloud computing resources. There are several approaches, such as heuristics, and game theory. Recently, machine learning is emerging [1], [2].

As today, most of jobs in high-performance computing are fixed, rigid, and non-preemptable batch jobs. The efficiency of job scheduler has always been a thought topic. The parameters utilized in the decision on what the computer need to do depending on the situation make hard to have a perfect solution. With the evolution of deep learning, researchers are constantly working on the efficiency of job scheduler. In this article, Dr. Zang wants to contribute by showing that deep reinforcement learning agents can learn how to schedule HPC batch jobs, proposing a set of design choices and optimizations in RLScheduler to stabilize and showing the ability to adapt to different optimization goals.[3]

To schedule high-performance computing jobs is critical assignment for computing systems. Optimization of tasks is a continuous topic of research since the computer need to be efficient on the execution of the tasks assigned to it. The algorithms used by the job scheduler need to be adapted to the new types of tasks and keep optimizing the task execution using deep learning. In this article, Dr. Zang wants to contribute by showing the application of the A2C deep reinforcement to the data center resource management. His second point is to show the benefit of using the multi-step Temporal-difference (TD) method to speeds up the training process. Last, he want to show the performance of their approach on both real-world and simulated datasets.[4]

The major issue for data centers has been the optimization of the job scheduling time for years. Resource management is the topic that researchers has been working on to find better solutions. The incoming jobs require different CPU and memory units, and span different number of time slots. This paper shows how the combination of deep learning and the usage of multiple clusters could improve the resource management. Chen, Dr. Xu and Dr. Wu reveals that deep reinforcement learning method can perform better traditional resource allocation algorithms in a variety of complicated environments.[5]

II. BACKGROUND

A. HPC batch job

HPC batch job has multiple parameters to take in consideration such as user id, group id, requested nodes, maximal execution time, and the submitted queue. So Dr. Zang investigated on the parameters that can help improve RLScheduler thought investigation. Those parameters are listed below:

- The id of a job id.
- The submission time of a job st.
- The requested computing nodes of a job nc.
- The requested execution time of a job rt.

The following equation is an optimization equation that help optimizing the job slowdown:

$$bsld = \max(wt + rt_{\max}(rt), 1)$$

metric bounded slowdown= bsld wt is the time that a job waits in the job queue; is a constant, with a typical value of 10s to prevent small tasks from having excessively large slowdown values. For a set of jobs, the goal is to minimize the average bounded slowdown.

B. Deep Learning

Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network. Common application of deep learning are Self Driving Cars, News Aggregation and Fraud News Detection, Natural Language Processing or Virtual Assistants like Siri, Alexa or Google voice.

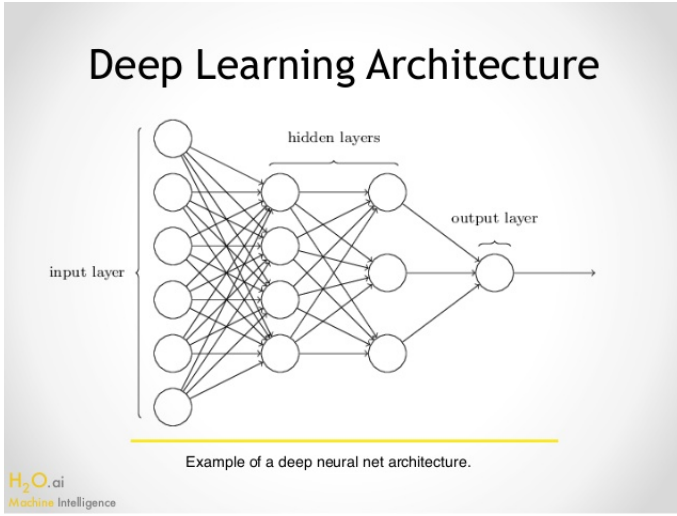


Fig. 1. Example of a deep learning architecture

C. Reinforcement learning

Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning

Reinforcement learning (RL) called is a type of machine learning technique that enables an agent to autonomously learn in an interactive environment by trial and error using feedback from its own actions and experiences.

III. METHOD

My research focus will be on the Reinforcement learning for the Job scheduler. My Design based on Dr. Zang research will be divided in three parts :

A. RLSCHEDULER AGENT

Fig. 2 displays a general structure for RL. At any t -step, The corresponding state S_t is observed by the agent, and takes an Action A_t . The operation will then move the state of the world from S_t to S_{t+1} and the agent will obtain The $R_t + 1$ reward.

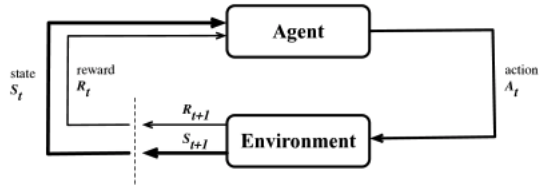


Figure 1: General framework of reinforcement learning.

Fig. 2. General framework of reinforcement

B. SCHEDGYM ENVIRONMENT

The state is gathered from the environment and created from it. The behavior will be applied to the setting and therefore

produce the next state and a reward. The agent learns from his acts and the corresponding rewards by steps.

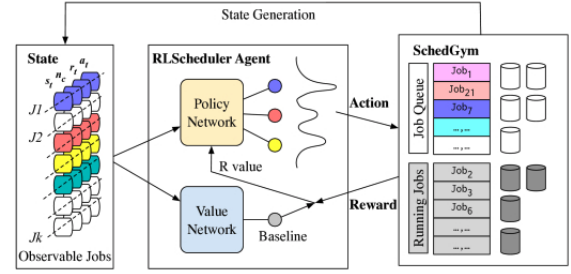


Figure 2: The overall architecture of RLScheduler.

Fig. 3. Example of a figure caption.

C. RLSCHEDULER OBSERVABLE STATE

IV. RESULTS

In order to check the efficiency of the RLScheduler, some test cases need to be run. In this research, we will analyse the GitHub repository posted by Hongzi Mao. He is a Ph.D. student in Electrical Engineering and Computer Science at Massachusetts Institute of Technology. His PHD thesis is Network System Optimization with Reinforcement Learning.

A. How to run an Iteration using the RLscheidungler repository

The first step is to find a repository to run the code. For this case, the repository can be downloaded by the following link: <https://github.com/hongzima/deepm>.

The second step is to install the prerequisites libraries. For this repository, the code is written in Python and the key libraries are NumPy, Theano, Lasagne and matplotlib. NumPy is needed to for the manipulation data in arrays. Theano is a Python library and optimizes compiler to manipulate and evaluate mathematical expressions, particularly matrix-assessed expressions. Lasagne is a light library designed to build and train neural networks in Theano. Matplotlib is the Python programming language plotting library and its NumPy numerical mathematics extension. It is widely used in data analysis and data visualization.

All these libraries can be obtained by setting up an anaconda environment. The last step is to launch an experiment by using the python file launcher.py which the main file.

B. List of the experiments

Below is the list of the experiments with their functions.

```

--exp_type <type of experiment>
--num_res <number of resources>
--num_nw <number of visible new work>
--simu_len <simulation length>
--num_ex <number of examples>
--num_seq_per_batch <rough number of samples in one batch update>
--eps_max_len <episode maximum length (terminated at the end)>
--num_epochs <number of epoch to do the training>
--time_horizon <time step into future, screen height>
--res_slot <total number of resource slots, screen width>
--max_job_len <maximum new job length>
--max_job_size <maximum new job resource request>
--new_job_rate <new job arrival rate>
--dist <discount factor>
--lr_rate <learning rate>
--ba_size <batch size>
--pg_re <parameter file for pg network>
--v_re <parameter file for v network>
--q_re <parameter file for q network>
--out_freq <network output frequency>
--ofile <output file name>
--log <log file name>
--render <plot dynamics>
--unseen <generate unseen example>

```

list of the experiments with their functions

C. Example of sample experiments

Here are some experiments that can be launched by running the command below:

```

python launcher.py --exp_type=pg_su --simu_len=50 --num_ex=1000 --ofile=data/pg_su
--out_freq=10

python launcher.py --exp_type=pg_re --pg_re=data/pg_su_net_file_0.pkl --simu_len=50 -
-num_ex=10 --ofile=data/pg_re

python launcher.py --exp_type=test --simu_len=50 --num_ex=10 --pg_re=data/
pg_re_10.pkl --unseen=True

```

Fig. 4. Job slowdown Test Case

1. The first line of the code function is to launch supervised learning for policy estimation
2. The second line of the code function is to launch policy gradient using network parameter just obtained
3. The third line of the code function is to launch testing and comparing experiemnt on unseen examples with pg agent just trained

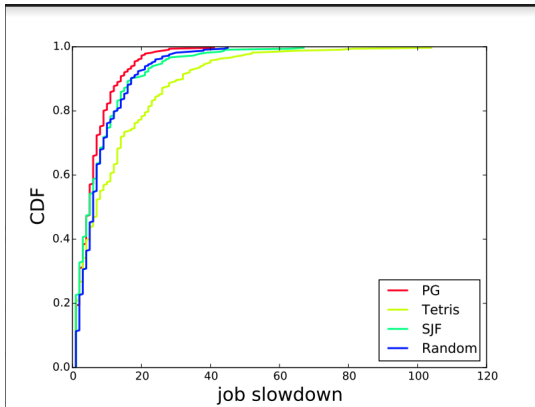


Fig.6 Job slowdown CDF

The fig.6 shows The CDF of the PG, Tetris,SJF and random in function of the slowdown metric.

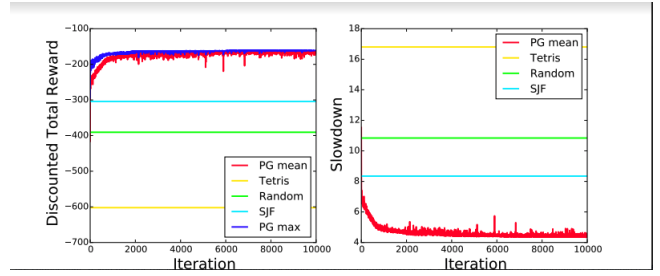


Fig.7 Job comparison data

The fig.6 shows The CDF of the PG, Tetris,SJF and random in function of the slowdown metric.

V. CONCLUSION

In this research paper,RLScheduler, a deep reinforcement Job scheduler based on learning. The goal was to get familiar with the RLScheduler job scheduler and how to run iterations using python. Overall, we are were able to get pick up some notions and add our inputs to this challenging topic which is really famous in the research industry.

REFERENCES

- [1] H. Miao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *HotNet '16: Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, 2016, pp. 50–56.
- [2] W. Chen, Y. Xu, and X. Wu, "Deep reinforcement learning for multi-resource multi-machine job scheduling," *CoRR*, vol. abs/1711.07440, 2017. [Online]. Available: <http://arxiv.org/abs/1711.07440>