

# CSx55: DISTRIBUTED SYSTEMS [INTRODUCTION]

## The Road Ahead

Winds from ports to packets  
and thence onto sockets

Over which all data must traverse

Intricacies of threads unraveled

To prod CPUs along roads less traveled

Consign the bane of deployments wobbly and stale

By writing code that will scale

Of programs adept at coordination

With nary a dash of intervention

SHRIDEEP PALICKARA

Computer Science

Colorado State University



# Topics covered in this lecture

- Introduction
- Course overview and expectations
- HW



The background is a dark blue gradient with a digital theme. It features a faint line graph with a red line and a bar chart with red bars. Scattered throughout are binary digits (0s and 1s) in a light blue color, some of which are slightly blurred or faded. The overall aesthetic is high-tech and data-oriented.

# **DISTRIBUTED SYSTEMS**

## **QUICK OVERVIEW**

# What is a distributed system?

- *A distributed system is one in which hardware and software components located at networked computers communicate and coordinate their actions only by passing messages.*

Coulouris, Dollimore, Kindberg and Blair

- *A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.*

Leslie Lamport



# Why Distributed Systems?

- Your hard-drive's primacy has been eroding
- Data and programs are delivered over the network
  - ▣ No single hard drive can hold all the data you need
- Services themselves are distributed
  - ▣ Google search is backed by a massive distributed cloud



# Distributed systems builds on a diverse set of areas

- Networking
- Concurrency
- Machine learning
- Algorithms and Graph Theory
- Information Retrieval
- Transactional Systems
- Cryptography
- Failure recovery and consistency models
- Probability theory



# Distributed Systems: CHALLENGES

[1 / 2]

- **Scale** with increases in data and users
- **Responsiveness**
  - ▣ Regardless of data size, responses must be prompt
- **Intelligent**
  - ▣ Correlate all sorts of information



- Dealing with system conditions
  - ▣ Murphy's Law
  - ▣ Malicious Users
  - ▣ Byzantine failures
- Security
  - ▣ Detection
  - ▣ Privacy and Accountability
  - ▣ Authorizations





# ABOUT ME



# About me ...

- I do research in the area of large-scale computing systems, Big Data, and GeoAI
- My research has been funded by agencies in the United States and the United Kingdom
  - ▣ These include the National Science Foundation, the Department of Homeland Security (including the *Long Range* program), the Environmental Protection Agency, the Department of Agriculture, the National Institute of Food & Agriculture, the National Endowment for the Humanities/Teagle and the U.K's e-Science program
  - ▣ Recipient of the National Science Foundation's CAREER Award
  - ▣ I direct the Center for eXascale Spatial Data Analytics and Computing (XSD) @ CSU [<https://spatial.colostate.edu>]



# My research has been deployed in

- ❑ Urban sustainability
- ❑ Commercial internet conferencing systems
- ❑ Defense applications
- ❑ Precision Agriculture
- ❑ Earthquake sciences
- ❑ Epidemic modeling
- ❑ Healthcare
- ❑ Bioinformatics
- ❑ Brain Computer Interfaces
- ❑ High energy physics
- ❑ Visualizations



# **COURSE LOGISTICS, EXPECTATIONS, AND SUCH**



# You are not allowed to take away learning opportunities from other students

- You can use a laptop ONLY IF you have a University accommodation that expressly requests it
  - ▣ You will have to sit in the last row starting at the corners
  - ▣ Turn off wireless and use it only for taking notes
  - ▣ Students not using a laptop should avoid sitting in the last row
- Tablet use with a stylus/pencil is allowed
  - ▣ Tablet use in keyboard mode is not allowed
- When the class is in session, put away your cell-phones!
- Also, please no cross talking when the class is in-session



# Course webpage

- All course materials will be on the course webpage  
<http://www.cs.colostate.edu/~csx55>
  - Schedule
  - Lectures
  - Assignments
  - Infospaces, Syllabus
- We will use Canvas for submission of assignments and grade reporting
- For the **801** sections of CS455 and CS555
  - Canvas will be used to administer exams, quizzes, and release lecture recordings



# Office Hours

- Professor: **Shrideep Pallickara**

Office Hours: 1:00 – 2:00 PM Friday and via Zoom

- GTA: Office hours will be in-person and via MS-Teams

- ▣ An almost finalized schedule is posted on the course website

- Please send all e-mails to: **compsci\_csx55@colostate.edu**



# Course textbook

- This class has two **optional** textbooks

Distributed Systems: Principles and Paradigms. *Andrew S. Tanenbaum and Maarten Van der Steen*. 3rd Edition. Createspace, ISBN 9781530281756

Distributed Systems: Concepts and Design. George Coulouris, Jean Dollimore, Tim Kindberg, Gordon Blair. 5th Edition. Addison Wesley. ISBN: 978-0132143011





# When I make slides ...

- I usually refer to several texts
  - ▣ And technical papers and articles (with URLs)
- I always list my references at the end of every slide set



# Textbooks that I will refer to during the course include ...

(1/2)

- *Distributed Systems: Principles and Paradigms. Andrew S. Tanenbaum and Maarten Van der Steen. 2nd Edition. Prentice Hall. ISBN: 0132392275/978-0132392273.*
- *Distributed Systems: Concepts and Design. George Coulouris, Jean Dollimore, Tim Kindberg, Gordon Blair. 5th Edition. Addison Wesley. ISBN: 978-0132143011*
- *Distributed Computing: Principles, Algorithms, and Systems. Ajay Kshemkalyani and Mukesh Singhal. 1st edition. Cambridge University Press. ISBN: 0521876346/ 978-0521876346*
- *Computer Networks: A Systems Approach. Larry Peterson and Bruce Davie. 4th edition. Morgan Kaufmann. ISBN: 978-0-12-370548-8.*
- *Java Concurrency in Practice. Brian Goetz, Tim Peierls, Joshua Bloch, Joseph Bowbeer, David Holmes, and Doug Lea. Addison-Wesley Professional. ISBN: 0321349601/978-0321349606.*
- *Java Threads. Scott Oaks and Henry Wong. . 3rd Edition. O'Reilly Press. ISBN: 0-596-00782-5/978-0-596-00782-9*



# Textbooks that I will refer to during the course include ...

(2/2)

- *Hadoop: The Definitive Guide*. Tom White. 3rd Edition. Early Access Release. O'Reilly Press. ISBN: 978-1-449-31152-0
- *Concurrent Programming in Java(TM): Design Principles and Pattern*. Doug Lea. 2nd Edition. Prentice Hall. ISBN: 0201310090/978-0201310092.
- *Cloud Application Architectures: Building Applications and Infrastructure in the Cloud*. George Reese. 1st edition. O'Reilly. ISBN: 0596156367/978-0596156367.
- *Practical Cryptography*. Niels Ferguson and Bruce Schneier. 1st edition. Wiley Publishing. ISBN: 0-471-22894-X/0-471-22357-3.
- *Unix Systems Programming*. Kay Robbins & Steve Robbins, 2nd edition. Prentice Hall. ISBN: 978-0-13-042411-2.
- *Operating Systems Concepts*. Avi Silberschatz, Peter Galvin, Greg Gagne. 8th edition. John Wiley & Sons, Inc. ISBN-13: 978-0-470-12872-5.



# Infospaces (<https://infospaces.cs.colostate.edu>)

- **Knowledge repository** that we have been building to enhance learning
- All videos are designed to be less than 2 minutes
- Improving Infospaces
  - ▣ Let us know what you would like to see
  - ▣ If you'd like to contribute to this repository let us know!



# GRADING



# Grading breakdown

- Assignments: 40%
  - ▣ HW1: 10%; HW2: 10%; HW3: 10%; HW4: 10%
- Term project [report and presentation]: 10%
- Quizzes (10 best) : 10%
- Midterm: 20%
- Comprehensive final exam: 20%



# To allow for some room to trip-up (not that you should plan to!):

- Each programming assignment will be curved so that the average score is 80%
  - ▣ Only students who have received a minimum of 40% will be considered for (a) calculation of the average, and (2) receiving the extra points (if any are awarded, to preserve the 80% average)
- If you receive a 40% in each of these assignments; we will drop one assignment where you had your lowest score
  - ▣ That is, if you score at least a 40% in each of the 4 assignments, your top-3 assignments will account for programming component of the course grade (which incidentally is also 40%)



# Quizzes, midterm, and final exams

- I will only ask questions about what I teach
- If the concepts were covered in my slides
  - ▣ You should be able to answer the questions





# Grading Policy

[1/3]

- Letter grades will be based on the following standard breakpoints:
  - $\geq 90$  is an A,  $\geq 88$  is an A-,
  - $\geq 86$  is a B+,  $\geq 80$  is a B,  $\geq 78$  is a B-,
  - $\geq 76$  is a C+,  $\geq 70$  is a C,
  - $\geq 60$  is a D, and  $< 60$  is an F.
- I will not cut higher than this, but I *may* cut lower.



- Every assignment will be posted 4 weeks before the due date.
- Every assignment will include information about:
  - ▣ How much it will count towards the course grade
  - ▣ How it will be graded
- Late submission penalty: 7.5% per-day for the first 2 days
  - ▣ Submissions after the late submission period will have an automatic ZERO



- ❑ **Automatic grading** of assignments
- ❑ Once you submit your assignment solution, the framework:
  - ❑ Crawls Canvas
  - ❑ Retrieves your submission
  - ❑ Subjects your assignment to a suite of tests, and
  - ❑ Puts your score alongside comments into Canvas. In most cases, you will see your scores on the programming assignments in about 45-60 seconds
- ❑ Expect code-review interviews for 2-3 of the programming assignments



# ASSIGNMENTS



# Assignments: What to expect

- There will be no busy work
  - ▣ No GUI
- Complexity will not be through obfuscation
- You will be able to look back and feel good about them
  - ▣ Delayed gratification



# Assignments: Other aspects

- You will have about **4 weeks** to complete each assignment
- The assignments will include **milestones** that should be achieved for each week



# Term project

- The term project is a group effort
  - ▣ CS455 students can only be in a team with CS455 students
    - Similarly, CS555 students can only be in a team with other CS555 students
  - ▣ Team size = 2-3 for CS455 students and you can choose your teammate
    - CS555 students: Team size is 2
  - ▣ Please **respond** to your teammate's e-mails on time!
    - Make sure teammate has the e-mail that you check regularly
  - ▣ If you have problems finding a teammate, please let us know via e-mail



# Assignments: Logistics

- Programming assignments will be due on **Wednesdays at 8:00 pm**
- You are allowed to submit up to 2 days late
  - ▣ There is a **7.5%** deduction for each day that you are late
- All programming assignments are to be done **solo**
  - ▣ Java version 11; Gradle version 8.3





# WHAT IT TAKES TO SUCCEED



# What it takes to succeed

[1 / 3]

- You should plan to work **~12 hours** per-week outside of class
  - ▣ Coding and reviewing material from class
- If you miss a lecture
  - ▣ Add about 3 hours per missed lecture



# What it takes to succeed

[2/3]

- Work on the assignments **every day**
  - ▣ There is no such thing as waiting for inspiration to strike!
- **Reflect** about how you could have designed things differently for better performance
  - ▣ Even after you have submitted an assignment
  - ▣ It will improve the choices you make in the next assignment



# What it takes to succeed

[3/3]

- Work in bigger-sized chunks
  - ▣ Too many short bursts = Too many context switches
    - You will be busy doing nothing
- Document your code



# Other pitfalls

- Poor management of **course loads**
  - ▣ Plan the number and type of courses you take
  - ▣ Don't spread yourself so thin that you do not give yourself the opportunity to succeed
- Not attacking the problem and working on the fringes
  - ▣ Spend your time wisely on critical paths



# Interactions

- You can have discussions with me, the GTAs, and your peers
- There are **two constraints** to these discussions
  - ① No code can be exchanged under any circumstances
  - ② No one takes over someone else's keyboard
- Bumps are to be expected along the way
  - ▣ But you should get over this yourself
  - ▣ It will help you with the next problem you encounter



# TOPICS COVERED IN CSx55



# Topics covered in CSx55

[1 / 2]

- Threads: Safety and Concurrency [CS455 and CS555]
- Time & Logical Clocks [CS555]
- Architectures & Topology [CS455 and CS555]
- Peer-to-Peer Systems & Distributed Hashtables [CS455 and CS555]
- Replication, Consistency and Coherence [CS555]
- Programming models for Cloud Computing: MapReduce [CS455 and CS555]



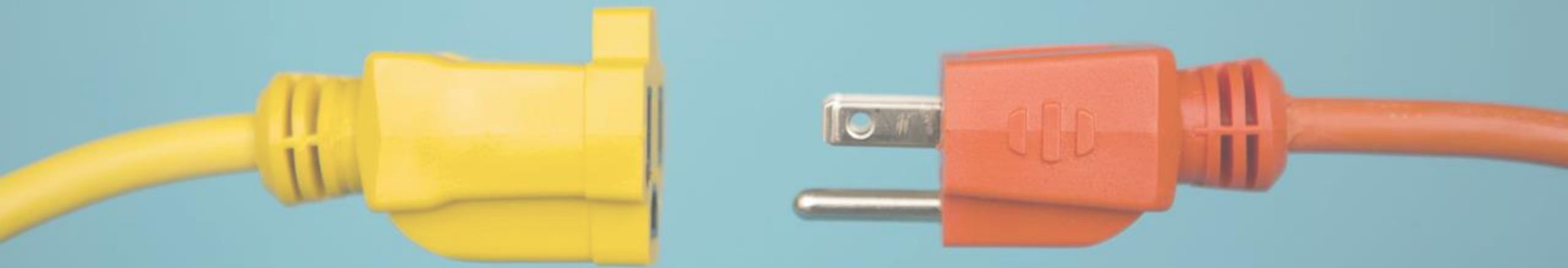


# Topics covered in CSx55

[2/2]

- Extreme Scale Distributed Storage Systems [CS555]
- Spark [CS455 and CS555]
- Spark Streaming [CS555]
- Distributed mutual exclusion [CS455 and CS555]
- Distributed election algorithms [CS455 and CS555]
- Google TensorFlow [CS555]





# SOCKET BASICS

# Example:

## Setting up connections to a server

- Programs open a **socket** to a server that's *listening* for connections
- To create a `Socket` you need to know the Internet host you want to connect to
- Servers don't know *who* will contact them
  - ▣ If it did, difficult to synchronize *when* this would happen



# An analogy

- Server is like a person sitting by the phone
  - ▣ Doesn't know *who* will call and *when*
  - ▣ When the phone rings?
    - Talk to *whoever* is on the other line



# Java provides a `ServerSocket` to enable writing servers

- `ServerSocket` runs on the server
  - ▣ **Listens** for *incoming* network connections on a particular **port** on the host that it runs on
- When a client socket on a remote host attempts to connect to that server port
  - ① Server **wakes** up
  - ② **Negotiates** a connection between the client and server
  - ③ **Opens** a regular `Socket` between the two hosts



# Some more about the two types of sockets

- `ServerSockets` **wait** for connections
- `Client Sockets` **initiate** connections
- Once the `ServerSocket` has set up the connection?
  - ▣ **Data always travels over the regular Socket**



# Using the `ServerSocket`

- Created on a particular **port** using the `ServerSocket(port)` constructor
- Listen for communications on that port using `accept()`
  - ▣ **Blocks until** a client attempts to make connection
  - ▣ Returns a `Socket` object that **connects** the client to the server
- Use the `Socket`'s `getInputStream()` and `getOutputStream()` to communicate



# Creating the ServerSocket

- `ServerSocket serverSocket =  
new ServerSocket(5000);`
  - ▣ Tries to create a server socket on port 5000
- `ServerSocket serverSocket =  
new ServerSocket(5000, 100);`
  - ▣ Can hold up to 100 incoming connections
- `ServerSocket serverSocket =  
new ServerSocket(5000, 100,  
InetAddress.getByName  
("address2.cs.colostate.edu"));`
  - ▣ On a **multi-homed** host, specify the network-address over which connections should be accepted





# Accepting network connections

```
ServerSocket serverSocket =  
    new ServerSocket(portNum);  
while(true) {  
    Socket socket = serverSocket.accept();  
    ...  
}
```



# Closing the client and server sockets

- ❑ Closing a `ServerSocket` **fre**es a port on the host that it runs on
- ❑ Closing a `Socket` **break**s the connection between the local and remote hosts



# We exchange byte streams over the socket

- **The `java.io` package contains the `DataInputStream` and `DataOutputStream` that lets you do this elegantly**
- `DataInputStream din =  
    new DataInputStream(socket.getInputStream());`
- `DataOutputStream dout =  
    new DataOutputStream(socket.getOutputStream());`





# HW1: DISCUSSION

# Why abstractions are important

- Abstraction is the key to managing **complexity**
- Good abstractions turn a difficult task into two manageable ones
  - ① Defining and implementing abstractions
  - ② Using abstractions to solve problem



# Other suggestions that will make life easier:

## A stitch in time ...

[1 / 2]

- **Comment** your code
  - ▣ This is especially true in places where you are performing bitwise manipulations
- **Name** your variables so that you can know what they are anywhere in the code
- Keep functions **short**
- Check for **invariant violations** in your code
- **Test** the functionality of the small pieces



# Other suggestions that will make life easier:

## A stitch in time ...

[2/2]

- Your code should run on CS department machines
- Use a version control system
  - ▣ Git, Subversion, Mercurial, etc.
  - ▣ Free hosted services: Github, Bitbucket, etc.
  - ▣ Commit Often!
  - ▣ Keep your repositories private
- Follow the guidelines from the beginning
  - ▣ build system, directory structure, etc.
- Follow the milestone plan as closely as possible

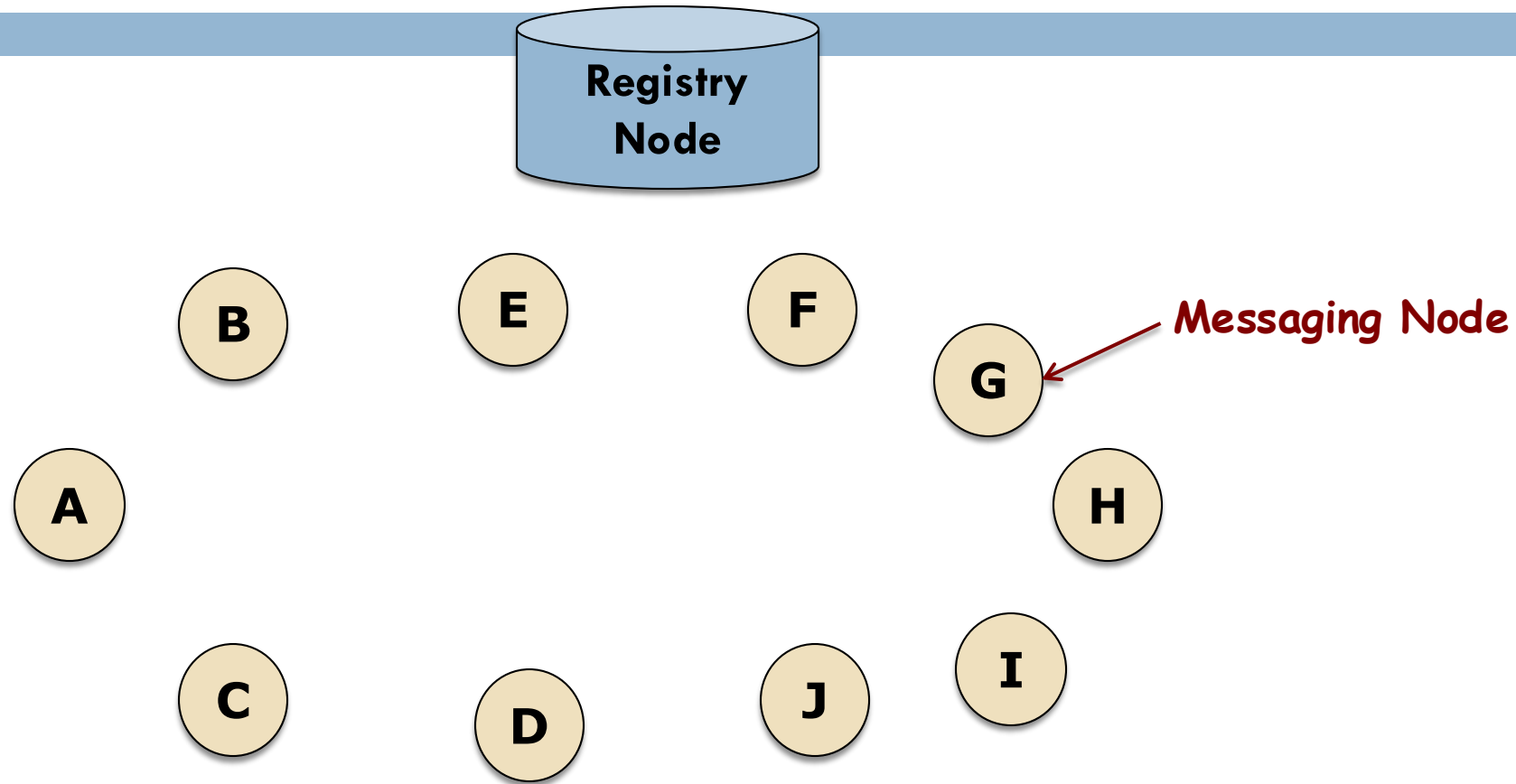


# ASSIGNMENT

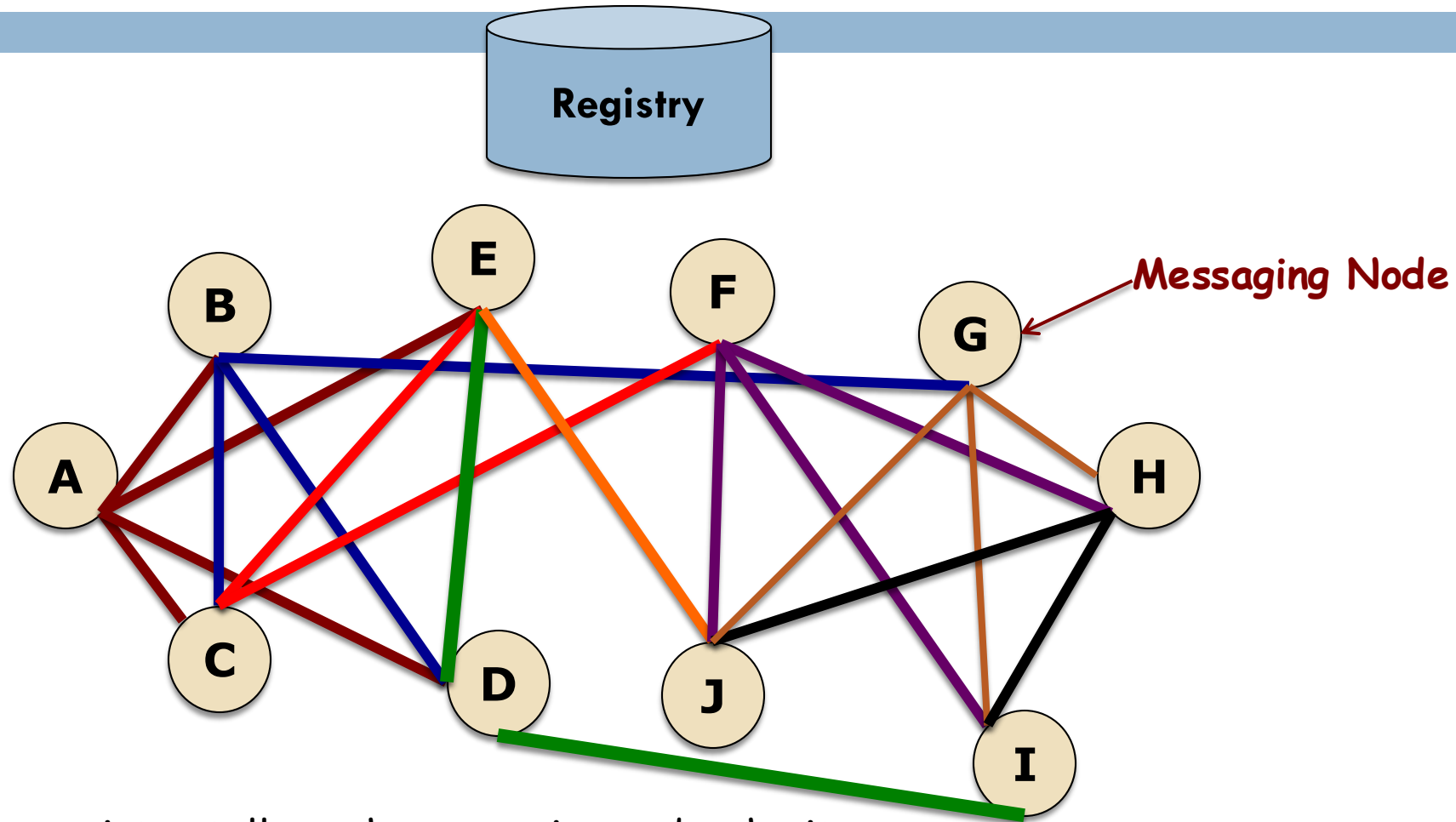




# Composition of the Overlay



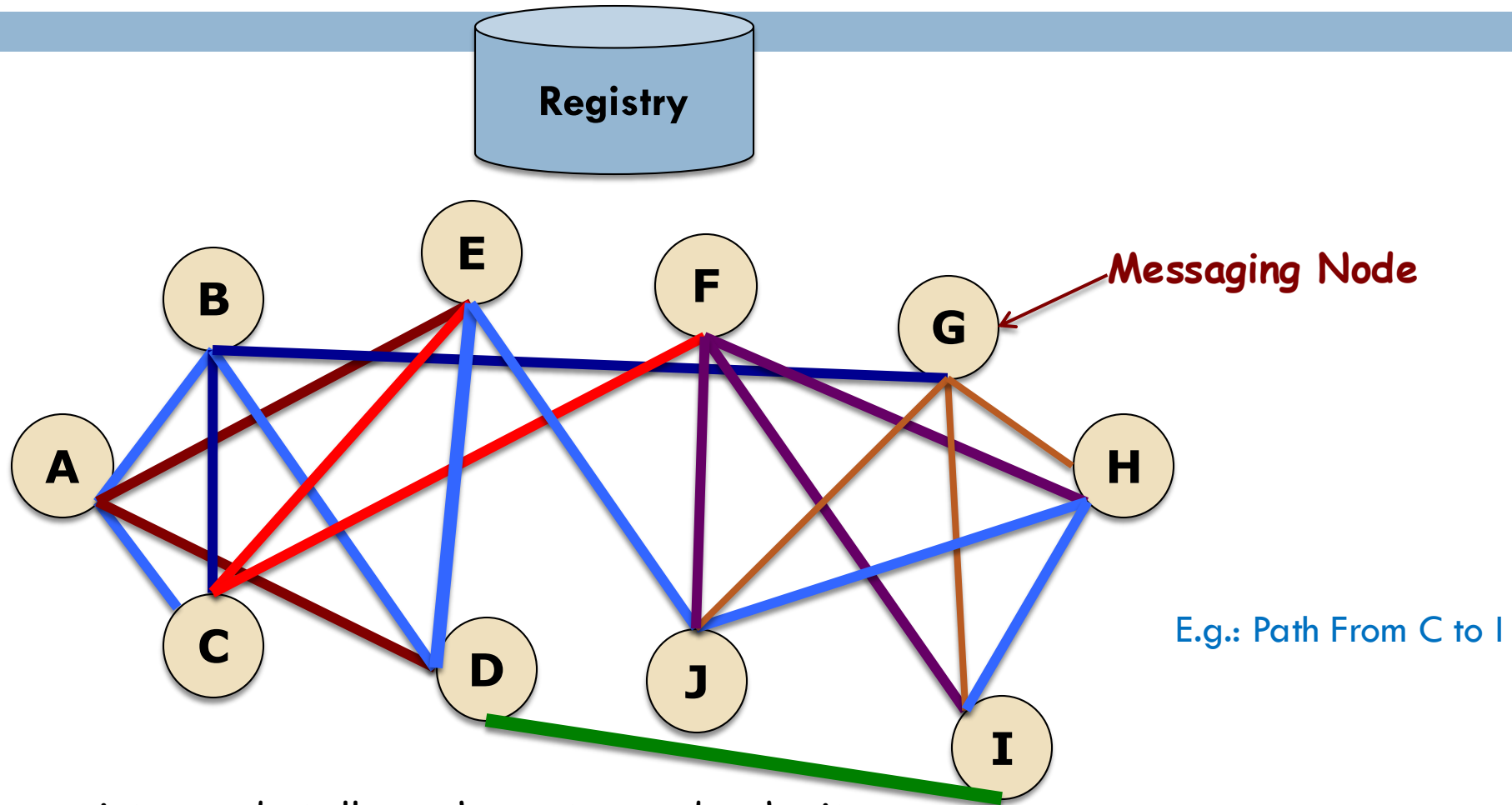
# Overlay topology set up with $C_R = 4$



The registry tells each messaging node who it should connect to via the `Messaging_Nodes_List`



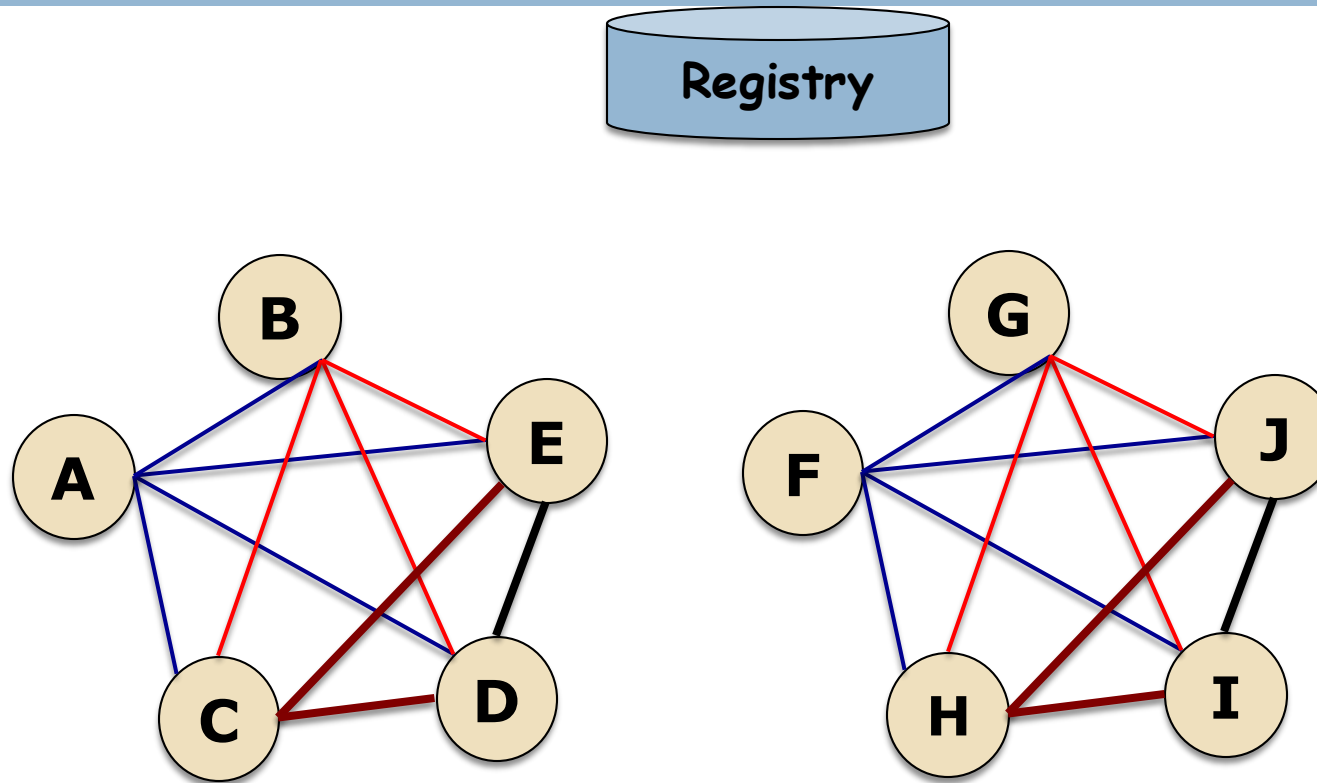
# Overlay topology set up with $C_R = 4$



The registry node tells each router node who it should connect to via the `Peer_Router_List`



# Looking at another overlay topology that could be set up with $C_R = 4$



This topology has a **partition**. The assignment asks you to **prevent this**. Nodes A, B, C, D and E have no way of communicating with F, G, H, I, and J.



# Avoiding network partitions

- Create a linear topology first, and then start making the required number of connections
  - ▣  $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow F \rightarrow G \rightarrow H \rightarrow I \rightarrow J$
  - ▣ Starting off at the point ensures that partitions will not exist



# Other topological aspects

- Necessary and sufficient conditions for a  $k$ -regular graph of order  $n$  to exist?
  - ▣  $n \geq k+1$
  - ▣  $nk$  is even
- During testing we will only specify values where a solution exists



# What is the Minimum Spanning Tree?

- A way of connecting all vertices in a graph using the **smallest possible total edge weight**
- No cycles allowed
  - ▣ Every vertex is connected, but no loops are formed
- Think of it as the cheapest network i.e., it is the leanest way to link everything together
  - ▣ For e.g.: designing computer networks, road systems, or electrical grids



# Does the MST give you the shortest path between two nodes?

- A minimum spanning tree **doesn't guarantee** the shortest route between two nodes
- An MST connects all nodes with the **minimum total edge weight**
- Why not shortest?
  - ▣ An MST may choose a longer, indirect route if it helps keep the overall weight down
  - ▣ For e.g., even if a cheap direct edge exists, the MST might bypass it to minimize the total cost of the tree





# Your programs will be working with two different data representations

- In **memory**: This is where you have your **data structures** such as lists, arrays, hash tables, trees, etc.
- Data that you will sending over the **network**
  - ▣ You do this as a self-contained **sequences of bytes**
  - ▣ Do references or pointers make sense here?
    - No!
    - So, the sequence-of-bytes representation will look VERY different from data structures in memory



# So, we do need some translation between these representations

- Translation from in-memory to network-bound byte sequence
  - ▣ **Marshalling**
    - Also called serialization or encoding
- Translation from network-bound sequence to in-memory representation (i.e., restoration of data structure)
  - ▣ **Unmarshalling**
    - Also called deserialization or decoding



# Marshalling and Unmarshalling

- Marshalling
  - ▣ **Pack** fields into a byte array
- Unmarshalling
  - ▣ **Unpack** byte array and *populate* fields that comprise the wire format message



# Example: Data Structure

[1 / 3]

```
public class WireFormatWidget {  
  
    private int type;  
    private long timestamp;  
    private String identifier;  
    private int tracker;  
  
    ...  
}
```



# Example: Marshalling

[2/3]

```
public byte[] getBytes() throws IOException {
    byte[] marshalledBytes = null;
    ByteArrayOutputStream baOutputStream = new ByteArrayOutputStream();
    DataOutputStream dout =
        new DataOutputStream(new BufferedOutputStream(baOutputStream));

    dout.writeInt(type);
    dout.writeLong(timestamp);

    byte[] identifierBytes = identifier.getBytes();
    int elementLength = identifierBytes.length;
    dout.writeInt(elementLength);
    dout.write(identifierBytes);

    dout.writeInt(tracker);

    dout.flush();
    marshalledBytes = baOutputStream.toByteArray();

    baOutputStream.close();
    dout.close();
    return marshalledBytes;
}
```



# Example: Unmarshalling

[3/3]

```
public WireFormatWidget(byte[] marshalledBytes) throws IOException {  
    ByteArrayInputStream baInputStream =  
        new ByteArrayInputStream(marshalledBytes);  
    DataInputStream din =  
        new DataInputStream(new BufferedInputStream(baInputStream));  
  
    type = din.readInt();  
    timestamp = din.readLong();  
  
    int identifierLength = din.readInt();  
    byte[] identifierBytes = new byte[identifierLength];  
    din.readFully(identifierBytes);  
  
    identifier = new String(identifierBytes);  
  
    tracker = din.readInt();  
  
    baInputStream.close();  
    din.close();  
}
```



# How to send data

```
public class TCPSender {  
  
    private Socket socket;  
    private DataOutputStream dout;  
  
    public TCPSender(Socket socket) throws IOException {  
        this.socket = socket;  
        dout = new DataOutputStream(socket.getOutputStream());  
    }  
  
    public void sendData(byte[] dataToSend) throws IOException {  
        int dataLength = dataToSend.length;  
        dout.writeInt(dataLength);  
        dout.write(dataToSend, 0, dataLength);  
        dout.flush();  
    }  
}
```



# How to receive data

[1 / 2]

```
public class TCPReceiver implements Runnable {  
  
    private Socket socket;  
    private DataInputStream din;  
  
    public TCPReceiver(Socket socket) throws IOException {  
        this.socket = socket;  
        din = new DataInputStream(socket.getInputStream());  
    }  
  
    public void run() {  
        ...  
    }  
}
```





# How to receive data

[2/2]

```
public void run() {  
  
    int dataLength;  
    while (socket != null) {  
        try {  
            dataLength = din.readInt();  
  
            byte[] data = new byte[dataLength];  
            din.readFully(data, 0, dataLength);  
  
        } catch (SocketException se) {  
            System.out.println(se.getMessage());  
            break;  
        } catch (IOException ioe) {  
            System.out.println(ioe.getMessage());  
            break;  
        }  
    }  
}
```



# A simple breakdown of classes

[1 / 5]

- csx55.overlay.wireformats
  - Protocol
  - Event [This is an interface with the getType() and getBytes() defined]
  - EventFactory [Singleton instance]
  - Register
  - Deregister
  - MessagingNodesList
  - LinkWeights
  - TaskInitiate
  - Message
  - TaskComplete
  - TaskSummaryRequest
  - TaskSummaryResponse



# A simple breakdown of classes

[2/5]

- csx55.overlay.spanning
  - ▣ MinimumSpanningTree
  - ▣ RoutingCache



# A simple breakdown of classes

[3/5]

- csx55.overlay.util
  - ▣ OverlayCreator
  - ▣ StatisticsCollectorAndDisplay



# A simple breakdown of classes

[4/5]

- csx55.overlay.transport
  - ▣ TCPServerThread
  - ▣ TCPSender
  - ▣ TCPReceiverThread



# A simple breakdown of classes

[5/5]

- csx55.overlay.node

- Node [Interface with the onEvent(Event) method]

- Registry

- MessagingNode

- Registry and MessagingNode should both implement the Node interface with the onEvent(Event) method in it



# The contents of this slide-set are based on the following references

- *Computer Networks: A Systems Approach*. Larry Peterson and Bruce Davie. 4th edition. Morgan Kaufmann. ISBN: 978-0-12-370548-8. [Chapter 1, 2]
- *Java Network Programming, Third Edition*. Elliotte Rusty Harold. O'Reilly. ISBN-10: 0596007213 / 978-0596007218. [Chapter 7]

