Chapter 4 – Part 2

Big Data Computing Technology Platforms & Cloud Security - *On Cloud Data Security & Storage*

Contents:

- Security & Privacy in Cloud Platforms
- Data Security – Components and issues
- Data Integrity, Confidentiality, Availability, Privacy
- Commonly used data encryption algorithms in Clouds
- Distributed Data storage – Basic solution

Service Models in Cloud Platforms

Three well-known and commonly used service models in the cloud paradigm are:

- Software as a service (SaaS) - In SaaS, *software with the related data* is deployed by a Cloud Service Provider (CSP), and users can use it through the web browsers;

- Platform as a service (PaaS) - In PaaS, a CSP facilitates services to the users by providing certain cloud components to certain software that can solve the specific tasks; Example - Programmers can create specific applications for a specific platform using proprietary APIs and make those applications available to any user of that platform;

Service models

- Infrastructure as a service (IaaS) - In IaaS, the CSP facilitates services to the users with *virtual machines and storage* to improve their business capabilities;

Major issues in the cloud computing, include:

- Resource security
- Resource management
- Resource monitoring

Cloud Categories - Public / Private / Hybrid Clouds

- Public clouds - Common way of deploying cloud resources (servers and storage) to users; Owned and operated by a third-party Cloud Service Provider (CSP) and delivered over the Internet

- Advantages of Public Clouds:
  - Low-Cost & Scalability (*pay-as-you-use*)
  - High reliability (*one of the QoS parameters*)
  - No maintenance on User's side

Public / Private / Hybrid Clouds

Private cloud characteristics:

- Computing and Storage resources used exclusively by one specific business / organization;
- They can be physically located at your organization's on-site datacenter, or it can be hosted by a third-party service provider;
- All equipment and resource deployment entirely depend on their local policies and norms;

Advantages of Private Clouds:

- Scalability (but low-cost cannot be assured)
- Highly secured storage and access
- Custom-driven local environments can be created as per the needs of the organization

Public / Private / Hybrid Clouds

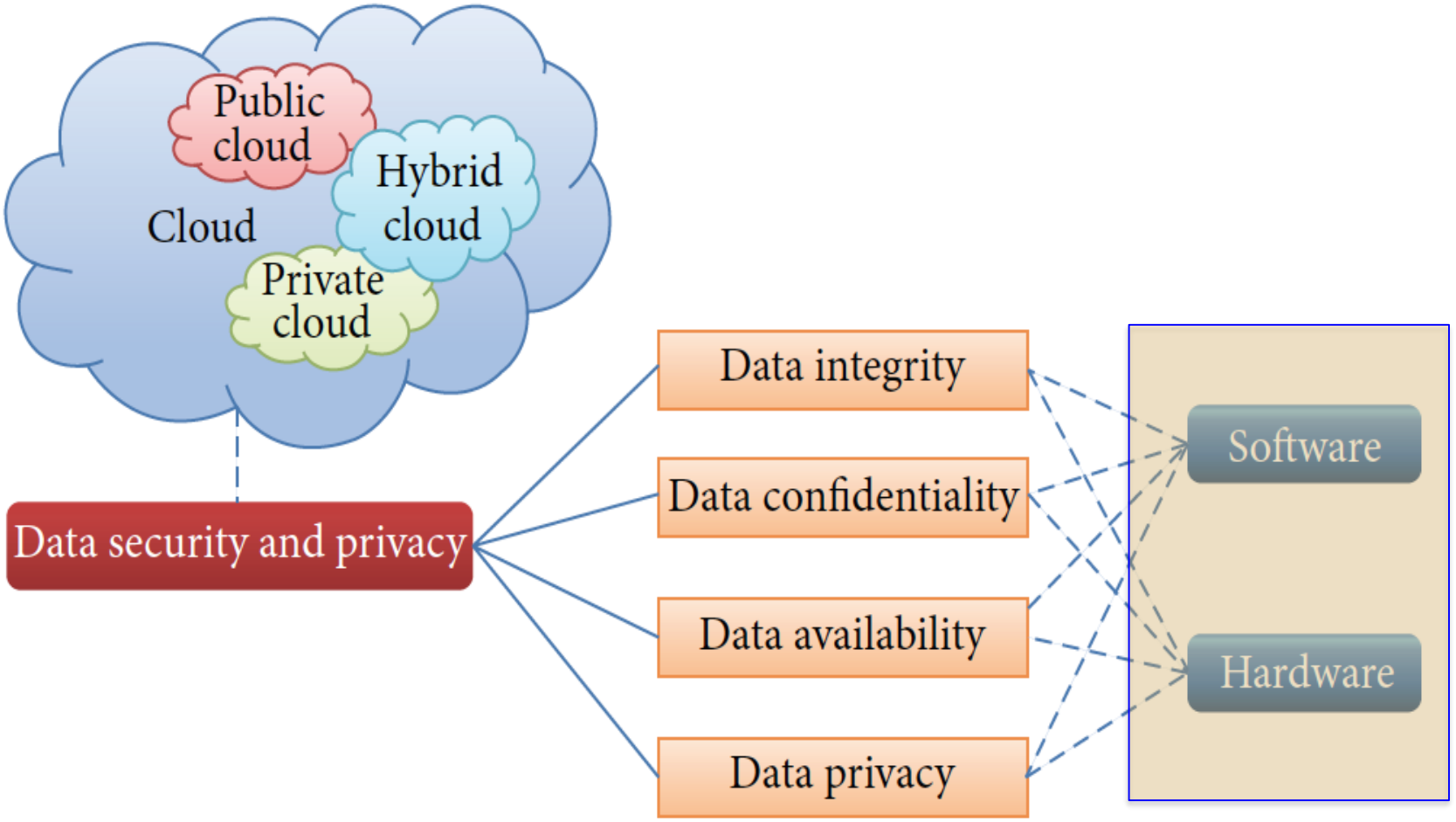
- **Hybrid** clouds - Combination of Public + Private - Attempt to take advantage of secured on premise infrastructure and at the same time trying to use public cloud services;

- Data and applications can migrate between Private and Public clouds for greater flexibility and more deployment options.

- Use of Hybrid Cloud:

- Within a Hybrid Cloud, we can use Public cloud features for high-volume data handling;
- Lower-security needs - web-based email, web-page hosting, etc, whereas, the Private cloud can be used for sensitive, mission-critical operations.

Security & Privacy in Cloud Platforms



Security & Privacy in Cloud Platforms

Security, pertaining to handling (storage, transportation, giving access, etc) data on cloud, is the combination of:

- Data Integrity
- Data Confidentiality
- Data Availability
- Data Privacy

Remember like this - iCAP!

All the above collectively attempt to address the following:

- Prevention of the unauthorized disclosure of information
- Prevention of unauthorized withholding of information
- Prevention of the unauthorized amendment or deletion of information

(A) Data integrity (DI)

- One of the most critical elements in any information system. *Data integrity (DI) is the basis* to provide cloud computing service such as SaaS, PaaS, and IaaS;
- *What does DI mean?* DI means protecting data from *unauthorized deletion, modification, or fabrication*;
- Managing entity's admittance and rights to specific enterprise resources *ensures that valuable data and services are not abused, misappropriated, or stolen*

(c) Bharadwaj V Dept of ECE NUS (2021)

Data integrity (DI)… Cont'd

- Data storage - Databases - DI is easily achieved in a standalone system with a single database;

How?

DI in the standalone system is maintained via database constraints and *transactions*, which is usually taken care by a database management system (DBMS). Transactions should follow *ACID (atomicity, consistency, isolation, and durability)* properties to ensure DI.

- Most DBs follow ACID model – DWHs/DataLakes

Data integrity (DI)… Cont'd

- *ACID - Atomicity, Consistency, Isolation, & Durability*

***Atomicity** – All operations are treated as atomic/single and if a failure in transaction happens it needs to start all over again or rollback to an earlier saved state!*

***Consistency** - Date needs to be consistent when a transaction is enabled w.r.t any other constraints or rules set within the database systems to keep data in a consistent state; For multiple copy scenario (replication and storing), consistency mechanism must enforce the rules across all nodes storing the data;*

Data integrity (DI)… Cont'd

- *ACID - Atomicity, Consistency, Isolation, & Durability*

Isolation *– In a multi-user scenario or when data is stored in a fragmented way, access to data can happen concurrently; This is decided by the CSP if it allows such concurrent access. If allowed, then isolation is defined as a characteristic of transactions to perform the operations in an isolated way without affecting the other nodes.*

Durability *– Guarantees that data is saved safely after a transaction amidst failures while updating; How? Data is locked until a transaction is completed; Results are first written into local transaction logs and once the work is done they are written as an actual database entry*

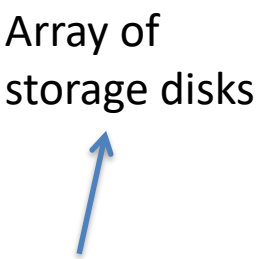
Data integrity (DI)… Cont'd

- *ACID-based transactions are one of the most domineering properties of DWH & DataLakes!*
- *When using a DataLake, if ACID–based transactions are implemented, then it allows users to see consistent views of their data even while new data is being modified in real-time, because each write is handled as an isolated transaction that is recorded in an ordered transaction log which can be used for analysis later!*

Data integrity (DI)… Cont'd

- *With ACID-based transaction in place you can trust your stored data!*
- *So, a DE making use of DataLake/DWH to perform ETL on his/her data can always be sure of data he/she is using!*

Data integrity (DI)… Cont'd



Q: How to achieve DI?

- *DI can be obtained by techniques such as RAID-like strategies and digital signature (Out of scope of this course!)*

Q: How to ensure data integrity as a Cloud envt. has several access points?

- *By avoiding the unauthorized access, organizations can achieve greater confidence in data integrity.*
- *Monitoring mechanisms offer the greater visibility into determining who or what may have altered data or system information, potentially affecting their integrity.*

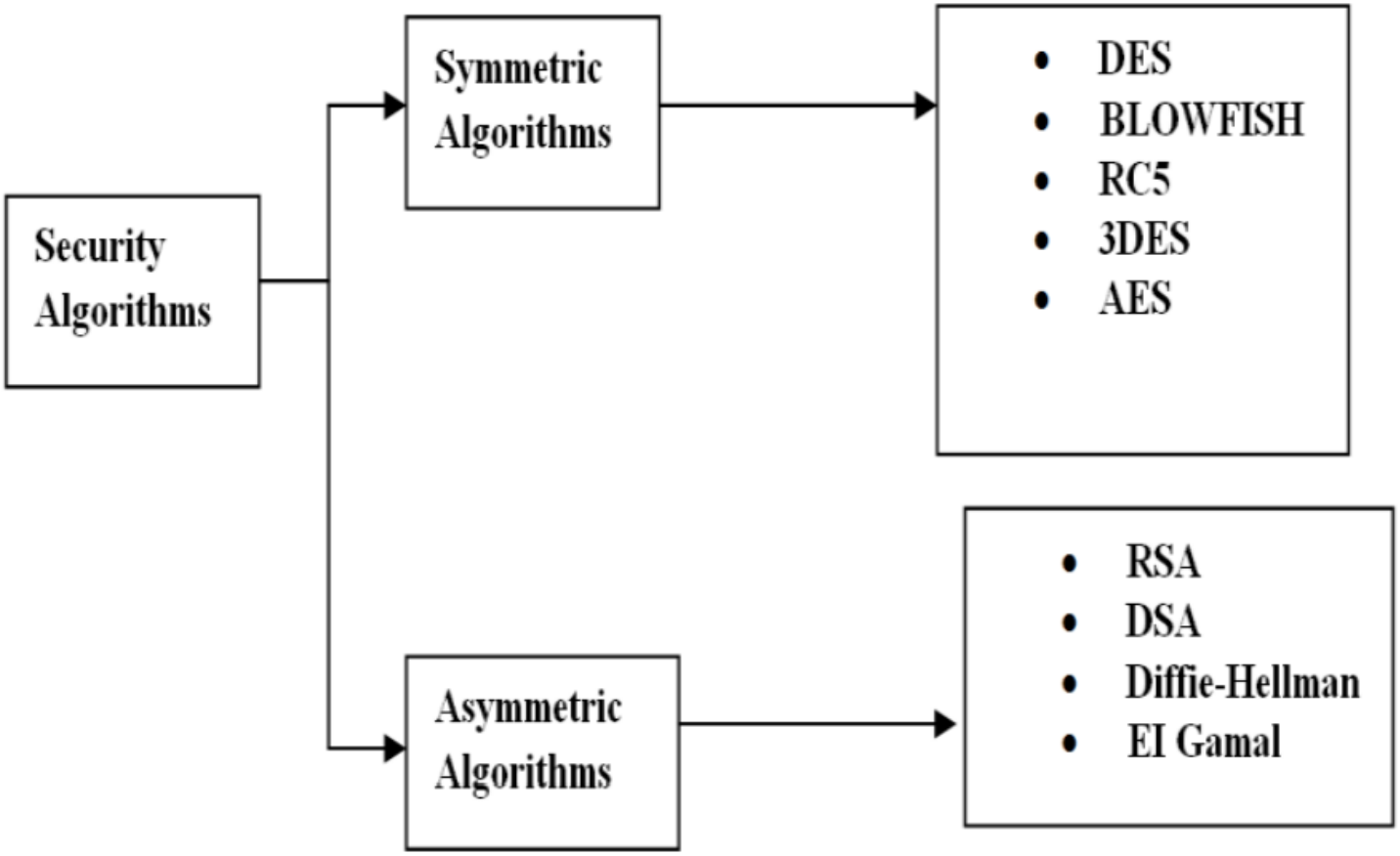
(B) Data Confidentiality (DC)

- Facilitates storing users' private or confidential data in the cloud. *Authentication & access control strategies are used to ensure data confidentiality*.
- No trust on CSPs; Can't store sensitive data; Simple encryption methods are not sufficient; Insider attacks may be overwhelming!
- In a cloud setting, a CSP could store the data based on service subscription levels – *Users can pay to store their data in a more trustworthy way!*

(B) Data Confidentiality (DC) ….cont'd

- Distributed storage - To ensure the data integrity, one option could be to store data in multiple clouds or cloud databases; *(AWS does this by replicating your data)*
- The data to be protected from internal or external unauthorized access are divided into chunks and each chunk is then encrypted and stored in separate databases (uses a concept of data distribution over cloud). Because each segment of data is encrypted and separately distributed in databases over cloud, this provides enhanced security against different types of attacks;
- T-Coloring algorithm is one such example – we will see later on some details.

Data encryption/security algorithms



Dealing with security algorithms is out of scope of this module; Anyone interested, contact me for additional materials.

Commonly used security algorithms for data handling in Cloud Platforms

1. **AES** (Symmetric block cipher; no known attack against AES exists; Preferred encryption standard for governments, banks and high security systems around the world)
2. **RSA** (Asymmetric encryption type; Commonly used for securing communications between web browsers and eCommerce sites, several ecommerce applications, etc)
3. **Twofish** (Belongs to AES category; speed, flexibility, and conservative design; fastest across all CPUs; Fits on smart cards, can operate even with a couple of registers, a few bytes of RAM, and little ROM, fits in hardware in few gates)
4. **Blowfish** (Symmetric block cipher algo; Products that use blowfish - https://www.schneier.com/academic/blowfish/products.html)
5. **Triple DES** (earlier versions of Microsoft Office, Firefox and EMV payment systems; no longer in use and replaced with AES and its variants)

(c) Bharadwaj V Dept of ECE NUS (2021)

(C) Data Availability(DA)

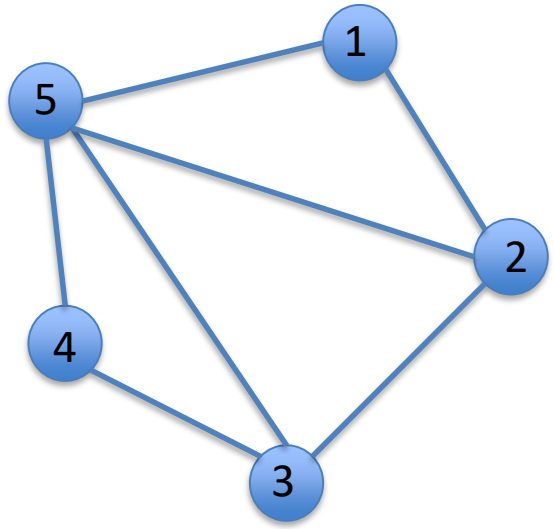
- Data availability means: When accidents such as hard disk damage, IDC fire, and network failures occur, the extent that user's data can be used or recovered and how the users verify their data by techniques rather than depending on the credit guarantee by the cloud service provider alone.

- Quickly & efficiently locating data can help users to increase their trust on the Cloud services.

- Algorithms such as T-coloring and its variants are used to ensure – data confidentiality & availability; Refer to the next slide on T-coloring application for data storage;

IDC – Initiating Device Circuit

T-coloring problem – Fragmenting and replicating data for high availability



Given a graph G=<V,E>, color the vertices of G using colors in such a way that no two adjacent nodes have identical colors.

Possible solution: (Note: not a unique soln)

(1,R), (2,G), (3,R), (4,G), (5,B)

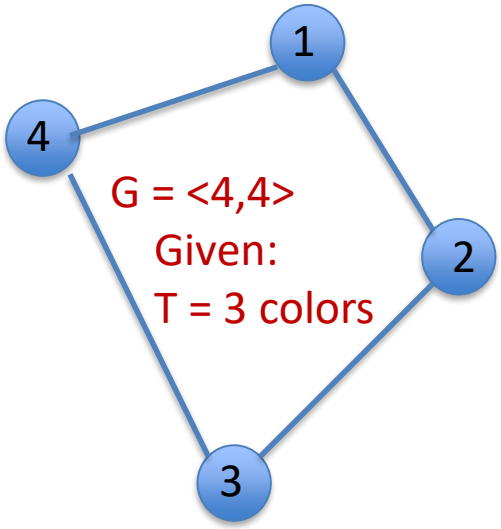
- Minimum # of colors needed: 3
- Chromatic number: 3

- Graph Coloring decision problem: Given a graph G=<V,E> and a chromatic number k, can we color the graph using k colors?

- Graph Coloring optimization problem: Given a graph G=<V,E>, what is the minimum number of colors needed to color the graph?

Backtracking

T-Coloring Optimization Problem



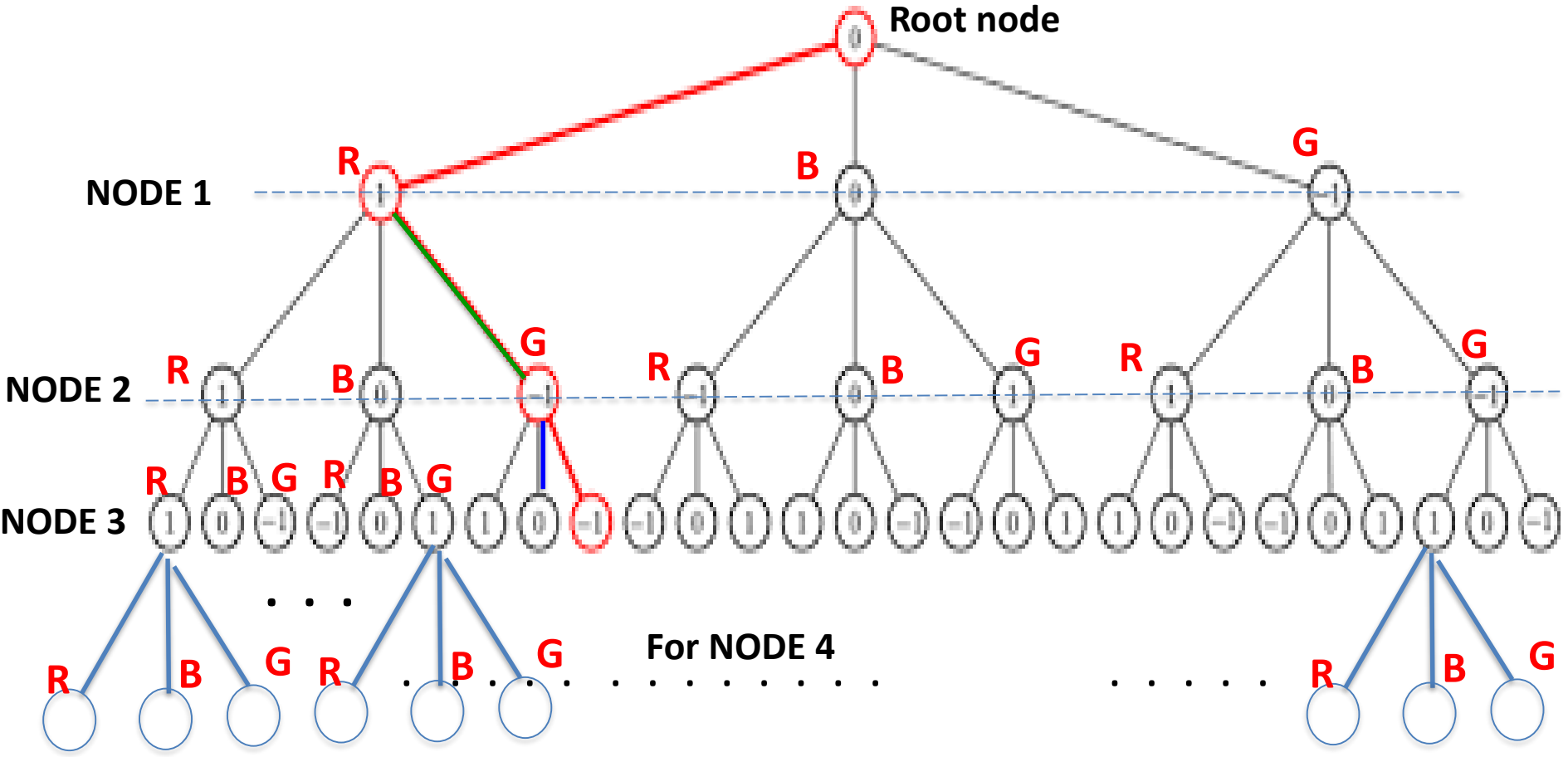
Solution Approach:

We can generate all possible coloring of nodes, referred to as state-space (*brute-force way!)* without checking adjacency color criteria. Let us get a feel of the *computational time complexity of this brute-force approach*!

Backtracking method – Does not reduce the complexity, but avoids certain coloring possibilities by the use of a bounding function.

Note: Python code available via Github repository

Brute-force approach – Exhaustive search!

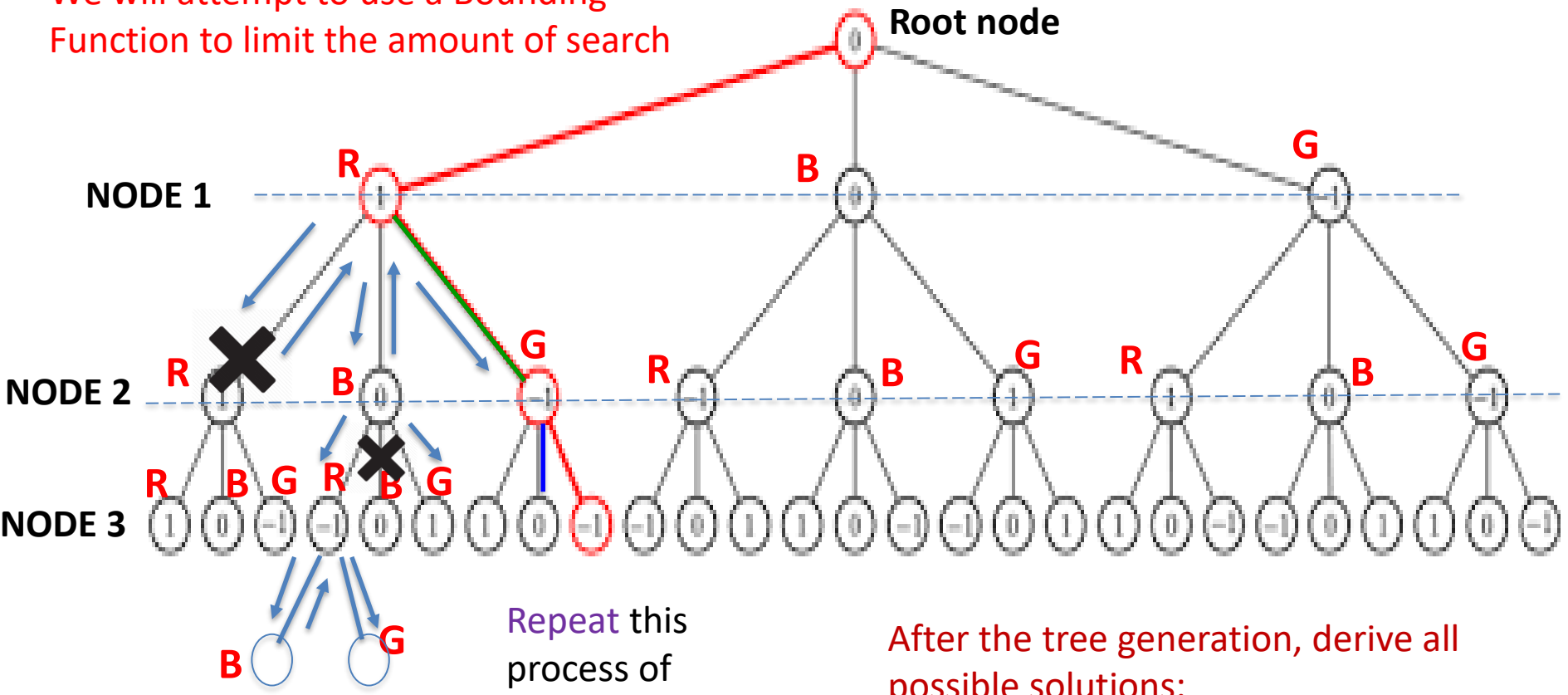


# of nodes generated for four nodes:

$(1 + 3 + 3^2 + 3^3 + 3^4) = (3^{(4+1)} - 1) / (3 - 1) = (3^{4+1} - 1) / 2$ **~= 3^{4+1}** = C^{n+1}

Bactracking approach

We will attempt to use a Bounding Function to limit the amount of search



Repeat this process of backtracking and bounding

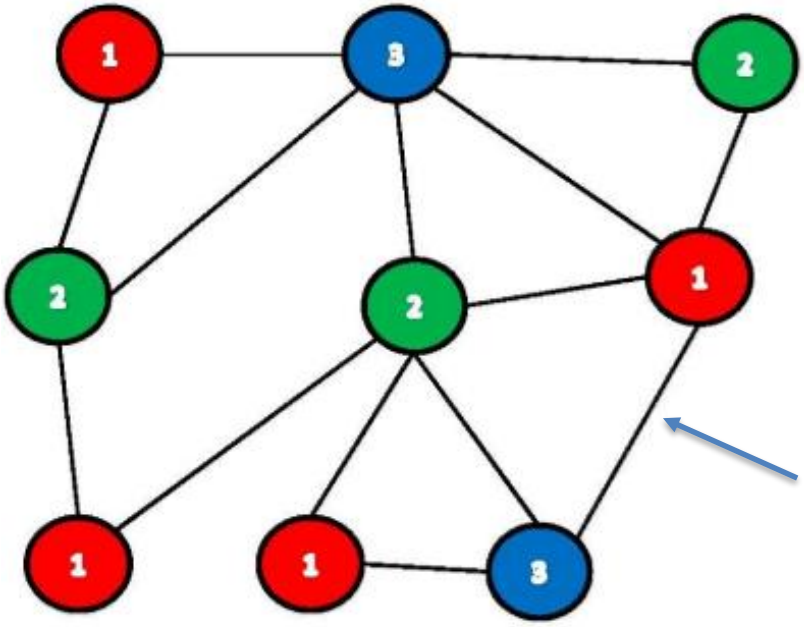
After the tree generation, derive all possible solutions:

1) R B R B
2) R B R G
3) R B G R

• • •

By applying a bounding function, search space is minimized; **In the worst case** the time taken will be: C^n

(C) Data Availability(DA)….cont'd



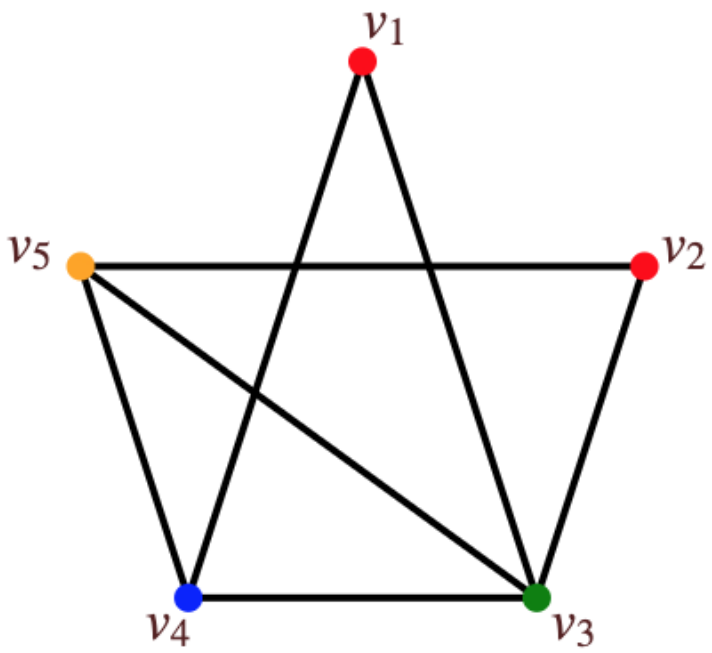
Data 1 – 8 Terabytes (4 chunks)
Data 2 – 7 Terabytes (3 chunks)
Data 3 – 3 Terabytes (2 chunks)

Challenges: Given such a distributed setting and the number of chunks for a data:
(i) What is the minimum number of colors needed to store the data? (*m-color optimization problem*)
(ii) If the network is fairly large, we also need to replicate each chunk for guaranteeing high availability of data for a user;

Note: All t-coloring algorithms generate solution addressing (i); No optimal solution exists (NP-hard to determine the chromatic number) and hence all the algorithms are heuristic;

Remark: By and large, size of a chunk (block) is kept constant

(C) Data Availability(DA) ….cont'd



Example:

In this network, chromatic color is 4.
Nodes: (1,2), 5, 4, 3

Q: Is it possible to reduce further the chromatic color ?

Q: Suppose node 1 is the access point for this network. An attacker disrupts the network and attacks node 3. How many paths exist to access data on nodes 2, 4, and 5?

Note that if a node is attacked, we cannot use any direct links from a node that leads to that attacked node.

(D) Data Privacy (DP)

- **Privacy** is the *ability of an individual or group to seclude themselves or information about themselves* and thereby reveal them selectively;

- In the cloud, the privacy means when users visit their sensitive data, the cloud services can prevent potential adversary from inferring the user's behavior by the user's visit model;

- Researchers have focused on *Oblivious RAM (ORAM) technology*. ORAM technology visits several copies of data to hide the real visiting aims of users. *ORAM has been widely used in software protection and has been used in protecting the privacy in the cloud as a promising technology.*

Concluding remarks on Cloud Security

- *Modern day Cloud Service Providers like AWS deliver security via a number of plug-and-play service components*

- *To enhance trust on CSPs data replication is attempted*

- *CSPs use different service options to users:*
  Cost effective solutions: Platinum / Gold / Silver…
  Data Availability & Confidentiality are assured

Thank you!