T-401-ICYB
Summary

Stephan Schiffel
stephans@ru.is
Reykjavik University, Iceland

11.12.2025

Outline

Intro

The Attack Surface

"The sum of all potential vulnerabilities in a system where an attacker could try to subvert the intended purpose of the system and organisation or person who is using it."

The Attack Surface (2)

- Email - can contain viruses, malware, links to bad sites
- All network points of access (Wired, Wifi, Bluetooth, ...)
- USB, CDROM, Hard drives, (Floppy drives)
- Downloaded viruses and malware
  - May be embedded in legitimate documents or software
  - Free games, personality tests, ...
- SMS messages
- Software Distributions, External Software, Bios, Chips, ...
- etc.

i.e. any form of input or control over software or machine

Defense in Depth

Defense as Risk Management

$$Risk = Threat \times Vulnerability \times Cost$$

Strategies for Handling Risk:

- **Mitigate/Reduce:** Reduce likelihood or impact to an acceptable level.
- **Avoid:** Discontinue the risky activity (e.g., "We will not store credit card numbers").
- **Accept:** The cost of the fix > impact. Management signs off.
- **Transfer:** Move risk to a 3rd party (Cyber Insurance, Cloud Provider).

Secure Design Principles

- Least Privilege
- Fail-Safe Defaults
- Economy of Mechanism
- Complete Mediation
- Open Design
- Psychological Acceptability

The Layered Approach

No single control is infallible. If one layer fails, the next must catch the threat. For example:

1. **Physical:** Locks, cameras, guards.
2. **Technical:**
   - **Perimeter/Network:** Firewalls, DMZ, VPN, Intrusion Detection
   - **Host/Endpoint:** Antivirus, Monitoring
   - **Application:** Input validation, secure code.
   - **Data:** Authorization, Encryption (at rest/transit), Hashing, Backups.
3. **Administrative:**
   - **People:** MFA, Password policies, Training
   - **Technology:** Patch Management, Risk Assessment
   - **Operations:** Principle of least privilege

Computer Networks

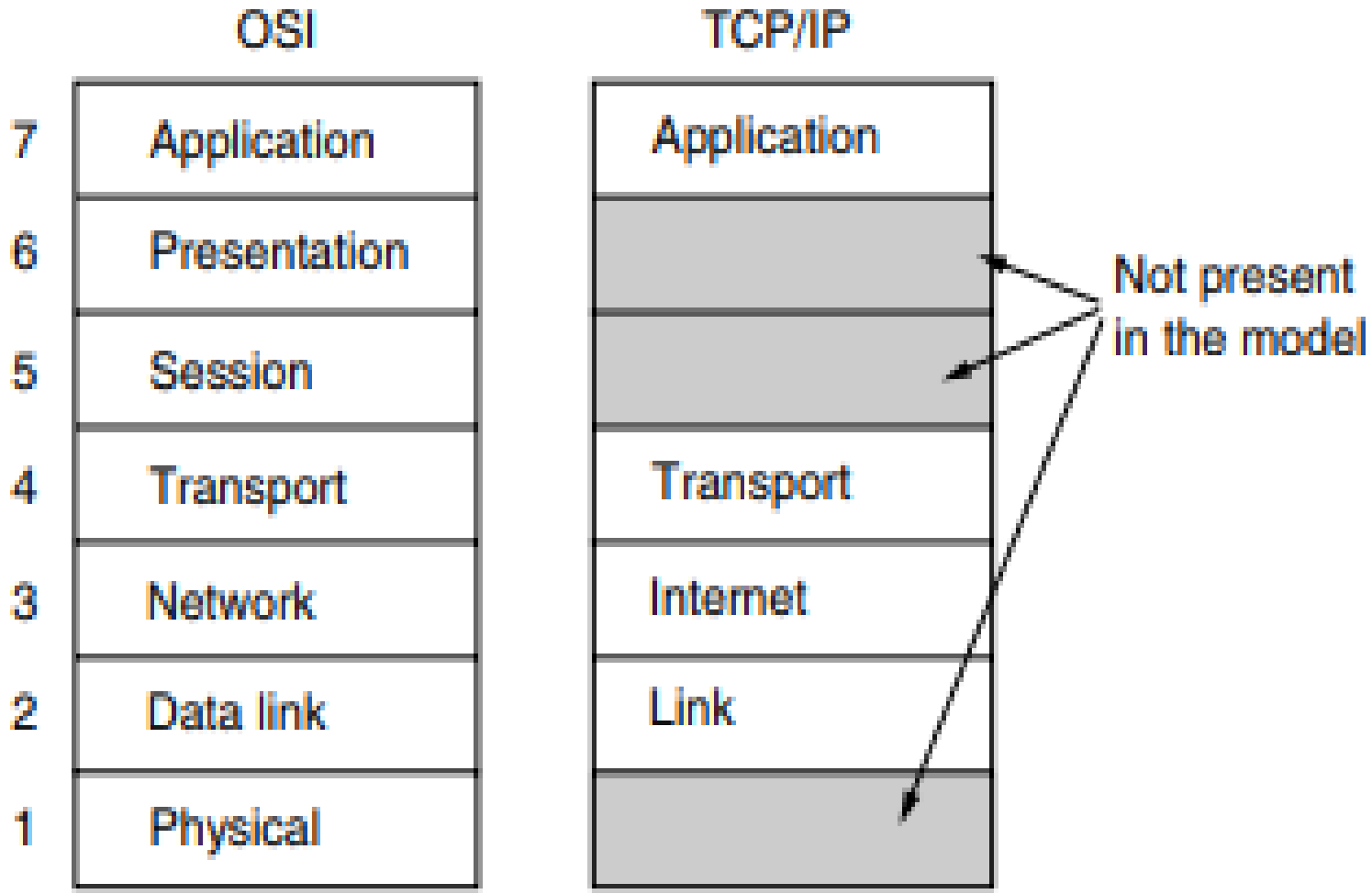
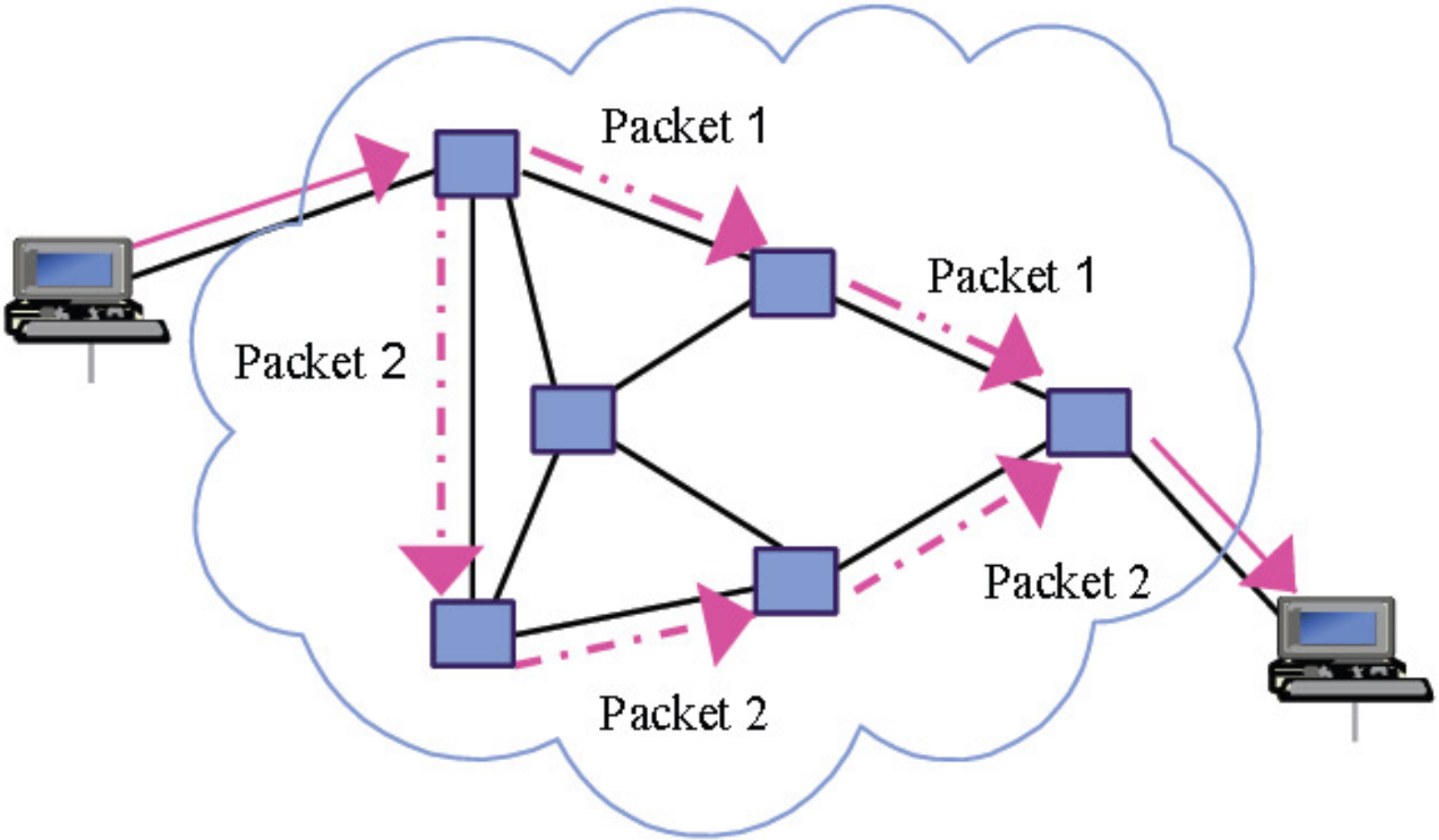


Figure 1-21. The TCP/IP reference model.

How is Data Delivered? - Packet Switching



Packets may take any path through the network, reassembled by receiver

Identity & Addressing

Problem: How do we identify a device among billions?

- **LAN: MAC Address (Physical)** (e.g., `00:1A:2B:3C:4D:5E`)
  - Hardcoded on the Network Interface Card (NIC).
- **Internet: IP Address (Logical)** (e.g., `192.168.1.15`)
  - Routable across the global internet (unless not).
- **Application Layer: Domain Names** (e.g., `canvas.ru.is`)
  - Human readable
  - Mapped to IP addresses using **DNS** (Domain Name System)

Potential Attacks

- **Spoofing**: Faking the source address (MAC or IP address in packets)
- Attacks on the mappings between addresses: ARP Poisoning, DNS cache poisoning, ...

Finding the Path

Problem: How does a packet know which wire leads to the destination?

- **Layer 2 (LAN): MAC address table / Switching table** in a switch, **Broadcast**, **Default Gateway**
- **Layer 3 (Internet):** Routers have **Routing Tables** that contain the next hop for each possible destination. Routing protocols (e.g., BGP) are used to compute routing tables.

Potential Attacks

- Attacks on the tables or devices, e.g., overloading the switch can lead to fallback to broadcast
- Attacks on the protocols that make the table (e.g., **BGP hijacking**)

Physical Layer

- Transmitting and receiving bits over a physical medium
- Synchronization between sender and receiver clocks
- Encoding & Signaling: defining how bits are represented (e.g., +5V vs 0V, Light Pulses, RF Waves)

Security Context: Physical Security

- **Wiretapping**: Copper emits electromagnetic usage; it can be tapped without cutting the wire.
- **Jamming**: Denial of Service against the physical medium (common in Wireless).
- **Rogue Devices**: Plugging an attackers device directly into a wall port circumvents the Firewall.
- **Rule #1**: If an attacker can physically touch the device/network, no software protocol can save it.

Link Layer

Node-to-Node delivery on the same network/medium

- **Ethernet (IEEE 802.3)**:
  - The standard for wired LANs.
  - In practice often Point-to-point (switches), but technically uses a shared medium (the wire).
  - Mechanism: Random access to medium with collision detection (CSMA/CD).
- **Wi-Fi (IEEE 802.11)**:
  - Uses a shared medium. Every packet is broadcast to everyone within range.
  - Mechanism: Random access to medium with collision avoidance (CSMA/CA).
- ...

Network Layer

Delivery of packets to devices anywhere in the network.

This requires

- **Addressing:** Each device is assigned a unique IP address
- **Packetization**: Divide data into manageable packets
- **Routing**: Direct packets across the network from source to destination

Important: IP protocol, NAT, Routing

Transport Layer: TCP & UDP

Application Multiplexing (ports)

Security Context:

- Open ports are part of the attack surface
- **TCP SYN Flooding**: For (D)DoS attacks

Application Layer:

Various Protocols:

- HTTP vs HTTPS
- TLS for CIA (confidentiality, integrity, authentication)
- SSH (vs Telnet)
- DNS
- DHCP
- Email: SMTP, IMAP, POP3
- File Sharing: FTP, SFTP, SMB

Network Security

- Firewall (packet filter)
- VPN

Operating Systems / CLI

What is an OS?

- Kernel vs. User Space
- Core Functionality:
  - Abstraction from raw hardware interfaces
  - Arbitration / Resource Management for multiple processes/users
  - Process Management / Scheduling
  - File Systems
  - Protection (kernel vs. user space, user A vs. B, process X vs. Y)
- Trusted Computing Base (TCB)
- Logging (for intrusion detection, forensics)

OS Security

Issues:

- Memory Safety: Buffer Overflows
- Race Conditions (in access control), e.g., TOCTOU
- Command Injection
- **Privilege Escalation**
- Rootkits

Important concepts:

- User (and process) permissions
- Root/Admin vs. User
- SUID on Unix

Virtualization

Types

- Different types:
  - Hypervisor / Virtual Machine
  - Container (e.g., Docker)
  - Emulation
- ... offer different
  - Isolation from each other
  - Separation from the host
  - Performance / cost
- Main security issues:
  - VM/Container escape
  - Attacks on other containers/images on the same host (e.g., side-channel attacks)

Binary Exploitation

Buffer Overflows

- Problem: writing in memory outside the designated buffer for the input
- Effect: execute code not intended by the author
- Goal: Start a shell or execute commands with privileges the attacker does not usually have
- Mitigation means: secure code, better programming languages, compiler and OS features

Web

Attack Surface

Server-Side:

- Visible Inputs (Forms, File Uploads)
- Hidden Inputs (URL Parameters, HTTP Headers, Cookies)
- API endpoints
- Supply Chain

Client-Side:

- HTML, CSS, ... → DOM (Document Object Model)
- Client-Side Storage (Cookies, Access Tokens, ...)
- (Third-Party) Scripts

Key Vulnerabilities

- SQL Injection / SQLi
- Insecure Direct Object References (IDOR)
- Broken Authentication (passwords, session ids, tokens)
- Cross-Site Scripting (XSS)

$\Rightarrow$ OWASP Top-Ten

AI

Attack Surface

Input Surface

- User Prompts
- Uploaded Documents
- Web Search results
- Attack Vectors: **Prompt Injection**, **Jailbreaking**

Model Surface

- Weights & Embeddings
- The neural network file
- Attack Vectors: **Extraction** of (private) training data, **Backdoors**

Agency Surface

- Plugins, API Calls, Code Execution, ...
- Attack Vectors: **Confused Deputy** (tricking the model into doing something it has the ability to)

AI-Enabled Cybercrime

GenAI lowers the barrier to entry for attackers.

- **Polymorphic Malware:**
  - Using LLMs to rewrite malicious code logic dynamically to evade signature-based antivirus (AV) detection.
- **Social Engineering at Scale:**
  - **Phishing:** Perfect grammar, localization, and context-awareness.
  - **Deepfakes:** Audio (Vishing) and Video for CEO fraud.
- **Vulnerability Discovery:**
  - Attackers interpret open-source code using LLMs to find zero-days faster than defenders.

Societal Impact

- Privacy
- Institutional Bias
- ”The Liar’s Dividend”
- Influence Public Opinion, Elections, ...
- Information Warfare
- Model Collapse
- Influence on Learning

?Social Aspects?

What is OSINT?

Definition

OSINT = **O**pen **S**ource **Int**elligence.
The practice of collecting, analyzing, and making decisions based on information that is **publicly available** and **legally accessible**.

Applications:

- Red Team: Penetration Testing
- Blue Team: Defense
- Law Enforcement & Intelligence
- Business Intelligence
- Journalism

What is OPSEC?

Definition

OPSEC (Operational Security) is the process of protecting individual pieces of data that could be grouped together to give away critical information (like your identity or location).

- Protecting Hardware & Software
- Hiding Your Location
- Hide your Identity
- Behavioral OPSEC (e.g., Avoid Cross-Contamination)

Phishing Website Detection

Giovanni Apruzzese, *PhD*
https://giovanniapruzzese.com

Outline of the talk – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
  - **Real attackers favor cheap tactics, which are often effective (hard to convince reviewers that these "cheap tactics" are interesting…)**
- Using ML to evade ML-based Phishing Website Detectors
  - **You can go crazy with sophisticated techniques to bypass state-of-the-art systems (but always consider how expensive they are…)**
- The viewpoint of human users in the above
  - **ALWAYS consider that humans are the ultimate target of phishing websites (attackers want to phish people–not evade systems!)**

Two goals:

Phishing Website Detection

Giovanni Apruzzese, *PhD*
https://giovanniapruzzese.com

Outline of the talk – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
  - **Real attackers favor cheap tactics, which are often effective (hard to convince reviewers that these "cheap tactics" are interesting…)**
- Using ML to evade ML-based Phishing Website Detectors
  - **You can go crazy with sophisticated techniques to bypass state-of-the-art systems (but always consider how expensive they are…)**
- The viewpoint of human users in the above
  - **ALWAYS consider that humans are the ultimate target of phishing websites (attackers want to phish people–not evade systems!)**

Two goals:

Information Warfare

- **Attack Surface:** Shift from broadcast media to Social Media
- **Actors:** State-sponsored vs. mercenary "Troll Armies" and decentralized cults (QAnon).
- **Tactics: Astroturfing** (Manufacturing consensus via bots and paid actors)
- **Detection:** Anomalies such as "Bot Holidays" (sudden drops in hate speech/activity) reveal coordinated automation.
- **Strategic Goal:** Unlike Cyber Warfare (infrastructure), Info-War hacks **perception**.
- **Defense:** Requires understanding that the "system" being hacked is now human perception, not infrastructure.

Privacy

- Privacy vs. Security (supporting and conflicting interests)
- Why does it matter?
- Potential attacks
- Techniques to protect privacy: k-Anonymity, Attribute Based Credentials, Differential Privacy, **limit data collection**

Up Next ..

Exam

Logistics:

- Friday 12.12. 13:00-15:00 (various rooms)
- meet in front of M106, at least 15 minutes before it starts
- Digiexam, but also need access to Internet to access your Knowledge Base

Types of Exam Questions

- Multiple-choice similar to Quizzes, some may be simple text input. E.g., "Which command line tool is often used to find which services are running on a server that you can not login to?"
- Questions about knowing how to use the tools you were supposed to use in the labs. E.g.,
  "Write a shell command or sequence of commands that finds all lines containing your username in the files in /var/log/."

Types of Exam Questions (2)

- Questions about understanding different kinds of **vulnerabilities**, how they can be **exploited**, what effect an exploit could have and how to **mitigate** the problem. E.g.,
  "The Morris Worm (1988) used two kinds of vulnerabilities. Briefly explain what those vulnerabilities were, how it allowed the worm to execute code and how similar exploits can be prevented."
- Questions about technologies we discussed, what their purpose is, roughly how they work, what security issues arise due to them, and which ones they potentially help mitigate.
- Understanding Defense in Depth and Design Principles for Secure Systems and how this ties into the other topics.

Knowledge Base

- Hand in a link today.
- Fix formatting, and finish after the exam.
- Don't forget the 1-2 page reflection document.