Διάλεξη 4

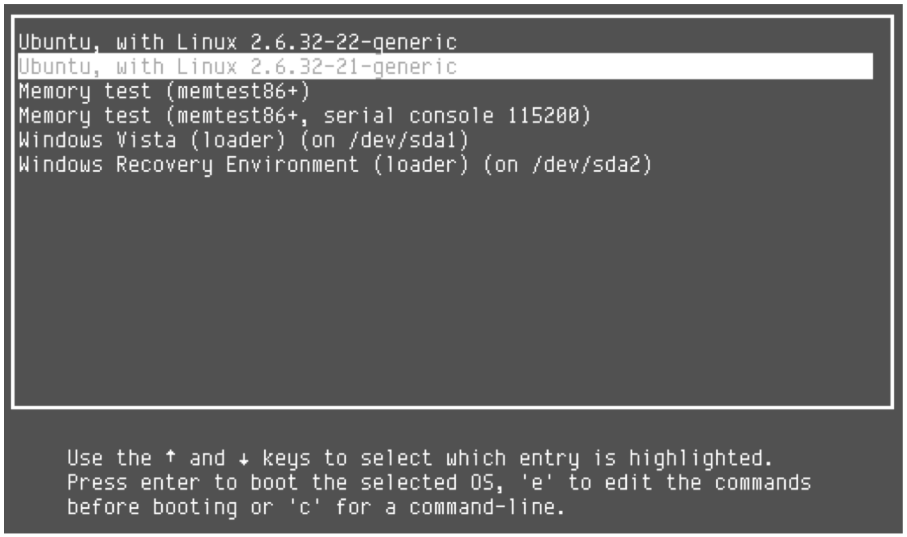
Διαχείριση Συστημάτων UNIX II

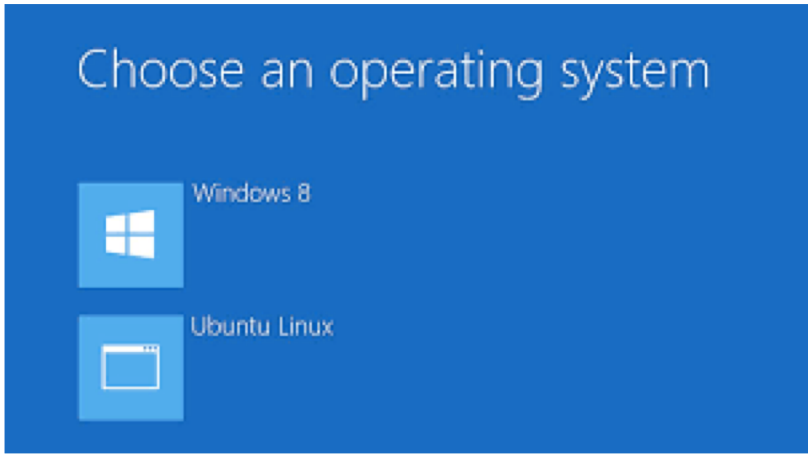
Δημήτρης Ζεϊναλιπούρ

Περιεχόμενο Διάλεξης

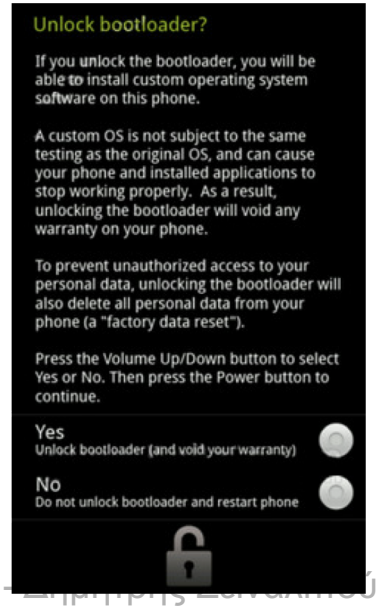
- ***Εκκίνηση & Εγκατάσταση Πακέτων:*** *grub, yum, apt-get, port, rpm, dpkg,*
- ***Δίκτυο:*** *ipTables, tcpdump, nmap, netStat, nslookup, ifconfig,*
- ***Ασφάλεια:*** *ssh-keygen/add, openssl, ssh @RaspberryPI*
- ***Ταυτότητες:*** *date, $$, $RANDOM, uuidgen, md5sum, uuencode/uudecode, base64*
- ***Ιστός / HTTP στο UNIX:*** *curl, wget*

Εκκίνηση Bootloader (grub)



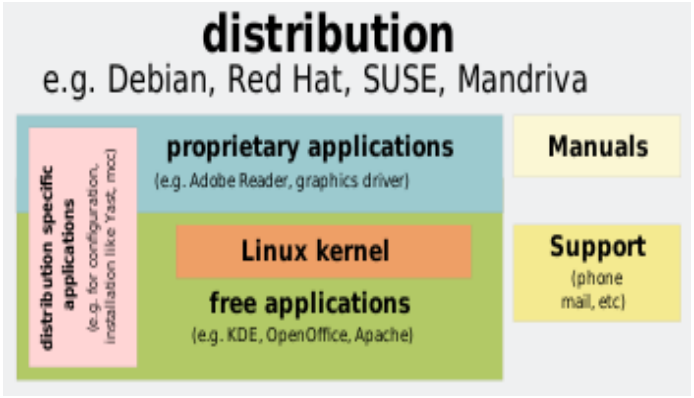


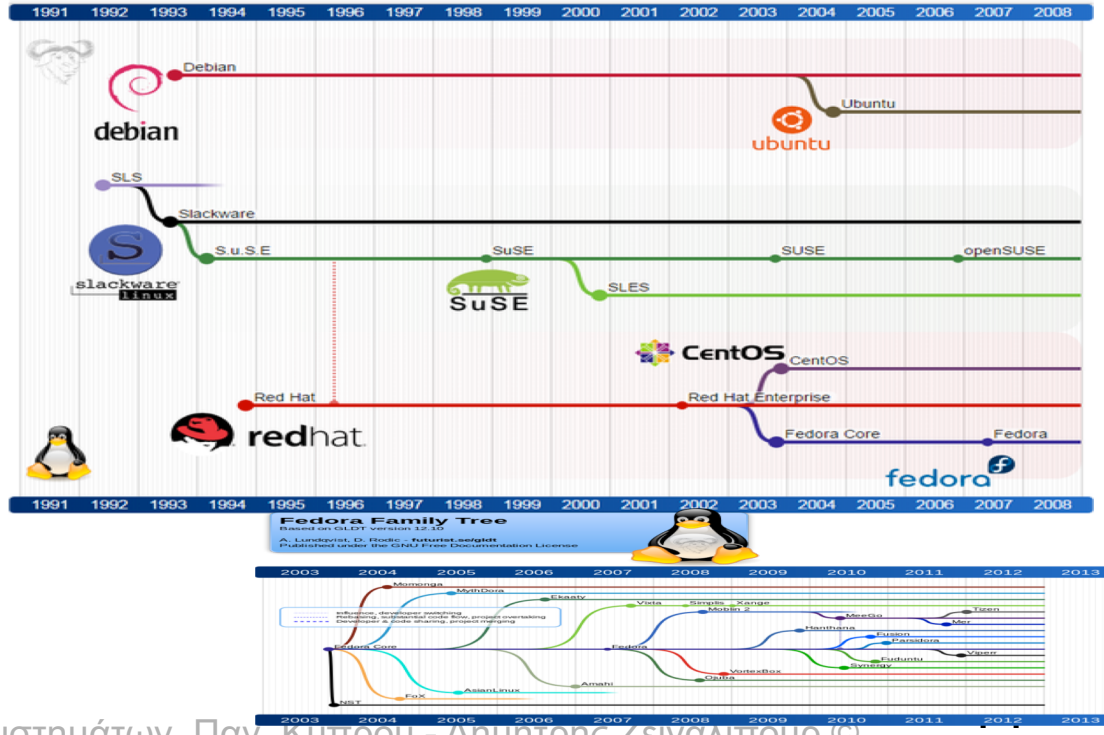
A **boot loader** is a computer program that loads an operating system or some other system software for the computer after completion of the power-on self-tests; it is the loader for the operating system itself.



Linux Distributions

- A **Linux distribution** (often abbreviated as **distro**) is an operating system made from a software collection, which is based upon the Linux kernel and, often, a package management system.





Find Distribution:

```
$ cat /etc/*-release
```

Find Kernel:

```
$ uname -a
```

Finding Distribution / Kernel

```
$ cat /etc/*-release
CentOS Linux release 7.4.1708 (Core)
NAME="CentOS Linux"
VERSION="7 (Core)"
ID="centos"
ID_LIKE="rhel fedora"
VERSION_ID="7"
PRETTY_NAME="CentOS Linux 7 (Core)"
ANSI_COLOR="0;31"
CPE_NAME="cpe:/o:centos:centos:7"
HOME_URL="https://www.centos.org/"
BUG_REPORT_URL="https://bugs.centos.org/"
CENTOS_MANTISBT_PROJECT="CentOS-7"
CENTOS_MANTISBT_PROJECT_VERSION="7"
REDHAT_SUPPORT_PRODUCT="centos"
REDHAT_SUPPORT_PRODUCT_VERSION="7"
CentOS Linux release 7.4.1708 (Core)
CentOS Linux release 7.4.1708 (Core)
```

```
$uname -a
Linux b103ws1 3.10.0-693.5.2.el7.x86_64 #1 SMP Fri Oct 20 20:32:50 UTC 2017 x86_64 x86_64
x86_64 GNU/Linux
```

Package Management (rpm, dpkg, wget)

Α) Από πηγαίο κώδικα

- Διαθέσιμο στον ιστό, CVS, Github, κτλ μεταγλωττίζεται με make ή αντίστοιχα **build automation software** (Apache Ant ή Maven / JAVA, GNU Build System / Autotools: Autoconf, Automake, Libtool)
- Συνηθέστερο σε εκδόσεις μη-διαδεδομενα UNIX (π.χ., HP-UX, AIX) αλλά και παλαιότερα στο Linux ή Linux/Android on ARM, κτλ.
- wget [program].tar.gz -> unpack -> ./configure -> make -> make install

Β) Από πακέτα εγκατάστασης:

•Σε RPM (Redhat, Fedora, Suse, Madriva, Oracle Linux, CentOS, Scient. Linux distributions)

- Ανάκτηση RPM (Red Hat Package Manager) από ιστό
- rpm –i installer.rpm # Install
- rpm –V installer.rpm # Verify (for conflicts before install)

•Σε non-RPM Linux Distributions

- **Debian Linux: dpkg --install foo_VVV-RRR.deb**
- **Slackwere Linux: installpkg [packagename].tgz**

Package Management (yum, apt-get, port)

Γ) Από βιβλιοθήκες πακέτων

- RPM Linux: **`yum search <package>`**
  - yum is an additional wrapper around rpm. It keeps its own database of rpm files available for your distribution, generally in online repositories.
- DEBIAN Linux: **`apt-get search <package>`**
  - On Debian systems, the equivalent repository and dependency-resolution tools are provided by Apt (apt-get and aptitude).
- MACOSX (Macports Project – requires Xcode/sudo):
  - **`sudo port search <package>`**
  - **`sudo port install <package>`**
  - sudo port select --set python python35
  - A more limited package manager for MacOSX is called homebrew (brew/ruby)

Παράδειγμα: Εγκατάσταση Πακέτων σε MacOSX (Command Line)

- Installing DMG or PKG from command-line on Mac, Apple TV
  - mounting dmg
    - hdid package.dmg
    - hdiutil attach package.dmg
    - diutil mount package.dmg
  - cd /Volumes/package/
  - Install pkg
    - sudo installer -verbose -pkg package.pkg -target /
    - sudo /usr/sbin/installer -verbose -pkg package.pkg -target /unmounting dmg
    - hdiutil detach /Volumes/package
    - hdiutil detach /Volumes/package –force

Τερματισμός Λ.Σ.
Εντολή **shutdown**

- **shutdown** -- close down the system at a given time
  - The **shutdown** utility provides an automated shutdown procedure for super-users to nicely (SIGTERM – Signal #2) notify users when the system is shutting down.
- `$shutdown –r now`
  - # Restart System now
- `$shutdown –r +number or yymmddhhmm`
  - # Restart at specific time
  - `+number`: brings system down in number minutes.
- `$shutdown –h now`
  - # Halt (Stop) system now (don't use on VPS as you won't have a way to restart.

Sleep/ Hibernate/ Shutdown

- **Hibernate Mode**: It writes all active data to the disk and then switches off the components as if the computer were fully turned off.
  - You can cut the the power of a system in hibernation, since it does not pose any risk to your data. Once the computer is powered back on it reads the data from disk and sends them back to RAM—this process can take few seconds to minutes. The data is restored to the point at which they entered hibernation. (good when boarding a plane or travelling).
  - pmset –a hibernatemode 0
    - Memory ON, no image on disk # sleep-mode!
  - **pmset –a hibernatemode 3 (default)**
    - Image on disk, Memory on => instant boot!
  - pmset –a hibernatemode 25 (best for battery)
    - Full Image on disk, **ALL Components OFF**

Setting Shutdown modes on MacOSX

```
$ pmset -g
System-wide power settings:
Currently in use:
 lidwake               1
 autopoweroff          1
 standbydelayhigh      86400
 autopoweroffdelay     28800
 standbydelaylow       10800
 standby               1
 ttyskeepawake         1
 hibernatemode         25
 powernap              1
 gpuswitch             2
 hibernatefile         /var/vm/sleepimage
 highstandbythreshold 50
 womp                  0
 displaysleep          60
 networkoversleep      0
 sleep                 60
 tcpkeepalive          0
 halfdim               1
 acwake                0
 disksleep             10
```

Sleep/ Hibernate/ Shutdown

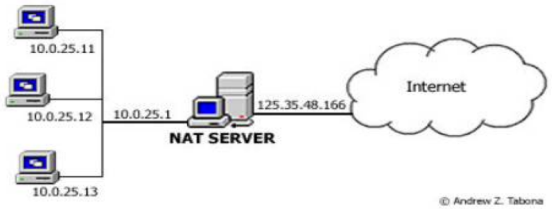
- **sudo pmset -a tcpkeepalive 0**
  - Warning: This option disables TCP Keep Alive mechanism when sytem is sleeping. This will result in some critical features like 'Find My Mac' not to function properly
- sudo pmset -a womp 0
  - womp wake on "magic" Ethernet packet, 1 to enable or 0 to disable
- More: https://en.wikipedia.org/wiki/Pmset

Setting Shutdown modes on MacOSX

```
$ pmset -g
System-wide power settings:
Currently in use:
 lidwake              1
 autopoweroff         1
 standbydelayhigh     86400
 autopoweroffdelay    28800
 standbydelaylow      10800
 standby              1
 ttyskeepawake        1
 hibernatemode        25
 powernap             1
 gpuswitch            2
 hibernatefile        /var/vm/sleepimage
 highstandbythreshold 50
 womp                 0
 displaysleep         60
 networkoversleep     0
 sleep                60
 tcpkeepalive         0
 halfdim              1
 acwake               0
 disksleep            10
```

Διαχείριση Δικτύου ipTables

- Εντολή ***ipTables***
  - Administration tool for IPv4 packet filtering. Provides means to setup a “firewall” on UNIX Systems.
  - It also allows Network Address Translation (NAT): a way to map an entire network (or networks) to a single IP address.
- **Iptables** is used to set up, maintain, and inspect the tables of IPv4 packet filter rules in the Linux kernel.



- **Several different tables may be defined.**
  - Each table contains a number of **built-in chains and may also contain user-defined chains.**

Διαχείριση Δικτύου ipTables

 # Start/Stop/Restart iptables service

•/etc/init.d/iptables start/stop/restart

 # LIST all configurations of INPUT chain (initially empty)

•sudo iptables -L INPUT

 # Adding & Removing Rules:

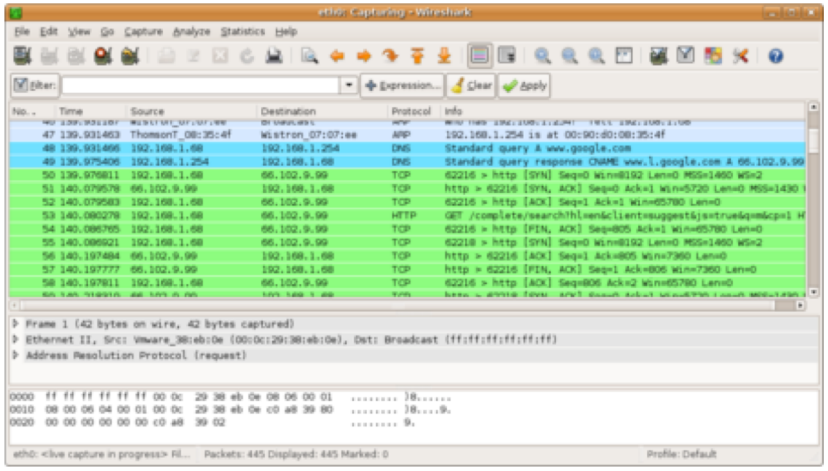
•sudo iptables -D INPUT # DELETE ALL RULES

•sudo iptables -L INPUT # LIST ALL RULES

•sudo iptables -A INPUT # ADD ALL RULES

Διαχείριση Δικτύου TCPDump

- **tcpdump** is a common **packet** analyzer that runs under the command line.
  - It allows the user to display **TCP**/IP and other packets being transmitted or received over a network to which the computer is attached.
- tcpdump uses the libpcap library to capture packets
- **Libpcap** is also used in Wireshark (prior Ethereal).
  - The port of tcpdump for Windows
- Requires **root** access to install it, as it is installed libpcap is installed very low in the OS stack (kernel).



Διαχείριση Δικτύου
Tcpdump Example

```
$ ifconfig | head
eth2      Link encap:Ethernet  HWaddr 52:54:00:7B:CA:99
          inet addr:10.16.1.101  Bcast:10.16.1.127  Mask:255.255.255.224
          inet6 addr: fe80::5054:ff:fe7b:ca99/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:214252728 errors:0 dropped:0 overruns:0 frame:0
          TX packets:148649576 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:156620098878 (145.8 GiB)  TX bytes:93799079896 (87.3 GiB)

lo        Link encap:Local Loopback
```

Receive packets flows on a particular port using tcpdump port

```
$ tcpdump -i eth0 port 22

19:44:44.934459 IP valh4.lell.net.ssh > zz.domain.innetbcp.net.63897: P
18932:19096(164) ack 105 win 71 19:44:44.934533 IP valh4.lell.net.ssh >
zz.domain.innetbcp.net.63897: P 19096:19260(164) ack 105 win 71
19:44:44.934612 IP valh4.lell.net.ssh > zz.domain.innetbcp.net.63897: P
19260:19424(164) ack 105 win 71
```

 # Capture packets for particular destination IP and Port

```
$tcpdump -w xpackets.pcap -i eth0 dst 10.181.140.216 and port 22
```

Διαχείριση Δικτύου
nMap Port Scanner

- **Nmap is a security scanner**
  - Most well known port scanner on Unix.
- **Features:**
  - **Host discovery** – Identifying hosts on a network. (e.g., listing the hosts that respond to TCP and/or ICMP requests or have a particular port open.
  - **Port scanning** – Enumerating the open ports on target hosts.
  - **Version detection** – Interrogating network services on remote devices to determine application name and version number.
  - **OS detection** – Determining the operating system and hardware characteristics of network devices.
- **Usage:**
  - Auditing, Find and exploit vulnerabilities, Generating traffic to hosts on a network, Network inventory, network mapping, maintenance and asset management.

Διαχείριση Δικτύου
Nmap Example

```
$ nmap  www.cs.ucy.ac.cy

Starting Nmap 5.21 ( http://nmap.org ) at 2016-02-11 22:06 EET
Nmap scan report for www.cs.ucy.ac.cy (194.42.17.135)
Host is up (0.00092s latency).
rDNS record for 194.42.17.135: clio.cs.ucy.ac.cy
Not shown: 997 filtered ports
PORT     STATE SERVICE
80/tcp   open  http
3128/tcp open  squid-http
8080/tcp open  http-proxy

Nmap done: 1 IP address (1 host up) scanned in 6.12 seconds
```

Διαχείριση Δικτύου
Ping (Host Latency)

- **Ping** is a computer network administration software utility used to test the reachability of a host on an Internet Protocol (IP) network and to measure the round-trip time for messages sent from the originating host to a destination computer and back.
- A host might not respond to ICMP (ping) messages!
  - ICMP is a subprotocol of IP.
- **Example:**

```
$ ping www
PING clio.cs.ucy.ac.cy (194.42.17.135) 56(84) bytes of data.
64 bytes from clio.cs.ucy.ac.cy (194.42.17.135): icmp_seq=1 ttl=64
   time=0.883 ms
64 bytes from clio.cs.ucy.ac.cy (194.42.17.135): icmp_seq=2 ttl=64
   time=0.942 ms
64 bytes from clio.cs.ucy.ac.cy (194.42.17.135): icmp_seq=3 ttl=64
   time=0.873 ms
```

Διαχείριση Δικτύου
Traceroute (Path Latency)

- **traceroute** is a computer network diagnostic tool for displaying the route (path) and measuring transit delays of packets across an Internet Protocol (IP) network.
- Quite similar to ping, but shows intermediate routers that respond to traceroute requests.

- **Example:**

```
$ traceroute www.cs.ucr.edu
```

- traceroute to thoth.cs.ucr.edu (169.235.30.15), 64 hops max, 52 byte packets
- 1 10.16.16.254 (10.16.16.254) 1.720 ms 1.182 ms 1.144 ms
- 2 cs-sw7.cs.ucy.ac.cy (194.42.17.65) 1.128 ms 0.836 ms 0.778 ms
- 3 194.42.0.139 (194.42.0.139) 1.046 ms 1.001 ms 1.012 ms
- 4 194.42.0.42 (194.42.0.42) 1.260 ms 1.109 ms 1.055 ms
- 5 ip6.vega2.ucy.ac.cy (194.42.13.150) 1.300 ms 1.370 ms 1.328 ms
- 6 82.116.192.190 (82.116.192.190) 1.335 ms 1.577 ms 1.463 ms
- 7 cynet-ap2.mx1.fra.de.geant.net (62.40.124.149) 58.767 ms 58.830 ms 58.739 ms

Διαχείριση Δικτύου
netStat (Network Statistics)

- **netstat** (***net**work **stat**istics*) is a command-line tool that displays network connections for the Transmission Control Protocol (both incoming and outgoing), routing tables, and a number of network interface (network interface controller or software-defined network interface) and network protocol statistics
- **Example:**

```
$ netstat
•  tcp4       0      0  cs.in.cs.ucy.49526 ec2-52-185-66.https CLOSE_WAIT
•  tcp4       0      0  cs.in.cs.ucy.49515 wk-in-f94..https ESTABLISHED
•  tcp4       0      0  cs.in.cs.ucy.49506 theano.cs.ucy.ac.imap  ESTABLISHED
•  tcp4       0      0  cs.in.cs.ucy.49473 ec2-52-2-24.https CLOSE_WAIT
•  tcp4       0      0  cs.in.cs.ucy.49471 17.172.232.9.5223      ESTABLISHED
•  tcp4       0      0  cs.in.cs.ucy.49379 17.110.225.84.5223     ESTABLISHED
•  tcp4      31      0  cs.in.cs.ucy.49373 45.58.74.129.https     CLOSE_WAIT
•  tcp4       0      0  cs.in.cs.ucy.49230 theano.cs.ucy.ac.imap  ESTABLISHED
```

Διαχείριση Δικτύου
Nslookup (DNS Resolution)

- **nslookup** is a network administration command-line tool available for many computer operating systems for querying the Domain Name System (DNS) to obtain domain name or IP address mapping or for any other specific DNS record.

Example:

```
$ nslookup www.google.com
```

```
$ nslookup www.google.com
Server:                 10.16.1.118
Address:        10.16.1.118#53

Non-authoritative answer:
Name: www.google.com
Address: 74.125.206.147
Name: www.google.com
Address: 74.125.206.106
Name: www.google.com
Address: 74.125.206.105
Name: www.google.com
Address: 74.125.206.104
```

Διαχείριση Δικτύου (Ifconfig - Network Settings)

- **ifconfig** is a system administration utility in Unix-like operating systems for network interface configuration. (Windows: ipconfig /all)
- **Example:**
- $ ifconfig

```
en2: flags=8863<UP,BROADCAST,SMART,RUNNING,SIMPLEX,MULTICAST> mtu 1500
    options=4<VLAN_MTU>
    ether 10:9a:dd:42:59:19
    inet6 fe80::129a:ddff:fe42:5919%en2 prefixlen 64 scopeid 0x4
    inet 10.16.16.188 netmask 0xffffff00 broadcast 10.16.16.255
    nd6 options=1<PERFORMNUD>
    media: autoselect (100baseTX <full-duplex,flow-control>)
    status: active
en0: flags=8863<UP,BROADCAST,SMART,RUNNING,SIMPLEX,MULTICAST> mtu 1500
    ether a8:66:7f:29:09:27
    inet6 fe80::aa66:7fff:fe29:927%en0 prefixlen 64 scopeid 0x5
    inet 10.16.4.248 netmask 0xfffffe00 broadcast 10.16.5.255
    nd6 options=1<PERFORMNUD>
    media: autoselect
    status: active
```

Public/Private RSA Keys (used for SSH)

- Generate Keys on PC

```
$ mkdir -p ~/.ssh # if not already created
$ chmod 700 ~/.ssh; cd ~/.ssh
$ ssh-keygen -t rsa # Generate rsa|dsa key
```

  - Enter file in which to save the key **(/home/user/.ssh/id_rsa):**
  - Enter passphrase (empty for no passphrase): Enter same passphrase again:
  - Your identification has been saved in **/home/user/.ssh/id_rsa.**
  - Your public key has been saved in **/home/user/.ssh/id_rsa.pub.**

- Transfer **id_rsa.pub** to SERVER.

```
$cat id_rsa.pub >> .ssh/authorized_keys; chmod 600
   .ssh/authorized_keys
```

- Add **ssh/id-rsa** to PC keychain

```
$ssh-add -K ~/.ssh/id-rsa
```

- **Troubleshooting!**

```
$ssh -vvv -l <user> <host> # 3 levels of verbose / debugging
   -v, -vv, -vvv, -l: different login name
```

Public/Private RSA Keys (~/.ssh/known_hosts)

When we connect to some node, we are requested to validate its authenticity. The given is recorded in the known_hosts file.

$ ssh b103ws6

```
The authenticity of host 'b103ws6 (10.16.6.243)' can't be established.
RSA key fingerprint is 01:9a:eb:42:02:ca:b4:cc:c0:c3:58:2c:49:85:45:e4.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'b103ws6,10.16.6.243' (RSA) to the list of known
hosts.
```

$tail ~/.ssh/known_hosts

```
b103ws6,10.16.6.243 ssh-rsa
AAAAAABIwAAAQEAtrjdSBK4Q60/7PtKRfotLLrxnqWG1QAMqLPtQFUZhV08fdQJANS4BoANYp9A
AvMPGME8tz1Ko0hIm9FkNFm5jDoXa3NkiUC/wbcqa8IwrW4kAI61m4PMpMYVDpPGk9/QvgzzBYK
cAvUHMMyfzHvWq2AQRHVcaeFafQEL9s343mUHlBhVe…
```

Why? If some attacker masquerates that IP/node, we will know as the RSA key fingerprint of the attacker won't match

Πιστοποιητικά ασφαλείας (Certificates / openssl)

- Στην Κρυπτογραφία, ένα **public key certificate,** γνωστό ως **digital certificate** ή **identity certificate,** είναι ένα ηλεκτρονικό αρχείο το οποίο χρησιμοποιείται για το **ownership** (ιδιοκτησία) ενός **public key.**
  - **OpenSSL** περιέχει μια ανοικτού πηγαίου υλοποίησή των πρωτοκόλλων **SSL** και **TLS** (κρυπτογραφημένη επικοινωνία μεταξύ διαδικτυακών κόμβων)
  - Παράδειγμα Εξέτασης certificate : `openssl s_client -showcerts -connect www.cs.ucy.ac.cy:443`
  - Δημιουργία του δικού σας Certificate & Εγκατάσταση στο Server σας: Συνήθως έχει κάποιο κόστος εφόσον απαιτεί κάποιο γνωστό Certification Authority.
  - Το https://letsencrypt.org/ σας δίνει τη δυνατότητα να δημιουργήσετε δωρεάν

Secure File Transfer **(SCP)**

- **scp** copies files between hosts on a network.
  - It uses ssh for data transfer, and uses the same authentication and provides the same security as ssh.
- Many protocols for File Transfer => the older were not unecrypted, but the newer introduced encryption.
  - e.g., FTP evolved to i) **FTP-SSL (FTPS);** ii) **SSH FTP (SFTP);** iii) **FTP over SSH** (i.e., tunneling FTP through an SSH connection - see next slide)
  - FTP originally had two channels (authentication and data transfer): encryption can apply to either channel or both.

Secure File Transfer (SCP)

- Here we focus on a single tool, i.e., scp, similar concepts with other tools as well.
  - Transfer Data from Production to Development server:
- `scp anyplace@ap.cs.ucy.ac.cy:/home/anyplace/anyplace_v3/floor_plans.tar.gz /home/anyplace/anyplace_v3/floor_plans.tar.gz`
  - Having the public/private key in place will circumvent the requirement of giving user/pass each time

SSH Port Forwarding (SSH Tunelling)

- SSH port forwarding is a mechanism in SSH for tunneling application ports from the client machine to the server machine, or vice versa.
- Usage:
  - *Adding encryption to legacy applications* e.g., you have a proprietary protocol that is not encrypted => you tunnel it over SSH to make it secure from eavesdropping!
  - *Opening backdoors* into the internal network from their home machines. (Dangerous as we bypass the Firewall)

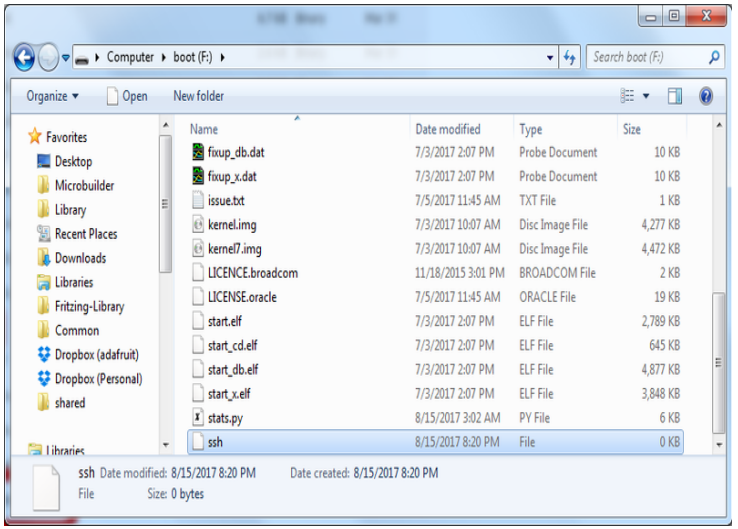
- How it works:
  - the SSH client listens for connections on a configured port, and when it receives a connection, it tunnels the connection to an SSH server

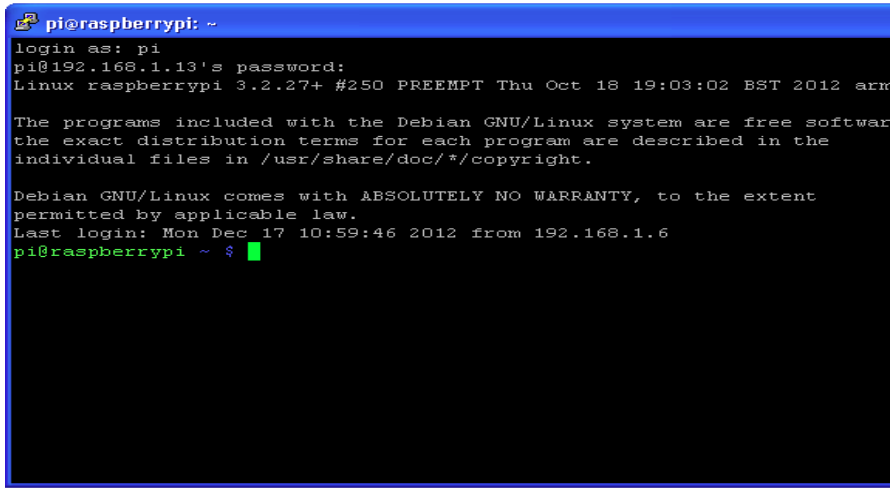
SSH Tunneling Example (Jump Server)

- Creating a Jump Server
  - E.g., CryptoAuditor can act as a jump server, record all sessions, and pass session contents to analytics for early warning of suspicious activity.
  - `ssh -L 127.0.0.1:80:web.example.com:80 jumpserver.example.com`
  - `(only local web service is permitted to be forwarded to jumpserver – no external traffic)`
  - More: https://www.ssh.com/ssh/tunneling/example

Raspberry PI SSH Server

- Add file named “ssh” to SDCard.
- Load the SD Card.
- Now ssh to node “pi” from PC





- From there it is all UNIX!
- Just type sudo raspi-config

Timestamp Identifier

- Many times we need to create an identifier (unique name) for separating data.
- **Use the machine clock time.**
  - Time since 1/1/1970 (epoch time). UNIX time!
  - Initially this time was since 1971-1-1.
  - A 32-bit *signed* integer using 1970-1-1 as its epoch can represent dates up to 2038-1-19.

```
$ date +"%s"          #seconds
1457034425
$date +%s%m #milliseconds
145703476903
```

- Nanoseconds is not available on all UNIXes

```
(ada)$ date +%s%N #nanoseconds
1457034579278836206
```

- date +"%Y%m%d_%H%M%S"
- 20180921_160726

Timestamp to Date:

Linux:

date -d @1267619929

MacOSX:

date -r 1267619929

>> Sat Apr 30 06:32:13 CET 48631

Other Identifiers

- Process Identifier:
  - $ echo $$ => 28835
- Random Identifier
  - $ echo $RANDOM => 23953
- Hostname Identifier
  - $echo $HOSTNAME => ada.in.cs.ucy.ac.cy

Cryptographic Identifier (**uuid RFC4122**)

- The **RFC4122 UUID** standard generates a **128-bit Unique Identifier** that is unique in space and time.
- The Result is usually printed in **Hexadecimal format** with or without dashes.

- **$uuidgen**
  - E.g., EEF45689-BBE5-4FB6-9E80-41B78F6578E2
- **$cat /proc/sys/kernel/random/uuid**
  - d6aa801c-6cd5-4c90-b16a-aaca0eeae1ec
- **$dbus-uuidgen** #dbus package on Debian
  - 52195bef65c5faab6ea13b4c0000b443

Cryptographic Identifier (**md5sum RFC1321**)

- The MD5 message digest is a way to compute a 128-bit sequence that is unique for the same sequence.
  - Widely used for **disseminating packages** on the internet (e.g., an ZIP, AVI, MP3 package has an accompanying MD5 digest to enable the downloader verify that the download was complete.
  - Not **cryptographically strong** and not used for encryption anymore, even though called a cryptographic hash function.

Download (only 249KB):

WinMD5 Freeware Download

WinMD5Free.zip MD5: 73f48840b60ab6da68b03acd322445ee

WinMD5Free.exe MD5: 944a1e869969dd8a4b64ca5e6ebc209a

- **$md5sum WinMD5Free.zip**

73f48840b60ab6da68b03acd322445ee WinMD5Free.zip

Binary-to-Text (uuencode / uudecode)

- **Uuencoding** is a form of **binary-to-text** encoding that originated in the **Unix** program **uuencode**, for encoding binary data for transmission over the **UUCP mail system**.
- **uuencode file.zip newname.zip > myfile.uue**
  - The purpose of the uuencode program is to translate a binary file that contains **unprintable (non-text) characters** into a format that is entirely readable.
  - This prevents mail, news, and terminal programs from **misinterpreting** non-text characters as special instructions.
  - Also helps with Endianess issues in transmission.

Binary-to-Text (uuencode / uudecode)

$ uuencode 01.pdf hi.pdf > encoded.txt

$ cat encoded.txt

```
begin 755 hi.pdf
M)5!$1BTQ+C4-"B6UM;6U#0HQ(#`@;V)J#0H\/"]4>7!E+T-A=&%L;V<O4&%G
M97,@,B`P(%(O3&%N9RAE;BU54RD@+U-T<G5C=%1R9652;V]T(#(P,B`P(%(O
M36%R:TEN9F\\/"]-87)K960@=')U93X^/CX-"F5N9&]B:@T*,B`P(&]B:@T*
…
…
…
4>')E9@T*,30Q-S(T,@T*)25%3T8`
`
end
```

$ uudecode encoded.txt

$ ls hi.pdf

hi.pdf

Binary-to-Text Conversion (base64)

- **base64** encodes/decodes Base64 data (RFC 4648): from non-printable to printable bytes.
  - 64 Characters are used in the output: A–Z (26) , a–z (26), and 0–9 (10) and + / (2)
  - Widely used in email attachments (IMAP & POP3)

```
X-mxHero-InitialSizeLimiter: rule=158
Sender: iss@ucy.ac.cy
X-Zimbra-DL: ucyacall@ucy.ac.cy

This is a multipart message in MIME format.
------=_NextPart_000_014D_01D17089.55100A00
Content-Type: multipart/alternative;
    boundary="----=_NextPart_001_014E_01D17089.55100A00"

------=_NextPart_001_014E_01D17089.55100A00
Content-Type: text/plain;
    charset="iso-8859-7"
Content-Transfer-Encoding: base64

xMnBxMnKwdPJwSDK0cHUx9PH0yDF0cPB09TH0cnZzSDHy8XK1NHPzcnK2c0g1dDPy8/DydPU2c0g
1MfTINXQ0w0KDQogDQoNCsjhIOjd6+Hs5SDt4SDz4flg5e3n7OXx/vPv9ezlIPz06SwgyvHc9Ofz
5yDF8ePh8/Tn8d/v9S/57SDH6+Xq9PHv7enq/u0NCtXw7+vv4+nz9P7tIOzw7/Hl3yDt4SDj3+3l
6SDh8Pwg7/Dv6e/k3vDv9OUg8PH84/Hh7OzhIPDl8ene4+fz5/Ig5Onh5Onq9P3v9Q0KKHdlYiBi
cm93c2VyKSDs3fP5IPTv9SAi0/3z9Ofs4fTv8iDK8eH03vPl+e0gxfHj4fP05/Hf+e0g9OfyINXQ
0yIg8/Tv7Q0K8/3t5OXz7O8gIDxodHRwOi8vd3d3LnVjeS5hYy5jeS9sYWJfcmVzZXJ2PiBodHRw
Oi8vd3d3LnVjeS5hYy5jeS9sYWJfcmVzZXJ2DQrelOrh6SDh8Pwg9O/tIPP97eTl8+zvIKvV8Ofx
```

Binary-to-Text Conversion (base64)

```
$ echo "UNIX rocks" > a.txt

# encode
$ base64 a.txt > a-b64.txt

# view encoded
$ cat a-b64.txt
VU5JWCByb2Nrcwo=

# decode
$ base64 -D a-b64.txt
UNIX rocks
```

Binary-to-Text Conversion (base64)

```
# Add red content to
# encoded.txt

# decode
$ base64 -D
  encoded.txt > a.jpg
```

----------050105040807090205020508Content-Type:
image/jpeg
Content-Transfer-Encoding:
base64Content-ID:part1.01080202.01010401@cs.ucy.ac.cy
/9j/4AAQSkZJRgABAQEAeAB4AAD/2wBDAAoHBwgHBgoICAgLCgoLDhgQDg0NDh0VFhEYIx8lJCIfIiEmKzcvJik0KSEiMEExND
k7Pj4+JS5ESUM8SDc9Pjv/2wBDAQoLCw4NDhwQEBw7KCIoOzs7Ozs7Ozs7Ozs7Ozs7Ozs7Ozs7Ozs7Ozs7Ozs7Ozs7Oz
s7Ozs7Ozs7Ozv/wAARCAA2ASUDASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAt
RAAAgEDAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY
3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDx
MXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6erx8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAt
REAAgECBAQDBAcFBAQAAQJ3AAECAxEEBSExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJyg
pKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHilmKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba
3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwD2Jcjk0yaYkLyEBRTyMD6frWH
rF8WlSGlCRA3zj0rnr1lTjc0pwc5WNaGWOYMYmBG7mp+Otc7ZXrQ35G3bA3UmuhB3Llenapw9f2yuwq03TdkDZ6ioUuYZH8tW
BNQajdm3gljwZW6KayPtMkQgaOHDk/NWFfFqE1EqFKUlc6UUtQ286zRBlOTUucDNd0ZqUU0YtOLaFoqOKYSqT0xVPVdYh0iN
ZbhGKOwUbad7biNCisy51tLS0N1NbSrCBkt6Vasr1b62W5RSqMMru7ijmixImimCVcbgwYUFwBncMevammgH0UwyKF3bgB6k8
UeYuM7hjPBz1o5kA+im+YmQN46+tBcA4J59BRdAOoppYLwzAZ6c0gkXOCwz2ANF0gH0U0yLjO4Y9c1XN/ELsWo5klz7Gpbim
Baopu4DgkZ/lShhzyDj3quZALRTQ4OBnnGeKQSKeAykjqAad0A+imlsKSuW9qzbTX7e8u5rZI3DQf60notS2gNSiqNnqi3ryCOF1R
Dje3AP0qQX8RvDaodzgZYnoBTUkBaoppdQwBcZ9M0eYvTcCewFHMgHUU0uucbgD6ZpN+OW+Udye1HMgH0U0MDk7hgelLu
HXcMU7pgLRSA8etFGgGfquopY2xPJkxgAVzEZ8+dXhZhIzZYGuyltoZiDIgYj1pn2K1B/wBUoPXIFebiMLUqy30OmlXjSj5nK3wbzy
1y2wJyFXvWtpWtJNbESHaU6ZqfVNKjukMirhlFczLINxxtZeMCvlrTqYOTZ204wrxt1L91dLfXDMX2FfutT5Y53hiRj8o6sKr6Npx1CTf
IT5SeneuqS1hCLH5fyL2NXh8LUxMXOXUipVjSfKjE0rUY4pjAqttzgE10C8gn2qNbS3ByIlB9QKmHAlxgYr2sLRnSjyyZwVJqTuVrLi
Ns8/NWJ42ZV0yEE8+cMe1bll/q3/3jS3Nta3CqLIUdQeN/rXRKNzMydfIUeFpWJDAxDt1rLlvJYtK0aGJmjtpsCVwPu/WuqltrZ4fLkRT
EB909MUi2lqYPJEamI9u34Vl7J9wOeWSNPF6W0MzNG8PKg8fWoYbi6sbu50SZnczNmBuvB7108dhawspSBQyjAbvipPJhabzjG
hkUY3dxVKD7gcxetLFrttYrlzRCPD7jgE+tP8QltlorSRTuzBxt2muhntLe4YedErt69xSGytTF5TQgx9geaHCQHNampibR2SSQec67
+farRleLxil1ZipgO1SeM1uta28gRXjUiPlc/wAJpWtoWlEpjUyD+PHNChK4HKWkhvdM1B7+4kW4jdsAHlfTFOZrn/hCTdSb0uQpww
PzEZrpnsbR5CzxJnv7/WpDHGyGJkBjPGMcYp8jbA56XZ/wiHnJMxkMWd2eQ1Z10j23h+x1BJ3+07IBb1HeuvWytVtzAlV8o9VoNl
btCIGiUx4+7jpSlTbAwb+5kh8Q6a3n7lpF/eAng8U3TroTeJ9ShWVymzjnj8KualobX2o2lyNoituDGepFaa2lrExkjhVXl5lqVCSA5KC
+vlvDl8YZHaaO4wSeqrVm/c21tpktjKzzyOAQDndnrmuiiis1MiRJHlz8y/3qWKxtYWVkgCkdB6VootATKTxu4bGSa4RY7uXxDqctm
4ZUky8XTeK7svGeGkU8Yxmoo7K0hl82KFVc87h1NJxdgMN9Ug1HQLmK2Rop4UO6EcMDVC7umbw7p8kUh+0l1B2/eP1rrBZ2y
zGVYVDt95gOv1pi6fZJMJVgUPnjilyOwGFfsT4nsIDl4DJlwD7U22dofEeox+a+2OLfGCeOldE1rbvOszxgy44fFMuLOKRXZI1EzoV
D4ocJAczZy/a/D099LcN9pBYrg9COgp0t/ezLpCXLPHFOv71sdT7+Iaui6INPszFcqkr7ywI6VqS20M0flyxqyjoMdKShIDBieWLxML
W3dpLR4D5nOdpqlZXMg8P6jHcXDCSORtm44brxXVQ20MLExRBGPVvWmSadaSymV4FL55z3qlCSAdpxY6dbliSTGM5oqcAA
YHAHQUVdmACkIpR0oxVJvcVrrUjlkWKJmbpiuLvGEs0kgOF3Gunv7G4u8qJtifzrOfw07KFE2a8DMqNavJKK0R34WpCk7sd4ZuE
8hkJAbNb59etc9B4duLaXzlrjBrehRki2yNuNd+D9pGHJJWMcQ4uXNFkg6elB6H6UDpS4yCK9BKxzlWy/1b/75rE8all02J43ZG80L
wa6CGIxIR6mqetaUNWsjb7tjDlWqKl2tAGXUeNF27jkRjmqdzq7aRZ2S+QZ/O4HtVoWV7JZrazTJsxywHWmanoz3ptBC+1bcg1ny
sBLfWpf7WGn3FuUd13l1V7XU2EupskbSNbnJX+9VyfSXl1qDUVlAES7SDRY6O1nqNzdNlrJc9VxRaSAji14XVjbXFvHuadtrJ3FR
yel1864it4zJ9mX58djUmmeH107U57lZd0b8pH/AHKZDocllqVxcWjlY7g7mjYd6PfAbdeJBb6dbXq2zSJO2wr3DU6512ayt4nu7Voz
K+1SOi/Wnalo8+ox2oEiQvDJ5mFHBq7d280+wfunQcMrjg0XmBRu9QdtluZpFyqfxleoqteaxdw2mmGzUFblgp3dali8OtFp97a+f/x9
HIHZPpSy6FPLp9lEkyrJaHlbFL3wLM2sNaatFZXEW1JR8svbNWbK8ku3lUx7Y0bCv61melYo7y1jspGb7UWG0qPzrYtbdbW2it1J
by1A+tWnJAUZdY3ao2n2sfmTRruc+gqo+ttdaZfxonk3Vuh3le3vViXSZItabUrWQb3Xa6H+KmQ6C0cF6fNBubsEM/8AdBqZOTYEf
hiIXGm297Kh83adrE/eFN8Tanc2jWlnbtsa6fDP6CtLSLKXTbBLWSRXMfCkCmaxpMOrwqjkplhzHlvY1dpWAcNJhMHllm3kcvnrVZ
NSXTtRi0uZWVGQlJm6EjtUkVtqyWv2dp42YDAfHNUNat1vYLbT2dnvlYFXUcL61D5kBauNcmg06a++zFo4mwB/e96jk8QParbXF
xDtt7nGG/u5o8QW62/heaFSRtUcjuaZDpU2qWNkLpgII0DbPU1LcmBJc67cR6sdOt7QySFPMVvaphrD/wBrQ6dNBteWPcT6Uf2T
KuujUVl+RYvL20X+kvNq8GpQSqrRLhgR1qryALbWvPub23aA77T0/iFN07XP7R2vGoKscMndaj07Rbuxury5a5V5LrpxwKF0FmvO
LyQpE8Z5EfR6ceYDcAlJ54x0p1IDnPtS1smwCiiimAYooopgxMUYooqVuCDFAUCiih7i6i0UUUxhRiiikAmKMUUUwDaAMdqWiim
AmB+NG0e/50UUAG0enajaD2oooAXFGKKKAGGKMuHKAsOhxzTgADnvRRQAFQTnv60AAdqKKAFxSbQM+hoooAXFN8tN+/a
N2MZxRRSsAPGkiIXUMp6g9KFjVF2qMAdBRRRZALgDPvS4FFFMBAMUuKKKADFFFFIAooooA//Z
----------050105040807090205020508--

Internet Bots (curl, wget)

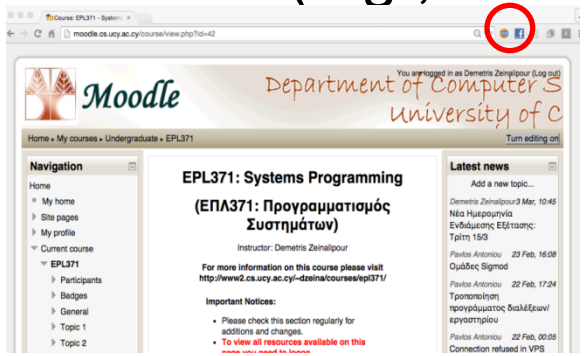
- An **Internet bot,** also known as **web robot**, WWW robot or simply bot, is a **software application** that runs **automated tasks (scripts)** over the Internet.
  - Typically, **bots perform tasks** that are both **simple** and **structurally repetitive**, at a much **higher rate** than would be possible for a human alone.
    - Think about a bot running on a dozen of UNIX machines (see Bash Programming)
- The largest use of bots is in web spidering, in which an automated script fetches, analyzes and files information from web servers at many times the speed of a human.
  - Wikipedia

Internet Bots (curl, wget)

- **curl is a tool to transfer data from or to a server, using one of the supported protocols**
  - DICT, FILE, FTP, FTPS, GOPHER, HTTP, HTTPS, IMAP, IMAPS, LDAP, LDAPS, POP3, POP3S, RTMP, RTSP, SCP, SFTP, SMTP, SMTPS, TELNET and TFTP.
- **Simple Example:**
  - `$ curl www.cs.ucy.ac.cy > index.html`
- **HTTP GET Example:**
  - `$ curl http://moodle.cs.ucy.ac.cy/enrol/index.php?id=42`
- **HTTP authentication (do only with SSL):**
  - `$` curl -u user:password https://example.org/
- **HTTP POST Example (e.g., do only with SSL) :**
  - `$` curl --data "user=<name>&pass=hi" https://www.example.com/login.php
- HTTP HEAD Example (e.g., find when a file was created!)
  - $ curl --head https://www2.cs.ucy.ac.cy/docs/prospectus.pdf

Cookie-based Crawling (wget)

- Most websites use session cookies for retaining authenticated users online.
  - HTTP Cookies are small pieces of data sent from a website and stored in the user's web browser.
  - Every time the user loads the website, the browser sends the cookie back to the server to notify the user's previous activity
- **How to Crawl a Site with Cookies?**
  - Fetch Cookie using Web Browser extension (e.g., cookies.txt in Chrome)

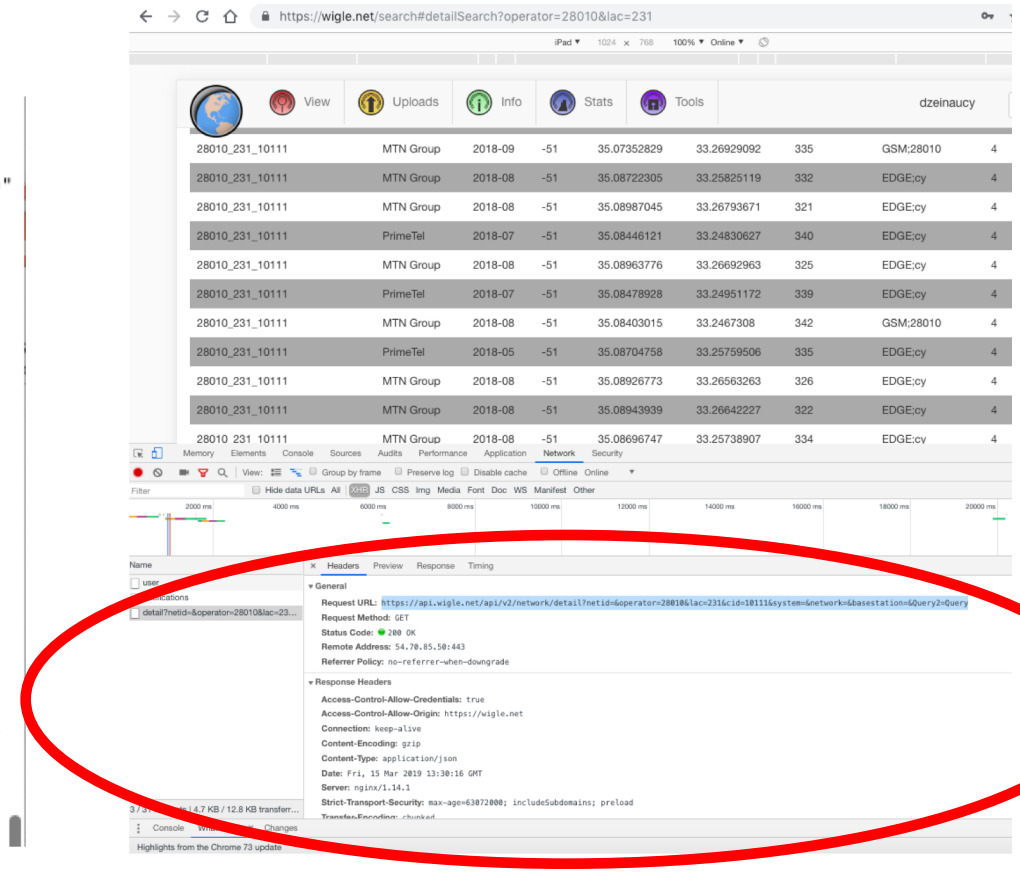


  - `$ wget -x --load-cookies cookies.txt http://moodle.cs.ucy.ac.cy/course/view.php?id=42`

Crawling AJAX Calls

Problem: No data ☹

```
<script id="searchResultsRows" type="text/template">
  {{#results}}
  <tr>
    <td><a href="javascript:void(0);" class="mapLink" map-lat="
{{trilat}}" map-lon="{{trilong}}" map-ssid="{{ssid}}" map-
netid="{{netid}}" title="click to view on map">map</a></td>
    <td><a href="javascript:void(0);" class="detailLink"
bssid="{{netid}}" title="click for detail">{{netid}}</a></td>
    <td>{{ssid}}</td>
    <td>{{name}}</td>
    <td>{{type}}</td>
    <td>{{firsttime}}</td>
    <td>{{lasttime}}</td>
    <td>{{networkIcon wep gentype}}</td>
    <td>{{trilat}}</td>
    <td>{{trilong}}</td>
    <td>{{channel}}</td>
    <td>{{bcninterval}}</td>
    <td>{{qos}}</td>
    <td>{{userfound}}</td>
    <td>{{free}}</td>
    <td>{{pay}}</td>
    <td netcomment="{{netid}}" class="commentcell"
id="commentcell-{{netid}}">{{comment}}</td>
    <td><input class="commentbtn" type="button"
id="comment{{netid}}" netid="{{netid}}"  value="add comment"/>
</td>
  </tr>
  {{/results}}
</script>
```



Chrome Developers Tools or Similar (e.g., Safari) can help us to find the underlying calls / HTTP headers upon which we can initiate the wget/curl commands

Crawl Complete Domain

- $ wget \ --recursive \ --no-clobber \ --page-requisites \ --html-extension \ --convert-links \ --restrict-file-names=windows \ --domains www.ucy.ac.cy \ --no-parent \ www.ucy.ac.cy/test/html/

- The options are:
- --recursive: download the entire Web site.
- --domains www.ucy.ac.cy : don't follow links outside www.ucy.ac.cy.
- --no-parent: don't follow links outside the directory tutorials/html/.
- --page-requisites: get all the elements that compose the page (images, CSS and so on).
- --html-extension: save files with the .html extension.
- --convert-links: convert links so that they work locally, off-line.
- --restrict-file-names=windows: modify filenames so that they will work in Windows as well.
- --no-clobber: don't overwrite any existing files (used in case the download is interrupted and resumed).

Διαχείριση Αρχείων XML / JSON (xmllint, jq)

- Στην εποχή των ανοικτών δεδομένων (Open Data) διατίθενται πλέον στον ιστό σωρεία δεδομένων προς κατανάλωση, π.χ.,
  - π.χ., δεδομένα κλινικών δοκιμών από το https://clinicaltrials.gov/ διαθέτει δεδομένα σε XML
  - Wikidata.org διαθέτει μια XML έκδοση της Wikipedia σε XML.
  - Οι πλείστες Web 2.0 υπηρεσίες (π.χ., Google, FB, Twitter, κτλ.) παρέχουν JSON APIs τα οποία επιτρέπουν την προσπέλαση σε JSON (lightweight XML) δεδομένα σε συνεχομένη βάση
- Τι είδους εργαλεία χρειαζόμαστε για να επεξεργαστούμε γρήγορα τέτοια δεδομένα;

Διαχείριση Αρχείων XML / JSON (xmllint, jq)

```
# Παρουσίαση περιεχομένου XML
$ xmllint --format 3178056.nxml
```

```
<ref id="B72">
  <label>72</label>
  <element-citation publication-type="journal">
   <person-group person-group-type="author">
    <name>
     <surname>Price</surname>
     <given-names>MN</given-names>
    </name>
    <name>
     <surname>Dehal</surname>
     <given-names>PS</given-names>
    </name>
    <name>
     <surname>Arkin</surname>
     <given-names>AP</given-names>
    </name>
```

Διαχείριση Αρχείων XML / JSON (xmllint, jq)

```
# Ανάκτηση και μορφοποίηση περιεχομένου JSON
$ curl -s
   'http://api.nytimes.com/svc/elections/us/v3/finances/2008/president/to
   tals.json?api-key=super-secret' | jq '.' | head
```


```
{
 "results": [
   {
     "candidate_id": "P80003338",
     "date_coverage_from": "2007-01-01",
     "date_coverage_to": "2008-11-24",
     "candidate_name": "Obama, Barack",
     "name": "Barack Obama",
     "party": "D",
```


Επιπλέον Εργαλεία για Data Science:

- **json2csv - convert JSON to CSV | xml2json - convert XML to JSON**
- **csvkit - suite of utilities for converting to and working with CSV**