Vladimír Veselý
veselyv@fit.vutbr.cz

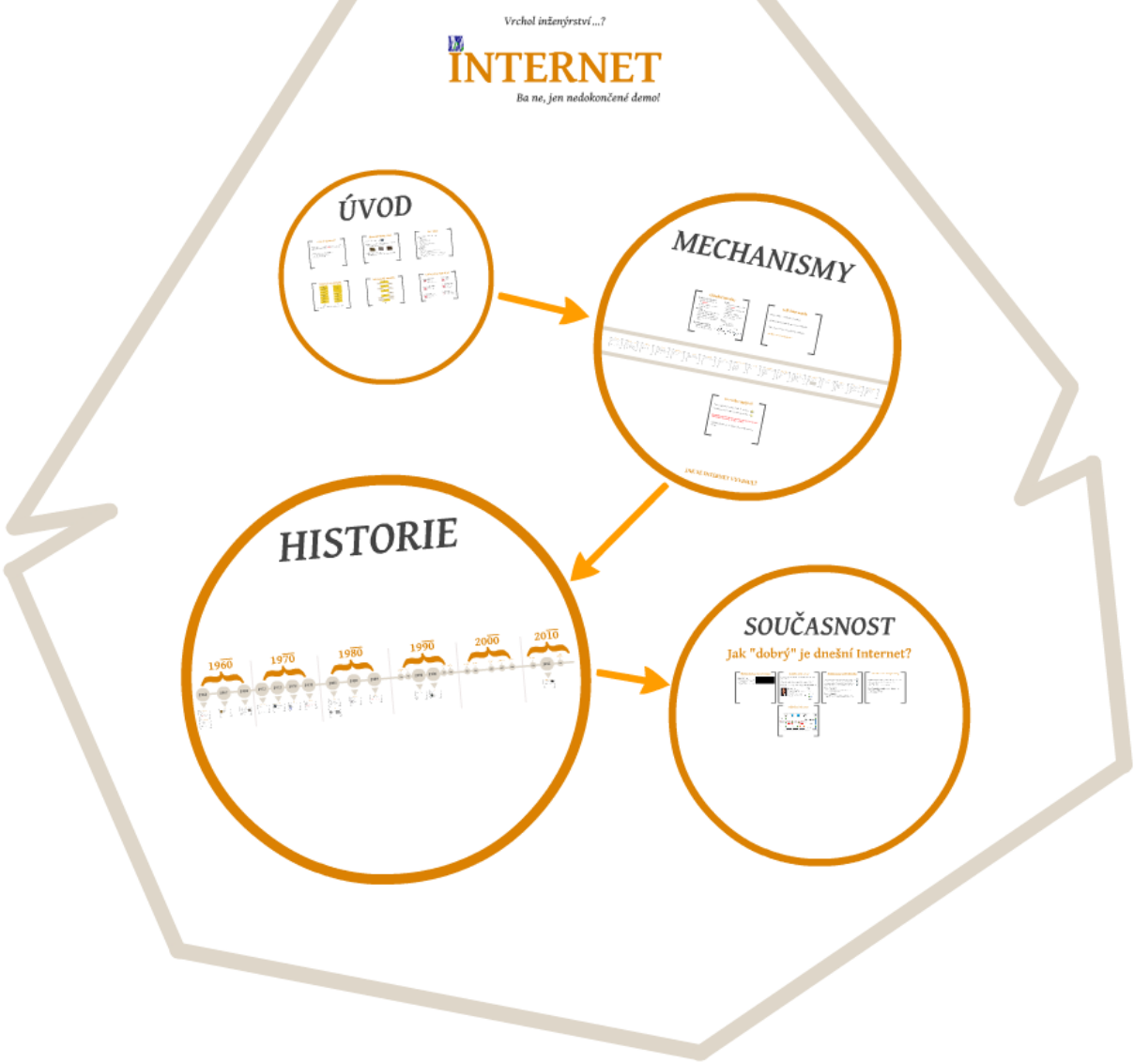
CCS 2017

Stručná historie internetové architektury

Vrchol inženýrství ...?

INTERNET

Ba ne, jen nedokončené demo!



Vladimír Veselý
veselyv@fit.vutbr.cz

CCS 2017

Stručná historie internetové architektury

Vladimír Veselý
veselyv@fit.vutbr.cz

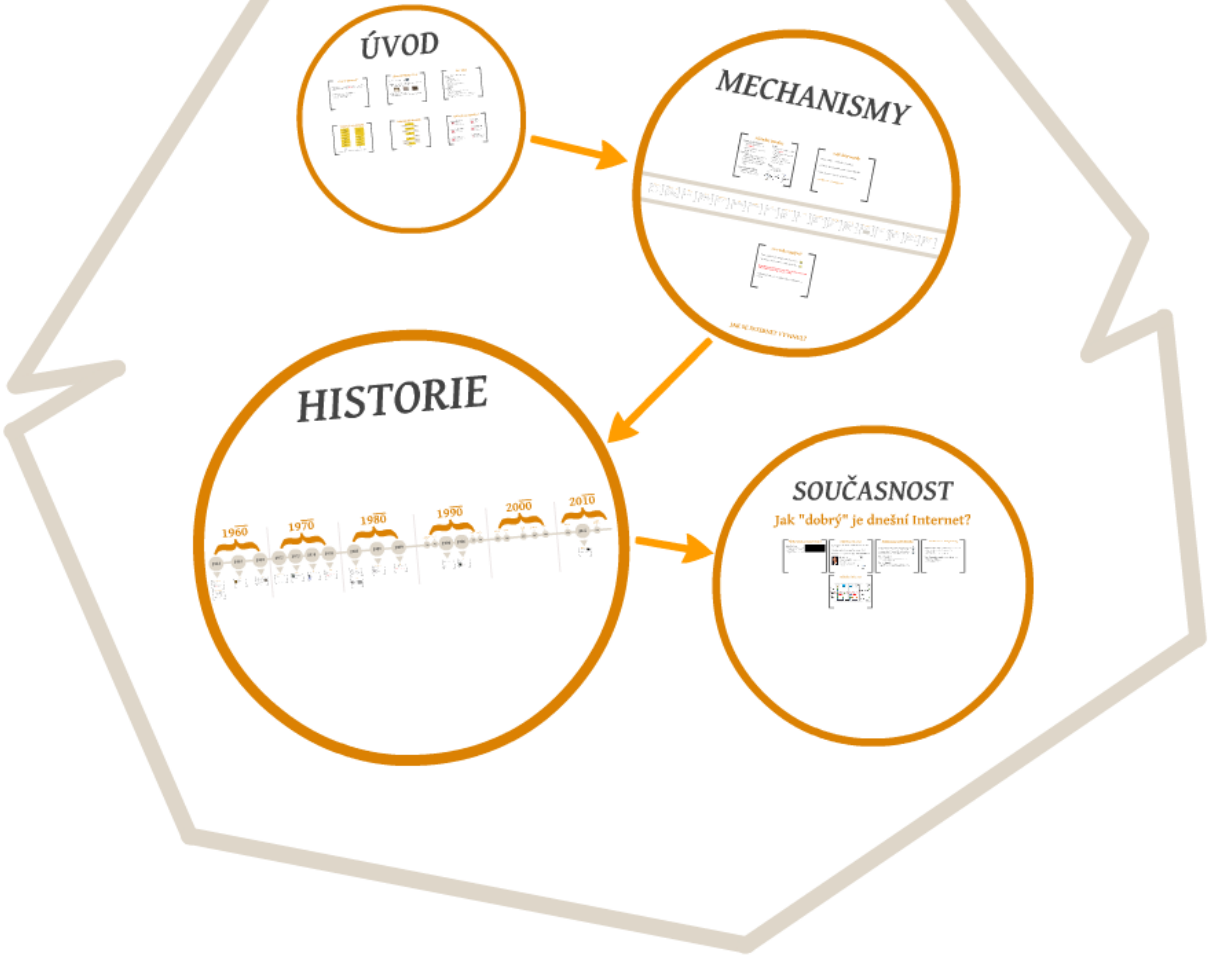
CCS 2017

Stručná historie internetové architektury

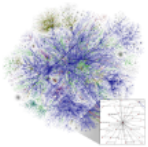
Vrchol inženýrství ...?

INTERNET

Ba ne, jen nedokončené demo!

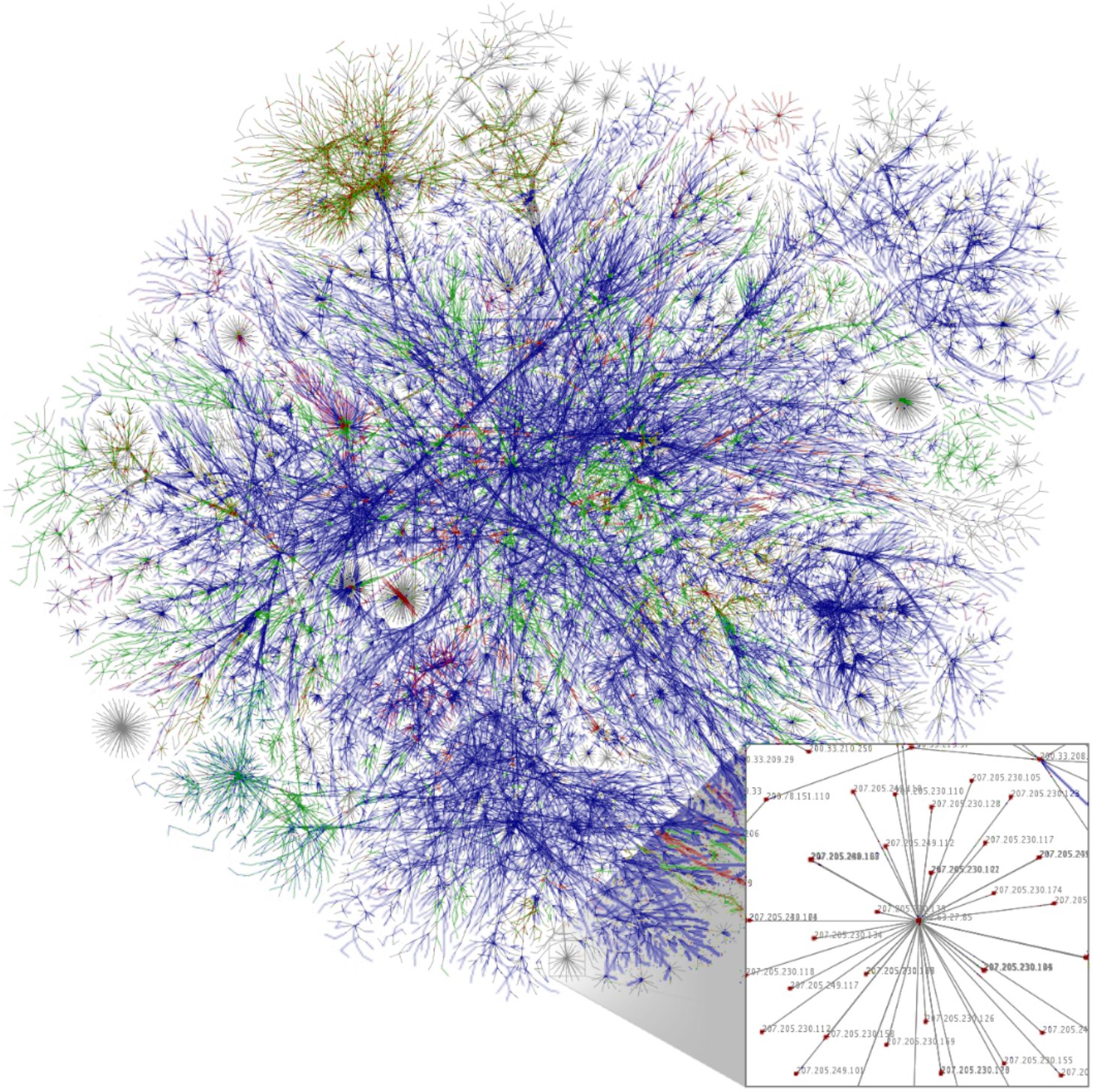


Vrchol inženýrství ...?



INTERNET

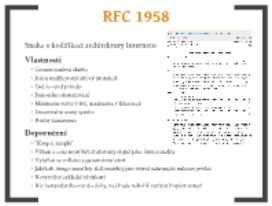
Ba ne, jen nedokončené demo!


207.205.230.105
207.205.230.110
207.205.230.128
207.205.230.117
207.205.230.102
207.205.230.174
207.205.230.134
207.205.230.118
207.205.249.117
207.205.230.126
207.205.249.101
207.205.230.155

ÚVOD


Co je to Internet?
"The Internet is a global system of interconnected computer networks that use the standard Internet protocol suite (TCP/IP) to serve several billion users worldwide"
Internet je kolekce zařízení, která komunikují
- některé části se komunikaci zároveň účastní
- jiné ji jen zprostředkovávají


Slovo ARCHITEKTURA
Rozpor v používání takoto slova !?
Architektura = množina různých pravidel, obsazených (možností) a omezení, která charakterizují cíl výsledku
ISO/IEC 7498-1 je příkladem architektury
= ISO/OSI RM je implementací architektury ve světě vrstvení


RFC 1958
Snaha o kodifikaci architektury Internetu
Vlastnosti
Doporučení


Protokolový zásobník
Application
Presentation
Session
Transportation
Network
Data-Link
Physical
ISO/OSI
Application
Transportation
Network
Data-Link
Physical
TCP/IP
"ISO/OSI je jak NOZ, TCP/IP je jak zažívací prach Co je lepší"


Jednoduchý zásobník
Application
Transportation
Network
Data-Link
Physical
TCP/IP


Takže jak to vlastně je?
Connectionless service
End-to-end princip
Best-effort doručování
Minimum stavu v síti...
Jeden síťový protokol
Hrubý konsenzus

Co je to Internet?

*"The Internet is a global system of interconnected computer networks that use the standard Internet protocol suite (**TCP/IP**) to serve several billion users worldwide"*

Internet je kolekce zařízení, která **komunikují**

- některé části se komunikace aktivně účastní
- jiné ji jen zprostředkovávají

Slovo ARCHITEKTURA

Rozpor v používání tohoto slova !!! 

Architektura = množina platných pravidel, obecných možností a omezení, která charakterizují styl výsledku



ISO/IEC 7498-1 je příkladem architektury

- ISO/OSI RM je implementací architektury se sedmi vrstvami

RFC 1958

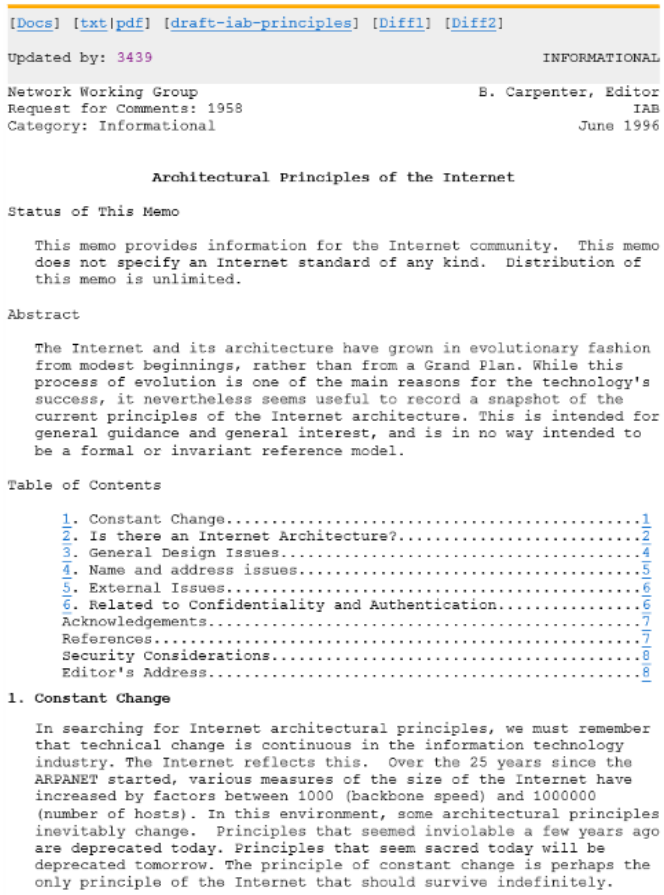
Snaha o kodifikaci architektury Internetu

Vlastnosti

- Connectionless služba
- Jeden unifikovaný síťový protokol
- End-to-end princip
- Best-effort doručování
- Minimum stavu v síti, maximum v klientech
- Decentralizovaný systém
- Hrubý konsenzus

Doporučení

- "Keep it simple"
- Výkon a cena musí být zvažovány stejně jako funkcionalita
- Vyhýbat se volbám a parametrizování
- Jakýkoli design musí být škálovatelný pro řešení zahrnující miliony prvků
- Nevytvářet cyklické závislosti
- Nic nestandardizovat do doby, než bude několik variant implementací



[Docs] [txt|pdf] [draft-iab-principles] [Diff1] [Diff2]

Updated by: 3439 INFORMATIONAL

Network Working Group B. Carpenter, Editor
Request for Comments: 1958 IAB
Category: Informational June 1996

Architectural Principles of the Internet

Status of This Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

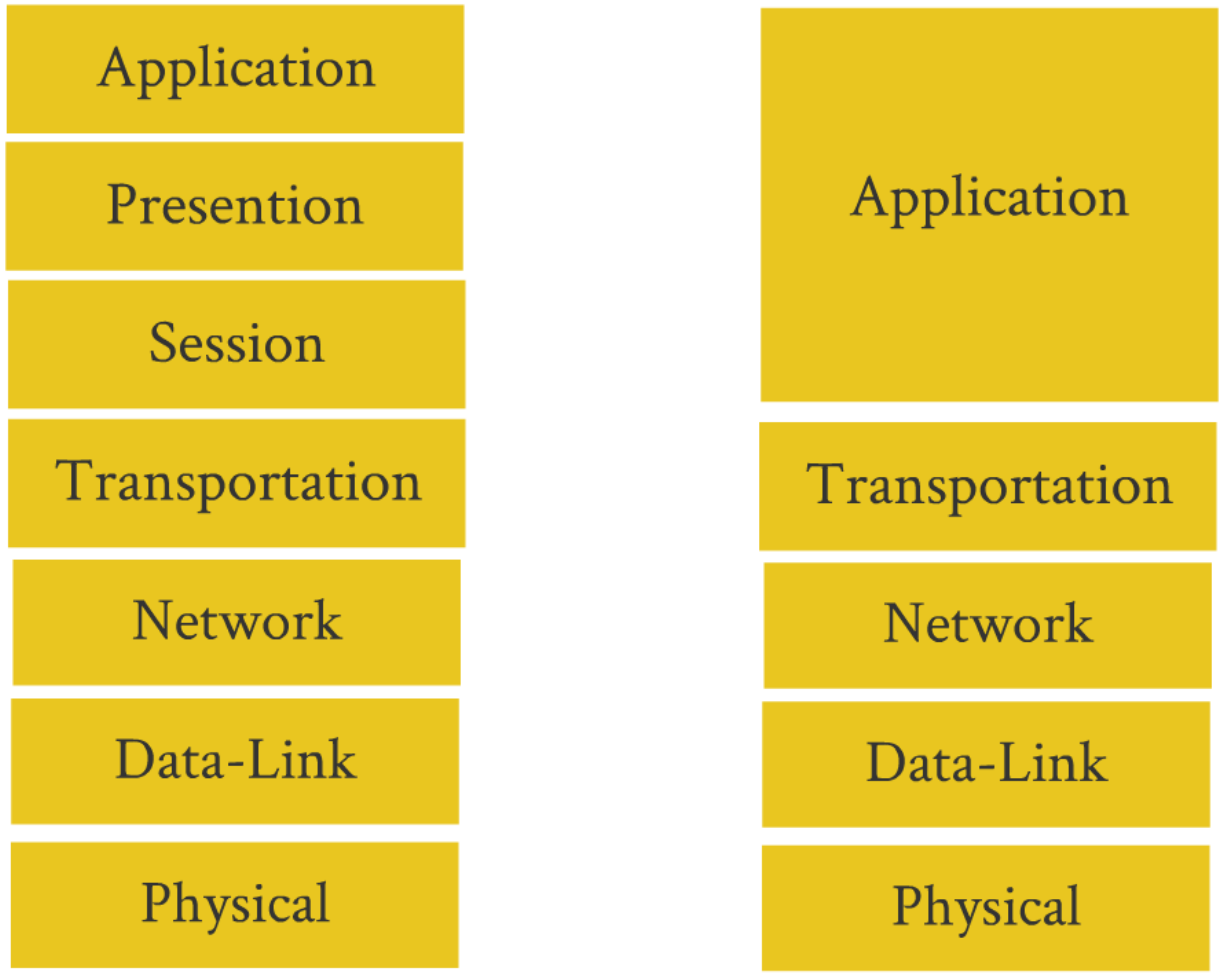
The Internet and its architecture have grown in evolutionary fashion from modest beginnings, rather than from a Grand Plan. While this process of evolution is one of the main reasons for the technology's success, it nevertheless seems useful to record a snapshot of the current principles of the Internet architecture. This is intended for general guidance and general interest, and is in no way intended to be a formal or invariant reference model.

Table of Contents

1. Constant Change

In searching for Internet architectural principles, we must remember that technical change is continuous in the information technology industry. The Internet reflects this. Over the 25 years since the ARPANET started, various measures of the size of the Internet have increased by factors between 1000 (backbone speed) and 1000000 (number of hosts). In this environment, some architectural principles inevitably change. Principles that seemed inviolable a few years ago are deprecated today. Principles that seem sacred today will be deprecated tomorrow. The principle of constant change is perhaps the only principle of the Internet that should survive indefinitely.

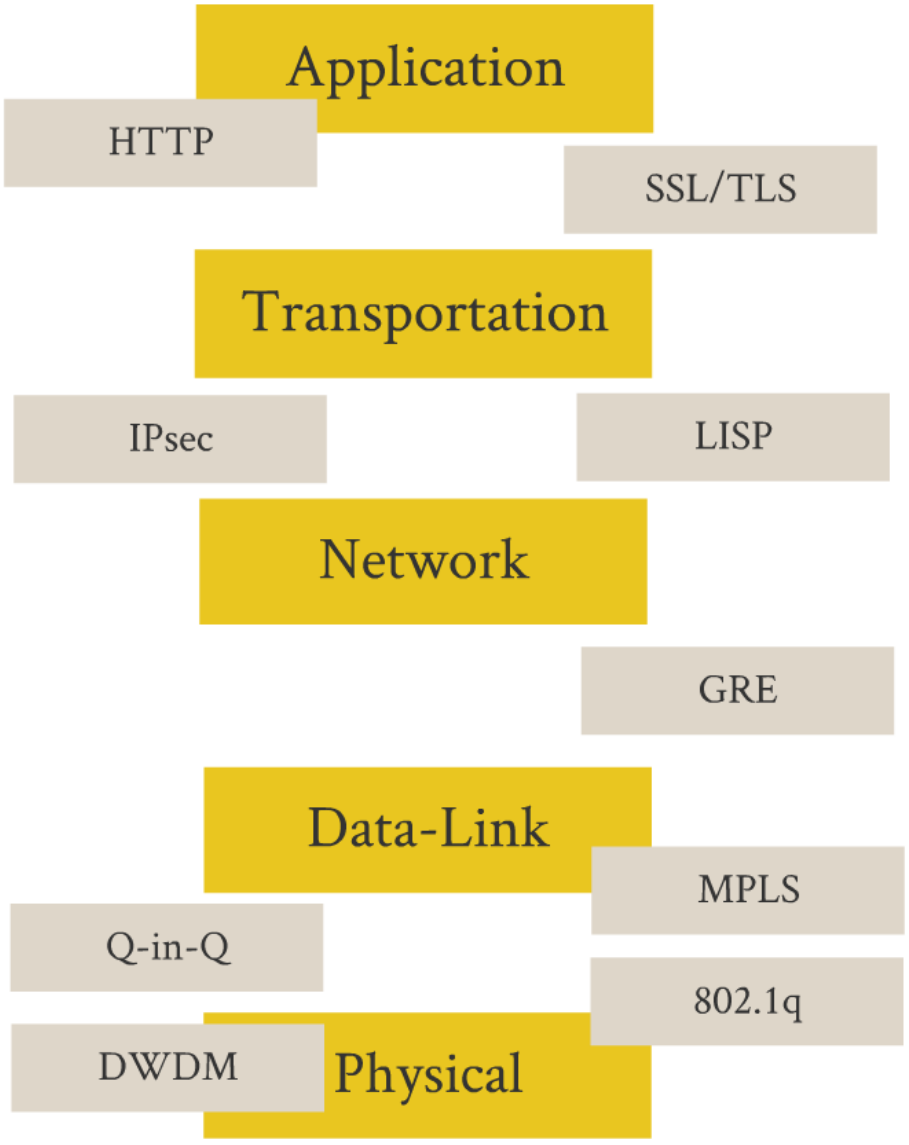
Protokolový zásobník



ISO/OSI ***TCP/IP***

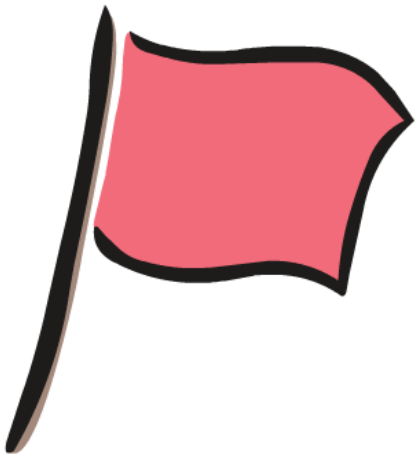
"ISO/OSI je jak NOZ, TCP/IP je jak zvykové pravo! Co je lepší?"

Jednoduchý zásobník



TCP/IP

Connectionless service



- only for UDP
- TPC je connection-oriented + point-to-point

Best-effort doručování

- nestačí
- požadavky na QoS

Jeden síťový protokol

- dual-stack IPv4 + IPv6
- flag-day je utopie
- migrace 16 let?

End-to-end princip

- NAT, CGN
- Shim6

Minimum stavu v síti...

- růst DFZ
- ACL pro zařízení
- BGP politiky

Hrubý konsenzus

- Jsou ICANN, IETF a IAB dostatečně flexibilní?

Takže jak to vlastně je?

Connectionless service

- only for UDP
- TPC je connection-oriented + point-to-point

End-to-end princip

- NAT, CGN
- Shim6

Best-effort doručování

- nestačí
- požadavky na QoS

Minimum stavu v síti...

- růst DFZ
- ACL pro zařízení
- BGP politiky

Jeden síťový protokol

- dual-stack IPv4 + IPv6
- flag-day je utopie
- migrace 16 let?

Hrubý konsenzus

- Jsou ICANN, IETF a IAB dostatečně flexibilní?

MECHANISMY

Základní termíny

Počítačová komunikace
- zařízení komunikují po síti, aby sdílela stav
- unicast – multicast – broadcast vs. [illegible]

Protokol
- [illegible] předpisů a procedur, které komunikující strany dodržují
- konečné kvantum informací = PDU
- popis pomocí KA či temporální logiky

Protokolový stroj (PM)
- implementace protokolu
- PM [illegible]

Vrstva
- [illegible] PM se stejným účelem
- [illegible]
- PDU se skládá z hlavičky + těla (+ [illegible])
- enkapsulace a dekapsulace
- lze vrstvit na sebe a vytvořit komplexní systém

Rank
- pozice v systému
- (N)-PM, (N)-PDU

Celé dohromady

Internet je kolekce zařízení, která komunikují

Zařízení využívají protokolů k synchronizování stavu PM

[illegible]

Co je třeba, aby se komunikovalo?

Co z toho vyplývá?

Katalog mechanismů je velký, ale přeci jen konečný...

Množství policy však může být prakticky neomezené...

Komunikační architekturu a její vlastnosti odráží kombinace PM mechanismů a policy, kterých využívá!

Architektura Internetu je jen tak dobrá, jako jsou PM, které má k dispozici...

JAK SE INTERNET VYVINUL?

Základní termíny

Počítačová komunikace

- zařízení komunikují po síti, aby *sdílela stav*
- unicast × multicast × broadcast vs. whatevercast

Protokol

- množina předpisů a procedur, které komunikující strany dodržují
- konečné kvantum informací = PDU
- popis pomocí KA či temporální logiky

Protokolový stroj (PM)

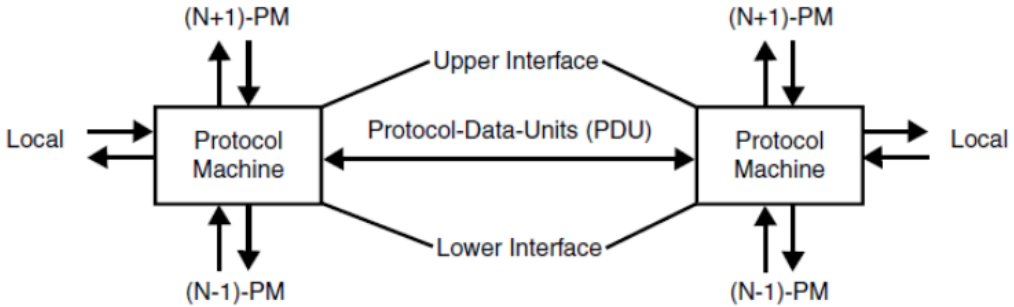
- implementace protokolu
- PM lze mezi sebou provázat

Vrstva

- sdružuje PM se stejným účelem
- omezuje dosah
- PDU se skládá z hlavičky + těla (+ ocásku)
- enkapsulace a dekapsulace
- lze skládat na sebe a vytvořit komplexní systém

Rank

- pozice v systému
- (N)-PM, (N)-PDU



Celé dohromady

Internet je kolekce zařízení, která komunikují

Zařízení využívají protokolů k synchronizování stavu PM

Málo vs. hodně PM se odráží v málo či hodně vrstvách

Co je třeba, aby se komunikovalo?

Delimiting

Na linkové vrstvě musíme být schopni zjistit hranice mezi rámci

Externí oddělovač = speciální bitová sekvence

- Jak zajistit, že se daná sekvence nevyskytne v Payloadu?

802.3 Ethernet frame structure

Preamble	Start of frame delimiter	MAC destination	MAC source	802.1Q tag (optional)	Ethertype (Ethernet II) or length (IEEE 802.3)	Payload	Frame check sequence (32-bit CRC)	Interframe gap
7 octets	1 octet	6 octets	6 octets	(4 octets)	2 octets	42[note 2]–1500 octets	4 octets	12 octets
	← 64–1518 octets (68-1522 octets for 802.1Q tagged frames) →							
← 84–1538 octets (88-1542 octets for 802.1Q tagged frames) →								

10101010 10101010 10101010 10101010 10101010 10101010 10101010 10101011

Interní oddělovač = je dopředu známa velikost PDU pomocí PCI (množství bitů, bytů, oktetů)

Počáteční synchronizace

Jakýkoli sdílený stav je potřeba na počátku inicializovat!

Čtyři formy spojení

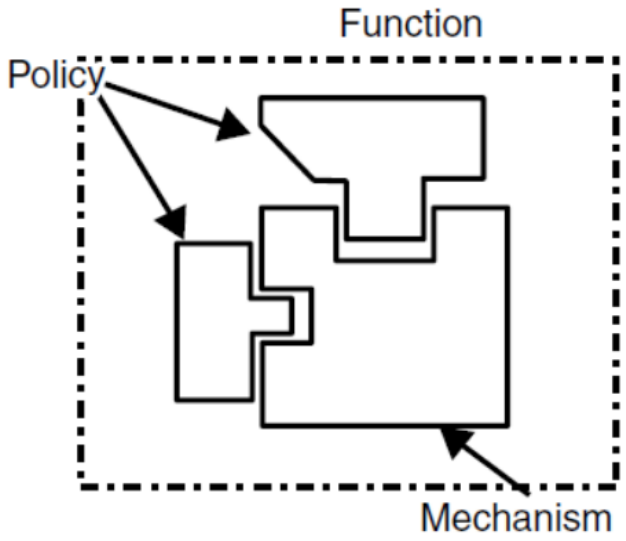
- **association** - minimální sdílený stav (např. UDP)
- **flow** - sdílený stav bez vázaných elementů reprezentovaný protokoly s two-way handshakem
- **connection** - sdílený stav s vázanými elementy reprezentovaný protokoly s three-way handshakem poskytujícím zpětnou vazbu (např. TCP)
- **binding** - totální synchronizace stavu na úrovni sdílení paměťového prostoru

Výběr policy

Jak měnit vlastnost mechanismu?

Mechanismus je ta část protokolu, která je neměnná

Policy = část protokolu, kteá je parametrizovatelná (např. jaký CRC polynom se bude používat na kontrolu chyb)



Adresování

Adresování objektu nelze bez identifikace objektu a naopak!

V případě multi-access prostředí (více potenciálních příjemců PDU), musí být protokol schopen specifikovat konkrétního příjemce

- v případěě point-to-point prostředí je to zbytečnost (např. PPP, HDLC)

Adresa = identifikátor, jehož jedinečnost je zaručena v daném dosahu

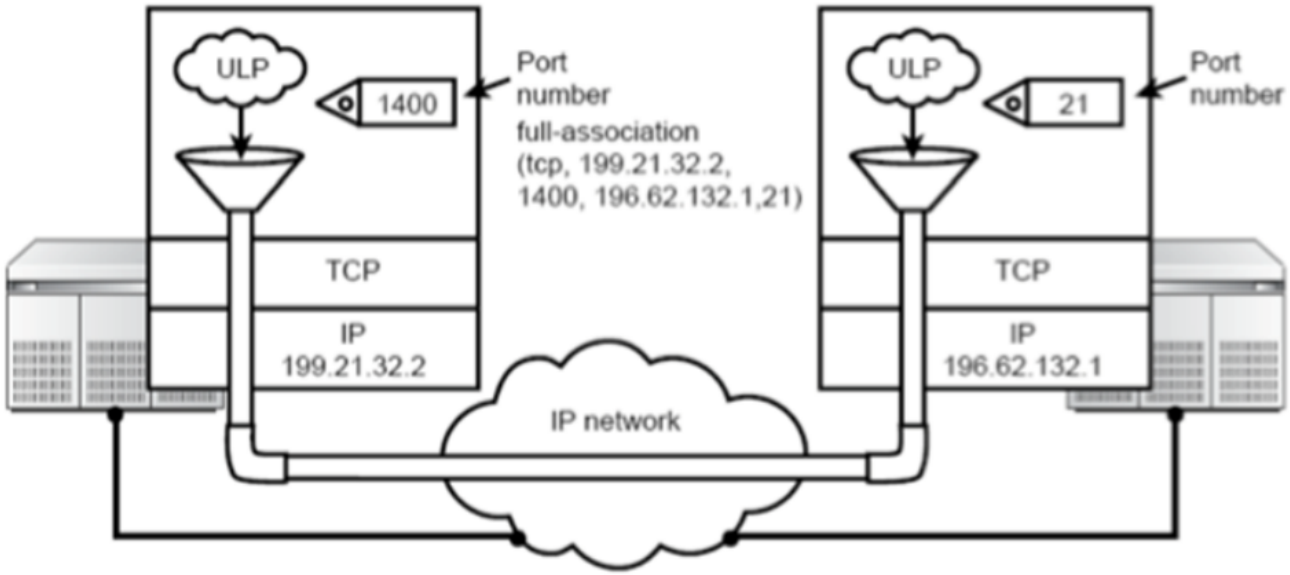
- hierarchické vs. nehierarchické
- location-dependent vs. location-independent

Identifikátor spojení

Každý protokol, který umožňuje provoz více svých instancí na jednom zařízení, potřebuje nějak mezi nimi rozlišovat

Port

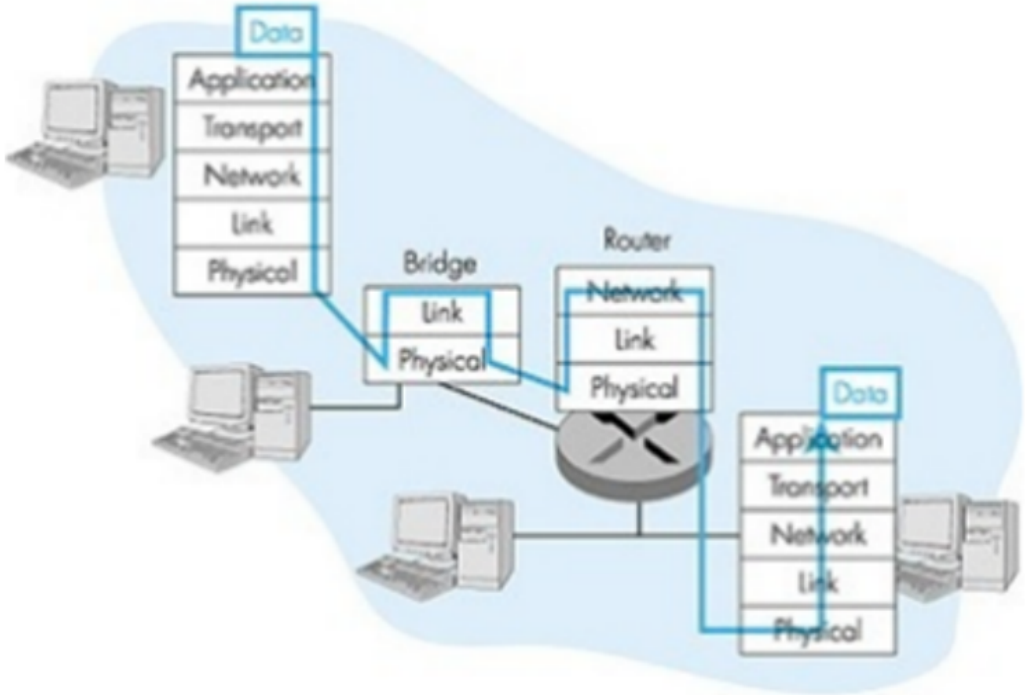
Dvojice zdrojového a cílového čísla portu je dostatečným identifikátorem spojení



Relaying

Relaying = Předávání PDU mezi dvěma PM

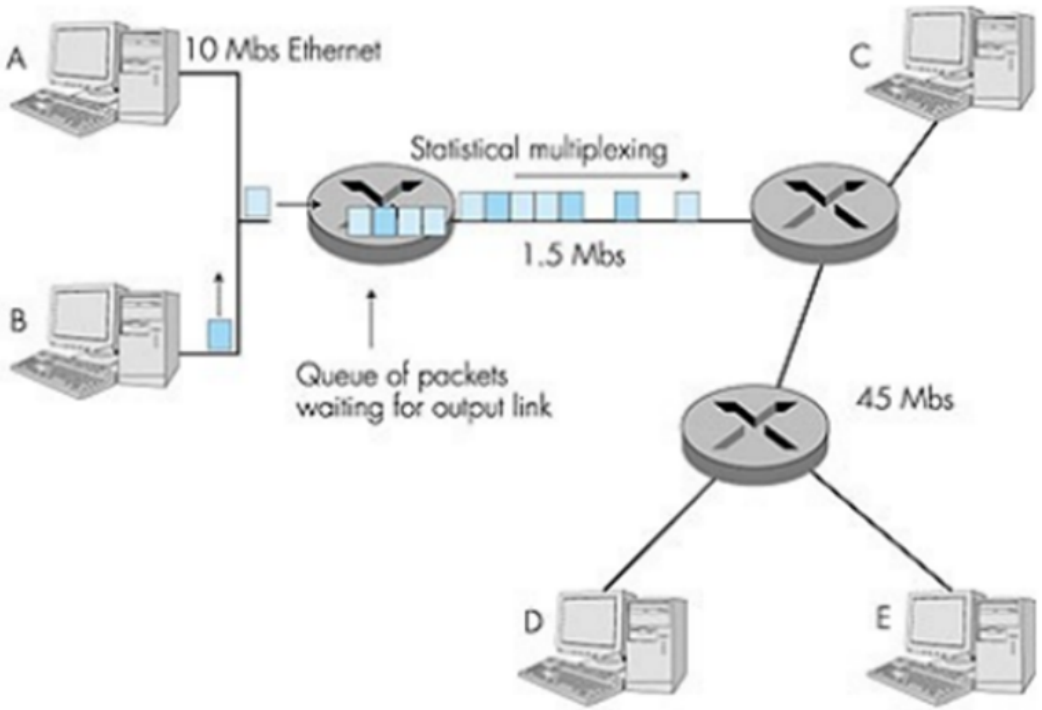
- předávání mezi (N)-PM a (N+1)-PM ve chvíli, kdy je (N+1)-PM příjemcem PDU
- **routing** = proces hledání a určení patřičného (N-1)-PM
- **forwarding** = proces předání (N)-PDU patřičnému (N-1)-PM, aby se dostal blíže k cíli



Multiplexing

Jak se vypořádat s přístupem ke sdílenému médiu?

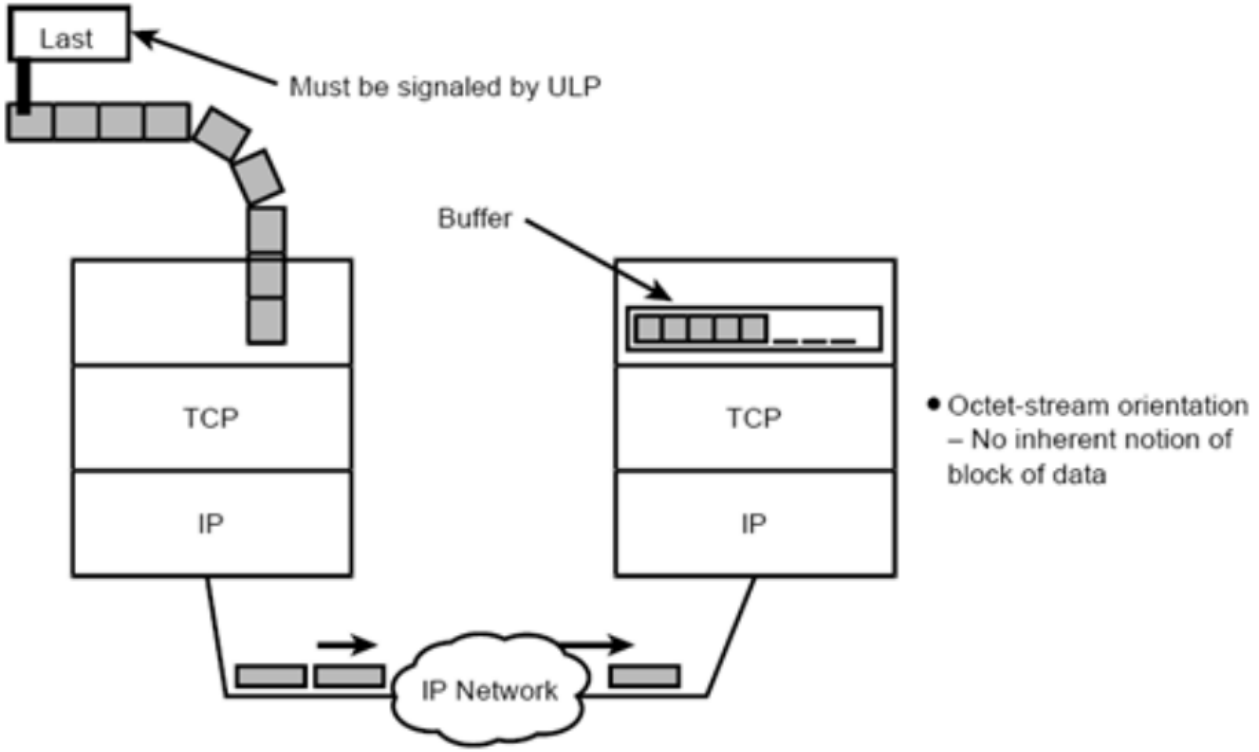
Multiplexing = proces komunikace několika různých (N)-PM skrz (N-1)-PM



Pořadí

Jak zajistit příjem paketů ve stejném pořadí, jako byly odeslány?

Sekvenční čísla = označkování kvant dat pořadovým číslem



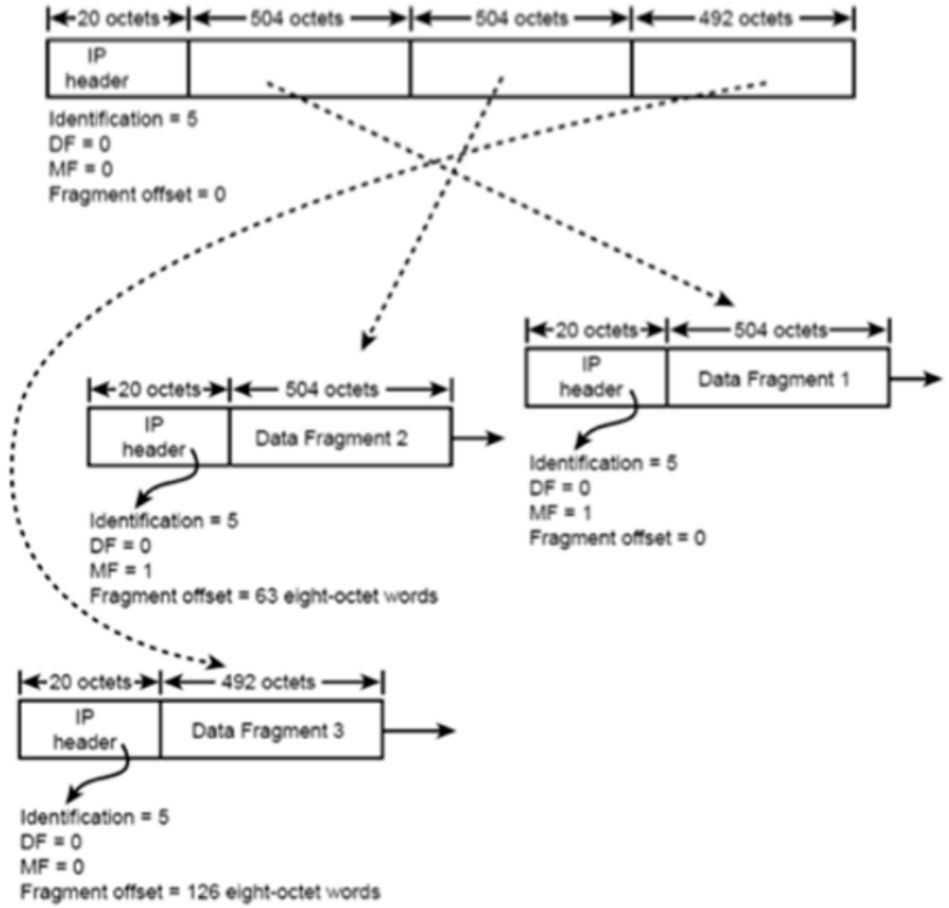
Fragmentace/Reassembling

Linkové technologie limitují velikost PDU. Co když je potřeba přenést větší množství informací?

Fragmentace (např. IP)
= rozdělení jedné velké (N)-PDU do několika menších (N)-PDU

Segmentace (např. TCP)
= rozdělení jedné (N)-PDU do několika menších (N-1)-PDU

Reassembling = proces opětovného složení dohromady

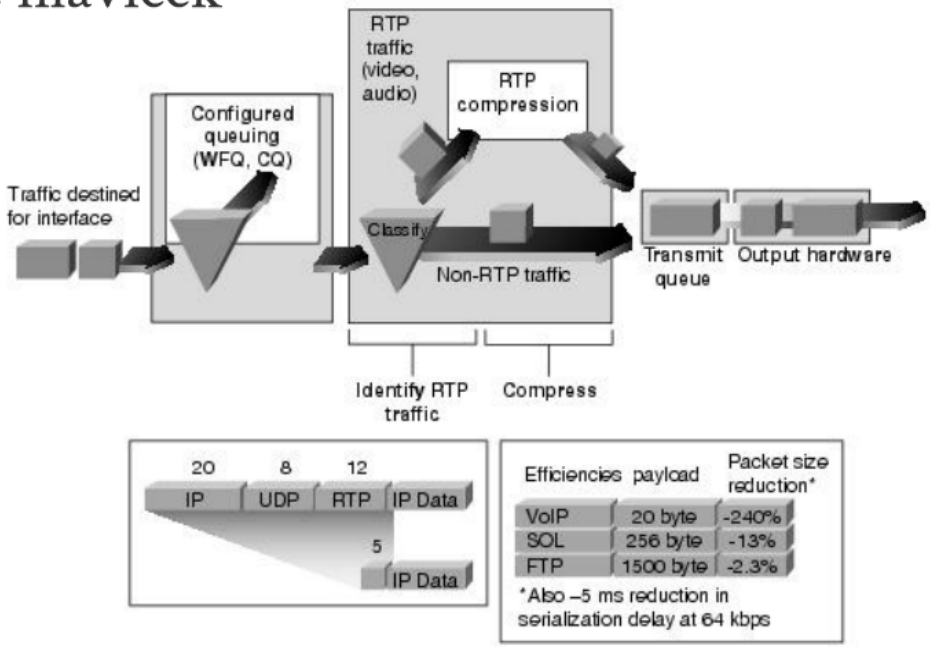


Komprese

Jak poslat linkou s omezenou propustností více dat? Jak ušetřit na protokolové režii?

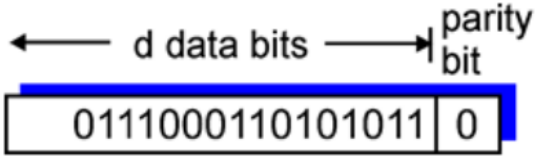
FTP komprese přenosu

RTP komprese hlaviček



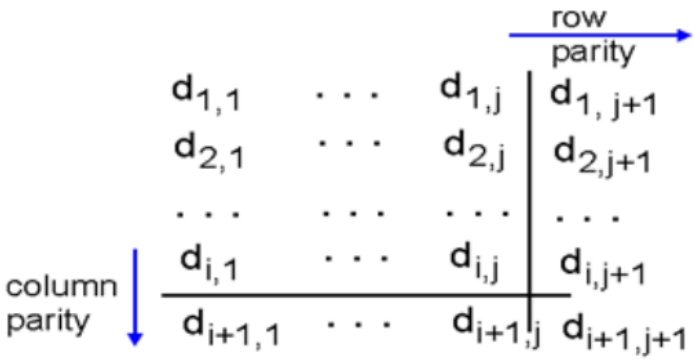
Detekce a korekce chyb

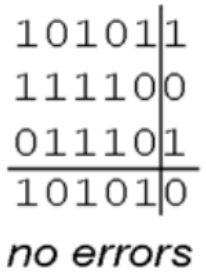
Přenosová média jsou nespolehlivá (podléhají např. EMI), a proto je nutné umět se vypořádat s bitovými chybami!

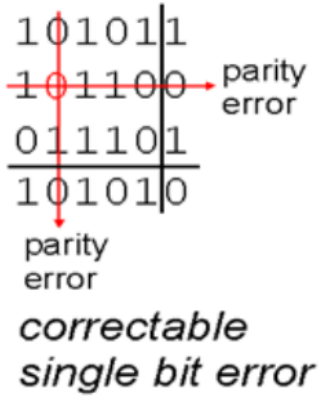


Error detection = schopnost detekovat chybu při přenosu (např. CRC)

Error correction = schopnost opravit chybu při přenosu (např. parita, Vitrebiho algoritmus)





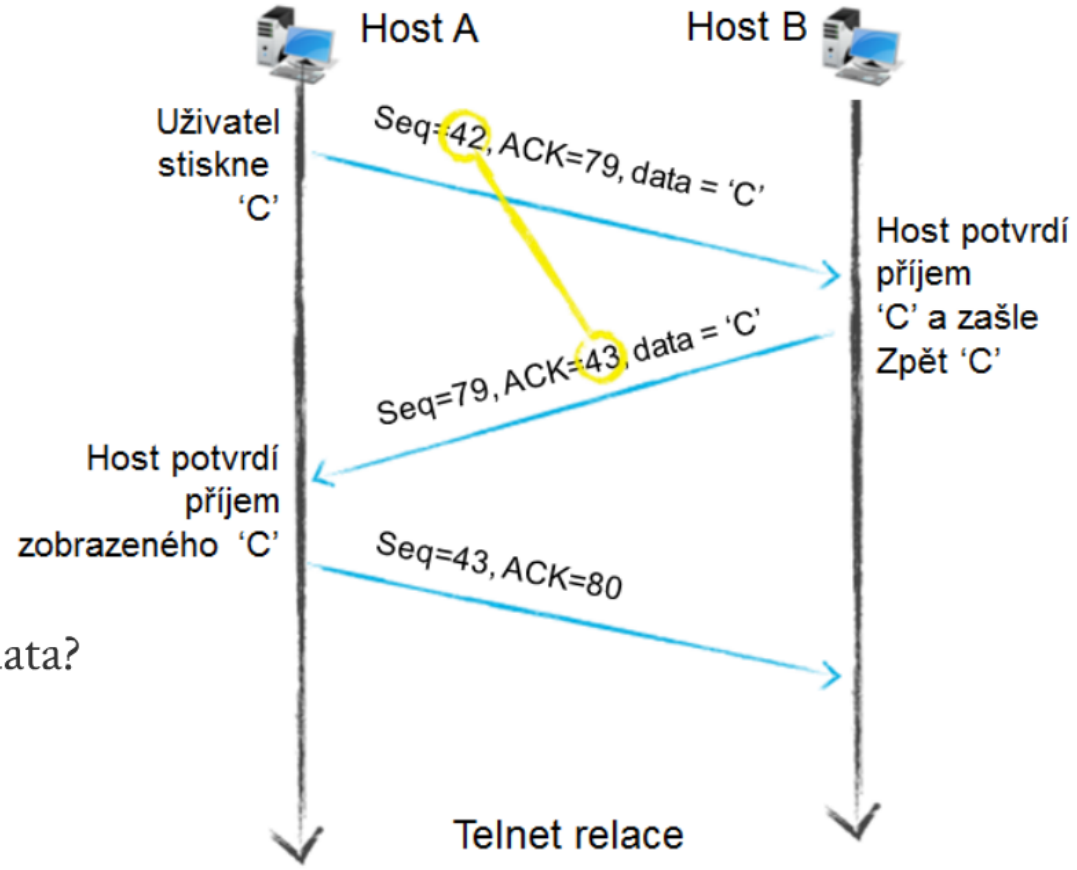


Duplicita, ztráta, potvrzování, znovuzaslání

Pakety v síti putují nedeterministicky, může dojít k jejich zdvojení (multicast) či ztrátě. Příjemce musí být schopen tyto stavy detekovat a vypořádat se s nimi.

Sekvenční čísla detekují "díry" v pořadí doručování **Ack** a **Nack** pak slouží k informování o vzniklé situaci

- Co když se ztratí data?
- Co když se ztratí potvrzení?
- Co když se ztratí znovuzaslaná data?



Řízení zahlcení

Protože používáme "hloupou" datagramovou síť, musí se o řízení zahlcení starat sami uživatelé...

Řízení zahlcení = aby odesilatel nepřetížil příjemce či samu síť daty

- Kreditové schéma, kdy příjemce říká odesilateli kolik dat naráz může maximálně zasílat (např. TCP)
- Pacing schéma, kdy příjemce říká explicitně jakou rychlostí se data mají posílat (např. leaking bucket)

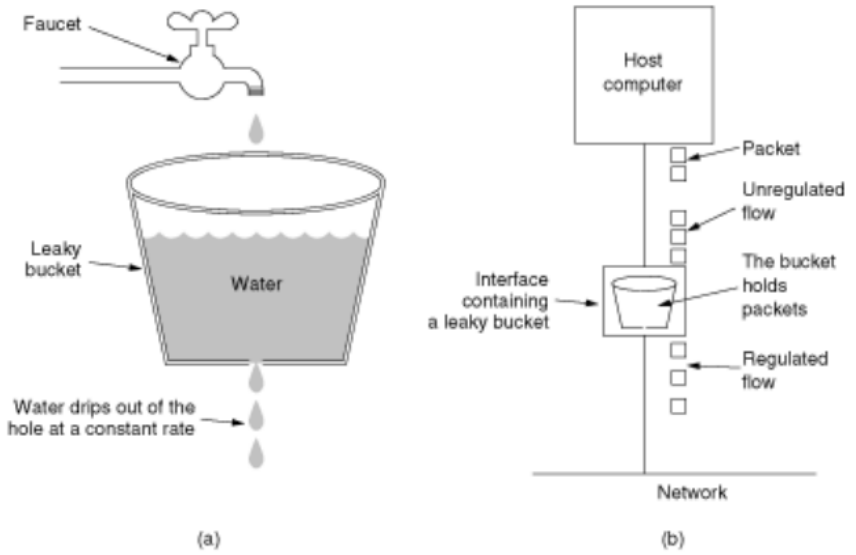


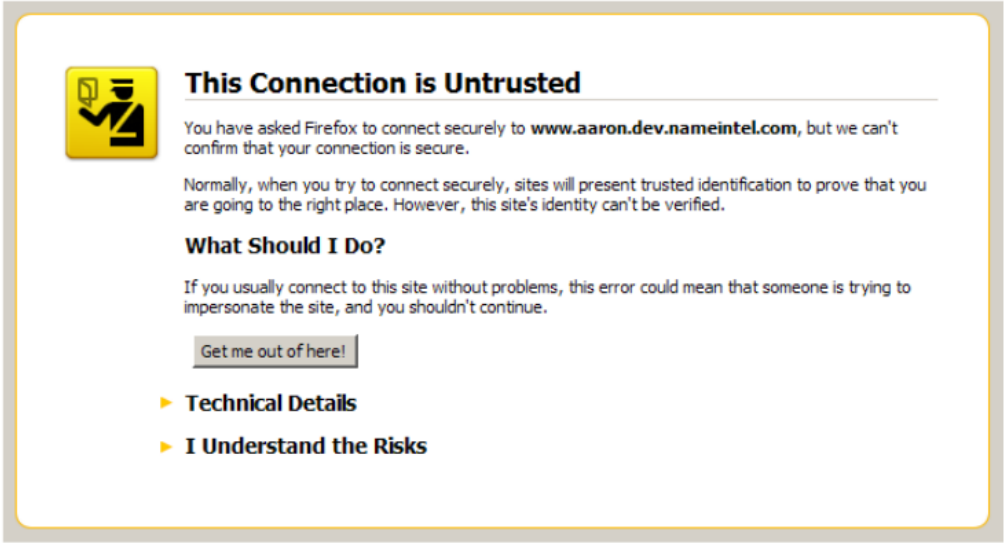
Fig. 5-24. (a) A leaky bucket with water. (b) A leaky bucket with packets.

Autentizace a řízení přístupu

Jak ověřit, že komunikující je opravdu tím, kým je?

Autentizace = určení identity (např. IKE, 802.1X)

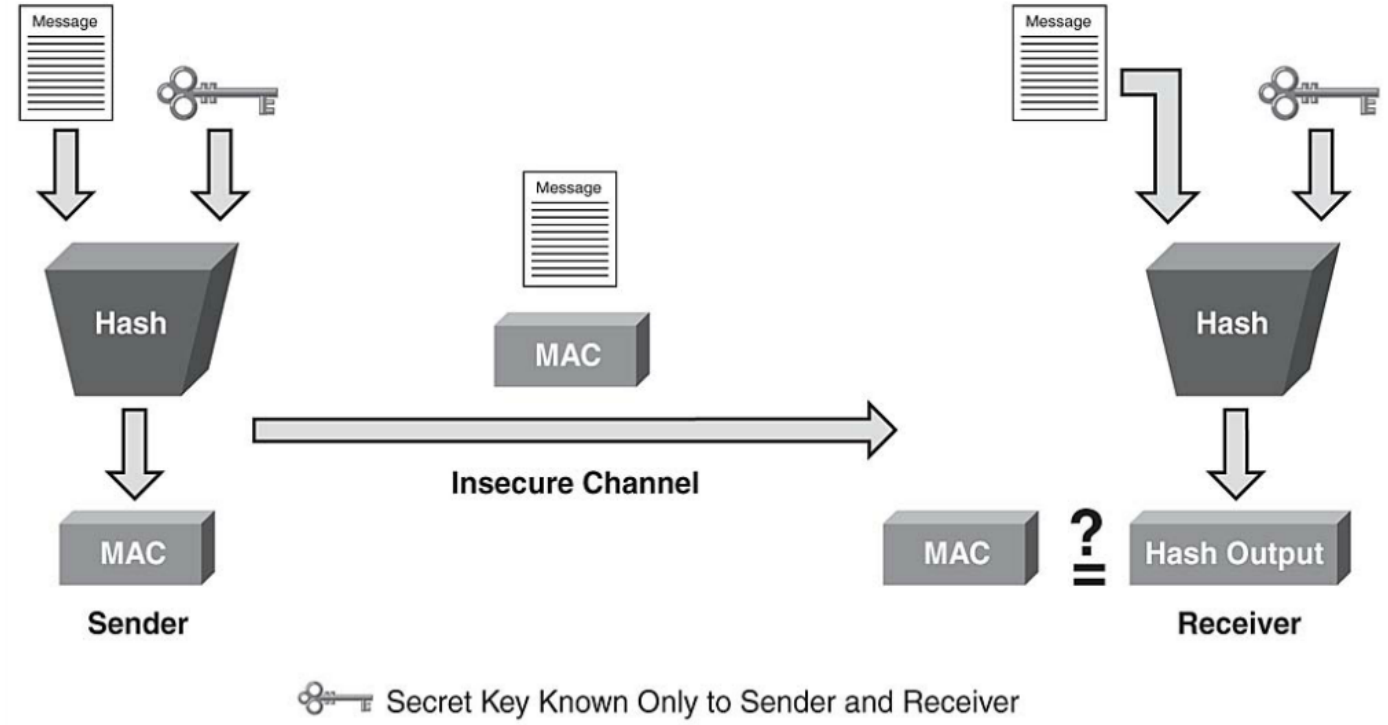
- username/password
- certifikát
- OTP



Integrita

Jak zabezpečit data proti manipulaci?

Integrita = mechanismus komunikace zabraňující změně dat při přenosu po nespolehlivém kanálu (např. HMAC)



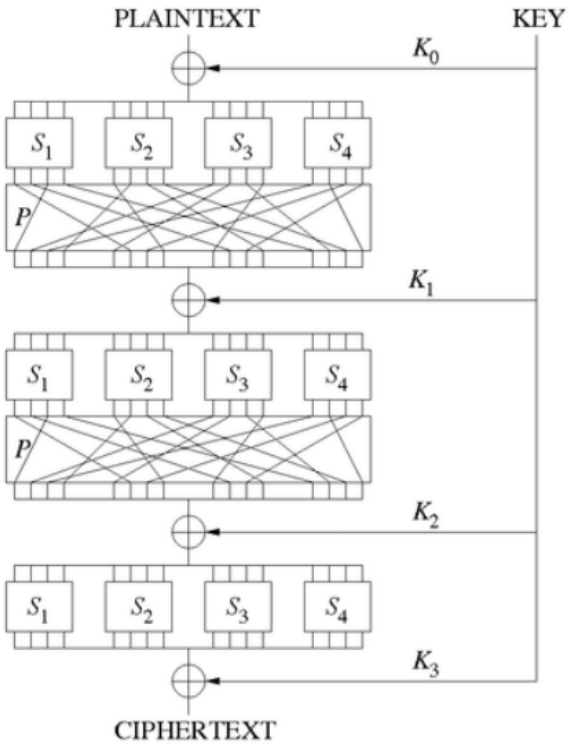
Důvěrnost

Jak zabezpečit data proti čtení cizími?

Důvěrnost (confidentiality) = mechanismus ochraňující komunikaci proti odposlechnutí

Šifrování komunikace

- Symetrická kryptografie (např. DES, 3DES, AES, RC4)
- Asymetrická (např. RSA, EC)



Nepopiratelnost a neopakovatelnost

Nepopiratelnost (nonrepudition) = mechanismus zabezpečující, že příjemce nebo odesilatel nemohou tvrdit, že komunikace neproběhla (že data neodeslali/nepřijali)

- účtování komunikace (NetFlow, data-retention)

Neopakovatelnost (nonreplayability) = mechanismus bránící odposlechnutá data (byť obrněná integritou a důvěrností) znovupoužít

- nonce

Keepaliveness

Jak se korektně vypořádat s komunikačním "tichem"?

Keepalive = mechanismus zajišťující kontrolu dostupnosti komunikujících stran

- Hello packety (např. směrovací protokoly)
- Blank packety (např. IPsec, GRE)
- Piggybacking (např. LISP)

Co z toho vyplývá?

Katalog mechanismů je velký, ale přeci jen konečný... 

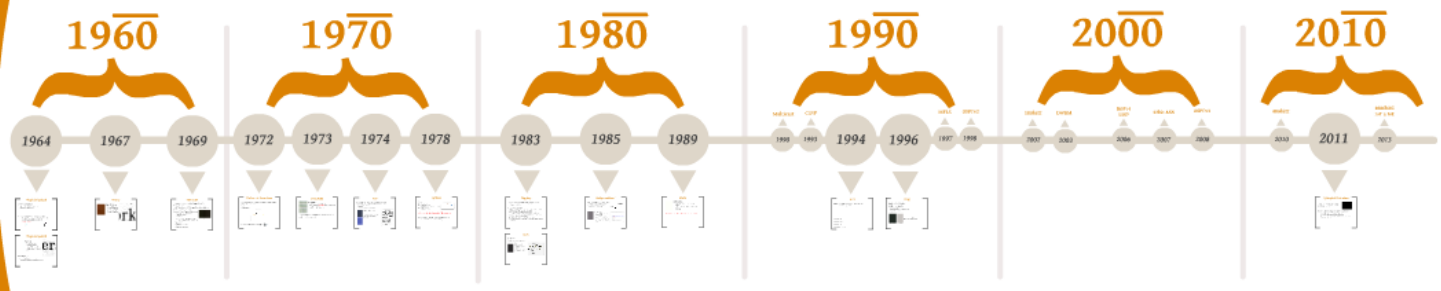
Množství policy však může být prakicky neomezené... 

Komunikační architekturu a její vlastnosti odráží kombinace PM mechanismů a policy, kterých využívá!

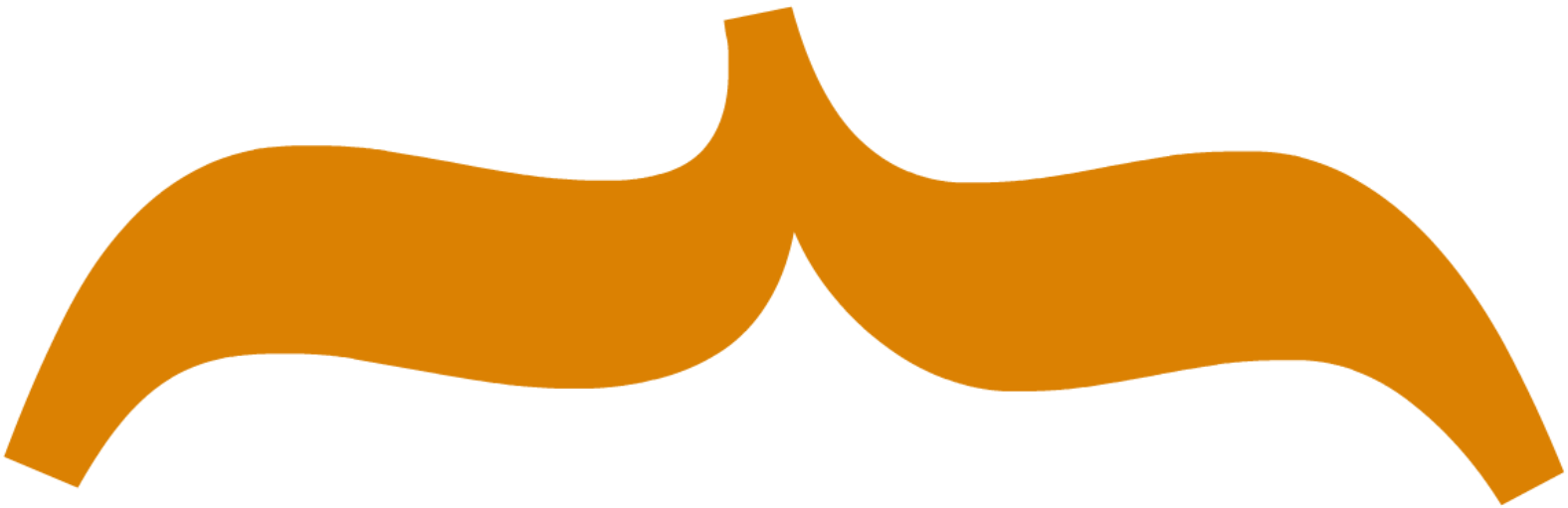
Architektura Internetu je jen tak dobrá, jako jsou PM, které má k dispozici...

JAK SE INTERNET VYVINUL?

HISTORIE


1960
1970
1980
1990
2000
2010
1964
1967
1969
1972
1973
1974
1978
1983
1985
1989
1994
1996
2011

1960



1964

1967

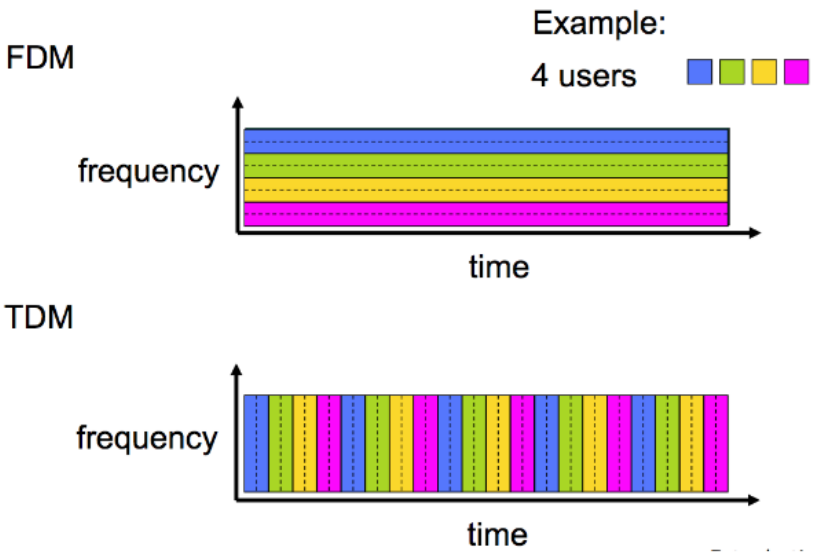
1969

1964

Přepínání paketů

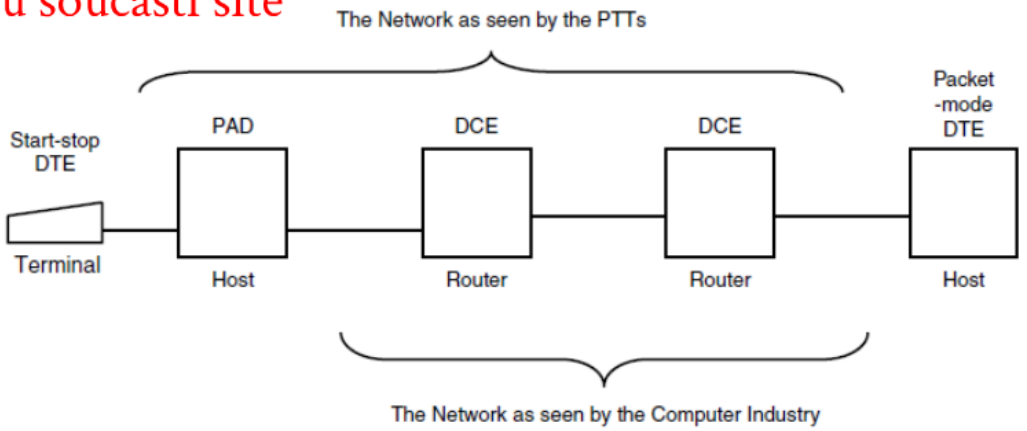
Connection-oriented služba

- telefonie je historicky spojově orientovaná
- TDM a FDM pro sílený přístup



Na konci 60.let měly na poli přenosu informací monopol IBM a PTT

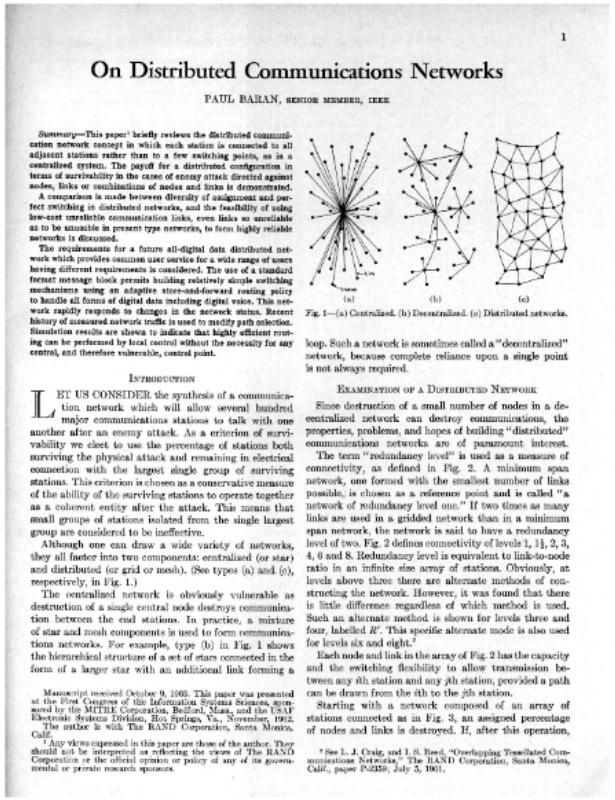
- v Evropě PTT státní, v Americe AT&T jako komerční subjekt
- IBM dodavatelem hierarchické SNA architektury
- služby jsou součástí sítě



Přepínání paketů

Paul Baran

- koncepce přepínaných sítí (a connection-less služby)
- data rozbity na malé části, které v síti mohou putovat nezávisle
- virtuální okruhy + datagramové sítě



1

On Distributed Communications Networks

PAUL BARAN, SENIOR MEMBER, IEEE

Summary—This paper[1] briefly reviews the distributed communication network concept in which each station is connected to all adjacent stations rather than to a few switching points, as in a centralized system. The payoff for a distributed configuration in terms of survivability in the cases of enemy attack directed against nodes, links or combinations of nodes and links is demonstrated.

A comparison is made between diversity of assignment and perfect switching in distributed networks, and the feasibility of using low-cost unreliable communication links, even links so unreliable as to be unusable in present type networks, to form highly reliable networks is discussed.

The requirements for a future all-digital data distributed network which provides common user service for a wide range of users having different requirements is considered. The use of a standard format message block permits building relatively simple switching mechanisms using an adaptive store-and-forward routing policy to handle all forms of digital data including digital voice. This network rapidly responds to changes in the network status. Recent history of measured network traffic is used to modify path selection. Simulation results are shown to indicate that highly efficient routing can be performed by local control without the necessity for any central, and therefore vulnerable, control point.

(a) (b) (c)

Fig. 1—(a) Centralized. (b) Decentralized. (c) Distributed networks.

INTRODUCTION

LET US CONSIDER the synthesis of a communication network which will allow several hundred major communications stations to talk with one another after an enemy attack. As a criterion of survivability we elect to use the percentage of stations both surviving the physical attack and remaining in electrical connection with the largest single group of surviving stations. This criterion is chosen as a conservative measure of the ability of the surviving stations to operate together as a coherent entity after the attack. This means that small groups of stations isolated from the single largest group are considered to be ineffective.

Although one can draw a wide variety of networks, they all factor into two components: centralized (or star) and distributed (or grid or mesh). (See types (a) and (c), respectively, in Fig. 1.)

The centralized network is obviously vulnerable as destruction of a single central node destroys communication between the end stations. In practice, a mixture of star and mesh components is used to form communications networks. For example, type (b) in Fig. 1 shows the hierarchical structure of a set of stars connected in the form of a larger star with an additional link forming a loop. Such a network is sometimes called a "decentralized" network, because complete reliance upon a single point is not always required.

EXAMINATION OF A DISTRIBUTED NETWORK

Since destruction of a small number of nodes in a decentralized network can destroy communications, the properties, problems, and hopes of building "distributed" communications networks are of paramount interest.

The term "redundancy level" is used as a measure of connectivity, as defined in Fig. 2. A minimum span network, one formed with the smallest number of links possible, is chosen as a reference point and is called "a network of redundancy level one." If two fitnes as many links are used in a gridded network than in a minimum span network, the network is said to have a redundancy level of two. Fig. 2 defines connectivity of levels 1, 1½, 2, 3, 4, 6 and 8. Redundancy level is equivalent to link-to-node ratio in an infinite size array of stations. Obviously, at levels above three there are alternate methods of constructing the network. However, it was found that there is little difference regardless of which method is used. Such an alternate method is shown for levels three and four, labelled *R'*. This specific alternate mode is also used for levels six and eight.[2]

Each node and link in the array of Fig. 2 has the capacity and the switching flexibility to allow transmission between any *i*th station and any *j*th station, provided a path can be drawn from the *i*th to the *j*th station.

Starting with a network composed of an array of stations connected as in Fig. 3, an assigned percentage of nodes and links is destroyed. If, after this operation,

Manuscript received October 9, 1963. This paper was presented at the First Congress of the Information Systems Sciences, sponsored by the MITRE Corporation, Bedford, Mass., and the USAF Electronic Systems Division, Hot Springs, Va., November, 1962. The author is with The RAND Corporation, Santa Monica, Calif.

[1] Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors.

[2] See L. J. Craig, and I. S. Reed, "Overlapping Tessellated Communications Networks," The RAND Corporation, Santa Monica, Calif., paper P-2359, July 5, 1961.

Nesmiřitelný souboj!

- zákopová mentalita
- "Je lepší korekce chyb hop-by-hop nebo end-to-end?!

1967

Vrstvy

Edsger W. Dijkstra

- rozdělení komplexního systému do menších částí
- hlavní inspirace pro budoucí vývoj počítačových sítí
- služby jako součásti klientů

THE STRUCTURE OF THE "THE" - MULTIPROGRAMMING SYSTEM[1)]

door prof. dr. E. W. Dijkstra

Lezing gehouden voor het Nederlands Rekenmachine Genootschap op 27 oktober 1967 te Utrecht.

Summary

A multiprogramming system is described in which all activities are divided over a number of sequential processes. These sequential processes are placed at various hierarchical levels, in each of which one or more independent abstractions have been implemented. The hierarchical structure proved to be vital for the verification of the logical soundness of the design and the correctness of its implementation.

Introduction

Papers „reporting on timely research and development efforts" being explicitly asked for, I shall try to present a progress report on the multiprogramming effort at the Department of Mathematics at the Technological University, Eindhoven, the Netherlands.

Having very limited resources (viz. a group of six people of, on the average, half time availability) and wishing to contribute to the art of system design — including all the stages of conception, construction and verification — we are faced with the problem of how to get the necessary experience. To solve this problem we have adopted the following three guiding principles:

1) Select a project as advanced as you can conceive, as ambitious as you can justify, in the hope that routine work can be kept to a minimum; hold out against all pressure to incorporate such system expansions that would only result into a purely quantitative increase of the total amount of work to be done.
2) Select a machine with sound basic characteristics (e.g. an interrupt system to fall in love with is certainly an inspiring feature); from then onwards try to keep the specific properties of the configuration for which you are preparing the system out of your considerations as long as possible.
3) Be aware of the fact that experience does by no means automatically lead to wisdom and understanding; in other words, make a conscious effort to learn as much as possible from your precious experiences.

Accordingly, I shall try to go beyond just reporting what we have done and how, and I shall try to formulate as well what we have learned.

I should like to end the introduction with two short remarks on working conditions, remarks I make for the sake of completeness. I shall not stress these points any further.

The one remark is that production speed is severely degraded if one works with half time people who have other obligations as well. This is at least a factor four, probably it is worse. The people themselves lose time and energy in switching over, the group as a whole loses decision speed as discussions, when needed, have often to be postponed until all people concerned are available.

The other remark is that the members of the group (mostly mathematicians) have previously enjoyed as good students a university training of 5 to 8 years and are of Master's or Ph. D. level. I mention this explicity because at least in my country the intellectual level needed for system design is in general grossly underestimated. I am more than ever convinced that this type of work is just difficult and that every effort to do it with other than the best people is doomed to either failure or moderate success at enormous expenses.

The Tool and the Goal

The system has been designed for a Dutch machine, the EL X8 (N.V. Electrologica, Rijswijk (ZH)). Characteristics of our configuration are:

1) core memory cycle time 2.5 mms., 27 bits; at present 32K.
2) drum of 512K words, 1024 words per track, rev. time 40 ms.
3) an indirect addressing mechanism very well suited for stack implementation
4) a sound system for commanding peripherals and controlling of interrupts
5) a potentially great number of low capacity channels; ten of them are used (3 paper tape readers at 1000 char/sec; 3 paper tape punches at 150 char/sec; 2 teleprinters; a plotter; a line printer)
6) absence of a number of not unusual awkward features.

The primary goal of the system is to process smoothly a continuous flow of user programs as a service to the University. A multiprogramming system has been chosen with the following objectives in mind:

1) a reduction of turn around time for programs of short duration
2) economic use of peripheral devices
3) automatic control of backing store to be combined with economic use of the central processor
4) the economic feasibility to use the machine for those applications for which only the flexibility of a general purpose computer is needed but (as a rule) not the capacity nor the processing power.

The system is not intended as a multi-access system. There is no common data base via which independent users can communicate with each other: they

1) This paper has been presented at the ACM Symposium on Operating System Principles, held at Gatlinburg, Tennessee, oktober 1-4, 1967.

1969

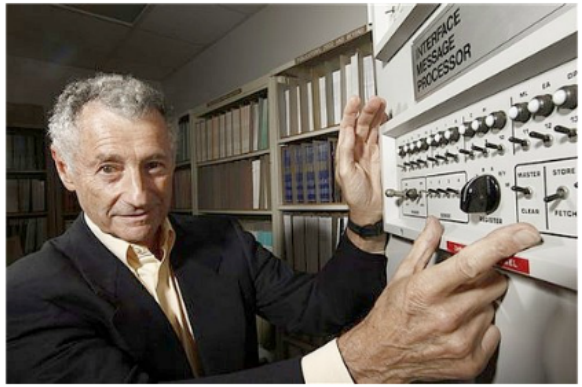
ARPANET

1963: memorandum o "Intergalactic Computer Network" od J.Licklidera z BBN

1968: B.Taylor jako vedoucí ARPA poptává kontrakt na výrobu sítě, kterou vyhrálo BBN Technologies

Interface Message Processor (IMP) = router

- 4 porty pro hosty, 5 portů pro IMP
- linky o rychlost 50 kbit/s
- 24 kB paměti



L.Kleinrock and 1st IMP

29. října 1969: První ARPANETový přenos

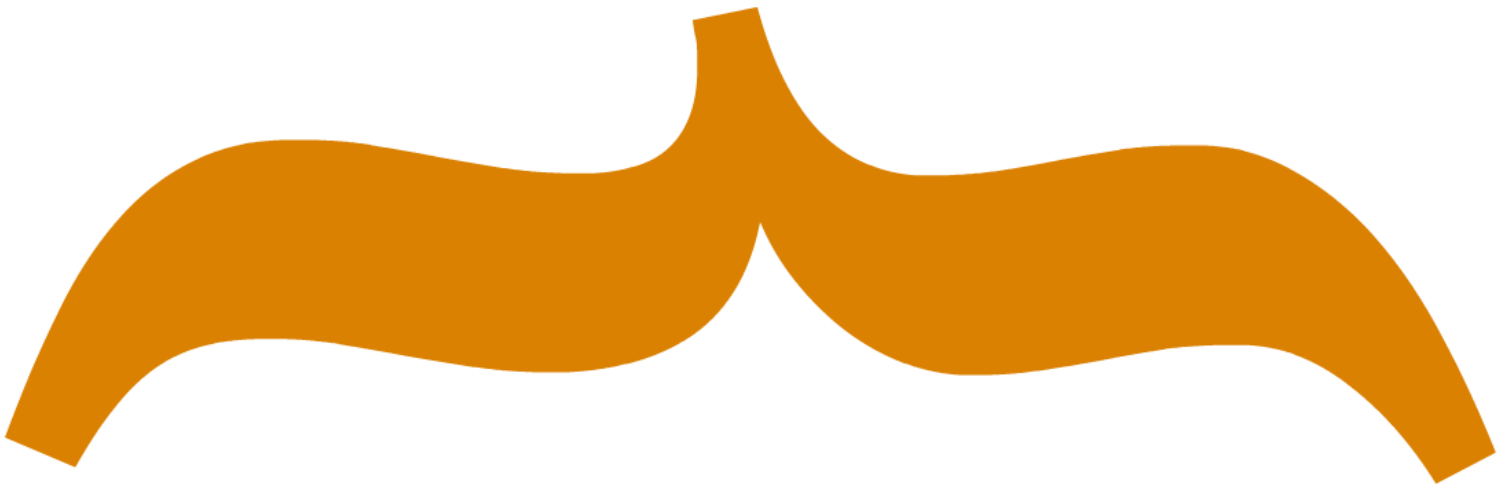
- 4 lokality (UCLA, SRI, UCSB, University of Utah)

1973: NORSAR

1975: prohlášen za funkční

1983: oddělený od MILNET

1970



1972

1973

1974

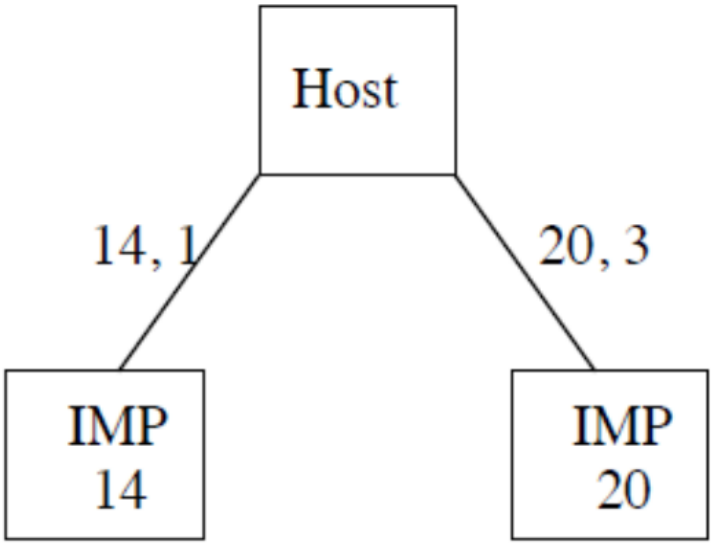
1978

1972

Tinker Air Force Base

Tinker Air Force Base v Oklahomě se rozhodla mít záložní připojení pro případ výpadku linky

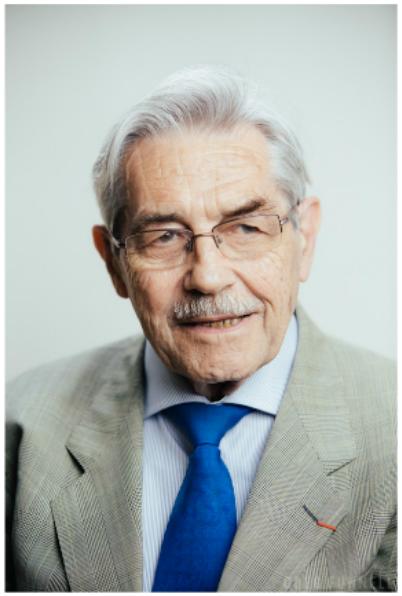
Multihoming



Adresa hosta odpovídá bodu jeho připojení!

1973

CYCLADES



Louis Pouzin

- francouzská alternativa ARPANETu
- čistě connection-less datagramová síť s hosty zodpovědnými za spolehlivé doručování
- CIGALE (L3) a TS (L4) protokoly

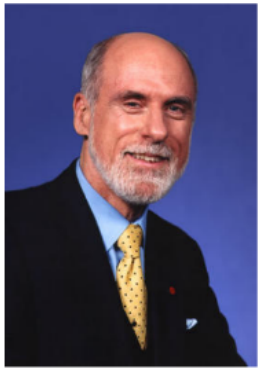
1976 příliš velkou konkurencí státnímu Transpacu (X.25 connection-oriented přístup) a vývoj pozastaven

1974

TCP

Host-to-Host a.k.a. **Network Control Program (NCP)**

- ustavení/ukončení spojení
- řízení toku
- dva simplexní kanály (liché/sudé porty)



Robert Kahn, Vint Cerf

Transmission Control Protocol (duplexní spojení, spolehlivá komunikace)

A Protocol for Packet Network Intercommunication

VINTON G. CERF AND ROBERT E. KAHN, MEMBER, IEEE

INTRODUCTION

IN THE LAST few years considerable effort has been expended on the design and implementation of packet switching networks [1]-[7],[14],[17]. A principle reason for developing such networks has been to facilitate the sharing of computer resources. A packet communication network includes a transportation mechanism for delivering data between computers or between computers and terminals. To make the data meaningful, computer and terminals share a common protocol (i.e. a set of agreed upon conventions). Several protocols have already been developed for this purpose [8]-[12],[16]. However, these protocols have addressed only the problem of communication on the same network. In this paper we present a protocol design and philosophy that supports the sharing of resources that exist in different packet switching networks.

After a brief introduction to internetwork protocol issues, we describe the function of a GATEWAY as an interface between networks and discuss its role in the protocol. We then consider the various details of the protocol, including addressing, formatting, buffering, sequencing, flow control, error control, and so forth. We close with a description of an interprocess communication mechanism and show how it can be supported by the internetwork protocol.

Even though many different and complex problems must be solved in the design of an individual packet switching network, these problems are manifestly compounded when dissimilar networks are interconnected. Issues arise which may have no direct counterpart in an individual network and which strongly influence the way in which internetwork communication can take place.

A typical packet switching network is composed of a set of computer resources called HOSTS, a set of one or more *packet switches*, and a collection of communication media that interconnect the packet switches. Within each HOST, we assume that there exist *processes* which must communicate with processes in their own or other HOSTS. Any current definition of a process will be adequate for our purposes [13]. These processes are generally the ultimate source and destination of data in the network. Typically, within an individual network, there exists a protocol for communication between any source and destination process. Only the source and destination processes require knowledge of this convention for communication to take place. Processes in two distinct networks would ordinarily use different protocols for this purpose. The ensemble of packet switches and communication media is called the *packet switching subnet*. Fig. 1 illustrates these ideas.

In a typical packet switching subnet, data of a fixed maximum size are accepted from a source HOST, together with a formatted destination address which is used to route the data in a store and forward fashion. The transmit time for this data is usually dependent upon internal network parameters such as communication media data rates, buffering and signalling strategies, routing, propagation delays, etc. In addition, some mechanism is generally present for error handling and determination of status of the network components.

Individual packet switching networks may differ in these implementations as follows.

1) Each network may have distinct ways of addressing the receiver, thus requiring that a uniform addressing scheme be created which can be understood by each individual network.

2) Each network may accept data of different maximum size, thus requiring networks to deal in units of the smallest maximum size (which may be unpractically small) or requiring procedures which allow data crossing a network boundary to be reformatted into smaller pieces.

3) The success or failure of a transmission and its performance in each network is governed by different time delays in accepting, delivering, and transporting the data. This requires careful development of internetwork timing procedures to insure that data can be successfully delivered through the various networks.

4) Within each network, communication may be disrupted due to unrecoverable mutation of the data or missing data. End-to-end restoration procedures are desirable to allow complete recovery from these conditions.

1978

Aplikace

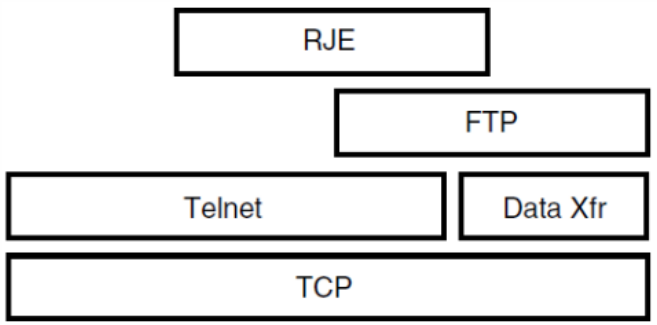
Telnet

File Transfer Protocol

- Email byl přenos speciálně označeného souboru

Remote Job Entry

- vzdálené spouštění úloh



Očekávala se vždy jen jedna instance běžící aplikace na počítači!

První setkání OSI v jak se ukáže později marné snaze sjednotit connection-oriented a connection-less tábory

1983

Flag Day

1.1.1983: Flag Day pro přechod na nový transportní protokol

- kdo nepřešel, měl smůlu...

Čtyři potenciální kandidáti:

- **TCP** – pořadí bytů, dynamické klouzavé okno, jen jedno PDU s kontrolními bity pro změnu sémantiky, piggybacking potvrzování
- **XNS Sequence Packet** a **CYCLADES TS** – pořadí paketů, dynamické klouzové okno, různá PDU na ustavení/uvolnění spojení, potvrzení a řízení toku, separace transportní a síťové vrstvy
- **Delta-t** – Radikální idea, že k zabezpečenému přenosu stačí mít zesynchronizované časovače, různá PDU pro potvrzení a řízení toku, separace transportní a síťové vrstv

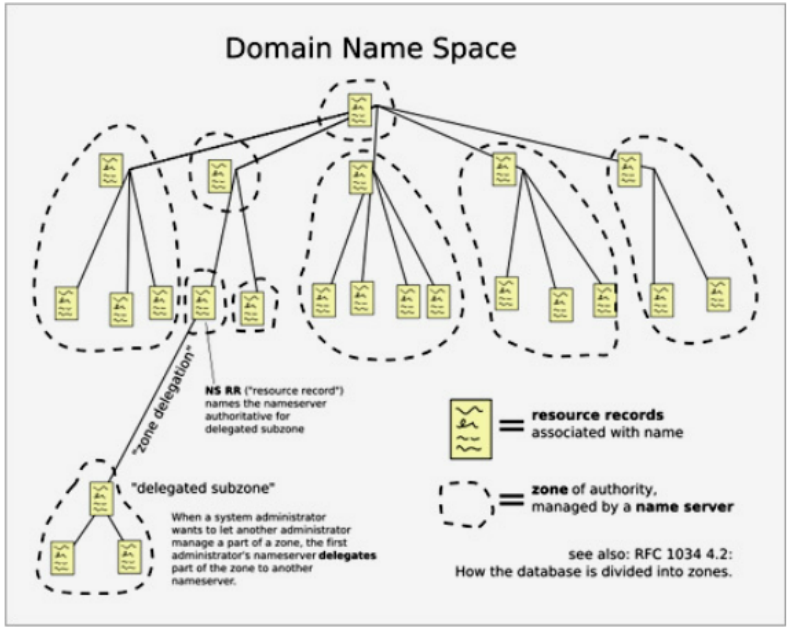
DNS

HOSTS.TXT

- seznam všech připojených zařízení k Internetu
- udržován SRI a pravidelně aktualizován

Paul Mockapetris

- hierarchická databáze jmen (aliasů) k IP adresám
- návrh DNS a první implementace
- 1985: BIND
- 1987: IETF standardizace

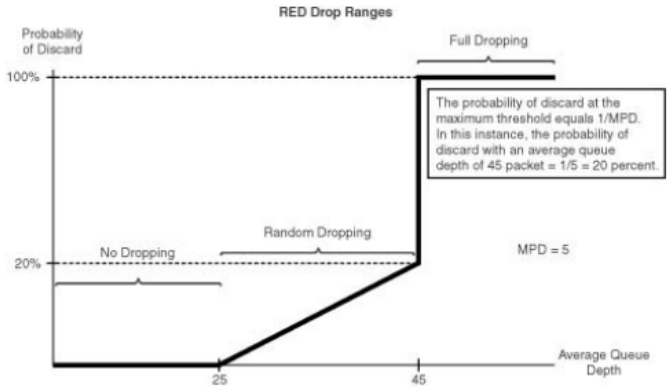


1985

Kolaps zahlcení

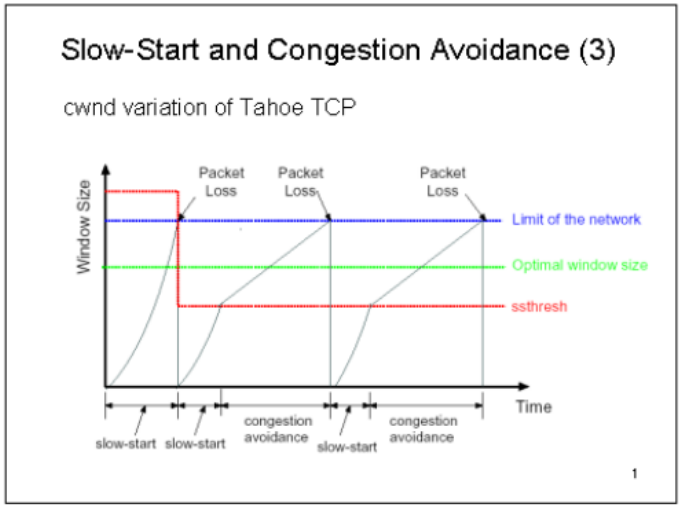
říjen 1986: první kolaps zahlcením linek

- pakety přicházejí rychleji, než se stíhají odbavovat
- zaplění bufferu vede automaticky k zahazování
- později předcházení pomocí RED a WRED



Van Jacobson

- úprava TCP zajišťující ochranu před zahlcením
- IF ztráta paket THEN zmenši window ELSE zvětši window
- TCP Tahoe, Reno, Vegas, New Vegas



1989

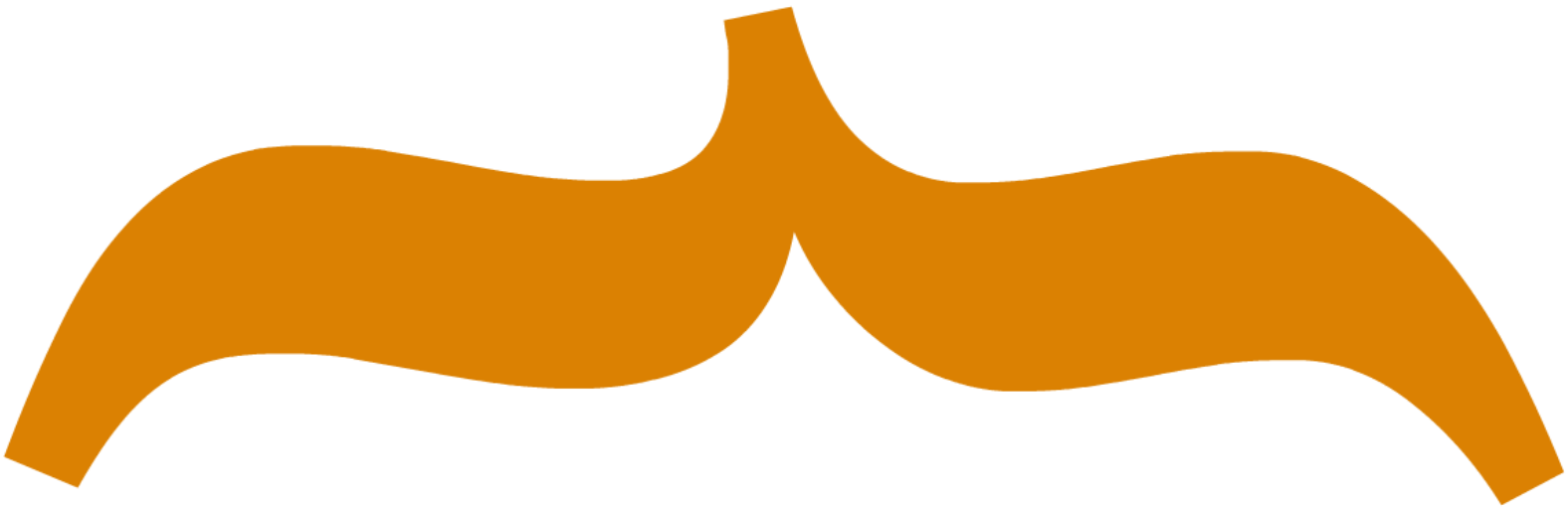
WWW

Tim Berners-Lee

- v CERNu pro výzkum linkování dokumentů jaderného výzkumu
- projekt World Wide Web (Hypertext documents)
- HTTP protokol
- HTML jazyk

Designově úplně nová aplikce - na jednom stroji více instancí téhož PM!!!

1990



Multicast

CLNP

MPLS

OSPFv2

1990

1993

1994

1996

1997

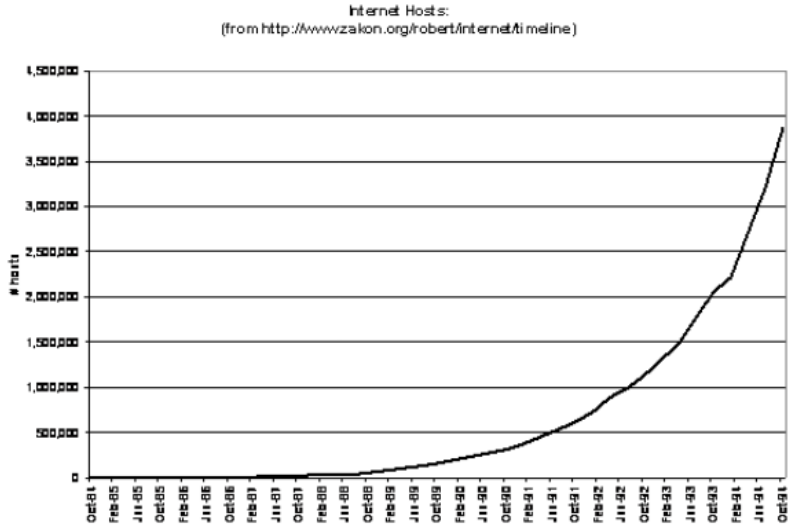
1998

1994

NAT

Evidentní neudržitelnost IPv4 při aktuální růstu klientů

- hlavní problém, že chybí adresy



RFC 1817 - CIDR

RFC 1876 - VLSM

RFC 1914 - NAT

RFC 1918 - privátní adresy

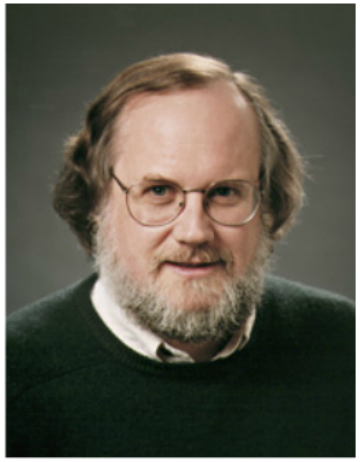
1996

IPng

Potřeba nové verze IP protokolu

Několik návrhů, vyhrál IPv6

- více adres (2^128)
- jedno rozhraní má víc než jednu IPv6 adresu
- unicast, multicast, anycast
- SLAAC

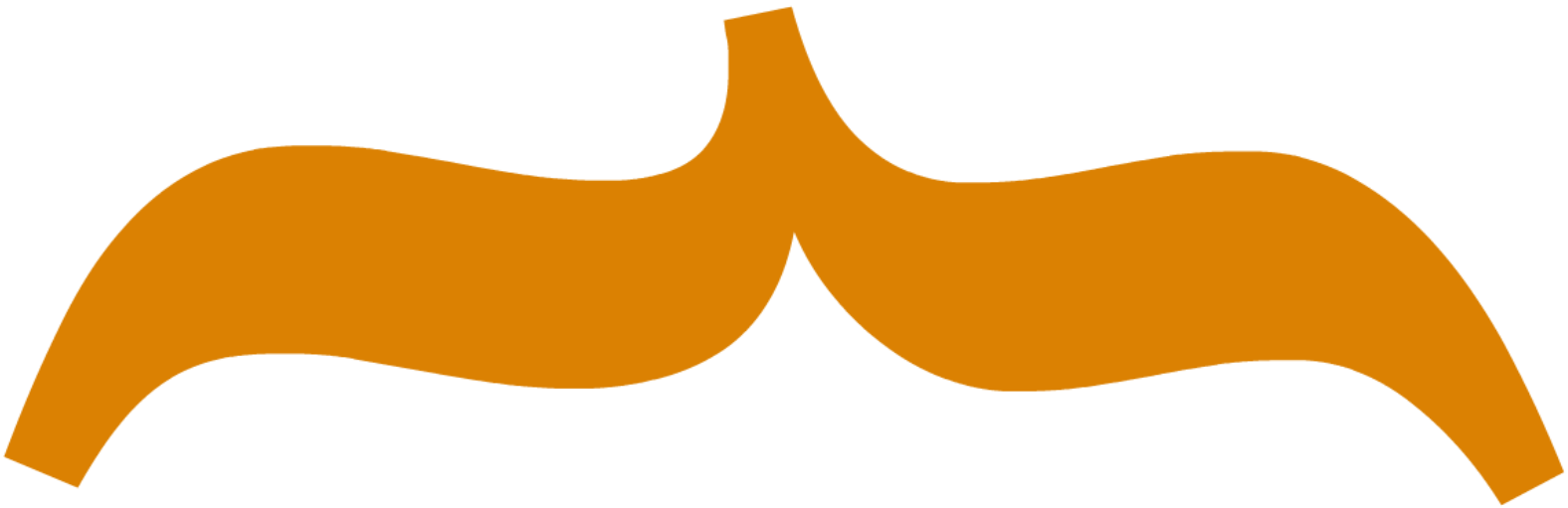


Steve Deering

Robert Hinden

prosinec 1998: **RFC 2460**

1990



Multicast

CLNP

MPLS

OSPFv2

1990

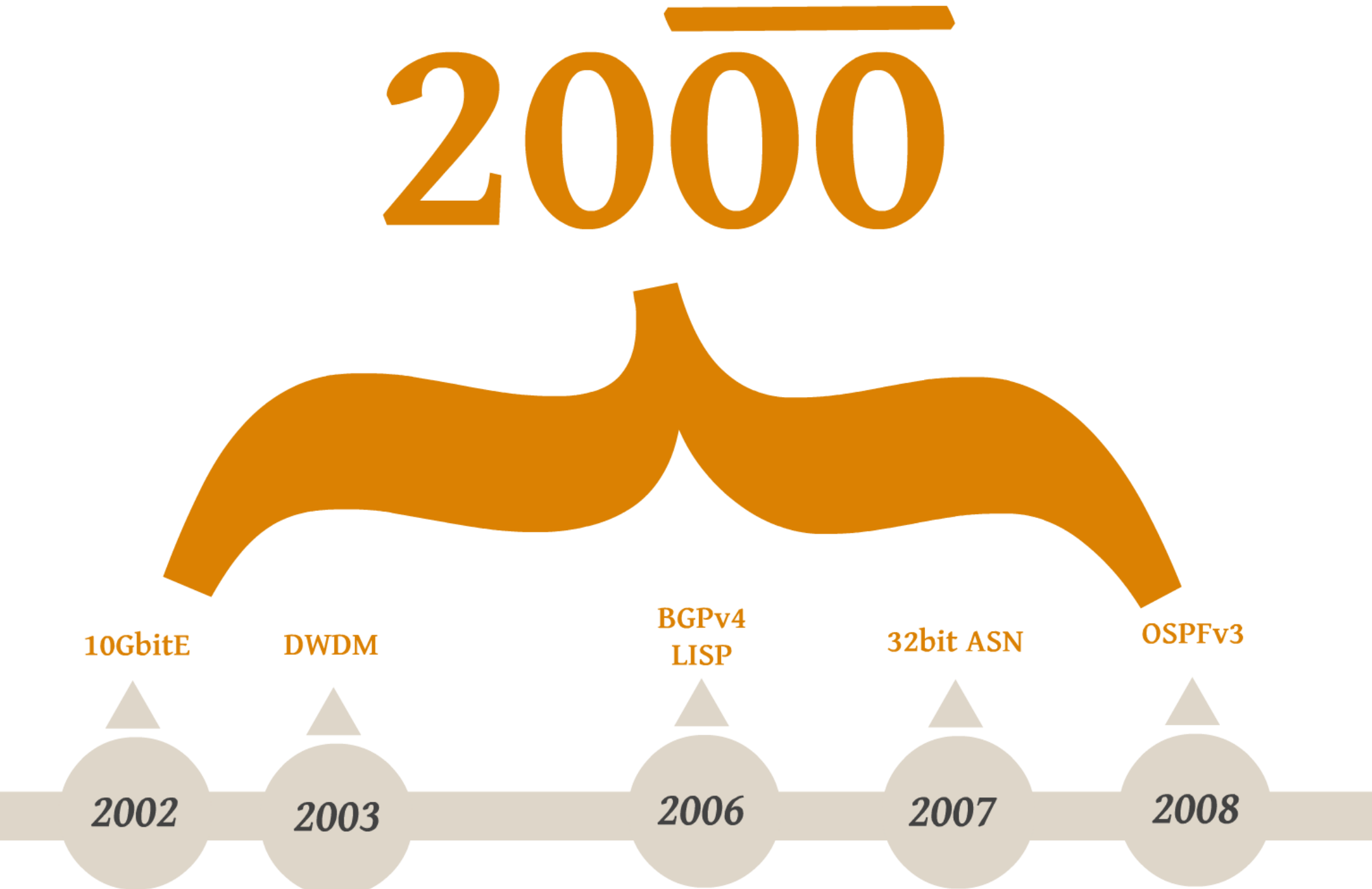
1993

1994

1996

1997

1998


2000
10GbitE
DWDM
BGPv4
LISP
32bit ASN
OSPFv3
2002
2003
2006
2007
2008


2010
40GbitE
100GbitE
IoT a IoE
2010
2011
2013

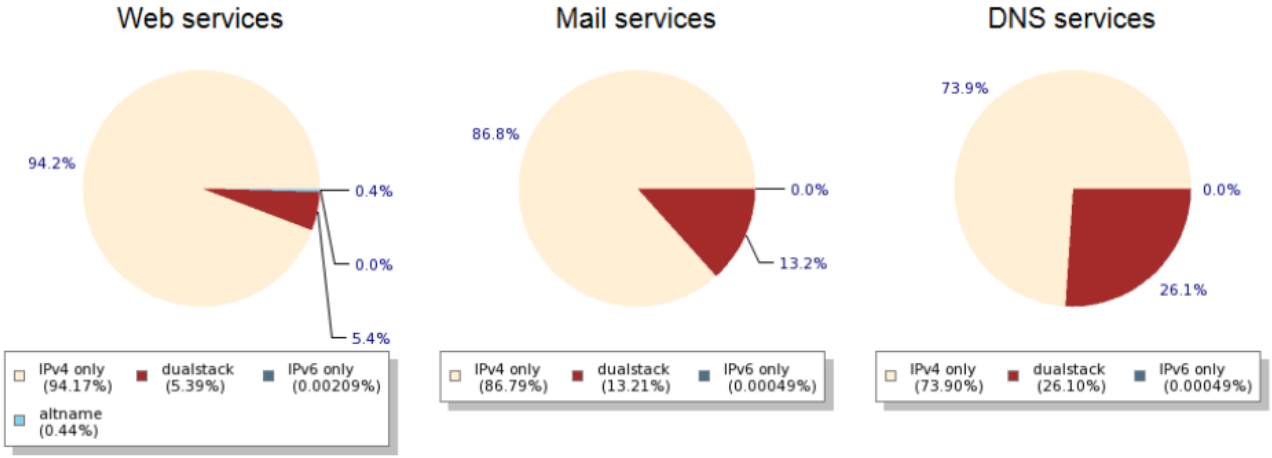
Vyčerpání IPv4 adres

31.ledna 2011: ICANN přidělil poslední nepřidělený blok adres

Dnes se IPv6 dostala ve "vyspělých" zemích na 4% penetrace

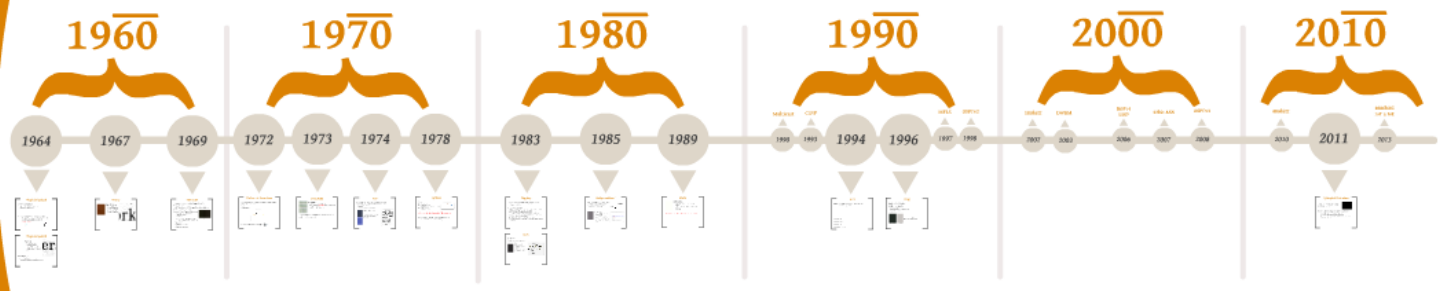
- spousta implementačních a provozních problémů
- na FIT VUT se dlouhodobě zabýváme výzkumem IPv6

Můj dojem je, že se tým na FIT VUT rozhodl, že IPv6 nemá rád, a proto místo řešení hledají problémy. Já myslím, že je pochopitelné, že implementační úroveň je zatím na horší úrovni (či je naprosto nesmyslně licencovaná jinak než IPv4), ale to přece neznamená, že je "Děravá IPv6". Osobně mne tenhle bulvární styl diskreditace IPv6 také velmi mrzí...



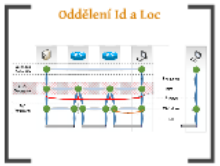

2010
40GbitE
100GbitE
IoT a IoE
2010
2011
2013

HISTORIE


1960
1970
1980
1990
2000
2010
1964
1967
1969
1972
1973
1974
1978
1983
1985
1989
1994
1996
2011

SOUČASNOST

Jak "dobrý" je dnešní Internet?



Škálovatelnost směrování

Default Free Zone

- páteř Internetu, která musí znát cestu do každého autonomního systému
- superlineární růst
- Jak výkoná taková zařízení musí být?

route-views.routeviews.org - PuTTY

```
route-views>show ip bgp summary
BGP router identifier 128.223.51.103, local AS number 6447
BGP table version is 2553610571, main routing table version 2553610571
476241 network entries using 62863812 bytes of memory
13297503 path entries using 691470156 bytes of memory
2144240/80726 BGP path/bestpath attribute entries using 360232320 bytes of memory
1880346 BGP AS-PATH entries using 74812950 bytes of memory
55473 BGP community entries using 4318864 bytes of memory
158 BGP extended community entries using 4552 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 1193702654 total bytes of memory
Dampening enabled. 9366 history paths, 10493 dampened paths
BGP activity 2567218/2074573 prefixes, 201917073/188486374 paths, scan interval 60 secs
```

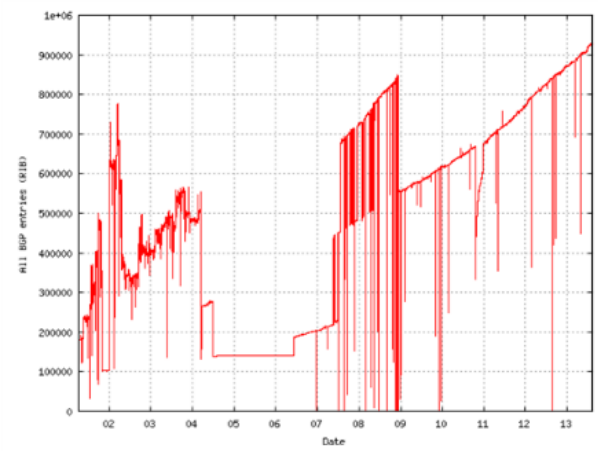


Fig. 1: IPv4 – All BGP entries in RIB

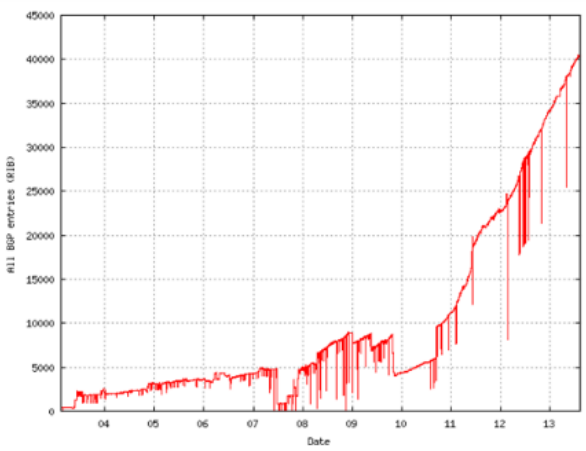


Fig. 1: IPv6 – All BGP entries in RIB

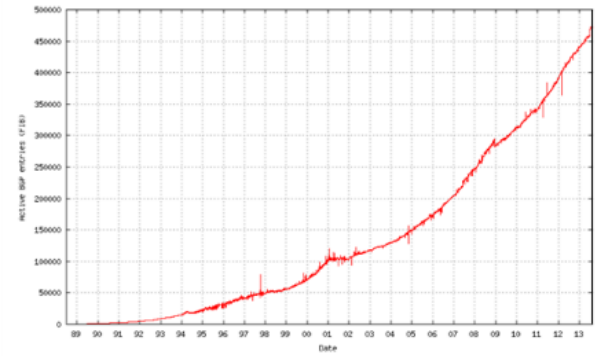


Fig. 2: IPv4 – Active BGP entries in FIB

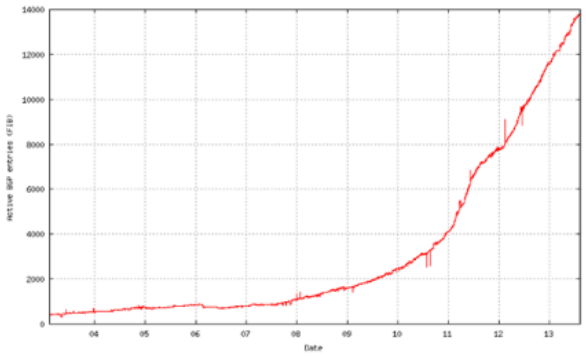


Fig. 2: IPv6 – Active BGP entries in FIB

Oddělení Id a Loc

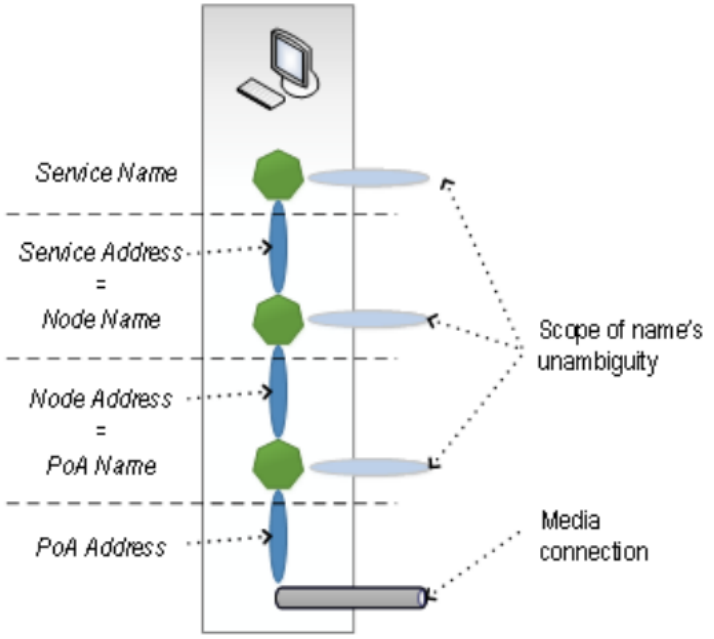
IP adresa přetěžuje svůj význam - poskytuje informaci o identifikaci i lokalizaci zaráz.

Co třeba jeden počítač se dvěma síťovkami, jaký je jeho identifikátor?

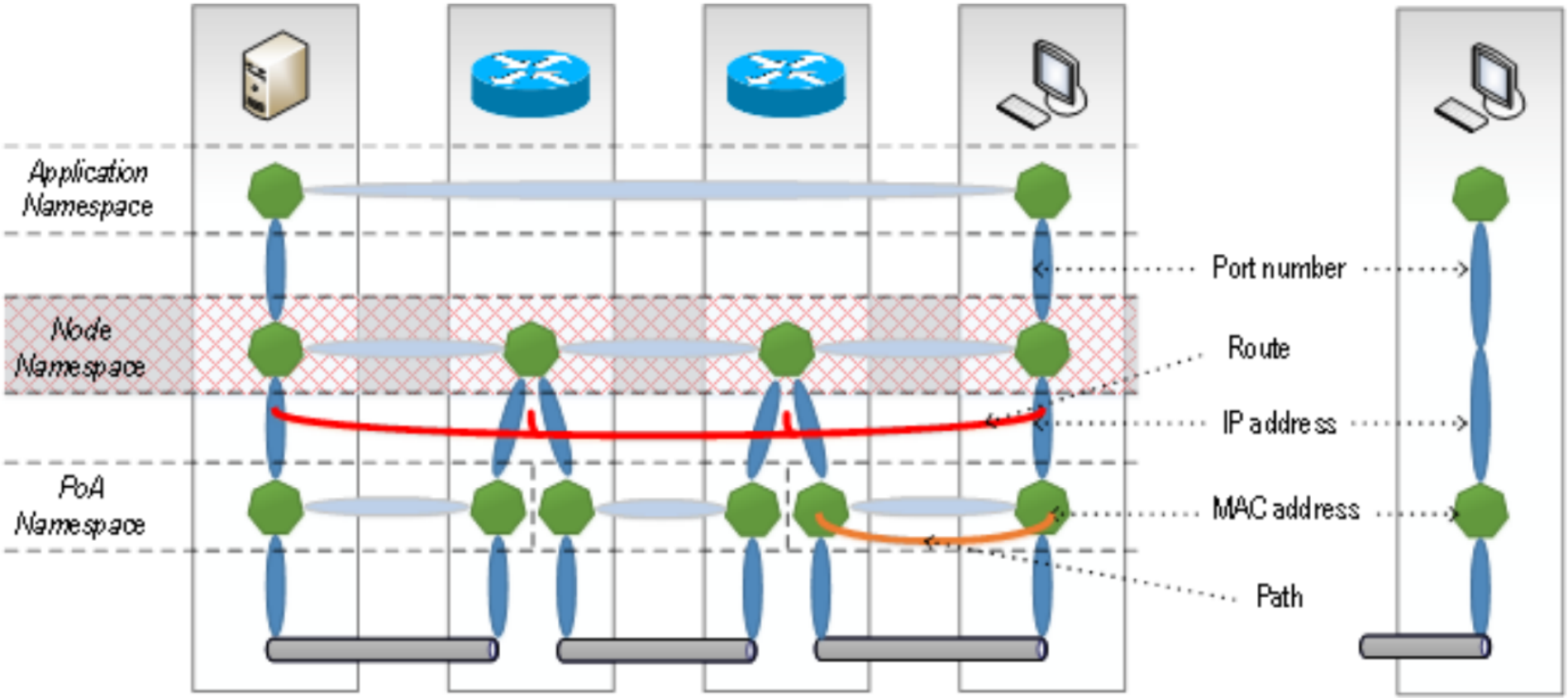
IP adresa není ničím víc, než bodem připojení (**Point of Attachment**)!

Jerome Saltzer

- v síti by se měli pojmenovat následující objekty: aplikace, uzly, PoA a cesty
- předpokládá se mobilita objektů
- adresovat bez identifikace nelze a vice versa



Oddělení Id a Loc



Multihoming a přečíslování

V řešení multihomingu, tedy záložního připojení k Internetu, jsme od dob Tinker Air Force Base před více než 30 lety nepostoupili...

Když chceme vyměnit ISP, musíme přečíslovat náš adresní prostor, protože dostaneme nový prefix! (**renumbering problem**)

Provider Aggregatable

- pokud máme PA adresy, provozuje se multihoming obtížně
- s PA adresou jsme rukojmím ISP

Provider Independent

- pro multihoming i odolnost proti přečíslování je nyní potřeba ASN
- když máme ASN šíříme prefixy do DFZ, přispíváme k jejímu růstu

Mobilita a Traffic Engineering

Mobilita = schopnost uzlu měnit své PoA bez výpadku komunikace

- Jak dynamicky změnit svou lokalitu bez změny identifikace?
- Jak oznámit nové PoA?
- Jak vytvořit tunel mezi starou a novou lokalitou?

Traffic Engineering = manipulace se směrováním provozu

- složité nastavování BGP politik
- masivně přispívá ke zvětšování DFZ
- inbound TE není nijak garantován

Vladimír Veselý
veselyv@fit.vutbr.cz

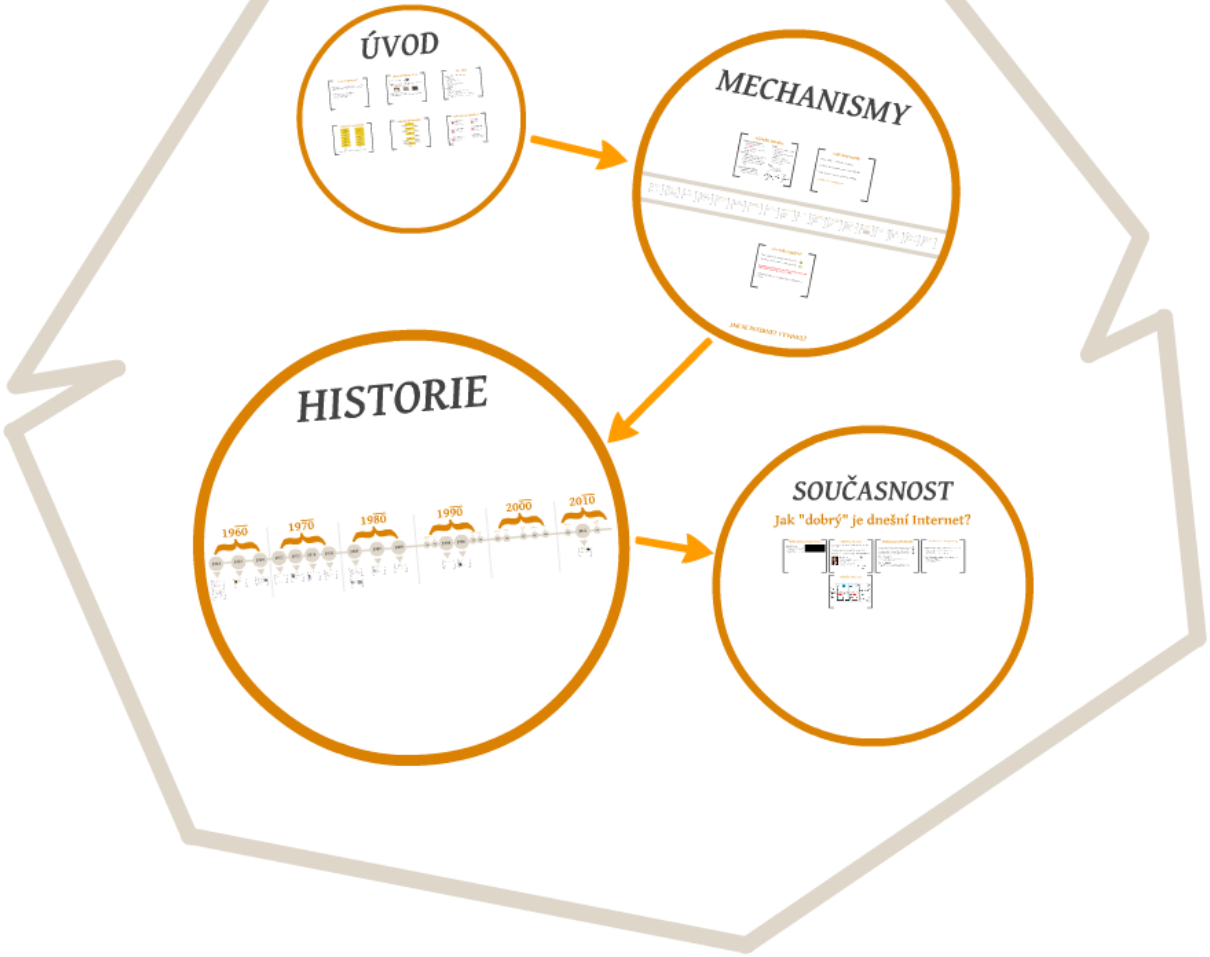
CCS 2017

Stručná historie internetové architektury

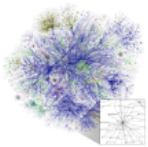
Vrchol inženýrství ...?

INTERNET

Ba ne, jen nedokončené demo!



Vrchol inženýrství ...?



INTERNET

Ba ne, jen nedokončené demo!