



新一代互联网的若干关键技术

苏金树 教授
国防科技大学

1



- 一、互联网面临的问题
- 二、发展网络体系结构技术
- 三、深度融合光电技术
- 四、充分应用并行技术

2



一、互联网面临的问题

3



□ 互联网取得巨大成就

- 是20世纪最伟大的工程成就
- 2005年互联网四位奠基者获得?? 奖

□ 针对“新一代”探索已经近十年

- “新一代”含义在逐步演变，可分为三个阶段
 - 提速阶段
 - 融合阶段
 - 创新阶段

4



“新一代”网络

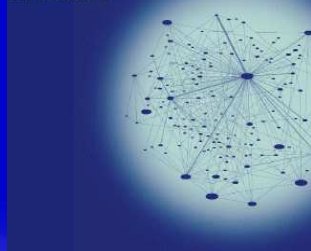
□ 提速阶段 (1997-2000)

- vBNS/NGI,
 - DOD DOE NSF等联合提出
 - 骨干网速度提高100倍和1,000倍
- Internet2,
 - 100多所美国大学联合开展
 - 重点创建灵活的边缘网络能力,发展新型网络应用
- CA*NET3/4.....

5



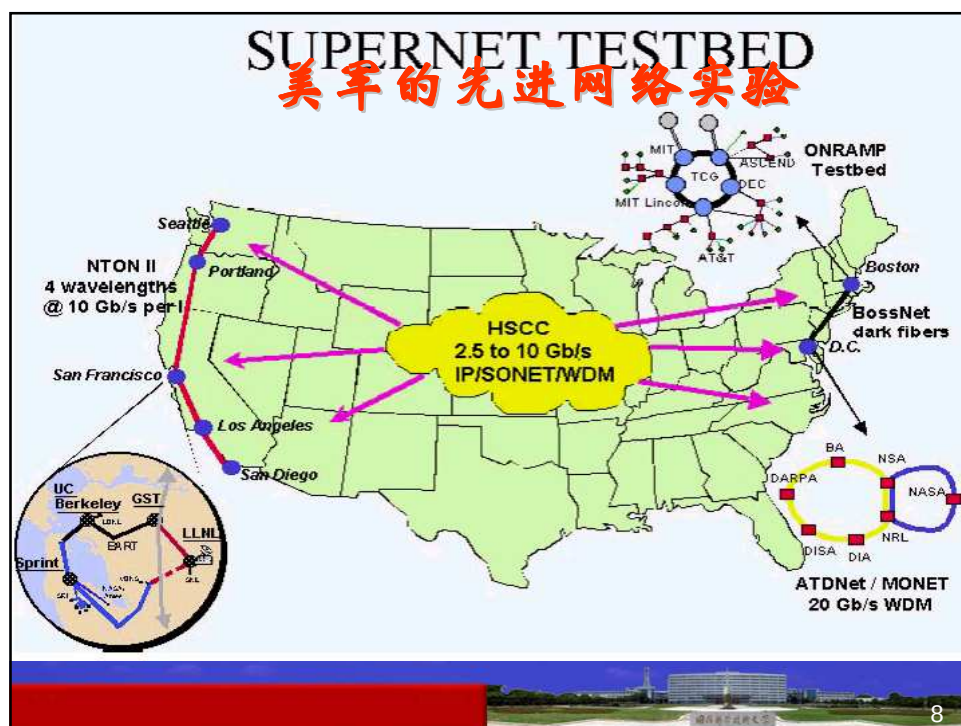
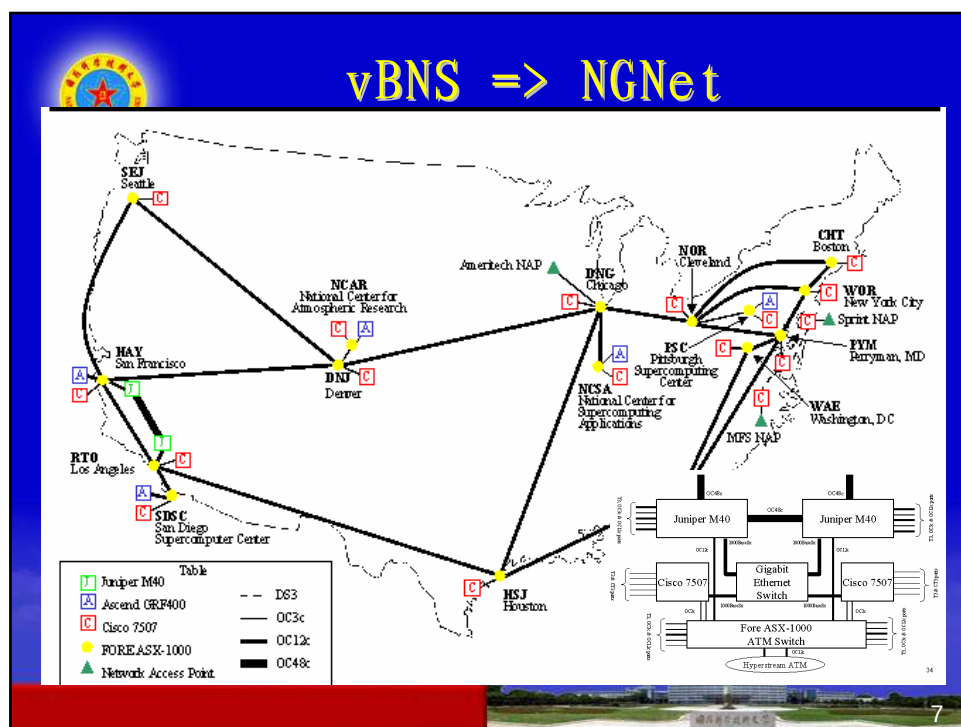
RESEARCH CHALLENGES FOR THE NEXT GENERATION INTERNET



- B.1 Applications
- B.2 Middleware
- B.3 Quality of Service
- B.4 Internet Traffic Engineering
- B.5 Security
- B.6 Architecture

1997年

6





□ 融合与拓展阶段 (2000-2005)

- ITU 提出 NGN
 - 放弃经营多年的 BISDN/ATM;
 - 转向研究 IP 技术
- 中国 CNGI
- 美国 NSF 的 TeraGrid

9



□ 创新阶段 (2005-

- 美国 NSF: 100*100, FIND, GENI
- 美国国防部: AGN
- 欧洲 FP7 ICT (信息通信技术)
 - 2007-2013, 91亿欧元
 - 普适可信的网络与服务基础设施是八个内容之一

□ 阶段特点

- 建立新型实验验证环境
- “重新”设计的理念十分明显

10

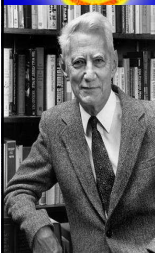


二、发展网络体系结构技术

11



ARPNT基础



□ Claude Shannon, 1948年在BELL System Technical发表论文



Photo by Louis Bachrach

□ Leonard Kleinrock提出分组交换概念



□ 1964年3月: Paul Baran发表"On Distributed Communications Networks", IEEE Transactions on Systems.

Reprinted with permission from The Bell System Technical Journal, 1948, 27, pp. 379-423, 623-656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

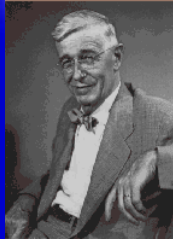
THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

Source: Livinginternet.com

12

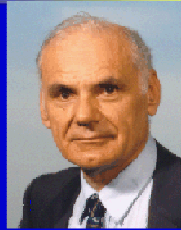


ARPANT基础 (2)



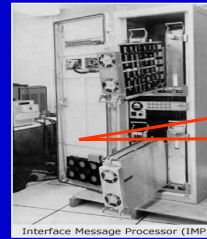
- Vannevar Bush 第一个提出信息技术潜在用处。
- 1940年美国国防研究委员会主席。

Source: Livinginternet.com

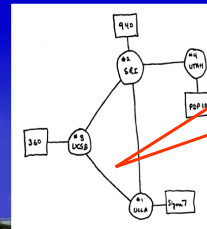


- Lawrence Roberts
- ARPANET项目经理

- Wesley Clark提出采用一个分离的小型机作为网络的接口。



接口消息处理器 (IMP)



1969年9月的设计草图与记录

13



新一代互联网重要理论研究进展

- 0) 总体思想
- 1) 编址
- 2) 路由
- 3) 能力管理和拥塞控制
- 4) 分发服务
- 5) 以内容为中心的结构
- 6) 安全服务
- 7) 分组与层次划分
- 8) 测量与监控
- 9) 增强的网络服务
- 10) 无线/移动网络
- 11) 传感网络

14



实例1:总体构思

□ 互联网的设计假设变化

- 互联网用户无法互相信任
- 需要考虑经济因素
- 用户基础产生变化.
- 技术与法律的共同演进

□ 重新思考当时理念

- 网络层次
 - 避免应用功能进入网络
 - 构建通用设备
- 应用层次
 - 应用应该足够可靠、足够安全等
 - 软件全部推到边缘
 - 或许要仔细考量服务器和可信任第三方的角色
- 引入认知平面。

Rethinking the Design of the Internet: The End-to-End Arguments vs. the Brave New World

MARJORY S. BLUMENTHAL
National Academy of Sciences
and
DAVID D. CLARK
MIT

ACM Transactions on Internet Technology, Vol. 1, No. 1, August 2001, Pages 70–109.

15



实例2: 报文分组与层次划分

□ ISI的B. Braden, 在HOTNETS-12002.提出 “From Protocol Stack to Protocol Heap – Role based Architecture

- 作者认为层次化带来许多问题, 例如
 - 一个TCP SYN向80 (HTTP) 端口无法声明不被重定向到WEB Cache中;
 - 网络拥塞是路由器(网络层)发现的, 但速率控制是在流层次(传输层)
 - 网络中间箱(防火墙、NAT、代理, web CACHE, 隧道)等与分层存在冲突
- 提出无层次的设计方法, 基于角色的体系结构

16

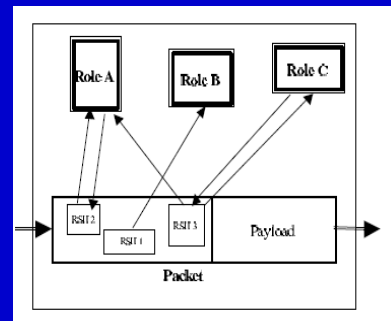


实例2：报文分组与层次划分（续）

□ RSH (Role-Specific Header)

□ 关键技术

- Role Matching
- Actor Execution Scheduling
- RSH Access Control
- Role Definition
- Role Composition



17



实例3：安全问题

□ 华盛顿大学T. Anderson和Intel/T. Roscoe (HotNets-II, 2003)

- 从重新设计互联网的角度，提出“目的端发放令牌”方法，控制从源到目的的链路和目的主机的有限资源。
 - 为了可行性，设计了Request-To-Send (RTS)服务器和“verification points” (VPs)，
 - RTS Server和VP位于BGP Speaker附近。
 - 设计了系列Hash函数对报文进行认证
- 跳出报文过滤的思路
 - Source Address Filtering
 - Traceback and Pushback
 - Overlay Filtering

18



实例4：资源命名

□ Berkeley 大学 Hari Balakrishnan 等 (SIGCOMM, 2004.)

□ 提出层次化命名体系结构

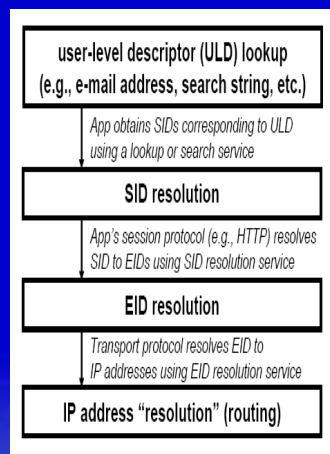
- 用户级描述符，服务标识符，端点标识符

□ 优势

- (1) 服务和数据可以被直接持久命名
- (2) 无缝容纳移动性和多宿主
- (3) 将防火墙、NAT等中件箱 (middlebox) 集成到互联网体系

□ 强调

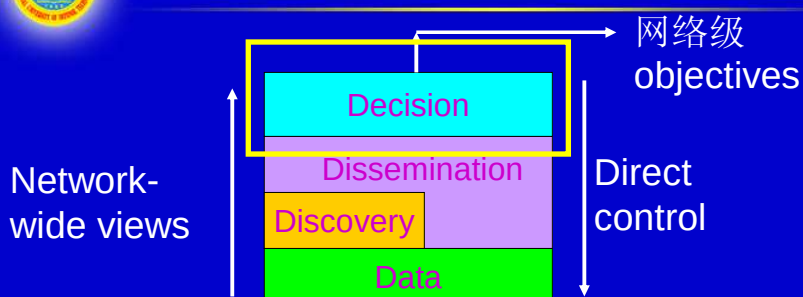
- 服务和端点标识符采用扁平化名字
- DHT是一种可选择的扁平化方案



19



实例5:4D 体系结构概要



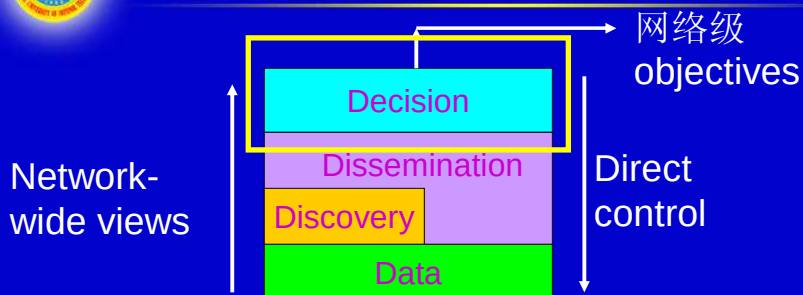
决策平面:

- 所有管理逻辑在中央服务器执行全部决策
- 决策要素采用views计算满足 objectives 的数据平面状态，直接将状态写到路由器中。

20



实例5:4D 体系结构概要



分发平面:

- 为每个路由器提供可靠通道，可靠性是唯一的目标。
- 可能与数据同一链路上，但是逻辑上是分离的，独立控制。

发现平面:

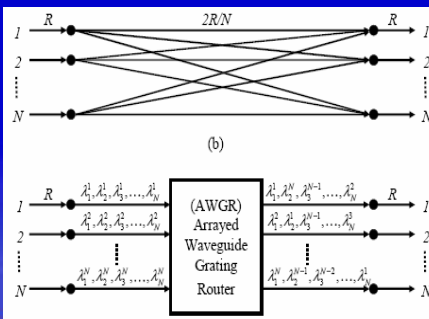
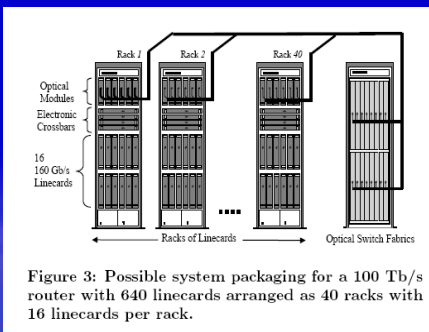
- 每个路由器发现自身资源和局部环境

21



100Tb容量/160G线卡路由器

- Stanford大学Nick McKeown等 **Scaling Internet Routers Using Optics (SIGCOM2003)**



22



三、深度融合光电技术

23



IP网络和光设备融合走向

B-ISDN

IP/ATM

IP/SDH

IP/Optical

光分组

IP路由器

ATM设备

IP/ATM设备

IP路由器

SDH/SONET设备

SDH/SONET设备

SDH/SONET设备

IP路由器

Optical设备

Optical设备

Optical设备

光设备

光分组交换

设备价格及网络运维费用

24



光技术如何影响互联网

□ 以Stanford大学 Nick McKeown为代表

— 电路交换(DCS):

- 优势

- 适合光技术

- 容量大

- 交换简单

- 弱点

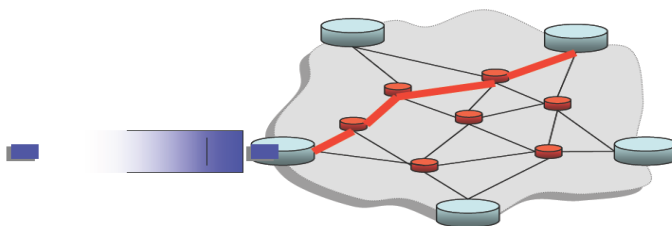
- 灵活性差

— 光电集成:价格和功耗的优势

— 光分组交换:采用创新体系设计路由器

25

DCS: Capacity on demand between border routers



Rule of thumb

Predict the need for capacity by monitoring how quickly new flows are created, rather than waiting for the buffer to fill

骨干网将在10年内部分采用光 DCS

26



光电集成的性价比优势

□ 对最后一公里接入的影响

□ 对路由器体系结构的影响

- 芯片内部, 插件, 插箱, 机柜互连
- 密度更高, 功耗更低

□ 总体影响

- 2005年的光子集成程度仅与1965年的集成电路相当

27



大缓冲是路由器的特点

□ 路由器基于报文交换

- 长途链路资源十分昂贵
- 统计复用可以有效共享长途带宽

□ 路由器设计理念

- 1. 必须处理报文分组头部,
- 2. 选个报文分组路由
- 3. 拥塞时, 缓冲报文, 报文丢失会产生严重后果

□ 为此需要很大的缓冲

- 幸运的是, 电的缓冲十分便宜

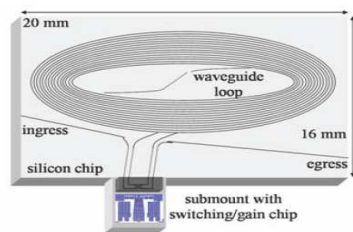
28



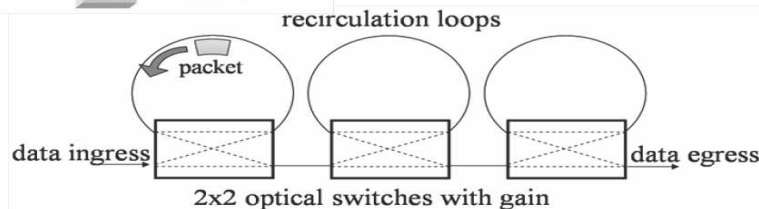
当前技术水平:光缓冲实验装置

Integrated optical buffers

[Burmeister and Bowers, UCSB]



Think: 10-50 packets on a chip



13



创新的设计理念设计路由器

□ 以DARPA DOD-N为代表的研究结论

- 1.Process headers:
 - 报文头部小,用电处理;
- 2.Switch packets:
 - Valiant Load Balancing (VLB) avoids packet-by-packet switching [Sigcomm03]
- 3.Buffer packets:
 - 20-50 packets might be enough in the backbone [CCR 05];
 - will be feasible with integrated optics [Bowers 05]

(IRIS and LASOR projects)



30



光电结合：两种大容量路由器

□ IRIS : Integrated Router Interconnected Spectrally

- 吞吐速率100Tb/s
- 采用高密度集成光路
- 经费1250万美元

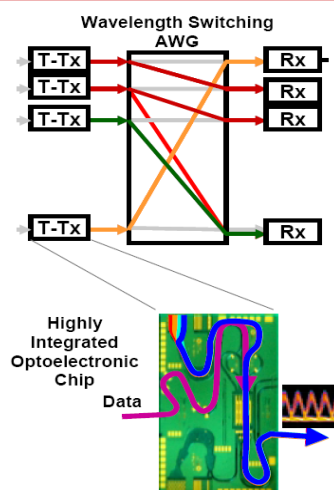
31



IRIS

IRIS Optical Data Router Project

- Program to develop Technologies for High Capacity Optical Packet Router Scalable to >100Tb/s
 - 4 year DARPA-MTO Funded Program
- Systems Research
 - Load Balanced Optical Networks
 - Optimization of distributed control and scheduling
- Hardware Research
 - Distributed Switching using Fast Nanosecond Wavelength Tunable Transmitters (T-Tx)
 - Highly Integrated Optoelectronic Chips handling multiple 40Gb/s data streams per chip



Nov 6, 2005

APOC 2005

32



□ LASOR: Label Switched Optical Router

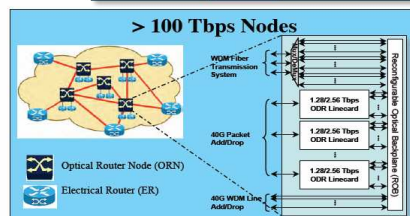
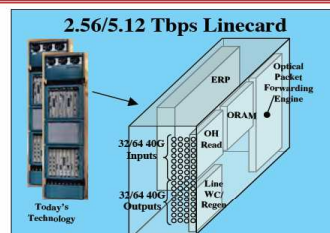
- 吞吐速率100Tb/s
- 研制全光波长变换器
- 经费1580万美元

33



Project Overview

- * Project Start Date: 4/8/04
- * Project Duration: 4 Years
- * Funding Agency: DARPA/MTO
- * Program: DoD-N
- * Program Manager: Dr. J. Shah
- * Technical Team
 - * UCSB
 - * Calient Networks
 - * Cisco Systems
 - * JDS Uniphase
 - * Stanford University
- * Collaborators
 - * USC (Willner)
 - * Telcordia



3

D. J. Blumenthal, ECOC 2005, Sept. 26-30, Glasgow, Scotland



GENI 实例

35



GENI

GENI Initiative

- Research Program
 - e.g., FIND program in NeTS solicitation
- Experimental Facility
 - proposal to MREFC program
 - ▶ jointly from CISE and the research community
 - funds (cutting-edge) facility construction, not research
 - we currently have a “Conceptual Design” document
 - will eventually require Congressional approval

Planning Group

- Tom Anderson, *Washington*
- Dan Blumenthal, *UCSB*
- Dean Casey, *Ngenet Research*
- David Clark, *MIT*
- Deborah Estrin, *UCLA*
- Larry Peterson, *Princeton* (Chair)
- Dipankar Raychaudhuri, *Rutgers*
- Mike Reiter, *CMU*
- Jennifer Rexford, *Princeton*
- Scott Shenker, *Berkeley*
- John Wroclawski, *USC/ISI*

36



GENI

□ 2005年7月12号到14号由美国自然科学基金 (NSF) 在卡内基梅隆大学 (CMU) 举办的第一次会议 (NGSI)

□ 研究目标和特色:

- 互联网在功能上取得实质性的进步
- 终极目标: 未来10年我们需要什么样的通信基础设施
- 不是对现有互联网改进
- 不考虑与当前互联网的兼容问题
- 大胆地进行创造性的设计并对设计进行评估
- 分析当前网络的优缺点, 探讨未来安全网络体系结构的设



National Science Foundation
WHERE DISCOVERIES BEGIN

37



为什么需要全新设计?

□ 可信通信基础设施及未来应用需求

- 高“弹性” (highly resilient)
- 可用性
- 可靠的通信环境

□ 脆弱的互联网:

- 蠕虫、病毒、垃圾邮件、DoS攻击.....
- 理想化的假定: 良性 (benign)、可信
- 最初的设计没有考虑安全问题



38



为什么需要全新设计？

□ 对现有互联网的不断改进就有可能偏离我们的长期目标

- 增加安全层不能满足安全要求
- 在现有协议中加入安全机制（复杂性、兼容性、功能限制……）

□ 全新设计的意义：

- 一个对未来互联网发展的整体设想
- 目标：安全灵活的网络体系结构

39



为什么需要全新设计？

□ 全新的设计：

- 避免把现有互联网体系结构的任何部分作为未来网络体系结构的必需
- 鼓励研究人员充分发挥其想象力而不受现有互联网体系结构思想的束缚
- 预测未来的需求和运行机制
- 从过去的互联网设计中吸取有价值的东西，提出新的方法，并为提出的新体系结构提供理论和技术支撑

推倒重来？

40



GENI(光传输+无线传输)

□ GENI设计概述

- GENI=PNS+GMF
- PNS: 物理网络“培养基”
- GMF: 全局管理框架

□ GENI底层物理网络构成

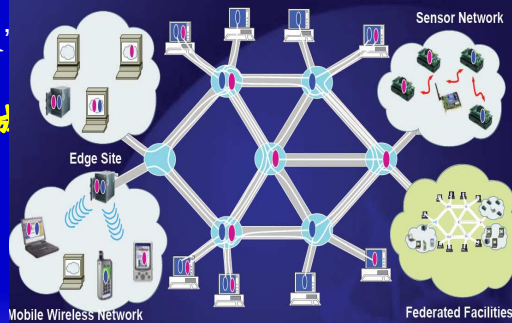
- 三种结点构件
- 两种链路构件
- 五种特定网络构件

□ 核心技术

- slicing
- 虚拟化技术
- 灵活,可编程(可定制)



Facility Design: Key Concepts



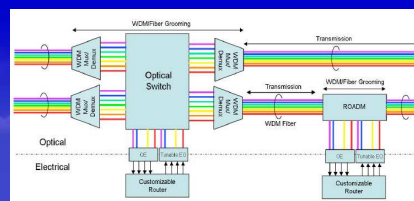
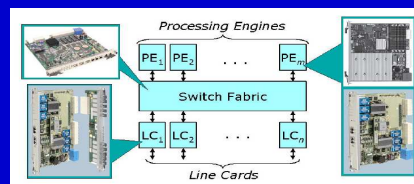
Slicing, Virtualization, Programmability

41



三种结点

- 灵活的边缘设备
- 可定制高速路由器
- 动态光交换和网络



42



两种链路

□ 国家光纤设施

- slice对应的带宽分配技术



□ 接入技术

- 物理链路直接接入(光纤等)
- MPLS
- Frame-Relay
- IP tunnels
- Etc...

43

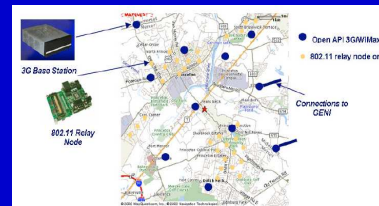


五种特定网络

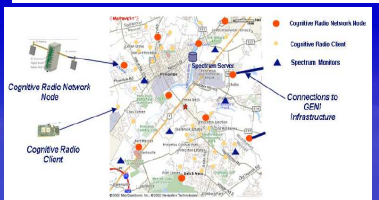
□ 基于802.11的城区 mesh/ad hoc网



□ Wide-Area Suburban 3G/WiMax-Based Subnet



□ Cognitive Radio Subnet



□ Application-Specific Sensor Subnet

□ Emulation Subnets

44



四、充分应用并行技术

45



互联网核心设备— 高性能路由器面临的挑战

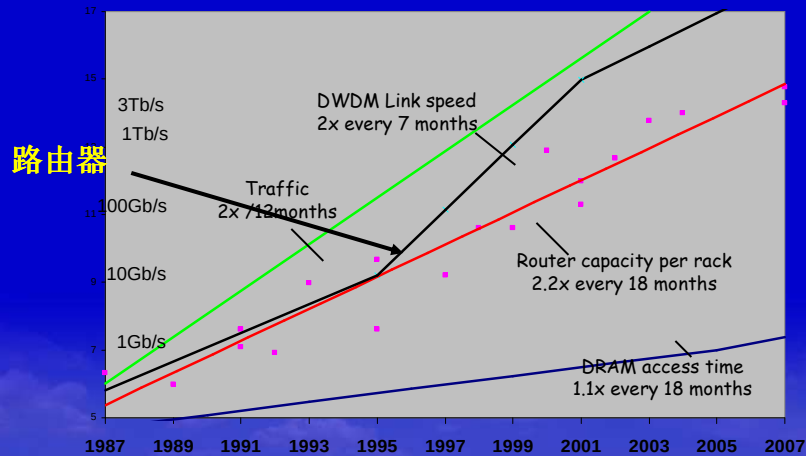
- 链路速率和Internet流量的迅猛增长
- DRAM技术的发展速度远远落后于摩尔定律
 - 商业DRAM $1.1 \times / 18 \text{ months}$
 - 报文转发
 - 报文交换
- FIB容量的快速增长

46



核心路由器的支撑技术

Backbone router capacity



47



FIB容量快速增长带来的问题

Faster 'Net growth rate raises fears about routers

By Carolyn Duffy Marsan
Network World, 04/02/01

RELATED LINKS BREAKING NEWS

SEND PRINT FEEDBACK REPRINT

Alarm bells are ringing in the Internet engineering community over an obscure statistic that indicates the 'Net is growing - in size and complexity - at a faster rate than today's routers can handle.

This recent finding has prompted the Internet Engineering Task Force (IETF) to embark on an overhaul of the communications protocol that handles routing across the Internet's backbone.

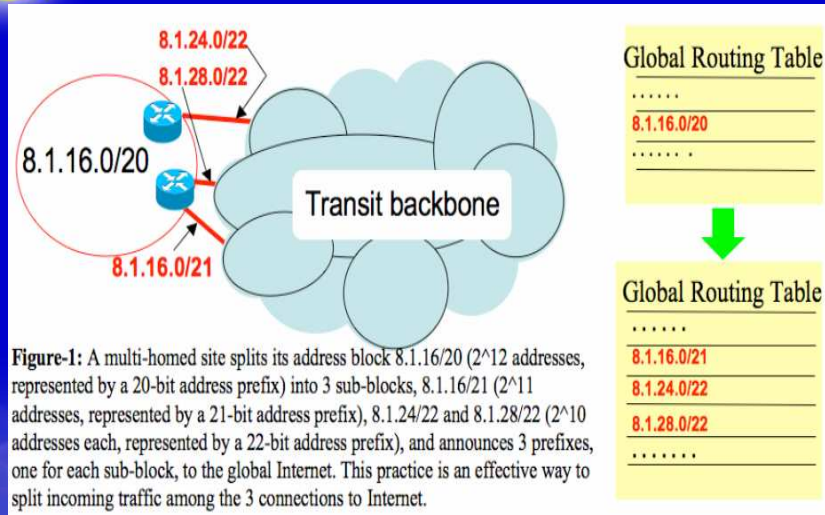
At stake for large companies may be the need to buy more powerful network gear earlier than originally intended.



48



多宿主是FIB增长的根本原因

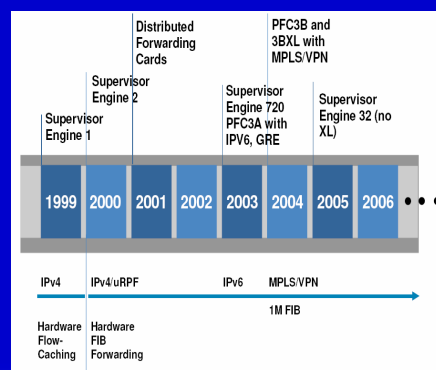
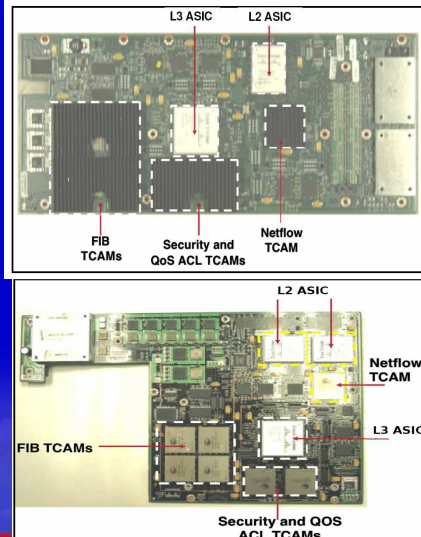


49



电处理接近极限

Separate Hardware Devices for L2, FIB, ACLs, Netflow Tables



- Cisco的麻烦
- 世界上最大的ASIC
- 184个32位CPU!

50



问题出路之一:并行

□ 解决方案

- 并行路由器体系结构

□ 更多的并行

- 并行缓冲
- 并行转发
- 并行交换

51



基本分类

□ 同构型并行路由器

- 报文交换单元硬件上同构
 - 斯坦福大学100T光路由器
 - 负载均衡路由器 (load-balanced switch) — 输入线卡、中间级输入线卡、输出线卡
 - Cisco CRS-1路由器
 - 三级Benes交换网络

□ 异构型并行路由器 (Crossbar-based parallel router)

- 每级交换单元采用不同硬件结构, 执行不同功能
 - 并行报文交换 (Parallel Packet Switch)
 - 中间级采用NxN CIOQ交换平面或OQ交换平面
 - 3级Clos交换网络—输入模块、crossbar、输出模块

52



谢谢!