Windows Vista Netzwerk-Stack

Innovativer Meilenstein oder Marketingblase

Hagen Paul Pfeifer
hagen@jauu.net

www.jauu.net

19. August 2006

Vortragsfahrplan

- Prolog
  - Gegenwärtige Situation
  - Herausforderungen
  - Rewrite
- Next Generation IP Stack
  - NDIS 6
  - QoS
  - Winsock Kernel Interface - WSK
  - Teredo
  - Firewallinterior
  - Krypto
  - Compound-TCP
  - TCP-Ingredienzien
  - SMB 2.0
  - Winsock-API

Kapitel 1
Prolog

Gegenwärtige Implementierung

- Dual Stack Implementierung
  - IPv4 und IPv6 als separate Module (TCPIP.sys, TCPIP6.sys)
- Technisch angegraut (Service Pack 2)

Herausforderungen

- Das Herausforderungs-Referenzmodell:
  - Anwendungsschicht: Triple Play ;-)
  - Transportschicht: Bindeglied - Abstimmungen notwendig
  - Internetschicht: Zunahme Hosts (China, Mobiltelefone)
  - Netzzugangsschicht: VDLS, VDSL2 und 1GB, 10GB Ethernet, 802.11n
- Säulenbildung:
  - Einhergehende Abstraktionsschicht
  - Streaming Media

Kapitel 2

Next Generation IP Stack

Microsoft Antwort: Rewrite

- Großer Teil des Stacks gegen neuen ausgetauscht
- Neuer Code, neue Bugs
- Network Stack ist von fundamentaler Bedeutung für Sicherheit des Betriebssystem - eine der größten Angriffsvektoren

NDIS 6.0

- NDIS - Network Device **Interface** Specification
- Kerntechniken
  - RSS - Receive Side Scaling
  - TCP Chimney Offload (Schornstein: oben (Transportschicht) rein - unten (Netzzugangsschicht) raus)
- RSS
  - MS schritt um SMP und DualCore CPUs Entwicklung Rechnung zu tragen
  - Früher: Interrupt CPU gebunden, NAPI-Eigenschaften
  - Heute:
    - parallel auf mehreren Kernen

- TCP-Verbindung auf Prozessor gebunden (Stichwort Cache-Trashing)
- Load Balancing: Verbindungen dynamisch verteilen
- Toeplitz Hash Funktion

- TCP Chimney Offload
  - TCP, IP und ARP/Neighbor Discovery und 802.X Offload
  - Chimney State Objects
  - Bulk Transfers
  - Sparta Software-Implementierung
  - Open-Source-Info: TCP-Offload mit vielen Patenten belegt
- vereinfachtes Treibermodell

Quality of Service

- Kriterien
  - Anwendung
  - IPv4/IPv6 Adressen
  - Ports
  - TCP/UDP
- DSCP oder/und Traffic Class-Feld Mangling
- Beschränkung Bandbreite (egress)
- Priorisierung von Netzwerk Verkehr
- Differenzierung für Anwendung und Benutzer möglich
- Pacer.sys(LWF): NDIS 6.0-Treiber (ehemals Psched.sys)

Transport Data interface - TDI

- Kernel-Mode Schnittstelle zwischen zwei Geräte-Treiber
- Integraler Bestandteil und Vorraussetzung für Implementierung von Netzwerkfunktionalität
- Kein Socket-Style Interface - eher abstrakter, funktional

  - TDI Provider:
    NDIS Protokol-Treiber (Transport Driver) welche die Grundimplementierung von Netzprotokollen enthalten (z.B. TCP/IP)
  - TDI Clients:
    Diese Kernel-Mode Treiber nutzen die Funktionalität der Provider. Ein tcpip Client kann dann beispielsweise Verbindungen aufbauen, nutzen und beenden.

- TDI Filter:
  Liegt logisch zwischen beiden und kann dadurch Verbindungen abfangen und bearbeiten.
  Anwendungsbeispiele: Emailscanner, Firewallprodukte

- Windows Socket sind beispielsweise als TDI-Clienten auf Kernelseite implementiert, eine Socket Emulation (diese kommunizieren dann mit der korrespondierenden dll im User-Space) (afd.sys, Anpassungen über Registry-Parameter)

Winsock Kernel Interface - WSK

- Ersatz für Transport Driver Interface (TDI)
- Socket-Style Interface
- Asynchrones IO möglich
- Netzwerk Module (Client Module oder Provider Module) implementiert Funktionen im Netzwerkstack (Data Link Interface, Transport Protokoll oder Netzwerk Applikation) - wie bei TDI
- Zur Zeit ist TDI Schnittstelle um mit den Netzwerkstack zu interagieren
- Schnittstelle für Entwicklung eigener Protokolltreiber
- Performancevorteile gegenüber TDI
- `http.sys` ist z.B. ein Kernel Mode HTTP Handler

- Für die „Schrauber“: `WskRegister()`

Teredo

- Tunnel Protokoll um IPv6 Konnektivität zu gewähren (NAT)
- Abstecher 6to4 und STUN (Simple Traversal of UDP Through Network Address Translators)
- IPv6 Konnektivität ohne Kooperaation des LAN's (schnell mal Piraten-IRC-Server™ starten, oder so . . .)
- IPv6 verpackt in UDP/IPv4 Paketen
- RFC 4380 - „Teredo: Tunneling IPv6 over UDP through Network Address Translations (NATs)"
- Netzkomponenten: Client, Server und Relays
- Linux/BSD User-Space Implementierung: u.a. Miredo (Beta)
- Oder simpler:

http://linide.sourceforge.net/nat-traverse/

Netwzwerk Zugangsschutzsystem - Firewalling

- Startschuß: Windows XP Service Pack 2
- Vista Erweiterungen:
  - IPv6 Filterung
  - Filterung von ausgehenden Paketen
  - Regeln für Quell- und Zieladresse sowie Portbereiche
  - IPSec Unterstützung
  - Benutzereinschränkungen möglich
  - Firewall Profile
- Ausprägungen:
  - User-Mode Filterung:

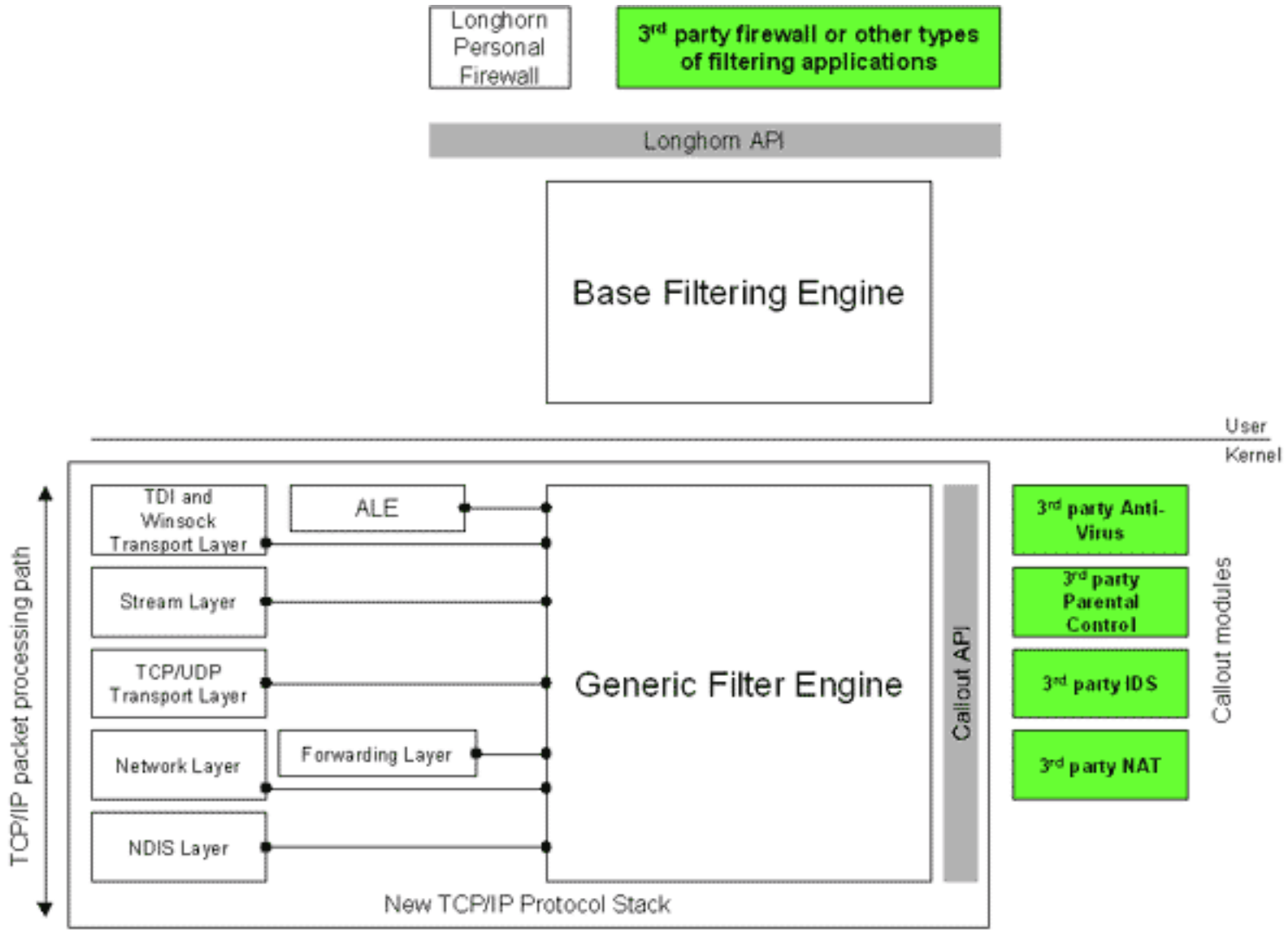
- Winsock Layered Service Provider (LSP)
- Windows 2000 Packet Filtering Interface (iphlpapi.dll)
- Winsock Replacement DLL (Linux Pendant: `dlopen(3)`)

- Kernel-Mode Filtering:
  - Transport Data Interface (TDI) Filter Driver
  - NDIS Intermediate (IM) Driver
  - WPF - Windows Filtering Platform

- Anwendungsspezifische Filterung:
  - Anwendungen können an dem Namen, Pfad oder einer Prüfsumme erkannt werden
  - Kernel oder Benutzerprozeß

- Linux (Aber: SMP Probleme):
  - `ipt_owner.c:match_pid()`
  - `ipt_owner.c:match_comm()`

WPF - Windows Filtering Platform



- hakt sich in die Paketverarbeitungspipeline
- Interessant für Firewall, Antivirus und Diagnosesoftware
- WFP stellt eine Basis Filterung Engine bereit - Logik in Modulen
- Erlaubt ein- sowie ausgehende Pakete zu analysieren und zu bearbeiten
- Möglichkeit, die Verbindung zwischen Applikation und Paket herzustellen
- Besser dokumentiert und endlich saubere Schnittstellen!

Compound TCP

- Staukontrollalgorithmus
- Zielsetzung:
  - Effizienz
  - TCP Fairness
- Entwicklung von MS-Research (mit Hilfe von NS-2 ;-)
- Mischung aus `loss-based` und `delay-based` Algorithmen
- Kontrolle über Verzögerung und Paketverlust
- Aggressiver Slow-Start (unter Umständen!)
- Alle verfügbaren Untersuchungen sehen vielversprechend aus – aber es fehlt an weiterführenden Studien
- Apropos: NewReno (RFC 2582; Fast Recovery

Anpassungen) findet Verwendung

TCP Ingredienzien

- Receive Window Auto-Tuning
  - Info Generalis:
    - Empfangsseitige Begrenzungsmaßnahme
    - Window Scaling, ein muss bei „normalen“ Verbindungen (kontinental) ($2^n * window$)
    - Default zwischen 1 und 2 (0 entspricht 64K)
    - `setsockopt()`
    - Middle Boxen
  - Windows Server 2003 und Windows XP:
    - 8KB (ausreichend für 10Mbps)
    - Defaultwert in Abhängigkeit des Links

- Unterstützung von Window-Scaling bis 1GB
- Applikationsspezifische Anpassung möglich
- Kann Manuell angepasst werden (Registry) - Neustart erforderlich ;-)

- Vista:
  - Weg von statischen Werten, hin zu DRS
  - Durch Window Scaling bis 16MB möglich
  - Bestimung durch Messungen des BDP und Applikations-Lese-Frequenz

SMB 2.0

- Neue Version des Filtetransferprotokoll
- Mehrere Aktionen in einem Request möglich
- Feste Headergröße und größere Typfelder
- SMB 2.0 Magic: `0xFE 'S' 'M' 'B'`
- Teilweiser dissector support in Ethereal

Kryptografie

- Diffie-Hellman Gruppe 19 und 20 Unterstützung (elliptischer Kurvenalgorithmus)
- DES, 3DES, AES128, AES192, AES256

Winsock-API

- Ursprung in BSD Sockets
- `-DIPV6STRICT` - IPv4 spezifische Strukturen und Aufrufe werfen Fehler
- WSAConnectByName()
  - Verbindet mit Peer, welcher mit grösster Wahrscheinlichkeit passt
  - Linux: `getddrinfo()`, glibc und /etc/gai.conf
  - RFC suchen (kommt von MS, `WSAConnectByName()` wird mit größter^{TM} Wahrscheinlichkeit matchen

NAP - Network Access Protection

- Abstrakt: Sicherheit in Netzwerken erhöhen
- Kontkret: Netzwerkzugriff erst wenn Sicherheitsanforderungen erfüllt
- Policy Enforcement Platform
- Überprüfung des Systems mit eventueller Anpassung
- System Health Agent und Quarantine Enforcement Clients (QECs)
- System Health Agent:
  1. Firewall ist aktiviert für alle Interfaces
  2. Antivirus ist aktiviert und up-to-date
  3. Antispyware ist aktiviert und up-to-date

4. Automatisches Update ist aktiviert und up-to-date

- Erweitbar: öffentliche API (Dateiversion, Registryeinträge, . . .)
- Client: XP und Vista; Server: Longhorn
- Longhorn Server wird mit aktivierter Firewall ausgeliefert - per default kann es sich nicht mit Network Policy Server verbinden

Was vergessen?

- Strong Host Model
- Netzwerk Profile
- 802.11 - WLAN
- ESTATS

Weiterführende Informationen

- Compound TCP (CTCP)
  - www.slac.stanford.edu/cgi-wrap/getdoc/slac-tn-06-005.pdf
- Transport Driver Interface - TDI
  - http://www.pcausa.com/resources/tdifaq.htm
  - http://www.codeproject.com/system/driverdev5asp.asp
- Windows Treiber Entwicklung
  - http://www.osronline.com/index.cfm
- Treiber Signierung in Vista
  - http://www.microsoft.com/whdc/system/platform/64bit/kmsigning.mspx
  - http://www.osronline.com/article.cfm?article
- Verschiedenes:
  - www.symantec.com/avcenter/reference/ATR-VistaAttackSurface.pdf
- Allgemeine Informationen zu den neuen Netzwerkstack:
  - www.microsoft.com/germany/technet/itsolutions/network/evaluate/new_network.mspx
- Elliptische Kurvenalgorithmen:
  - http://www.certicom.com/index.php?action=ecc_tutorial,home

- Routing:
  - IPv6: http://www.sixxs.net/tools/grh/dfp/all/
  - Tabellengröße: http://bgp.potaroo.net/
  - BGPlay: http://www.ris.ripe.net/bgplay
- SMB 2.0
  - http://www.ethereal.com/docs/dfref/s/smb2.html
  - http://samba.org/ftp/unpacked/samba4/source/libcli/smb2/

FIN

- Fragen/Anregungen/Bemerkungen?
- Falls Latenz der CPU zu gering oder der Mut zu klein war - einfach eine EMail schreiben
- hagen@jauu.net
  - Key-ID: `0x98350C22`
  - Fingerprint:
    `490F 557B 6C48 6D7E 5706 2EA2 4A22 8D45 9835 0C22`

MS Cookies

- „Ever wonder how the Windows shell is designed? Ever try and write a Windows shell extension? It gets easier in Vista!„
- „If an application is not running on the CPU that RSS has scheduled the receive traffic to be processed on, some cache optimizations may not occur“

Kapitel 3

Sicherungs- und Infofolien

Completion Ports

- Skalierbare Netzanwendungen mit Microsoft
- Implementiert als Warteschlagen mit fertigen Anforderungen
- Abarbeitung mit separaten Prozeß (Anzahl CPU's)
- Ab Windows NT verfügbar
- Windows select Implementierung skaliert NULL
- `CreateIoCompletionPort()`
- Unix Pendants: select, poll, kqueue, epoll, threads und kevent

DRS unter Linux

- DRS - Dynamic Right Sizing
- `/proc/sys/net/ipv4/tcp_window_scaling`
- Wächst von `default` nach `max`
- Speicher:
  - BDP: 52Mbps und RTT von 0.4s = 2.5MB Kernelspeicher
- `ip route add 23.23.23.23/32 via 10.8.0.1 win-dow 65535`

Code Signierung in Vista

- Driver Reliability Signature Program (DRS) (Driver Quality Signature, Kernel Mode Code Signing (KMCS))
- Betrifft Vista und Longhorn
- Nur Admin kann unsignierten Treiber (device drivers, filter drivers) installieren (x86)
- x64 nur signierte Treiber erlaubt
- Bei Berürung mit PMP (Windows Vista Protected Media Path) (z.B. PUMA) dürfen nur signierte Treiber gleichzeitig laufen (Softice, virtuelle Laufwerke, eigene Audio Treiber ade ;-)
- Zwei Signierungsausprägungen
  - Treiber Packet Signierung (CAT signing)

- Embedded Signierung (Image)

- Driver binaries that load at boot time must contain an embedded signature (deutliche DRM Ambitionen zu erkennen)
- Alles was über IE kommt, muss signiert sein (für Ausführung)
- Microsoft Quotes:
  - To improve Windows drivers reliability and stability
  - ... allow administrators and end users who are installing Windows-based software to know whether a legitimate publisher has provided the software package
- Um es doch zu erwähnen: Nebeneffekt ist auch der Schutz von Interessen gewisser Industriezweige vor dem Anwender

Linux kevent

- Hot Off The LKML-Press
- Patchset von Evgeniy Polyakov
- Generischer Mechanismus - select, poll, AIO, epoll, and inotify, netlink
- **Ein** Interface für Programmierer