TCP/IP Services for OpenVMS New Features and Performance Enhancements

Yanick Pouffary
Networks Technical Director - OSSG
July 2003

HP WORLD 2003
Solutions and Technology Conference & Expo

Agenda

- TCP/IP V5.4 Technical update
  - Failover (failSAFE IP)
  - Security SSHv2 (Secure Shell)
  - Secure POP Server with SSL
  - Performance
  - Scalability
  - TCPDUMP
  - IPv6 support
- TCP/IP V5.5 (Itanium) Technical Update
- TCP/IP Roadmap

TCP/IP Services V5.4 Technical Update

- TCP/IP Services V5.4 (target date Oct 2003)
- Alpha only release
- Based on Tru64 UNIX® 5.1B, third party software and BSD based public domain software

TCP/IP Services V5.4 Features List (1)

- IP address fail over (failSAFE IP) capabilities within a host and/or a cluster
- Secure shell (SSH) client and server
- Secure Socket Layer (SSL) for POP
- BIND 9.2.1
- TCPDUMP Support to provide both dump analysis and packet capture
- Software update and new programming examples using IPv6 APIs

TCP/IP Services V5.4 Features List (2)

- Scalable Kernel to provide increased scalability for symmetric multiprocessing (SMP)
  - Require V7.3-2 (OPAL)
- Telnet Server performance and scaling enhancements
- NFS Server Performance enhancements
- INET driver performance enhancement of the SRI QIO interface
- Support for more than 10K BG devices (up to 32K)
- Fast BG device creation and deletion

TCP/IP Services
failSAFE IP - IP Failover

Existing TCP/IP High Availability Solution (1)

- IP Cluster Alias (ARP based Cluster Alias)
  - Presents single cluster IP interface to client
    - Cluster alias IP address configured on each member
  - One member at a time acts as impersonator
  - If impersonator becomes unresponsive, another host becomes impersonator
  - Addresses high availability
    - Not load balancing
  - NIC is the Single Point Of Failure (SPOF)

Existing TCP/IP High Availability Solution (2)

- Load Broker Server
  - "Load Balancing" scheme comprised of a Metric server and a Load Broker
  - Metric server on each cluster member tells Load Broker its "metric" - how busy it is.
    - Algorithm to calculate metric same as LAT
  - Load Broker makes list of IP addresses based on member load
    - Sends dynamic DNS update to name server
  - Addresses high availability with load balancing

New TCP/IP High Availability Solution failSAFE IP - IP Address Fail Over

- Failover of IP addresses and static routes across NICs
  - Addresses High Availability
  - Removes NIC as SPOF
- Load-balancing of outgoing connections
- Provides Higher throughput
- All NICs are active (no standby)
  - Independence of NIC type or speed

failSAFE IP

- Configuration Requirements
  - Address configured across multiple NICs, (within a node or across a cluster)
    - Only one instance of the address is active, others are standby
  - failSAFE service enabled (monitors health of interfaces)
- Failures Detected
  - Anything that stops the NIC receive counter from changing. For example a cable disconnect, NIC failure, switch failure, node shutdown, etc
  - Health of a NIC is determined by monitoring “Bytes received” counter.
    - On a quiet network failSAFE service generates MAC-layer broadcast messages

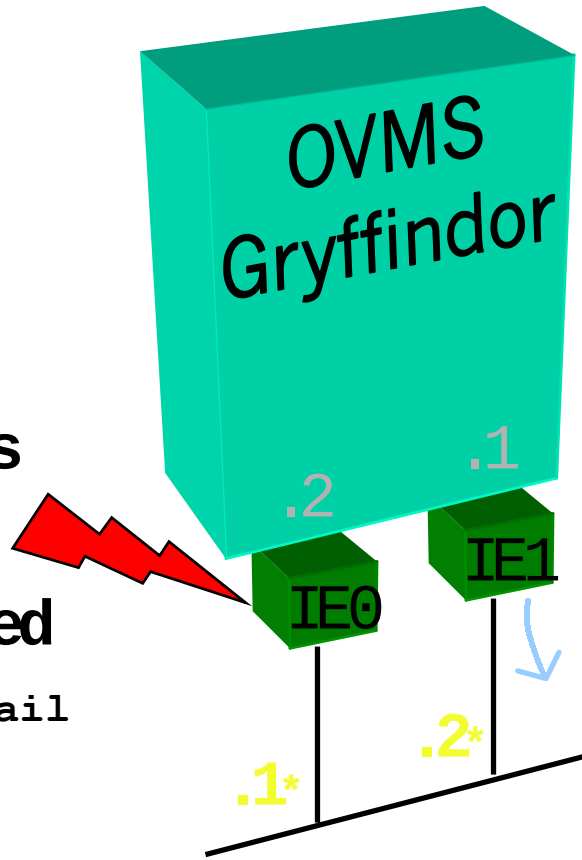
failSAFE IP Failure and Recovery

- Upon NIC Failure
  - IP addresses and static routes on failed NIC are removed
  - Standby IP addresses become active
    - Static routes created on any NIC where the route is reachable
  - Existing connections are seamlessly maintained if failover to NIC on same node
    - IP addresses preferentially failover to a NIC on the same node in an effort to maintain existing connections
- Upon NIC Recovery
  - IP addresses may be returned to the home NIC
    - IP addresses will not return to a home interface if it means connections will be lost

failSAFE IP
Avoid Phantom Failures

- In a quiet network with just 2 interfaces being monitored by the failSAFE service, a single NIC failure may result in a phantom failure of the other NIC
  - Because the surviving NIC is not able to keep its own “Bytes received” counter ticking over
    - MAC-layer broadcast messages which are received on every interface on the LAN, except for the sending interface
  - Advise to configure at least 3 interfaces monitored by failSAFE
    - In the event one interface fails, the surviving interfaces will continue to maintain the others “Bytes received” counter

Single Node Configuration
NIC Failure



Failure occurs

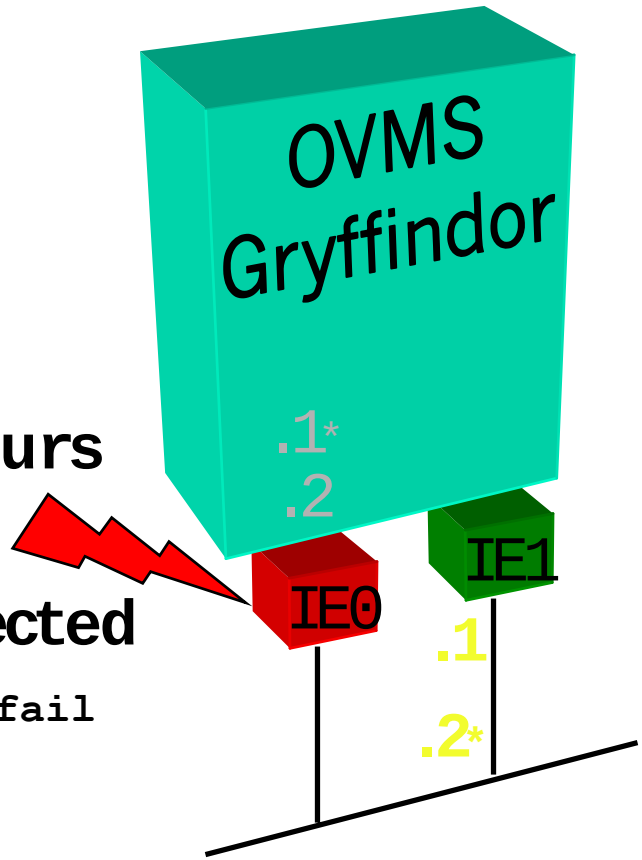
Failure detected

```
$ ifconfig ie0 fail
```

Gratuitous ARP packet is sent

*IP addresses home NIC

Single Node Configuration
NIC Recovery



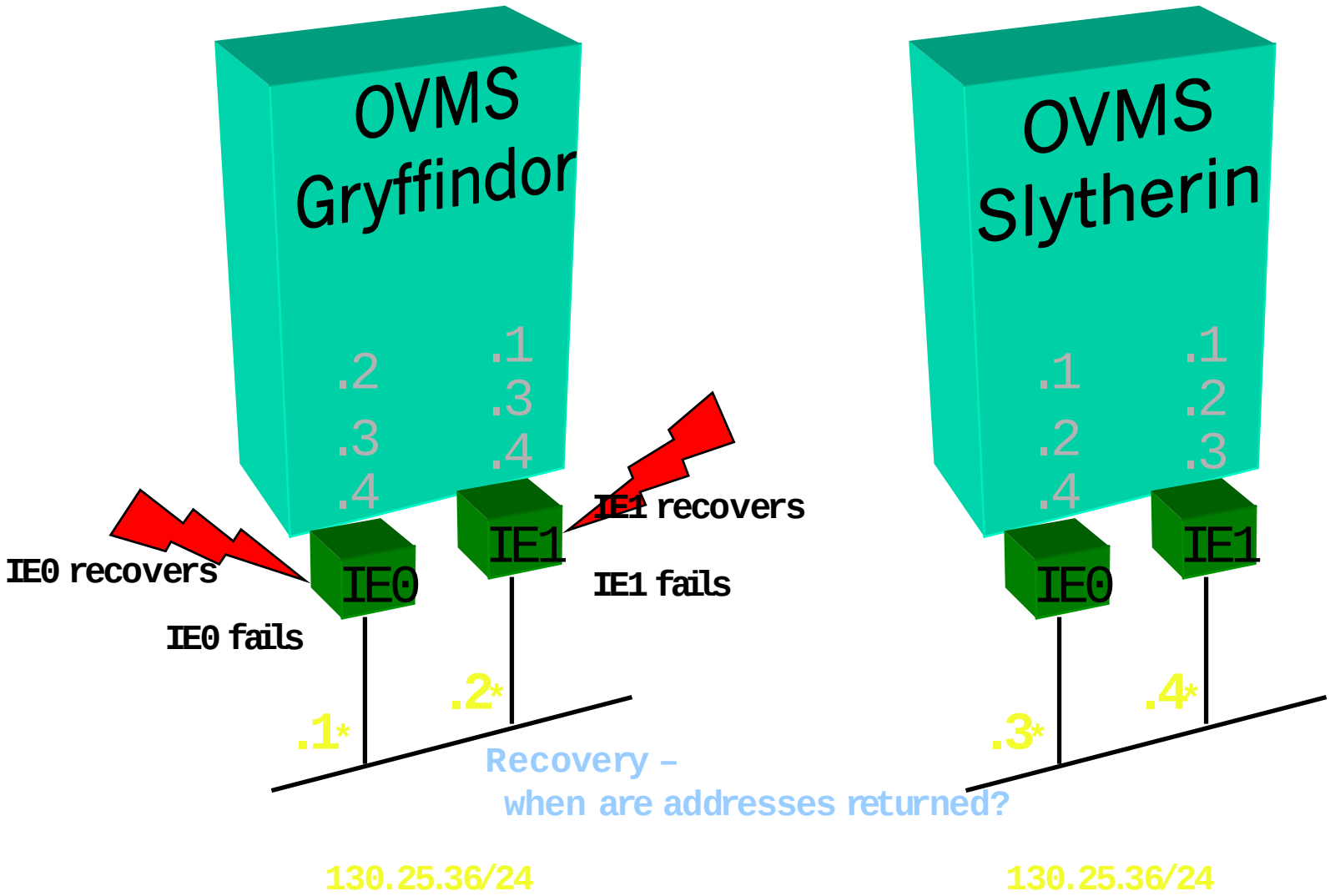
Recovery occurs

Recovery detected

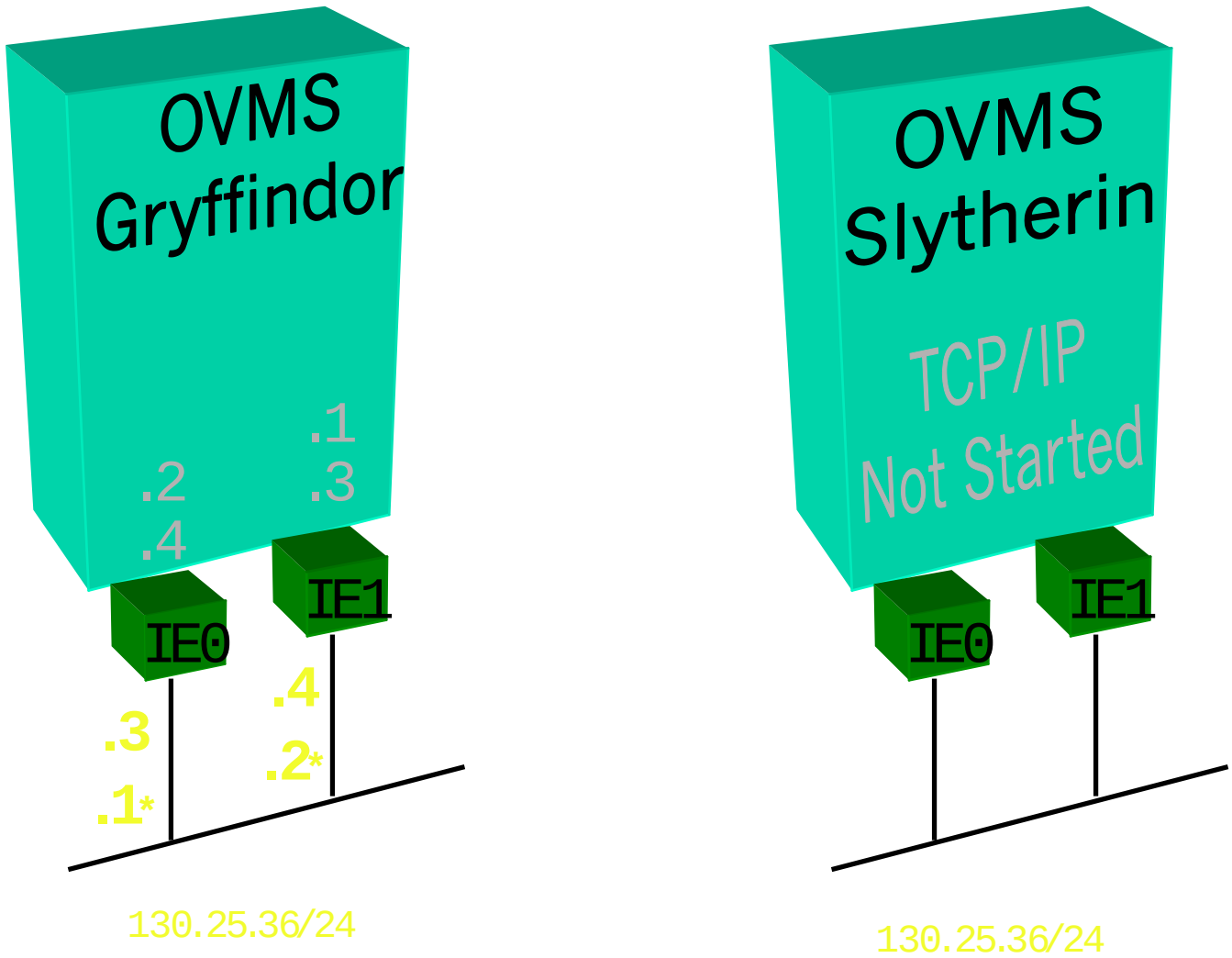
```
$ ifconfig ie0 -fail
```

*IP addresses home NIC

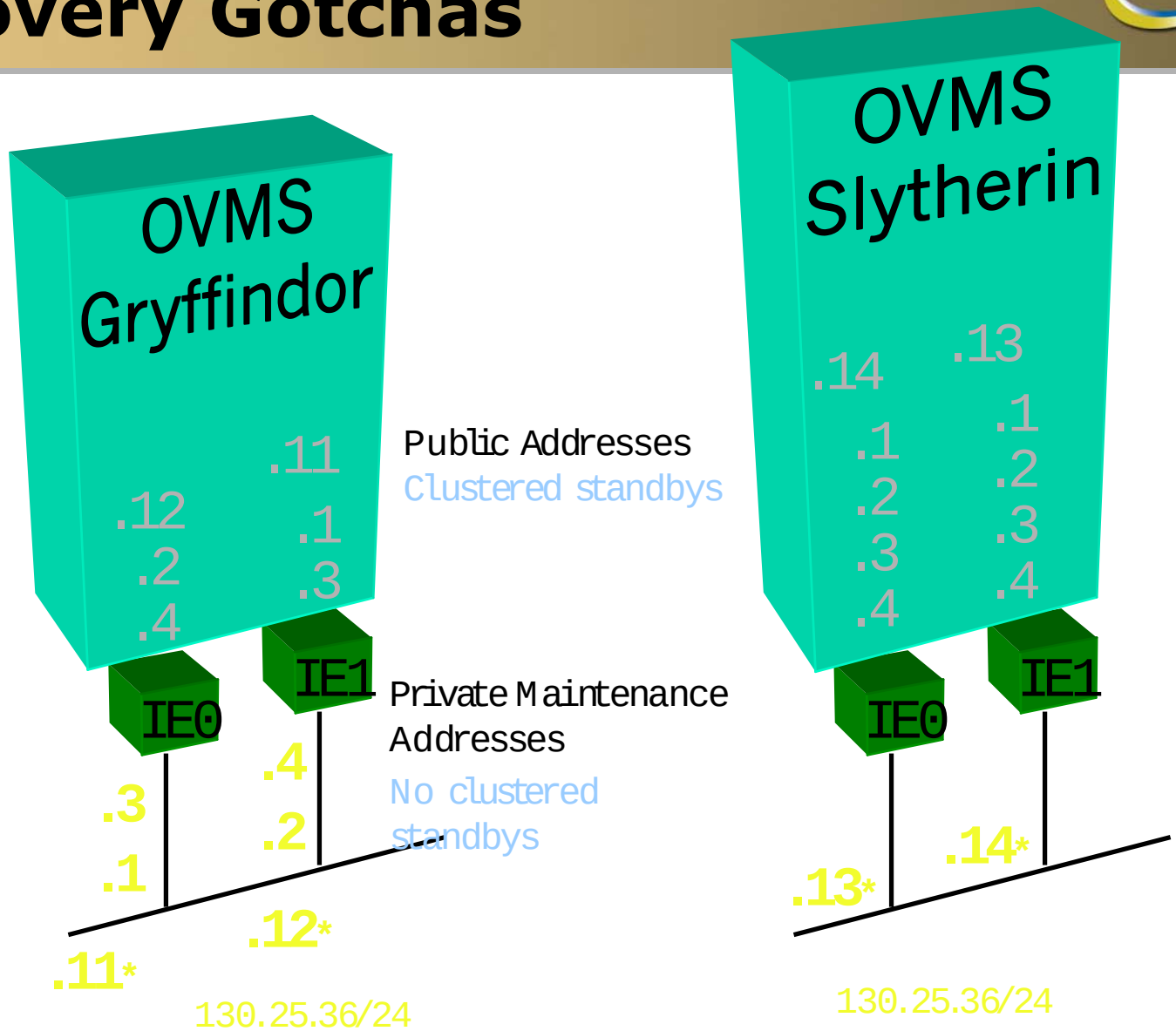
failSAFE Service Monitors Interfaces Effects Failover and Recovery



Startup and Recovery Gotchas



Avoiding Startup and Recovery Gotchas



Best Practices

- Validate the configuration
  - Networks are stable and rarely will the services of failSAFE be required. However, in the one critical moment where it is needed, you want to be sure you have validated the operation.
- Validate the time to effect a failover, time to recover
- Avoid cluster-wide standby's for private maintenance addresses
- Assign alias with home interface to maintain spread
- See OpenVMS Technical Journal (June 2003)

OpenVMS LAN Failover

- Multiple NICs form a LAN Failover Set
  - One NIC is active others remain idle
  - In event of failure, the MAC address migrates to standby NIC
  - Must be connected on same LAN
  - Supports all LAN client protocols
  - Support for DEGXA (GbE) and DE600 (FastEthernet)
  - Failover time is typically milliseconds for link disconnects

LAN Failover and failSAFE IP

Feature	LAN Failover	failSAFE IP
NIC usage	One active NIC, others are standby	All NICs active, load balancing
Devices Supported	DEGXA, DE600	Independent of device types
Protocols	LAN client protocols	IP client protocols
Failover time	Typically milliseconds	Typically a few seconds

TCP/IP Services Security Update

TCP/IP Security

- Internet Security Technologies Available
  - SSH (Secure Shell)
  - SSL (Secure Sockets Layer) - OpenVMS
  - Kerberos - OpenVMS
  - IPsec (future)
  - DNSSEC

Security and Cryptography Terminology

- Security
  - Authentication
    - to determine identity of users
  - Integrity
    - to guarantee that data are unaltered
  - Privacy
    - Encryption to provide confidentiality
- Cryptography
  - Secret-Key
    - Symmetric
    - Shared secret key
  - Public-Key
    - Asymmetric
    - Public and Private key
    - Digital Signatures

What is SSL?

- Used by secure (https://…) web servers
  - Widely deployed
- Certificate based
- SSL protects privacy and offers server authentication
- Server setup required
  - Easy, using self-signed certificates
  - Using CA-issued certificates boost security
- Requires no special client configuration
  - Does not provide mutual authentication
- Can secure other protocols (e.g., POP, IMAP, TELNET)

Secure POP Server with SSL

- New Feature in TCP/IP Services V5.4
- POP server accepts SSL connections on port 995
- Passwords and mail are no longer sent in the clear
- Many clients, including Outlook [Express], support this
- Requires OpenVMS SSL kit
- No need for a separate, unsupported Stunnel process

- Similar IMAP enhancement planned for future release
  - Note: IMAP SSL remains available via Stunnel
    - Stunnel allows to encrypt arbitrary TCP connections inside an SSL connection

What is SSH?

- Application-level Solution to Security
- SSH secures
  - Terminal sessions, File copy, and Remote command execution
  - Other protocols via port forwarding:
    - POP, FTP, X, SMTP, IMAP, even VPNs
- Consists of Client, Server, and Support Programs
- SSH is de-facto Standard
- Available on many platforms

SSH Components

- SSH Consists of:
  - SSH — Login client
  - SSHD — SSH2 server
  - SSH-KEYGEN — Key generation facility
  - SSH-AGENT — Holds keys in memory
  - SSH-ADD — Maintains keys inside agent
  - SSH-SIGNER — Digital key signer
  - SCP/SFTP — File transfer applications

SSH Capabilities

- Features in the first release
  - Remote logins (yes)
  - File transfer (stream_lf)
  - Remote Command Exec (yes)
  - Key generation and agents (yes)
  - Port forwarding (yes)
  - Authentication (password, host, key) (yes)
  - Multiple encryption algorithms (yes)

SSH Secure Shell for OpenVMS

- Ported from Tru64 UNIX® Version 5.1B
- Developed by SSH Communication Security, Inc.
- Secure Shell (SSHv2) – V5.3 EAK available now
  - Supports SSHv2 secure connections
  - Supports SCP secure file transfer
- Fully integrated in TCP/IP Services for OpenVMS V5.4:
  - Configurable using TCPIP$CONFIG
  - Managed through UNIX-style commands
  - Compatible with OpenVMS auditing and access control
  - Uses ASCII configuration files (same as UNIX)

Kerberos

- Scheme for mutual authentication and optional data encryption, developed at MIT
- TCP/IP Services provides “Kerberized” Telnet
- TCP/IP Applications under development
  - RLOGIN, FTP
  - NFS work deferred
- Support for encryption under development for Telnet

What is IPsec?

- Standards based IP-level Solution to Security
- IPsec secures everything above IP
- Provides:
  - ESP (Encapsulated Security Payload)
  - AH (Authentication Header)
  - IKE (Internet Key Exchange)
- Security policy dictates what is encrypted and what algorithms are available during IKE dialog
- When selected, can protect every packet
- Work for both for IPv4 and IPv6

IPsec Components

- Complex set of protocols, mechanisms, tools
  - Engine: processes incoming and outgoing packets in real time
  - Interceptor: interface to the engine
  - Policy Manager: maintains a security policy DB
- Applications:
  - Digital Certificate utilities
  - Cryptographic utilities
  - LDAP utilities
- ISKAMP/IKE: Security Association and Key Management

IPSec Implementation Plan

- Ported from Tru64 UNIX version 5.1B
- Developed by SSH Communication Security, Inc.
- Encryption obtained using OpenVMS CDSA
- See roadmap for delivery timeframe

IPsec vs. Secure Application Layer

- SSH, SSL/TLS
  - Built into each application
  - Controlled by the application
  - Only applies end-to-end
- IPsec
  - Applies to all network traffic
  - Controlled by the system administrator
  - Part of network infrastructure (VPNs)

DNS Security

- TSIG (Transaction SIGnature)
  - Signs DNS messages
  - Uses shared secret
- “Simple” to configure
  - Does not scale well
- Part of BIND 8.2

- DNSSEC is a set of protocols to authenticate the data returned by DNS name servers
- Also secures dynamic update transactions
  - Partial DNSSEC support: part of BIND v9 code

TCP/IP Services Performances Enhancements

TELNET/RLOGIN Driver Performance / Scalability Improvements

- On Multi-CPU systems
  - No longer uses IOLOCK8
  - Added support for multiple concurrent I/O thru TN devices
- Reduced timer maintenance overhead
- Amount of overhead required for maintaining the TN devices has been reduced

NFS Performance Improvements (1)

- Support for a name cache for faster lookups (ODS-2/5)
  - Reduces the number of QIO operations required by the NFS server to look up files by name
  - Logical name establishes the size of the cache TCPIP$CFS_NAME_CACHE_SIZE
- Support for a file system directory cache (ODS-2/5)
  - Retains information about sequential files that will require record format conversion
  - Logical name establishes the size of the cache TCPIP$CFS_ODS_CACHE_SIZE
- These caches increase the virtual memory requirements of the NFS server. See release notes for details

NFS Performance Improvements (2)

- Enable NFS server to take advantage of the directory and name caches by using the NFS attribute ovms_xqp_plus_enabled in SYSCONFIGTAB.DAT
  - This attribute is specified as a bit mask
- Internal performance improvements
  - New hashing algorithms, call reduction, code streamlining
  - Buffer alignment for faster moves
  - Increased number of NFS threads
- Ability for NFS to run on its own CPU on Multi-CPU system

BG devices handling

- Support for More Than 10,000 BG devices
  - Useful on very busy systems like web servers
  - Enabled thru sysconfig by setting ovms_unit_maximum greater than 9999 (< 32K)
  - Alpha only
- Faster UCB creation and deletion
  - Support systems where large numbers of BG devices are continuously being created and deleted or where the number of BG devices has been increased above the 10,000 device unit limit
  - Enabled thru sysconfig by setting ovms_unit_fast_credel
  - This attribute can affect the amount of virtual memory used
    - See release notes for details

TCP/IP Services Scalable Kernel

Scalable kernel

- Design Goals
  - Scale close to linearly over large number of CPUs
  - Reduce MPSYNCH to near zero
  - No longer use IOLOCK8

- Design Considerations
  - Synchronization of internal data must be done with little or no CPU contention (I.e. no MPSYNCH)
  - Transmits and Receives must proceed in parallel
    - Overwhelming majority of CPU cycles consumed in Transmit and Receive

Design Key: Reduce CPU Contention

- No longer use IOLOCK8 thereby removing CPU contention with other OpenVMS IOLOCK8 users
  - Use dynamic spinlock to lock main internal database
  - Use several “mini” spinlocks to lock small subsets of database for small numbers of cycles
- Direct all processing requiring locking of internal data to a single designated “TCP/IP” CPU thereby removing CPU contention with other TCP/IP users
  - For optimal performance ensure that the LAN processing is done on a different CPU

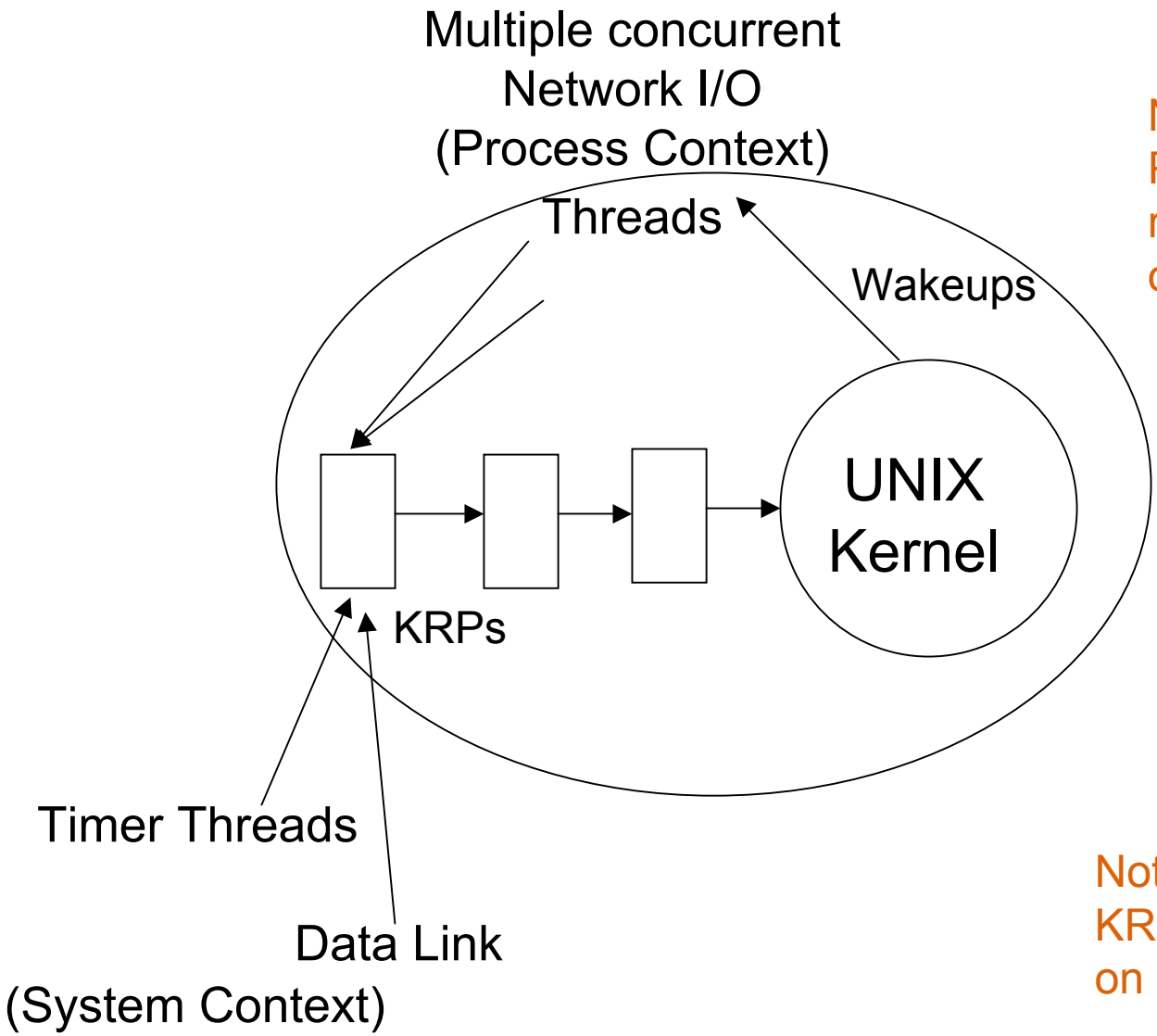
Design Key: Allow Parallelism in Transmits and Receives

- Total redesign/reorganization of the TCP/IP Kernel to break Send and Receive paths into 2 parts
  - Process context based operations done in parallel
    - Prepares (among other things) the work that specifies what must be done in the second part
    - Queues the kernel request to run asynchronously on the TCP/IP CPU
    - Completes user I/O request
  - UNIX Kernel context based operations done under a single dynamic spinlock
- Turns network I/O into asynchronous transaction-based operations

Steady State Multi-CPU Operation

- Many CPUs running TCP/IP applications
  - User applications scheduled and running in parallel on several CPUs
    - They generate TCP/IP kernel requests
    - The only thing that slows them down when they are ready to run is lack of a CPU
- The “TCP/IP” CPU servicing queue of TCP/IP Kernel Request Packets
  - Runs as a fork thread at IPL 8

Transactions based System



Note:
Process context threads
running concurrently
on different CPUs

Note:
KRP queue processed
on a specific TCP/IP CPU

Scaling Limits

- As number of “user” CPUs grows the number of TCP/IP Kernel Requests requests grows
  - The amount of load each “user” CPU puts on the “TCP/IP” CPU is application dependent
  - Function of amount of application processing per TCP/IP Kernel Requests
- Adding more CPUs to the configuration of the system scales almost linearly until TCP/IP CPU approaches saturation

Scalable Kernel

- V5.4 feature with V7.3-2 (OPAL)
  - To enable the scalable kernel, add the following lines to SYS$MANAGER:SYLOGICALS.COM before the command to start TCPIP$STARTUP.COM

```
$ DEFINE/SYSTEM/EXECUTIVE
  TCPIP$STARTUP_CPU_IMAGES "PERF=ALL"
```

  - If TCP/IP Services has already been started, you must reboot the system

TCP/IP Services
TCPDUMP

Packet Tracing - TCPDUMP

- TCPDUMP
  - Provides native packet tracing and file based tracing
  - Native tracing in copyall mode
    - It only sees what the TCP/IP kernel sees
    - No promiscuous support (yet)
  - Boolean based filter expression
    - This example shows tracing of both the ftp control session and ftp data session
    - $ tcpdump ip host lassie and (port 21 or port 20)
  - See release notes

TCPDUMP versus TCPtrace

- TCPDUMP
  - Standard UNIX packet trace analysis tool
  - Binary file
    - Compatible with Tru64 UNIX TCPDUMP
    - Readable from other libpcap applications like Ethereal
- TCPtrace
  - Traditional VMS style command interface with ‘dump’ style output
  - Continued support for current users familiar with this tool

TCP/IP Services
IPv6 Technical Update

OpenVMS IPv6 History & Roadmap

1997- 2000

Early Adopter Kits

- IPv4/v6 Dual Stack
- IPv6 base protocol and addressing
- Stateless Address Autoconfiguration
- Neighbor Discovery
- 32 / 64 bits support
- Ethernet / FDDI
- Basic IPv6 services

2001

TCP/IP Services V5.1

- Add to EAK
- Update to latest RFCs
- RIPng
- Basic socket APIs
- Applications: BIND8.*, SMTP, TELNET, FTP, RSH, RCP, REXEC, RLOGIN, Network Management
- Simple Transition Mechanisms: Tunneling and dual stack

2002

TCP/IP Services V5.3

- Add to Rel 1
- Advanced socket APIs
- Mobile IPv6 Correspondent Node (EAK)
- Generic tunneling IP-in-IP
- Transition Mechanism: 6to4 transition
- Apache web server (Separate Product)
- Mozilla IPv6 support (Separate Product)

2003

TCP/IP Services V5.3

- Add to Rel 2
- Update to latest RFCs
- BIND9
- Mobile IPv6 Home Agent (EAK)
- SCTP (separate product)
- Java IPv6 support (separate Product)

200x

Future Plan (**)

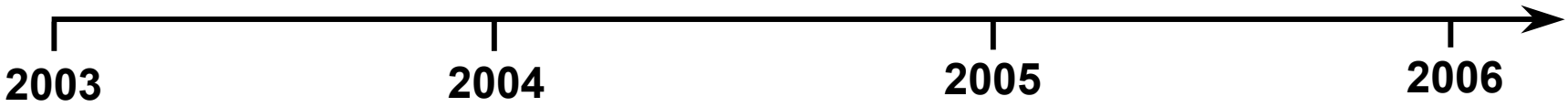
- Add to Rel 3
- Itanium support
- Update to latest RFCs
- Mobile IPv6 Correspondent Node and Home Agent
- More IPv6 Routing
- SSH IPv6 support
- IPSEC IPv6 support
- Additional RFCs as per customer demand
- Leverage of public domain BSD and hp-ux

Note:
Platforms: Alpha / VAX / Itanium (CY04)
Common TCP/IP code across Tru64 UNIX / OpenVMS / NSK
(**) Subject to change without any notice

TCP/IP IPv6 Direction

- Continue adding IPv6 support to product facilities
- Continue adding updated/enhanced IPv6 capabilities
- Support IPv6 Customer transition efforts

TCP/IP Services for OpenVMS Roadmap



TCP/IP V5.4 (Oct 2003)
Featuring IP security and performance enhancements
- SSHv2 client functionality
- failSAFE IP (IP fail over)
- Scalable kernel
- TCP/IP kernel updated
- Performance enhancements to Telnet server
- NFS Server performance enhancement
- 10k+ BG device support
- Bind 9.2.1 upgrade
- SSL POP security
- INETDRIVER performance
- TCPDUMP support

TCP/IP V5.5 on Itanium® (H2 2004)
- Support OpenVMS V8.2

TCP/IP (H2 2005)
Continued focus on performance & security
- BIND V9 Resolver
- More IPv6 Routing
- Standards
- Improved cluster support

TCP/IP (H2 2006)
- IPsec
- DHCPv6
- Standards
- Clustering
- Multi media
- SCTP

failSAFE IP PCSI EAK (Feb 2003)

SSHv2 PCSI EAK (Feb 2003)

IPSEC PCSI EAK

TCP/IP Itanium®

- Track OpenVMS release schedules
- Mako (Itanium® ISV) OpenVMS V8.0 - Release
- Jaws (Itanium® update) OpenVMS V8.1 H2 2003
- Topaz (Itanium®) OpenVMS V8.2 H2 2004
  - First full Customer release
  - TCP/IP Services V5.5
    - NFS server, IMAP and DHCP not committed

TCP/IP Services for OpenVMS Pointers and Contacts

- HP OpenVMS Network Transports Home Page:
  - http://www.hp.com/products/OpenVMS
- Contacts:
  - Product Management
    Lawrence.Woodcome@hp.com

Thanks for Listening!
....any questions?

General Feedback


hp
invent