z/TPF EE V1.1
z/TPFDF V1.1
TPF Toolkit for WebSphere® Studio V3
TPF Operations Server V1.2

IBM Software Group

TPF Users Group Spring 2006

TCP/IP Enhancements

Name : Mark Gambino
Venue : Communications Subcommittee

AIM Enterprise Platform Software
IBM z/Transaction Processing Facility Enterprise Edition 1.1.0

IP Trace Data Conversion - z/TPF APAR PJ30808 (PUT 2)

- Ability to convert TPF IP trace data to TCPDUMP format
  - Data can then be analyzed by open tools such as Ethereal
  - Ethereal is widely used in the industry
- Allows you to analyze IP trace data from TPF using the same tooling that you use for analyzing data from other platforms
- New parameter on the IPTPRT utility:
  - Converts the output to packet capture (PCAP) format
  - Write the converted data to the specified HFS file
- Can still use IPTPRT to select specific packets rather than all the packets on a tape
- Can put packets from multiple tapes into one file

LPAR to LPAR Communications using TCP/IP

- TPF can communicate with another LPAR (Linux, z/OS, another TPF system) in the server box sharing an OSA-Express adapter
  - Packets never leave the server box (do not flow over the Ethernet)
  - Packets flow from LPAR1's memory to OSA-Express memory to LPAR2's memory
- Output messages are sent:
  - If an output buffer becomes full (16 messages is a full buffer)
  - When a timer pop occurs (every 10 ms on TPF 4.1) and any messages exist in the current output buffer
- As the output message rate increases, response time and latency are reduced. For example, on TPF 4.1:
    - At 1000 messages/second, average latency is 5.00 ms
    - At 5000 messages/second, average latency is 1.60 ms
    - At 10000 messages/second, average latency is 0.80 ms

Reduce Latency LPAR to LPAR over TCP/IP
TPF 4.1 APAR PJ31168, z/TPF APAR PJ31198

- TPF now uses a dynamic blocking algorithm for output messages
- Designed to produce consistent and low latency for all message rates and message sizes
- Makes it more attractive to consolidate multiple operating systems onto the same server box and communicate via high speed memory
  - Can now use for time sensitive applications where message rate is low at certain times
  - Can use for transactions that do several message exchanges between LPARs
- Average round trip time (RTT) test results *
  - < 0.1 ms, 100-byte messages at 7500 messages/second
  - 0.5 ms, 500-byte messages at 25000 messages/second
  - 0.6 ms, 1400-byte messages at 23900 messages/second

* Message exchange between 2 native TPF 4.1 LPARs, each with 1 dedicated I-stream on z990. Your results may vary.

TCP Resets (RSTs) on IPL - z/TPF APAR PJ30720 (PUT 2)

- After a software IPL, send RST messages for TCP sockets that were active at the time of the IPL
  - Informs remote nodes that the old sockets no longer exist
- Enables faster application recovery after the IPL
  - For example, remote applications that are suspended from a socket read API waiting for data from TPF
  - Cleans up information in stateful routers and firewalls
- Sending of the RST messages is throttled
  - RST message is the last flow (no ACK to a RST)
  - If you flood the network and some RST messages are dropped, the end result is the same as if the RST was never sent

TCP Resets on IPL Example - Part 1

- z/TPF system environment:
  - 8 GB of memory
  - Dump buffer size is large enough to hold multiple control (CTL) dumps
  - 2000 active connected TCP sockets
- Sequence of events:
  1. Force software IPL (multiple CTL dumps)
  2. After the IPL when the OSA-Express connections are restarted, z/TPF sends a RST message for each of the 2000 sockets

TCP Resets on IPL Example - Part 2

```
CPSE0151T 16.09.42 IS-0001 SS-BSS  SSU-HPN  SE-005329 CTL-I000001 CATASTROPHIC
CPSF0010I 16.09.42 DISK QUEUES PROCESSED
CSMP0097I 16.09.42 CPU-C SS-BSS  SSU-HPN  IS-01
CPSF0014W 16.09.42 CRITICAL RECORD FILING COMPLETED    *** Cause dumps  ***
CSMP0097I 16.09.42 CPU-C SS-BSS  SSU-HPN  IS-01         *** to force a   ***
CPSF0013I 16.09.42 SOFTWARE IPL INITIATED              *** software IPL ***

CVRN0004I 16.10.03 RESTART COMPLETED- 1052 STATE  *** Back in 1052 state ***

ZDBAI DISP                                        *** Display dump buffer ***
CSMP0097I 16.10.05 CPU-C SS-BSS  SSU-HPN  IS-01
DBAI0001I 16.10.05 DUMP BUFFER AREA CONTENTS

SEQ NUM   SYSTEM ERROR   BLOCKS   DUMP SIZE       *** Dumps are being ***
005328*   CTL-I000001               25783K         *** written to tape ***
005329    CTL-I000001      817      25819K         *** after the IPL   ***
                      815
AVAILABLE BLOCKS - 2744
END OF DISPLAY
```

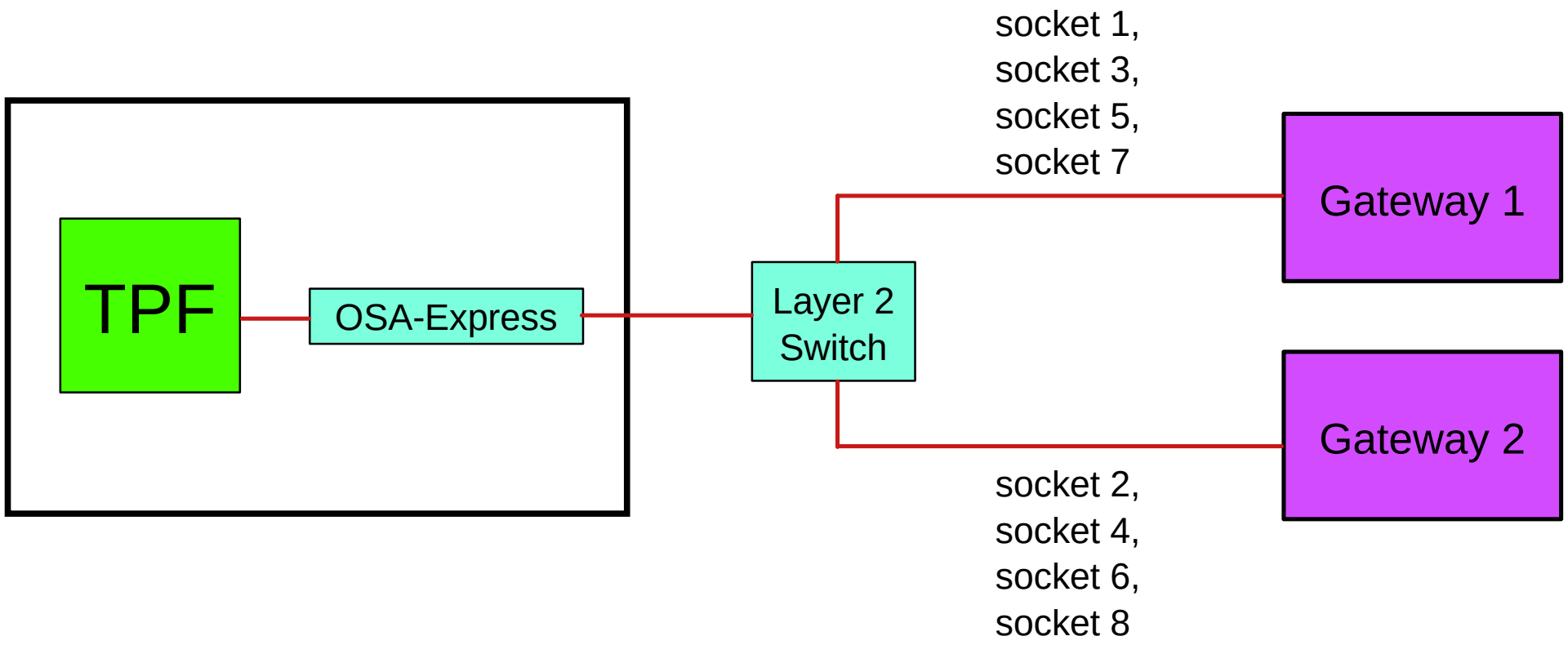
TCP Resets on IPL Example - Part 3

```
CVCX0001I 16.10.15 SS BSS  NOW IN NORM STATE

TTCP0060I 16.10.18 OSA-OSA1 ACTIVATED    *** Start sending 2000 RSTs ***

*** Sample entry in the offline IP Trace output ***
RWI-01  IPCCW-D1  SOURCE IP-9.57.9.198  DEST IP-9.57.13.12  LEN-40
  TOD-BE4459341781C4CA (Jan 25 16:10:18.234908)
  PROTOCOL-06 (TCP)  SOURCE PORT-9999  DEST PORT-1248
 SEQ-2308737483   ACK-2308727988  WINDOW-32767  URGENT OFFSET-0
  TCP FLAG BYTE-14 (ACK, RST)
  REASON CODE - SYSTEM IPL
 IP HEADER    45000028 592F0000 3C06FC5D 093909C6 09390D0C
 TCP HEADER   270F04E0 899C89CB 899C64B4 50147FFF D8E50000

IPTS0000I PROCESSED 3713 FILE RECORDS, SELECTED 2000 TRACE ENTRIES

          *** In this example, it only took 36 seconds ***
          *** to IPL and restart the network           ***
```

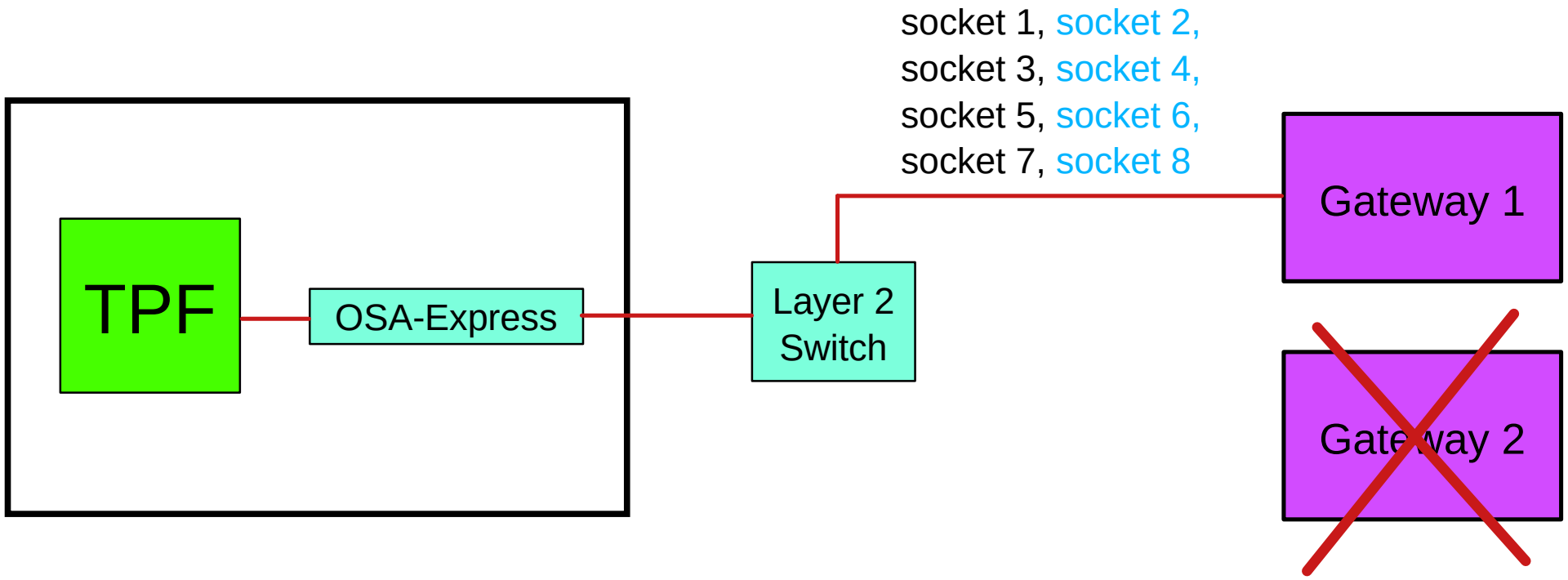
Rebalance Sessions when Failed Gateway Becomes Active Again - TPF 4.1 APAR PJ30781 (PUT 20), z/TPF APAR PJ30960 (PUT 2)

- Each OSA-Express connection can have two default gateways
  - New sessions are distributed equally across both gateways
- If one gateway fails:
  - All sessions now go through the one remaining gateway
  - Each gateway must be able to handle 100% of the load
- When a gateway (GATEWAY1) that failed becomes active again:
  - TPF now rebalances sessions across the two gateways
  - Sessions that were originally using GATEWAY1 are now moved back to GATEWAY1
  - Half the sessions that were started after GATEWAY1 failed (when only one gateway was active) are moved to GATEWAY1

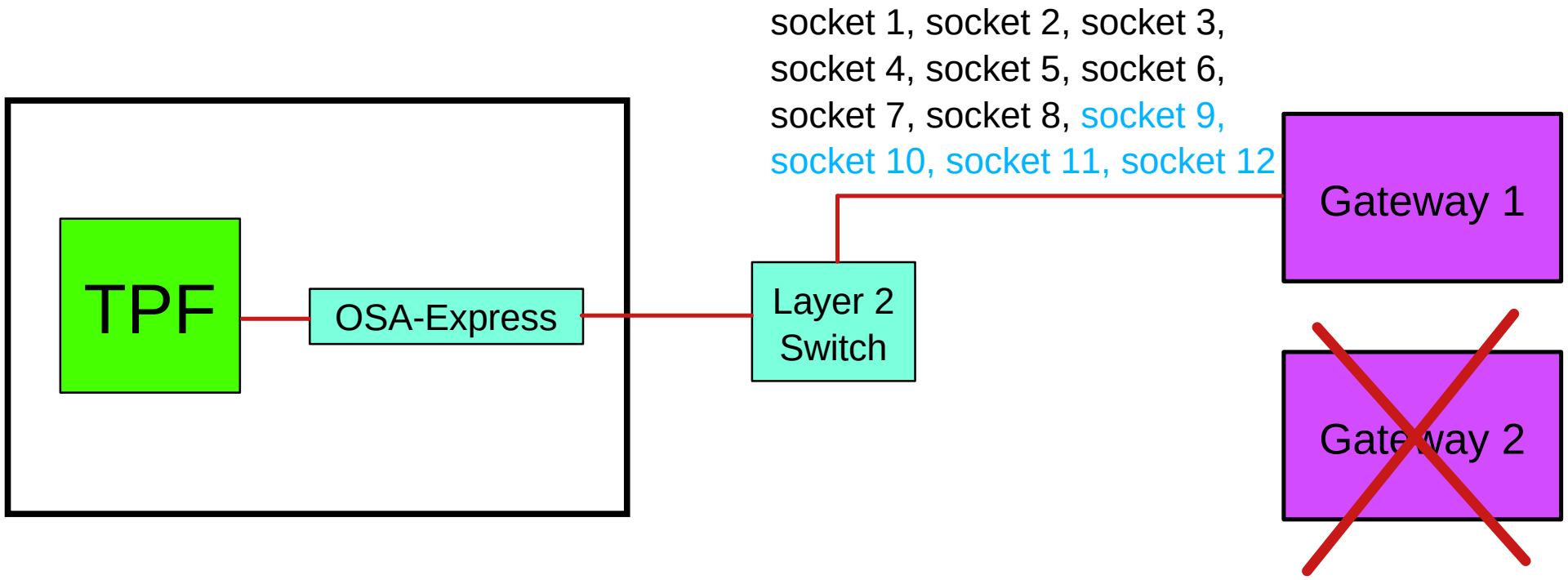
Example: Step 1 - Each Gateway has Some Sockets



Example: Step 2 - Gateway 2 Fails, Sockets Moved to Gateway 1

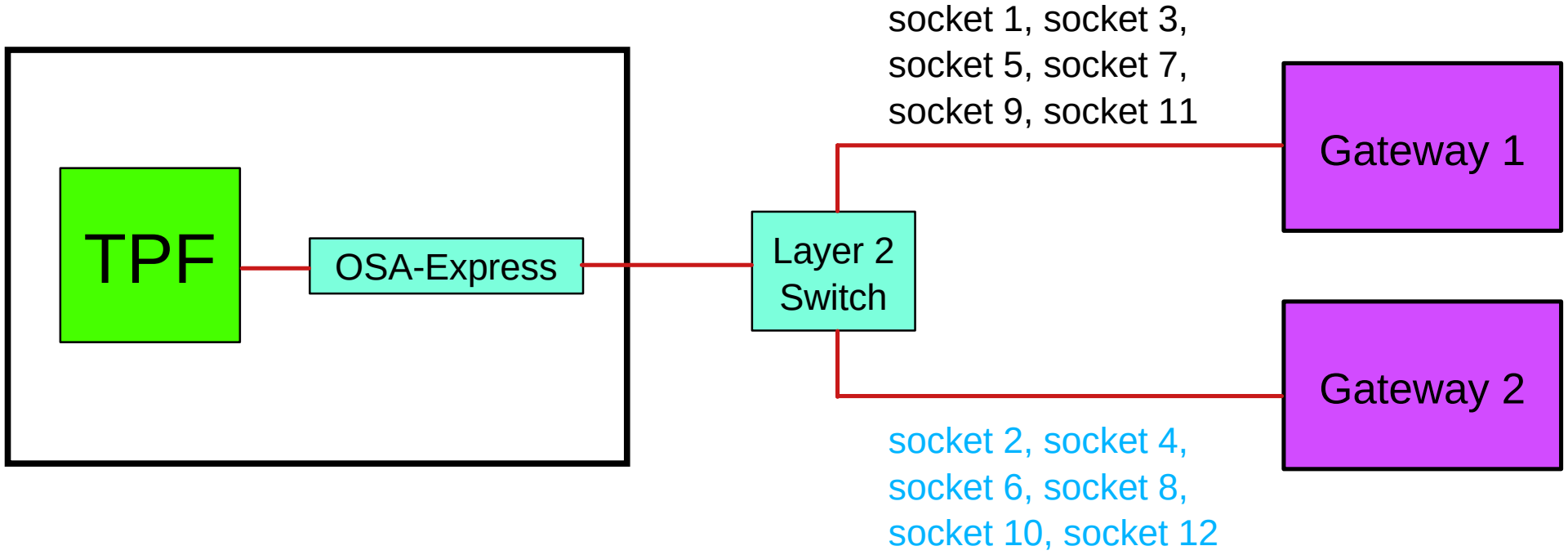


Example: Step 3 - New Sockets are Started



Example: Step 4 - Gateway 2 Becomes Active Again,

Sockets are Rebalanced



poll API support - z/TPF APAR PJ30767 (PUT 2)

- Added support for the *poll* API
  - Similar to the *select* API
- Allows you to monitor file descriptors (FDs) and check status:
  - POLLIN - is there any normal data to be read
  - POLLPRI - is there any priority data to be read
  - POLLOUT - can normal data be written without blocking
  - POLLWRBAND - can priority data be written
- Many application and middleware packages use *select*, *poll*, or both
  - Both APIs are now supported to make it easier to port code to z/TPF

Hardware Acceleration for Starting SSL Sessions

- Starting Secure Socket Layer (SSL) sessions requires significant CPU overhead if the RSA crypto operations are done in software
- PCI cryptographic accelerator (PCICA):
  - Hardware crypto adapter that does clear key RSA operations
  - Can start up to 1000 SSL sessions per second per PCICA card
  - Supported on the z900, z800, z990, and z890 processors
- TPF 4.1 and z/TPF added support for PCICA in 2005
  - RSA operations done in hardware (if installed) rather than in software
  - Enabled TPF to start thousands of SSL sessions/second

Crypto Express2 support - z/TPF APAR PJ30717 (PUT 2)

- Crypto Express2 on System z9 can be configured in one of two modes:
  - Crypto Express2 accelerator (CEX2A):
    - System z9 replaced PCICA with CEX2A
    - CEX2A only does clear key RSA operations
    - Can start up to 3000 SSL sessions per second per CEX2A card
  - Crypto Express2 coprocessor (CEX2C):
    - System z9 replaced the secure crypto card (PCIXCC) with CEX2C
- z/TPF now supports CEX2A
  - z/TPF does ***not*** support CEX2C

Summary of Recent Enhancements

- Ability to analyze TPF IP trace data using open tooling such as Ethereal
- Reduce latency communicating between LPARs
- Reduce application recovery time following an IPL
- Rebalance sessions when a gateway that failed becomes active again
- Support *poll* API to make porting code easier
- Support Crypto Express2 accelerator (CEX2A) to further increase the rate at which SSL sessions can be started

Trademarks

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Ethereal is a trademark of Ethereal, Inc.

Other company, product, or service names may be trademarks or service marks of others.

Notes

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.