z/TPF V1.1

TPF Users Group - Fall 2009

TCP/IP Enhancements

Name: Mark Gambino
Venue: Communications Subcommittee

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

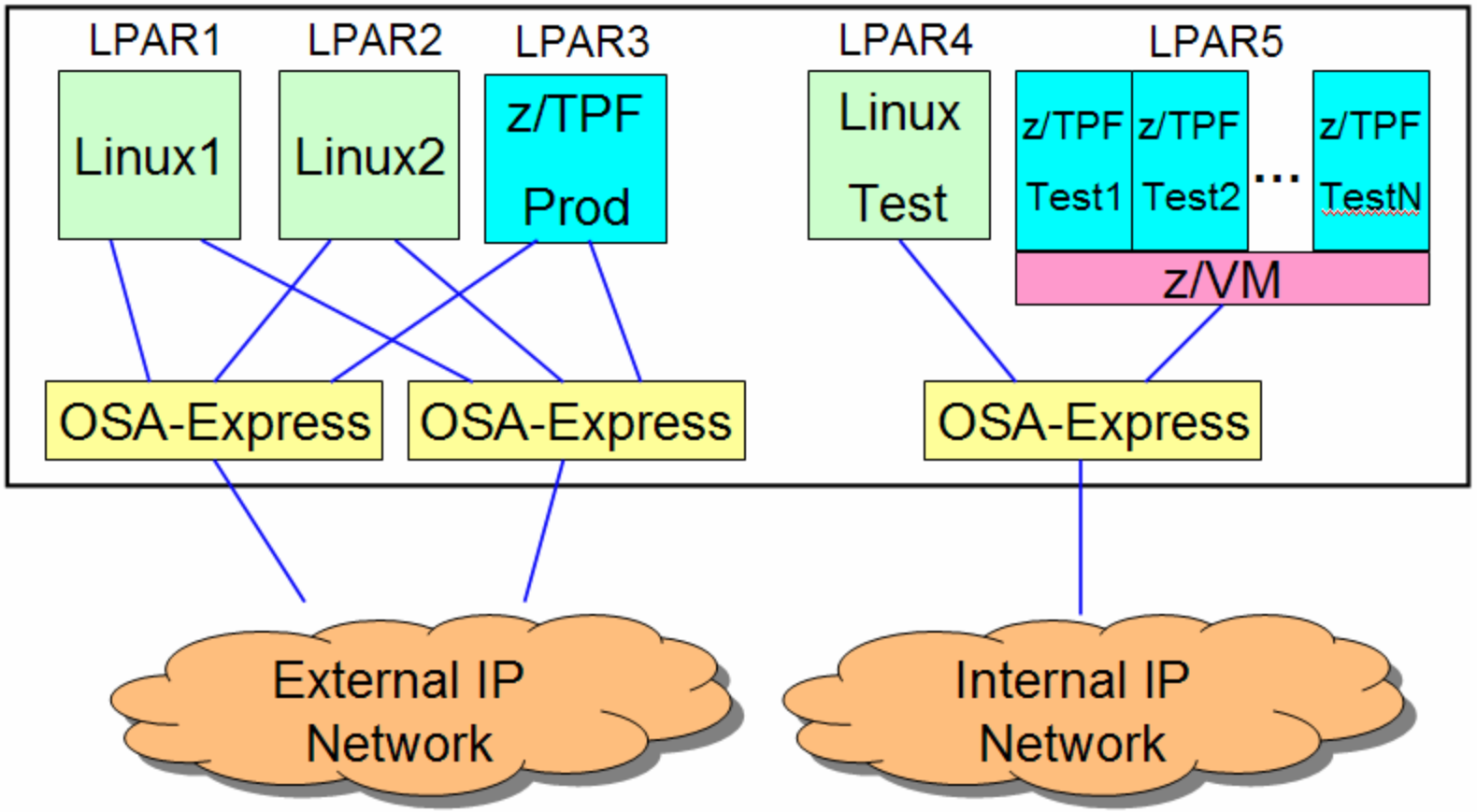
All Transport Protocols Were Not Created Equal

- **SNA is a connection oriented protocol**
  - Routers have knowledge about end user sessions
  - Routers use “store and forward” approach
- **IP is a connectionless protocol**
  - Routers have no knowledge of sessions
  - Routers make “best effort” to forward packets
    - If a router becomes overloaded, packet loss is likely
  - TCP layer at socket endpoints is responsible for detecting and retransmitting lost packets
  - Lost packets can significantly reduce throughput

Inbound IP Traffic

- **Traffic destined for a z/TPF system flows across the IP network and then to z/TPF via an OSA-Express adapter**
- **An OSA-Express adapter can be shared by multiple hosts (LPARs or VM guests)**
  - OSA routes a packet to the appropriate host based on destination IP address in the packet

Sample System z Server Network



Reading Data From OSA

- **A host communicates with OSA using the Queued Direct I/O (QDIO) architecture**
- **Host provides OSA with addresses of read buffers into which OSA will put input traffic destined for this host**
- **OSA sets flags in host memory when a given read buffer has data that should be processed by the host**
- **When the host has processed all the data in one read buffer, the host typically provides OSA with another read buffer**

Host is Slow Reading Data from OSA

- **If OSA posts multiple read buffers (marks them as having data available for the host to process) and the host is slow processing those buffers**
  - OSA generates a PCI interrupt to inform the host that it needs to process the read buffers
- **If the host is running native under shared PR/SM or is running under VM, the host might not be running for a period of time**
  - The (PCI) interrupt is also used as a mechanism to wake up (dispatch) the host

Host is Too Slow Reading Data from OSA

- **If a packet arrives from the network and there is no read buffer available for the target host**
  - OSA queues the packet internally until a read buffer is posted by the host
- **OSA has a finite number of internal buffers for queuing packets**
- **If a packet arrives from the network, there is no read buffer available for the target host, and all the OSA internal buffers are full**
  - Packet is discarded (standard IP router behavior)

Original OSA Support (TPF 4.1, PUT 13)

- **16 read buffers per OSA connection**
- **Timer interrupts (every 10 ms) set a flag that next time through the CPU loop would call OSA polling (which will process messages in read buffers)**
- **PCI interrupts from OSA also set the flag to invoke OSA polling next time through the CPU loop**
- **OSA polling was called in other instances as well**
  - Every time a write buffer filled up and was sent to OSA, polling was called to check for new input messages

Initial Customer Experience

- **Initial customer usage resulted in lost packets in certain environments**
- **Analysis showed TPF was not reading data fast enough**
  - All 16 read buffers had data waiting for TPF to process them and all OSA internal buffers were full
- **Example of the problem:**

1. Data arrives from network and OSA posts read buffers
2. OSA generates PCI interrupt but active ECB runs for another 120 ms
3. More data arrives causing all read buffers to fill up and OSA internal buffers to fill up, resulting in packet loss
4. Running ECB finally gives up control, CPU loop calls OSA polling (120 ms after the fact)

Updated OSA Polling Support (TPF 4.1, PUT 15)

- **Number of read buffers is now configurable**
  - Up to 64 read buffers per OSA connection
- **Timer interrupts and PCI interrupts**
  - No longer set a flag to call OSA polling next time through the CPU loop
  - Instead these interrupts call OSA polling directly when the interrupt is received

Updated OSA Polling Results

- **Good News**
  - Packet loss due to TPF not reading fast enough from OSA was virtually eliminated
- **Not So Good News**
  - Processing packets during external interrupt processing opened up timing problems running multiple I-streams where the interrupted I-stream was holding a core lock that packet processing also needs
    - Worst case causes CTL-571 catastrophic dump (fixed by subsequent TPF 4.1 APARs)

Looking Down the Road

- **Concern**
  - Even though known timing problems have been fixed, the concern is additional function added to the packet processing will use more system services that use core locks and could result in more CTL-571 dumps
- **Solution**
  - APAR PJ33919 – z/TPF no longer invokes OSA polling (no longer processes packets) during external interrupt processing

APAR PJ33919 – Timer (External) Interrupt Processing

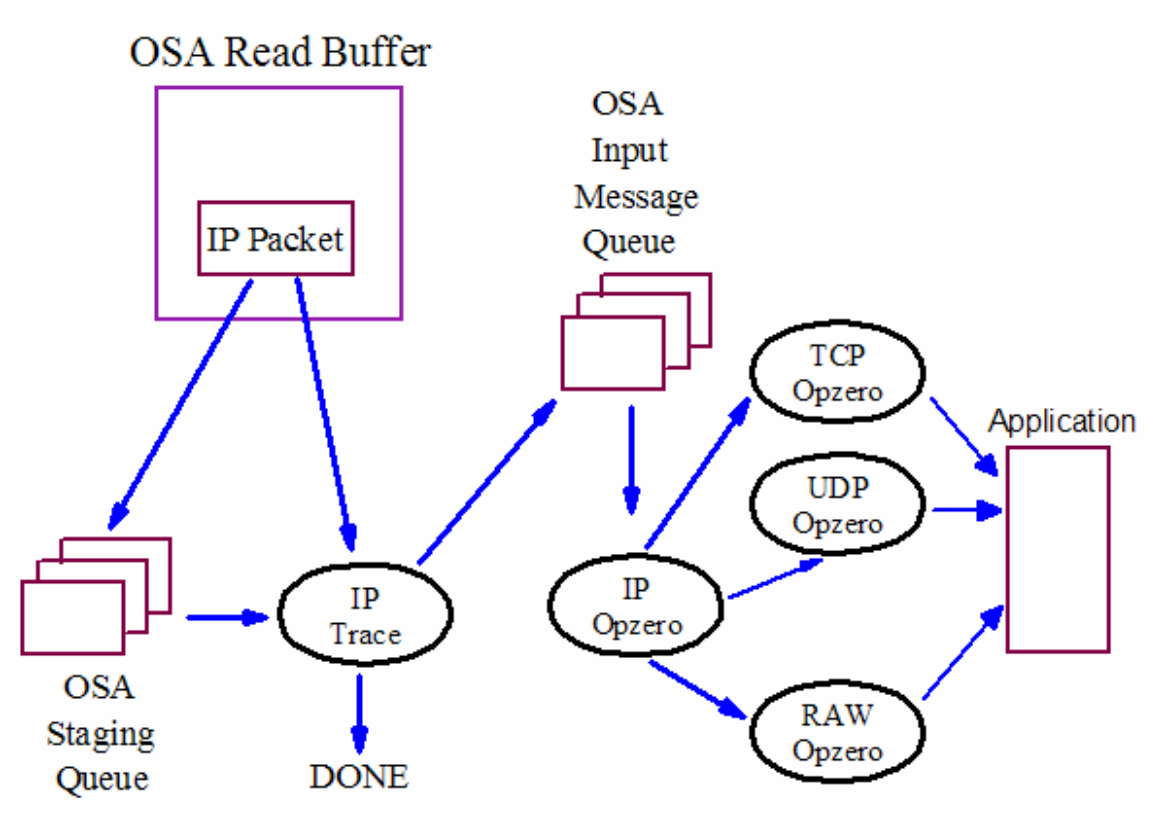
- **Set flag to indicate to do OSA polling next time through the CPU loop**
- **If OSA polling not called by the time the next timer interrupt occurs**
  - Special OSA polling routine is called directly during interrupt processing to move input packets from the read buffers to the OSA staging queue
  - Packets are not processed at all during external interrupt processing

What is the OSA Staging Queue

- **New core queue introduced by z/TPF GA**
- **During z/TPF dump processing**
  - Special OSA polling routine is called periodically to move packets from read buffers to OSA staging queue
    - This reduces the likelihood of OSA internal buffers filling up and causing packet loss
- **When z/TPF resumes after a dump, packets on the OSA staging queue are processed (in a throttled manner) as if they just came in from the network**
- **With APAR PJ33919, OSA staging queue is also used now if z/TPF is slow in processing read buffers**

*** There are actually two OSA staging queues; one for high priority input messages and the other for normal messages**

OSA Input Message Processing Flow



OSA-Express Connections

- **Each host to OSA connection requires 3 subchannels**
  - Write SDA, Read SDA, and Data SDA
- **First generation OSA-Express adapter**
  - 240 of 255 subchannels were available for host connections
  - Maximum of 80 connections per OSA adapter

QDIO Multiple Control Unit Facility

- **Introduced with OSA-Express2 adapter**
- **Allows multiple (up to 16) control units to be defined to one OSA adapter**
- **Allows up to 1920* subchannel addresses to be defined per adapter**
  - Up to 640* host connections per OSA adapter
- **Useful in VM environment if you have many VM guests (z/TPF test systems), each of which requires IP connectivity**

*** Refer to OSA hardware specifications for specific limits**

New z/TPF OSA Support

- **APAR PJ33919 adds support for QDIO multiple control unit facility to z/TPF**
- **Multiple control units are defined via definition of the OSA adapter in the IOCP**
- **No changes to definitions or displays on z/TPF**
  - An OSA connection is still defined to z/TPF (via the ZOSAE command) specifying its 3 subchannel addresses
    - z/TPF dynamically determines which OSA control unit a given subchannel maps to

How to Define Multiple Control Units for OSA in IOCP

- **CHPID Statement**
  - CHPARM=00 is the default. CHPARM=02 must be specified to enable higher number of subchannels (up to 1920) to be defined for this OSA
- **For each control unit for a given OSA**
  - Define CNTLUNIT statement with a unique CUADD value (0-15) and a corresponding IODEVICE statements

IOCP Example for Defining Multiple Control Units for OSA

```
CHPID PATH=(CSS(0),06),TYPE=OSD,SHARED,
      PARTITION=((TPF1,TPF2,LINUX5)),(=)),
      PCHID=111,CHPARM=02
```

```
CNTLUNIT CUNUMBR=0600,PATH=((CSS(0),06)),CUADD=0,UNIT=OSA

IODEVICE UNITADD=00,UNIT=OSA,ADDRESS=(1A00,075),CUNUMBR=0600,
         PARTITION=(TPF1,TPF2,LINUX5)
```

```
CNTLUNIT CUNUMBR=CA00,PATH=((CSS(0),06)),CUADD=1,UNIT=OSA

IODEVICE UNITADD=00,UNIT=OSA,ADDRESS=(CA00,075),CUNUMBR=CA00,
         PARTITION=(TPF1,TPF2,LINUX5)
```

```
CNTLUNIT CUNUMBR=CB00,PATH=((CSS(0),06)),CUADD=2,UNIT=OSA

IODEVICE UNITADD=00,UNIT=OSA,ADDRESS=(CB00,075),CUNUMBR=CB00,
         PARTITION=(TPF1,TPF2,LINUX5)
```

OSA SDA Restrictions Lifted

- **Before PJ33919**
  - Value of UNITADD had to match the last 2 digits of the first subchannel specified by ADDRESS
    - **IODEVICE UNITADD=00,UNIT=OSA,ADDRESS=(1A00,075)**
  - This also meant you could not remap OSA SDAs on z/TPF systems running under VM
- **With PJ33919**
  - z/TPF dynamically determines the unit address for OSA SDA
  - UNITADD does not have to match ADDRESS anymore
  - Can remap OSA SDAs when z/TPF is running under VM now

Questions

Trademarks

- **IBM is trademark of International Business Machines Corporation in the United States, other countries, or both.**
- **Other company, product, or service names may be trademarks or service marks of others.**
- **Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.**
- **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.**