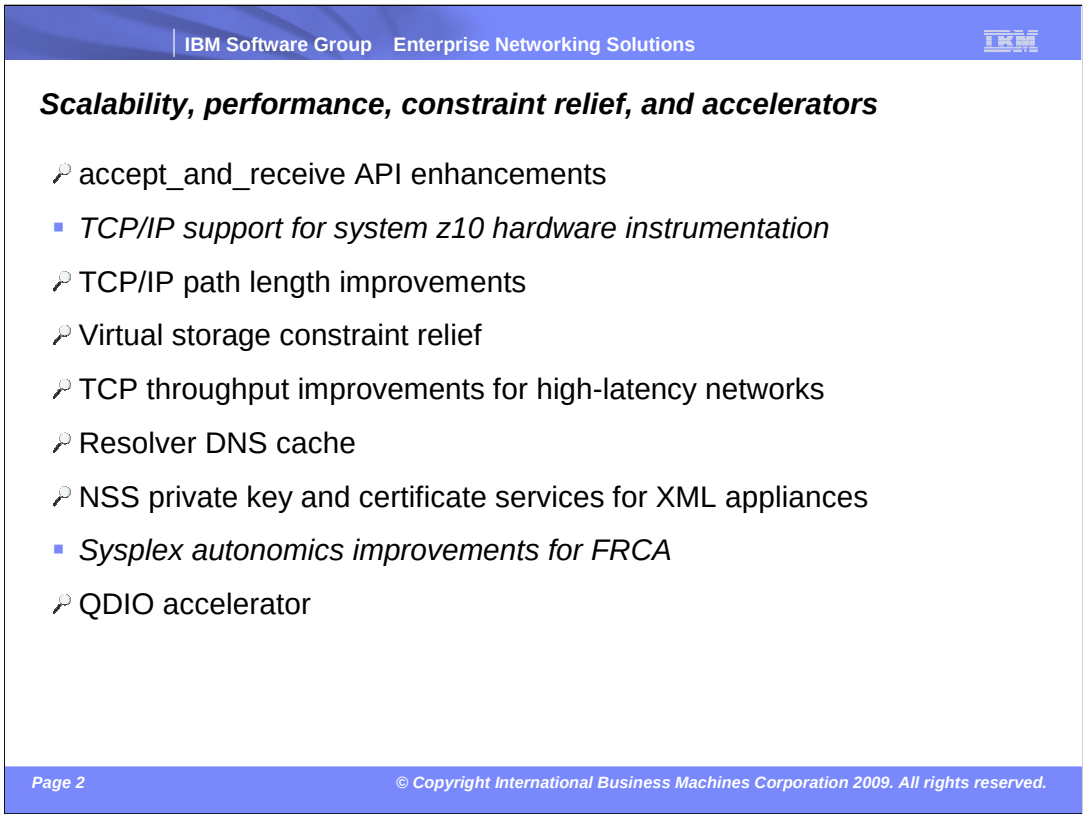
This presentation will give you an overview of the enhancements to the Communications Server in z/OS V1R11 for scalability, performance, constraint relief, and acceleration. This theme is a major area of enhancements in z/OS V1R11 Communications Server.

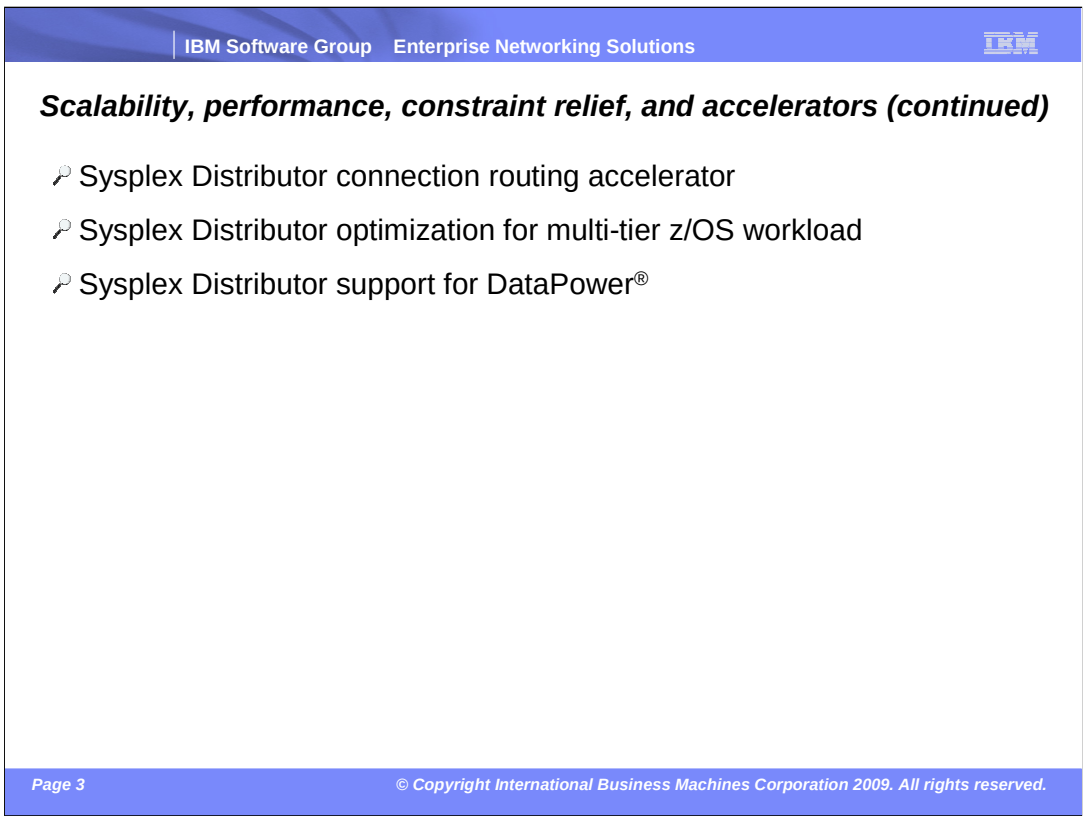


This overview presentation will touch on a few selected items in the list on this slide.

The new asynchronous version of the accept_and_receive sockets API is targeting high-volume servers on z/OS, such as WebSphere® Application Server. Such servers repetitively issue three socket calls for every new connection. This enhancement collapses those three socket API crossings into one – thereby saving processor resources and improving response time.

One of the sockets-related control blocks has been moved into 64-bit common storage from ECSA. This change provides virtual storage constraint relief (VSCR) on high workload systems.

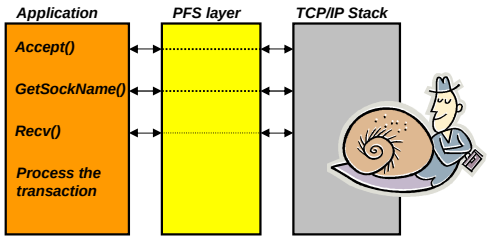
In addition to the general drive to reduce path-length, this release has specific enhancements aimed at improving performance when securing Enterprise Extender traffic with IPSec.



In addition, this overview will discuss the sysplex distributor enhancements in V1R11.

Asynchronous accept_and_receive sockets call background

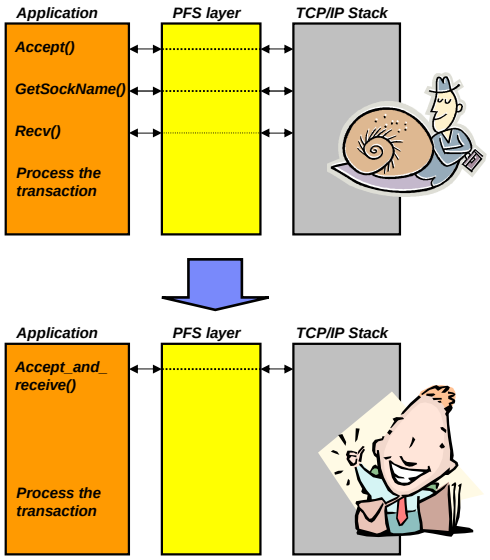
- The accept_and_receive call has existed for a few releases
  - BPX1ANR
- It combines three sockets API crossings into a single API crossing
  - Reduced latency and processor time for server applications that receive connections



The accept_and_recv call BPX1ANR was introduced in OS/390® R7 as a means of optimizing the performance of TCP server applications that immediately issue a recv call. The main idea here was that accept() and recv() can be combined into a single API call, reducing the overhead of traversing both the USS and the TCP/IP layer two times for each API. It also provided the ability to retrieve the local IP address associated with the new connection on the same call, eliminating the need for a getsockname() to be issued. The number of API calls was thus reduced from three to one, boosting performance.

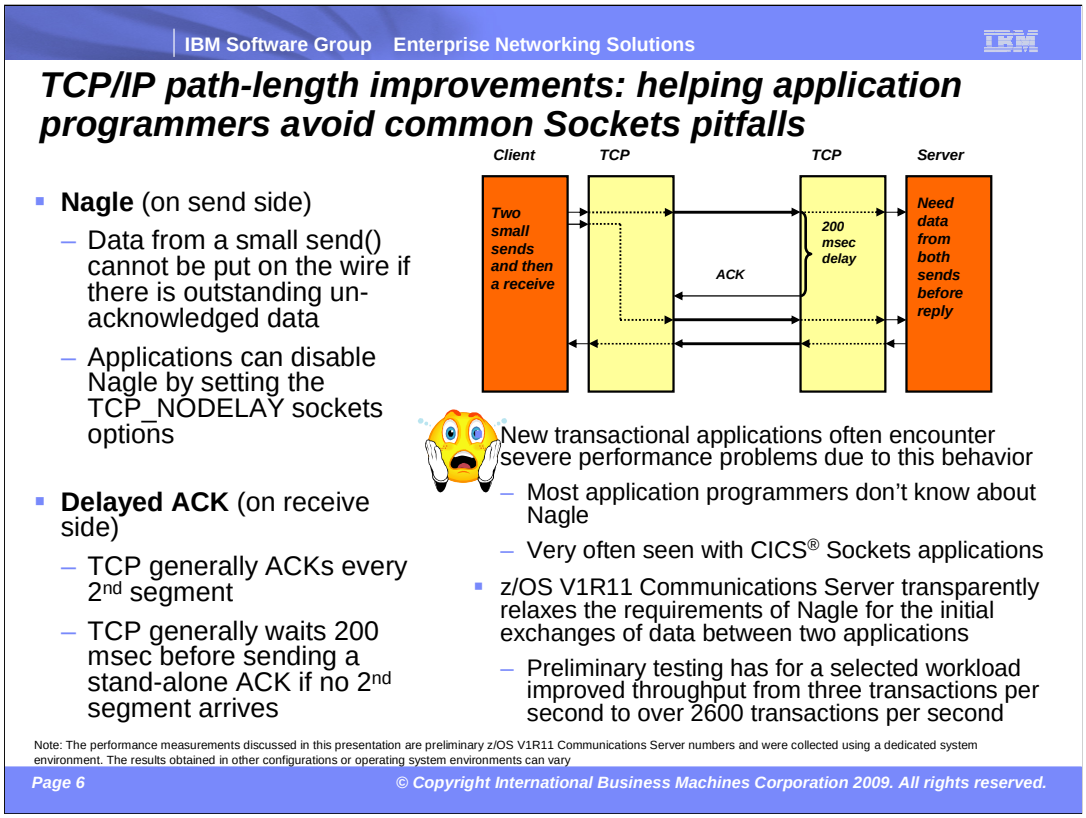
Asynchronous accept_and_receive sockets call in V1R11

- z/OS V1R11 adds these capabilities to the accept_and_receive call:
  - 64-bit support
    - BPX4ANR
  - Asynchronous support
    - BPX1AIO
    - BPX4AIO (64-bit)
- Is available to be exploited by all server implementations on z/OS



In z/OS V1R11, support has been added in BPX4ANR so that 64-bit applications can exploit accept_and_recv.

Support has also been added so that accept_and_recv processing can be done using asynchronous I/O by way of BPX1AIO or BPX4AIO.



Nagle and delayed ACK can have severe performance implications for transactional workloads.

Most application programmers do not know about Nagle and do not set the TCP_NODELAY option.

Nagle is there for a reason, so it cannot be totally ignored.

Even when the application does not set the TCP_NODELAY option, z/OS V1R11 Communications Server will transparently relax the requirements of Nagle for the initial exchanges of data between two applications.

Transactional applications should consider setting the TCP_NODELAY socket option since this 'automatic' behavior only applies to the initial sequences of data exchange.

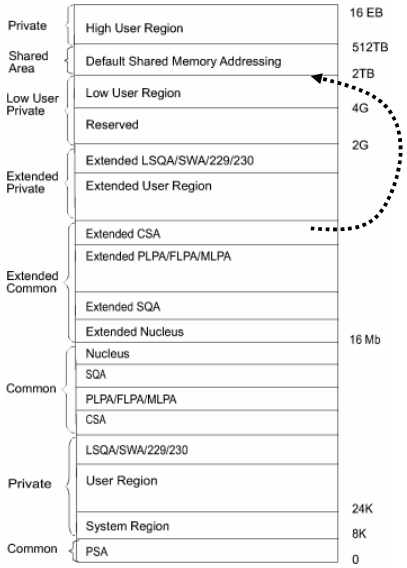
There are other performance improvements in z/OS V1R11 Communications Server, such as, avoid stalling an application that uses msg_waitall. This can occur if an application asks for 64K on a receive using the msg_waitall option, and the TCP receive buffers are smaller than 64K. In that case, the TCP layer will close the receiving window before 64K has been received and the application will basically be blocked forever.

Other performance related enhancements are related to the use of z 64-bit instructions, use of asynchronous cache line pre-fetching, and "scrubbing" of selected code paths for normal cases.

Virtual storage constraint relief

- Data areas that map socket connections were previously obtained from Extended Common System Area (ECSA) common storage
  - Each data area that maps a socket requires roughly 384 bytes of ECSA common storage
  - A large number of socket connections will result in a substantial amount of ECSA being used
- z/OS V1R11 Communications Server moves socket control blocks into 64-bit common storage
  - 64-bit common storage is allocated as common memory objects which are each one MB in size
  - Each common memory object will support 2,730 socket connections
- Helps relieve common virtual storage constraint for everyone on the platform

64-bit common resides just below the shared area that starts at 2TB. Specified by way of HVCOMMON in IEASYSxx



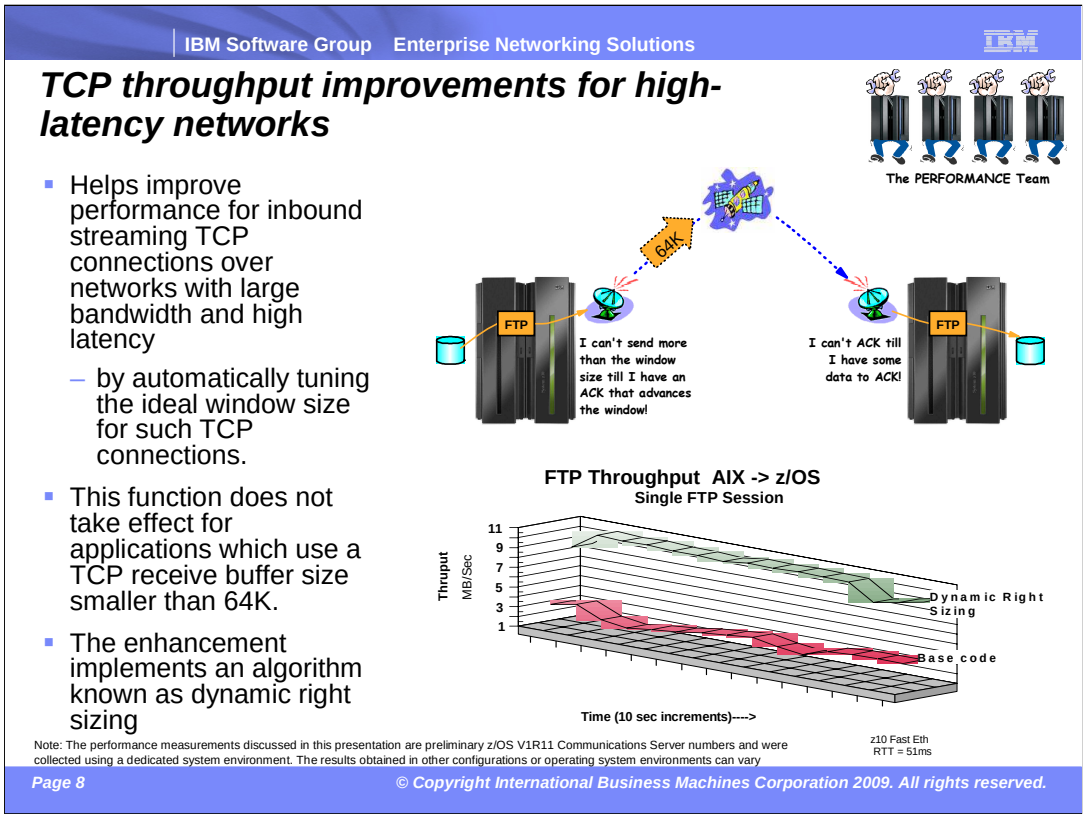
Both the TN3270 server and SNA/EE provide separate ECSA relief functions in this releases.

The single most important ECSA relief comes from moving the socket control blocks out of ECSA and into 64-bit common storage.

Each socket control block occupies 384 bytes. A system with just 2730 sockets occupies roughly one MB of ECSA for this purpose.

64-bit Common storage resides on a 2GB boundary, and the size is a 2GB multiple. The minimum size is 2GB. The maximum size is 1TB. The default size is 64GB.

The 64-bit Common storage size can be specified by way of the HVCOMMON keyword in IEASYSxx, or system parameter HVCOMMON=10G (in this example, the size of the 64-bit common area is 10GB).



Streaming workload over large bandwidth and high latency networks (such as satellite links) is generally constrained by the TCP window size. The problem is that it takes time to send data over such a network. At any point in time, data filling the full window size is 'in-transit' and cannot be acknowledged until it starts arriving at the receiver side. The sender can send up to the window size and then must wait for an ACK to advance the window before the next chunk can be sent.

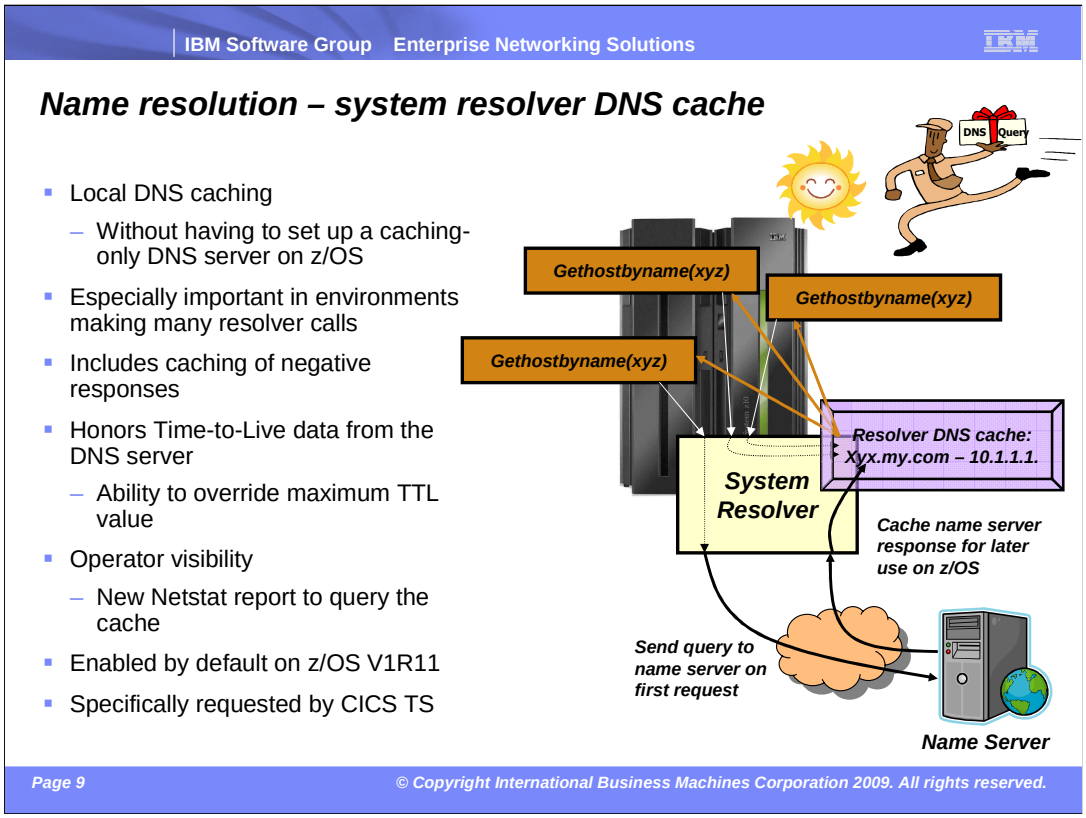
If it were possible to dynamically adjust the window size to what it takes to fill the network in-between the sender and the receiver, higher throughput might be achieved.

This support will, on the receiver side, dynamically adjust the window size upward (beyond 180K if needed) in an attempt to 'fill' the pipe between the sender and the receiver. The goal is that as soon as the sender has sent the end of its window, the sender receives an ACK from the receiver. That ACK allows the sender to advance the window and send another chunk onto the network.

The dynamic right sizing (DRS) algorithm is based on a paper that was published by Los Alamos National Laboratory. The goal of DRS is to keep the pipe full and prevent the sender from being constrained by the advertised window.

The window size can grow as high as two Mbytes. The TCP/IP stack will disable the function if the application doesn't keep up.

A netstat all report shows the DRS-adjusted receive buffer size.



z/OS V1R11 Communications Server implements a global cache in the system resolver. The resolver caches DNS responses without your having to set up a caching-only name server on all LPARs where the function is needed. The cache function is enabled on z/OS V1R11 by default.

This function is especially important if you make many resolver calls, such as outbound Web services requests by CICS, IMS Connect, or WAS. The support is assumed to be valuable in all environments. The ones mentioned here are the ones which are supposed to have the most benefit.

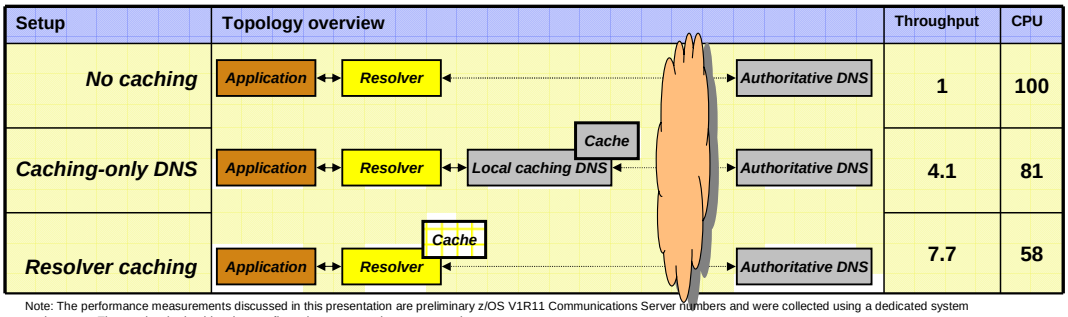
The cache will include the negative responses from name servers.

The cache will honor all time-to-live data from the DNS server – but with the ability for the installation to specify a maximum TTL value to honor.

There is operator visibility through a new netstat report to query the cache.

Resolver DNS cache benefits

- **The performance benefits of local name caching depend on**
  - Amount of calls to the resolver in general
    - Client application workload, Web services workload, some services that do reverse resolution of client IP address, and so on.
  - Amount of repetitive resolutions of the same host names or addresses
    - The more repetitive resolutions, the more cache hits
  - The time-to-live (TTL) values that are returned by the name server
    - TTL values of zero cannot be cached



Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments can vary

Caching of name server replies is especially beneficial for environments that generate a high rate of resolver calls, and where a high percentage of those calls are repetitive resolutions.

Before z/OS V1R11, the only way to provide name serving performance benefits was to configure and run a local name server in caching-only mode.

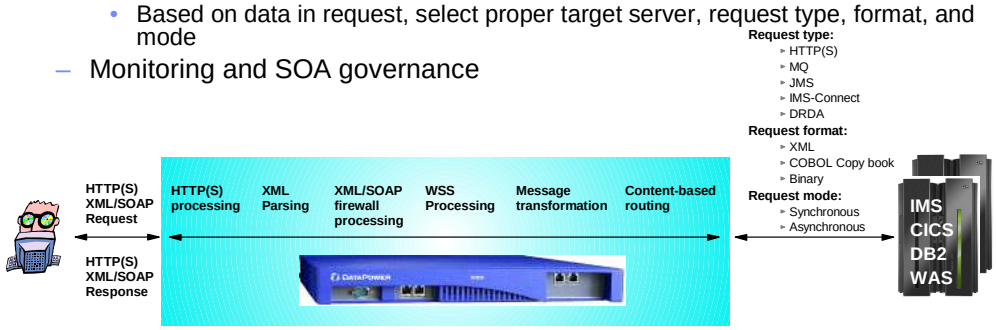
With z/OS V1R11, name server caching is built into the z/OS system resolver.

Preliminary performance results indicate the system resolver caching provides better performance than a local name server in caching-only mode.

For the resolver performance runs, all calls were gethostbyname(). 1000 different host names were used for the test. Thus, with the resolver cache, the number of queries was 1000 higher than the number of hits and a total of 1000 entries were cached. So the cache hit ratio was approximately 100%. So for this test, the first query for a particular name was looked up on the external Linux® DNS server and then was cached by the z/OS resolver. All subsequent lookups were resolved by the resolver cache.

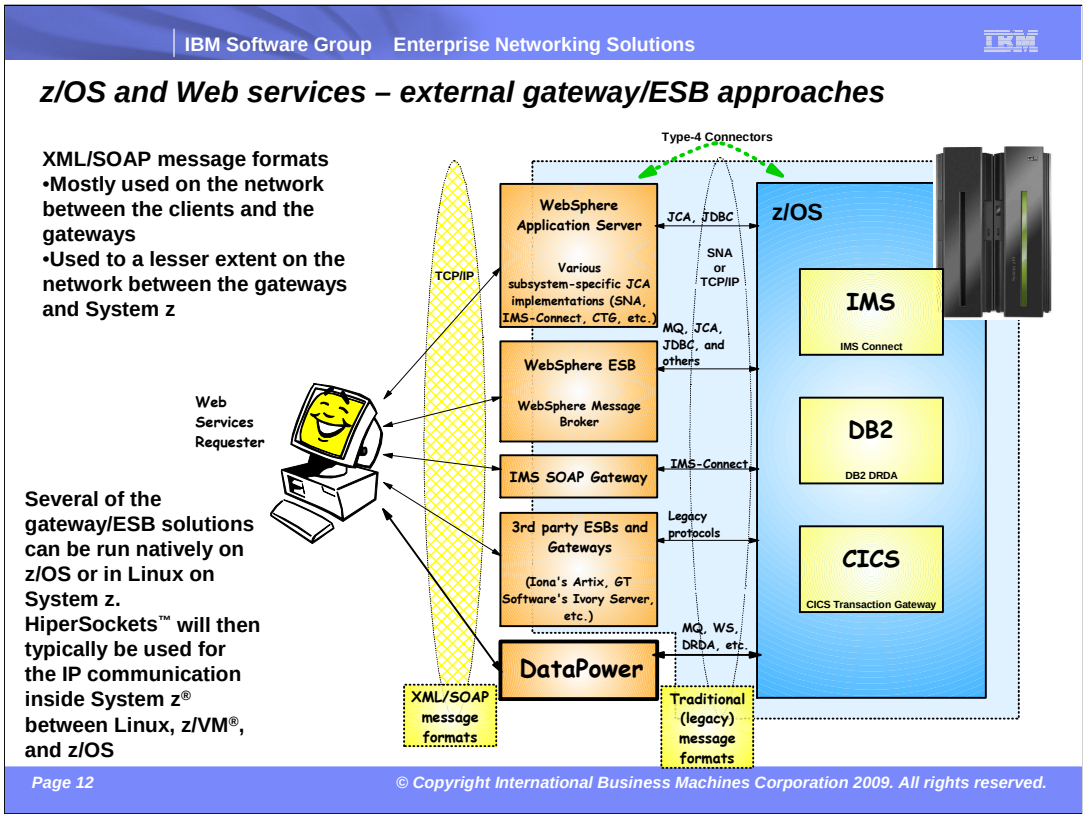
What is DataPower?

- DataPower can perform advanced Web services operations, contents-based routing, and transformation of requests to traditional z/OS applications
  - Security: XML/SOAP firewall capability, Web services security processing
  - XML offload: XML parsing on specialty device
  - Message transformation: transform XML/SOAP to traditional z/OS application data formats and interface with existing z/OS applications
    - Can use HTTP, MQ, JMS, IMS-Connect, or DB2® DRDA®
  - Contents-based routing
    - Based on data in request, select proper target server, request type, format, and mode
  - Monitoring and SOA governance



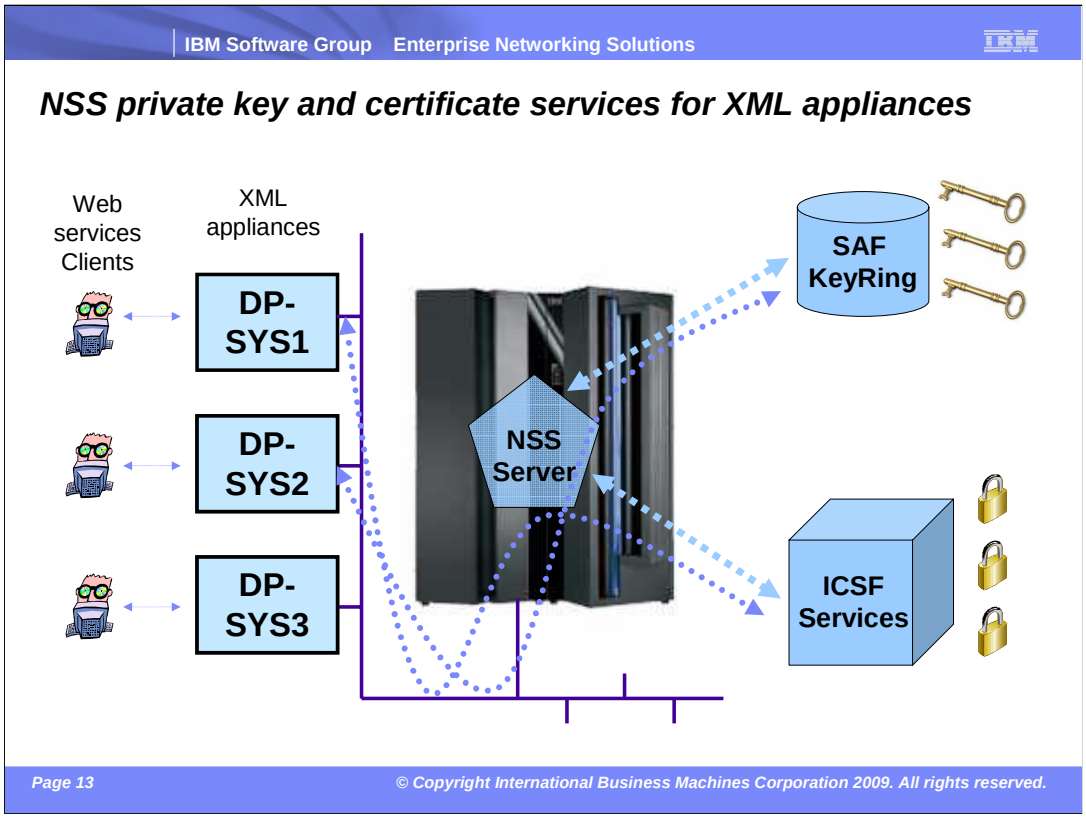
DataPower can be used to provide Web services access to other platforms, such as z/OS.

DataPower provides a full Web services protocol stack, including support for Web services security. DataPower can be customized to act as a Web services gateway to z/OS using traditional transaction interfaces to existing z/OS applications including MQ, IMS-Connect, DB2-DRDA, and others. In such a setup, DataPower can provide the ability to integrate existing z/OS transactions into a Web services environment.



Where do the Web services workload network requirements fit from a System z perspective? It depends on the overall solution architecture. Does the Web services traffic go to a series of front-end gateways that are external to System z (traditional network traffic between them and System z)? Or does the Web services traffic extend all the way into System z (to Linux on System z or z/OS itself)?

Note that z/OS provides full support for native Web services through multiple z/OS-resident components: WebSphere Application Server for z/OS, CICS native Web services, and IMS SOAP Gateway (IMS V10).



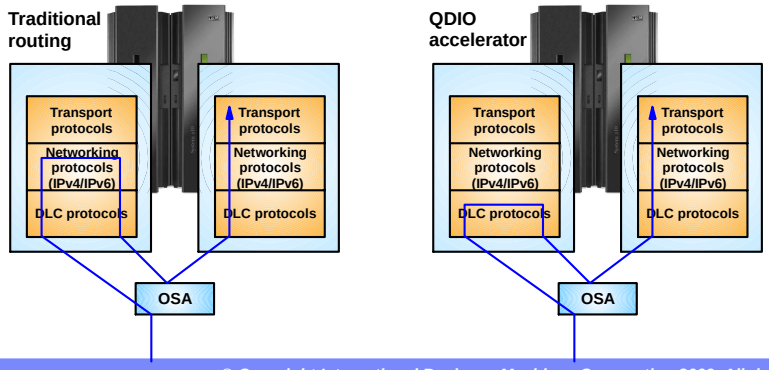
The XML appliance security integration effort is extended in z/OS V1R11 to support authorized sharing of x.509 Certificates and non-ICSF-protected RSA Private Keys between a z/OS SAF KeyRing and authorized DataPower clients.

In addition, the Network Security Services (NSS) server can perform RSA-based signature generation and RSA-based decryption operations using ICSF-protected RSA Private Keys for authorized XML appliance clients.

These new security features effectively extend the centralized certificate services available in the NSS server to XML appliances.

QDIO accelerator

- QDIO acceleration for routing between these network interfaces
  - QDIO-QDIO
  - iQDIO-QDIO
  - QDIO-iQDIO
  - iQDIO-iQDIO
- Based upon the existing HiperSockets Accelerator technology
  - But adds support for Sysplex Distributor connection routing
- Routing table pushed down into DLC layer
- Routing decisions being performed at the DLC layer
  - Less processor resources
  - Lower latency
- Cannot be used in combination with IP Security



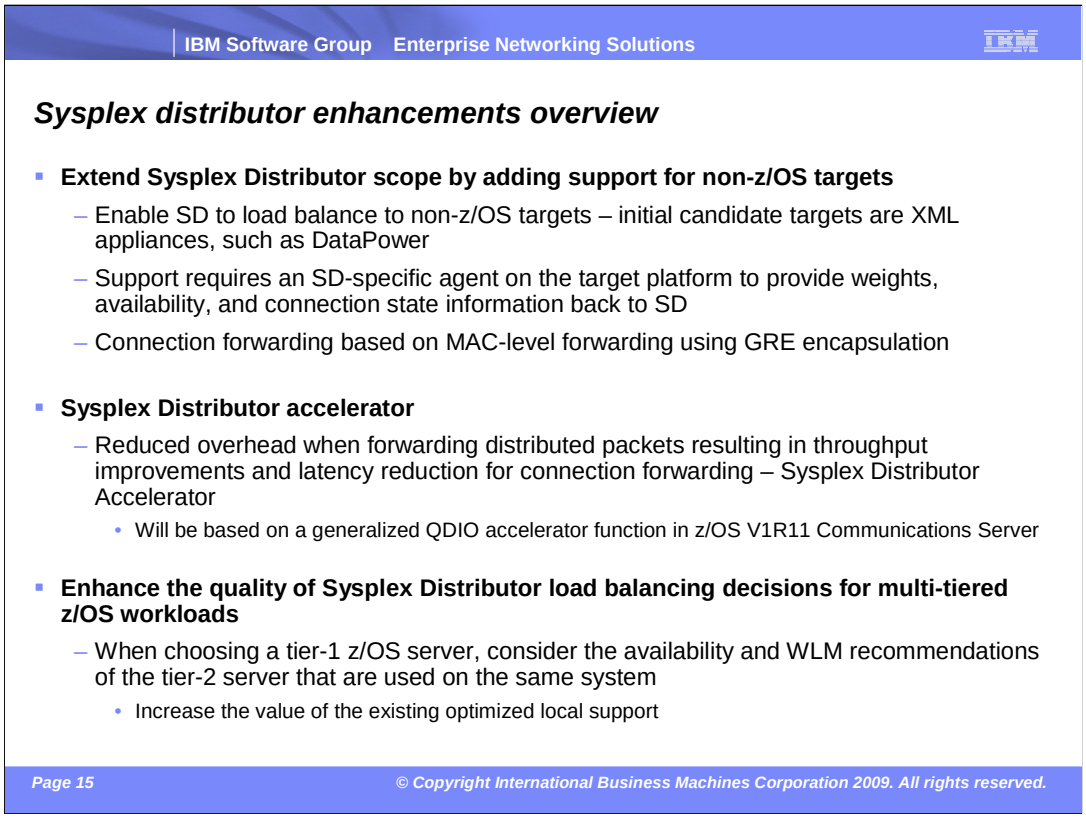
The QDIO accelerator function in this release extends the existing HiperSockets accelerator function. HiperSockets accelerator is able to accelerate general forwarding of IP packets from QDIO to HiperSockets (iQDIO) and from HiperSockets to QDIO. The QDIO accelerator function in this release will support similar acceleration between all combinations of QDIO and iQDIO network interfaces, such as QDIO to QDIO.

The concepts are the same for HiperSockets accelerator. Elements of the IP routing table are ‘pushed’ down into the DLC layer. This enables the DLC layer to make routing decisions between the supported interfaces without passing IP packets up through the IP layer.

HiperSockets accelerator did not support connection forwarding (done by SD for inbound packets belonging to distributed connections). The R11 QDIO accelerator function will also support sysplex distributor acceleration.

As with HiperSockets accelerator, QDIO acceleration cannot be used in combination with IPSECURITY enabled. When IP security is enabled, logic in the IP layer must examine IP packets and make various security-related decisions based on the IP security policies. These functions are not pushed down to the DLC layer.

The QDIO accelerator in z/OS V1R11 will support IPv4 forwarding only.



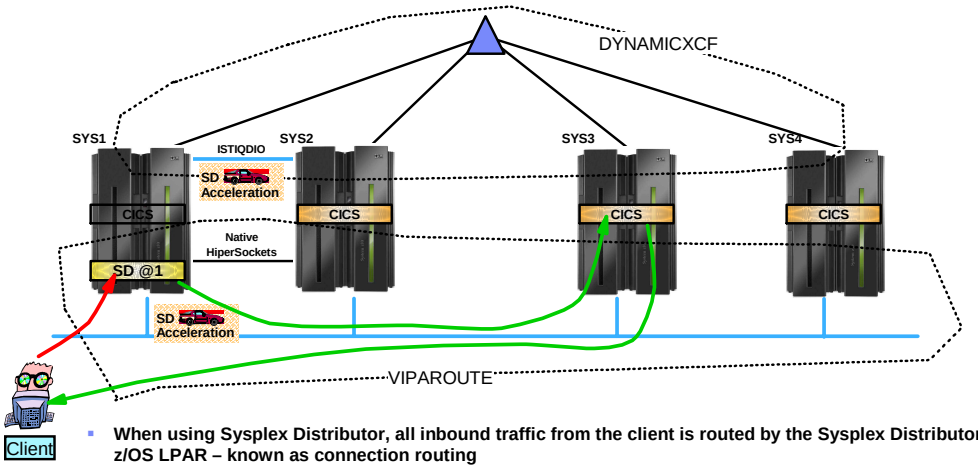
There are extensive enhancements to the sysplex distributor (SD) function in this release. This slide summarizes those enhancements. The next slides will review some of them in more detail.

SD in this release adds support for distributing workload (TCP connections) to target systems other than z/OS. Doing so also means that SD cannot rely on XCF communication with the target systems and it cannot rely on WLM to provide server weights for the targets. SD solves that by implementing support for an SD-specific agent on the target systems that uses an SD-specific protocol to communicate with SD. The agent provides metrics back to SD, which SD uses to determine availability and capacity of the target servers. In addition, the agent also provides SD with connection state information so SD can maintain its normal connection routing table information, and allow non-disruptive takeover of non-z/OS targets by a backup SD. Incoming IP packets to connections that have been distributed to a non-z/OS target are forwarded to the target using generic routing encapsulation (GRE). The initial target platform is IBM DataPower.

SD also implements accelerated connection routing – aimed at reducing processor overhead and latency for inbound routing through the distributing node.

Finally, SD improves optimized local processing. The distributing stack factors in metrics for both the tier-1 servers on each target TCP/IP stack and metrics for the tier-2 servers that are used as a target by that tier-1 server on the same TCP/IP stack.

Sysplex distributor accelerator for connection routing

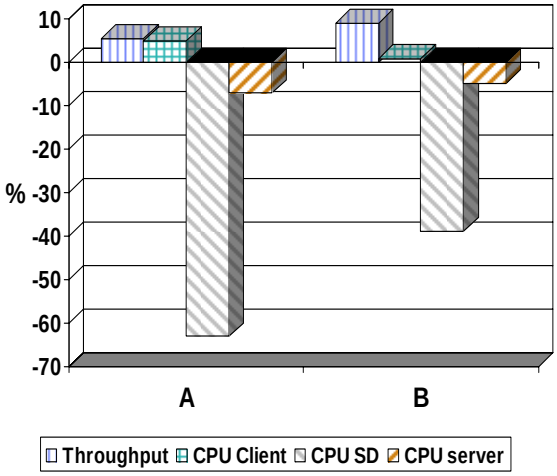


- **When using Sysplex Distributor, all inbound traffic from the client is routed by the Sysplex Distributor z/OS LPAR – known as connection routing**
  - Outbound traffic goes directly back to the client
- **When inbound packets to Sysplex Distributor use QDIO or iQDIO, Sysplex Distributor will perform accelerated connection routing when**
  - Outbound is a DYNAMICXCF iQDIO interface
  - Or when the outbound interface is a QDIO network interface
  - Helping reduce processor overhead and latency in the Sysplex Distributor LPAR (SYS1)

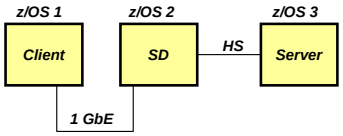
Sysplex distributor acceleration will benefit all existing distributed connections. This benefit is available as long as the SD routing is done over a DYNAMICXCF HiperSockets network interface (not XCF itself). The benefit is also available when using VIPAROUTE over a QDIO network interface (a shared LAN infrastructure). Note that SD acceleration will not occur when VIPAROUTE is used and the outbound interface is HiperSockets.

Sysplex distributor accelerator performance

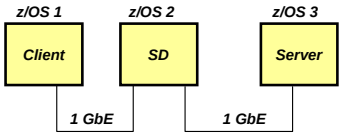
- ✓Intended to benefit all existing Sysplex Distributor users
- ✓Measurements with Interactive workload
- ✓Small data sizes (100 in, 800 out)
- ✓Percentages relative to no acceleration



Configuration A – Three z10 LPARs with OSA Express3 cards and HiperSockets between SD and server LPARs

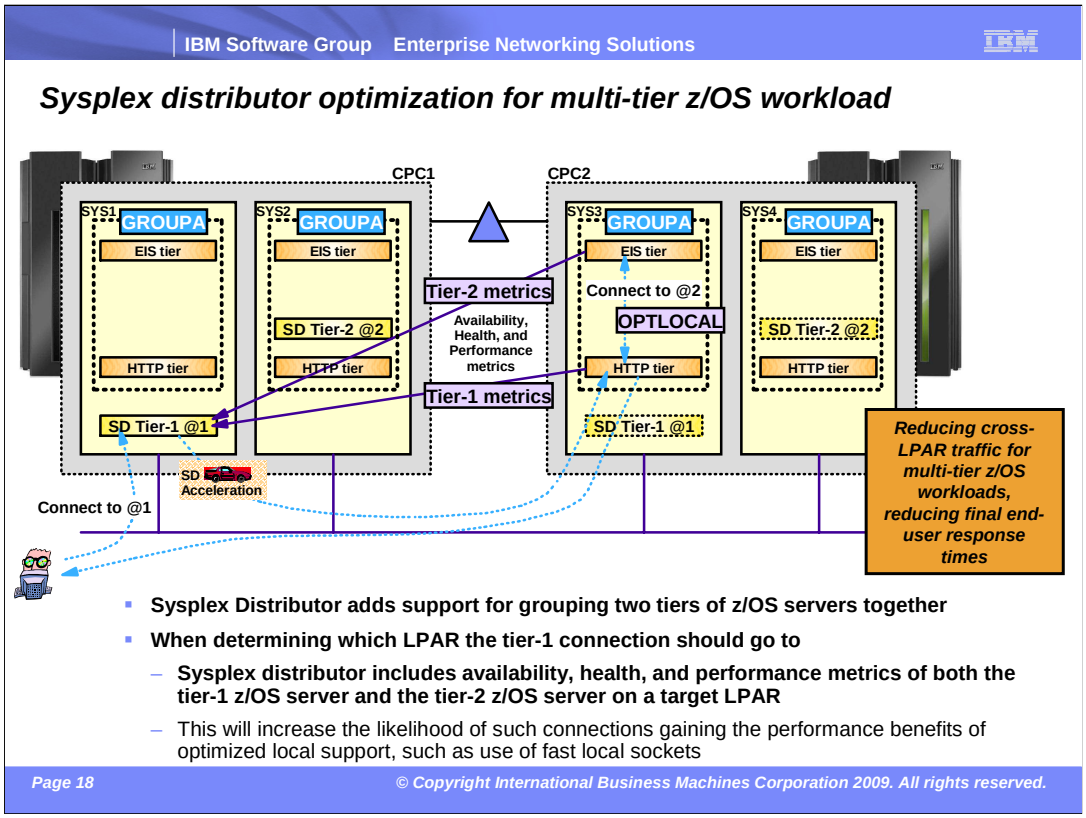


Configuration B – Three z10 LPARs with OSA Express3 cards



Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments can vary

Use of sysplex distributor accelerator is expected to provide a noticeable reduction in processor usage on the distributing stack that does the connection routing of inbound IP packets.

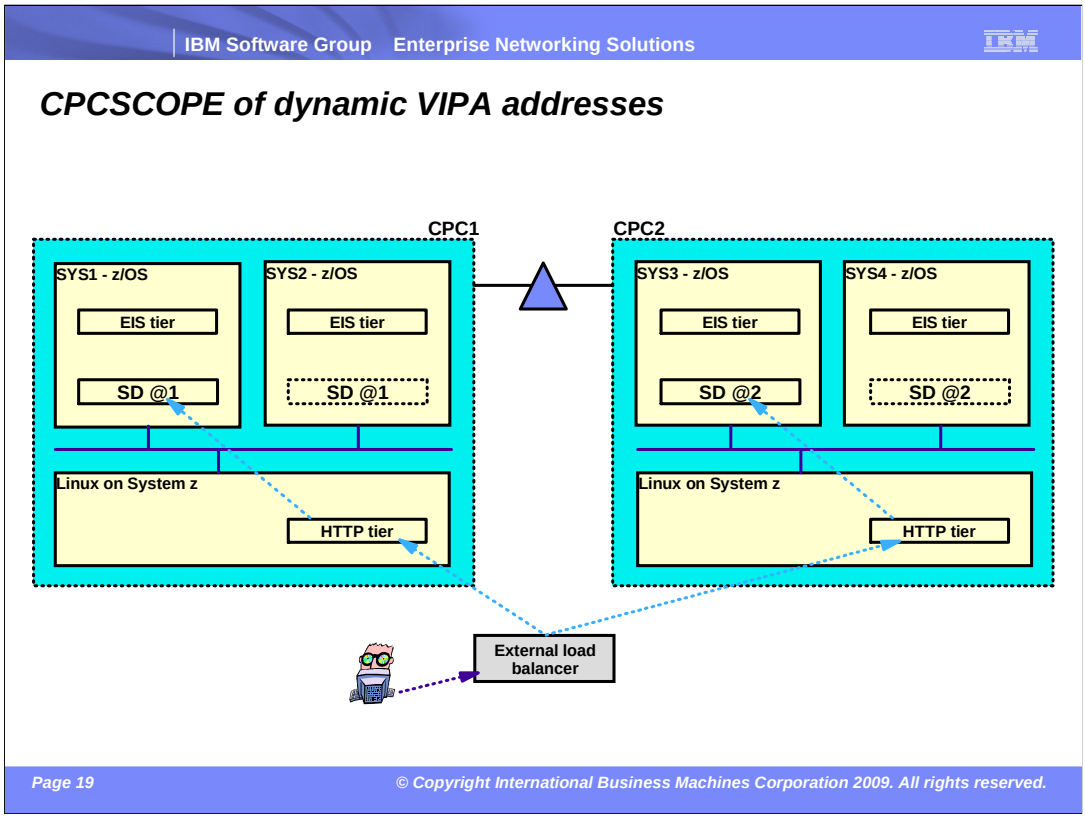


Multi-tier support for z/OS targets will provide an enhancement to the existing z/OS Communications Server optimized local support.

Sysplex distributor will include availability, health, and performance metrics of both the tier-1 z/OS server and the tier-2 z/OS server on a target LPAR when determining which LPAR the tier-1 connection should go to. This will increase the likelihood of such connections gaining the performance benefits of optimized local support, such as use of fast local sockets.

In the example on this slide, a connection request comes into SD on SYS1. SD consults with WLM and the TCP/IP stacks in the z/OS sysplex to determine availability, health, performance, and capacity of the target systems. This is done for both the HTTP tier server instances and the EIS tier server instances on each LPAR. When the chosen HTTP server connects to the tier-2 server and optimized local support in in effect, that second connection will stay on the chosen LPAR. When optimized local is configured with a default value of 1, the second connection will stay local if the WLM weight is greater than 0 and the server is healthy.

The support increases the likelihood that the optimized local option for the 2nd tier connection will indeed be effective.



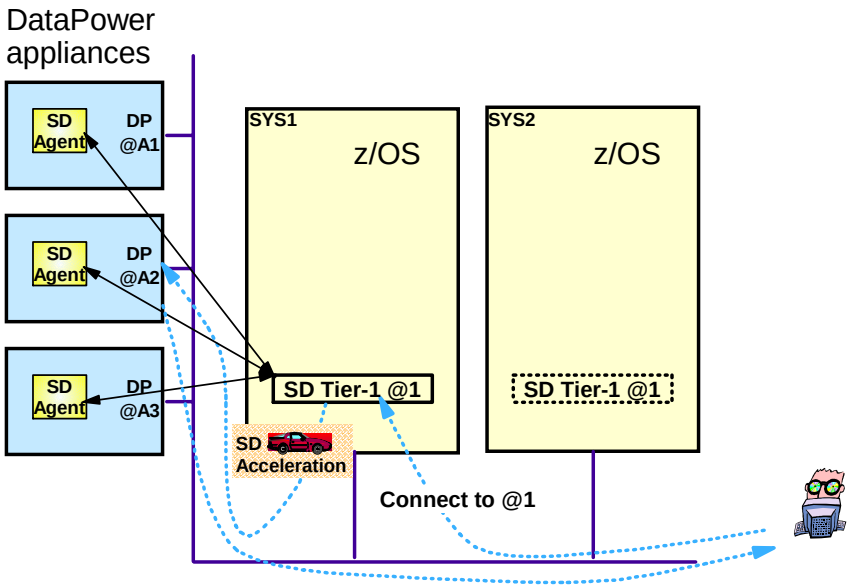
Dynamic VIPA addresses can be defined in z/OS V1R11 Communications Server with a scope of a CPC.

Such DVIPAs will never be activated on LPARs that reside in other CPCs than where the address is currently defined.

In scenarios where Linux on System z acts as the first tier server, CPCSCOPE DVIPAs can be used to keep tier-2 connections and traffic on the same CPC where Linux is running. The HTTP tier on CPC1 can be configured to connect to SD@1 for the EIS tier, while the HTTP tier on CPC2 can be configured to connect to SD@2 for the EIS tier.

This topology will keep traffic between the Linux systems and the z/OS systems on HiperSockets interfaces. The goal is to improve overall response times due to reduced cross-CPC traffic in mixed Linux on System z and z/OS workload scenarios.

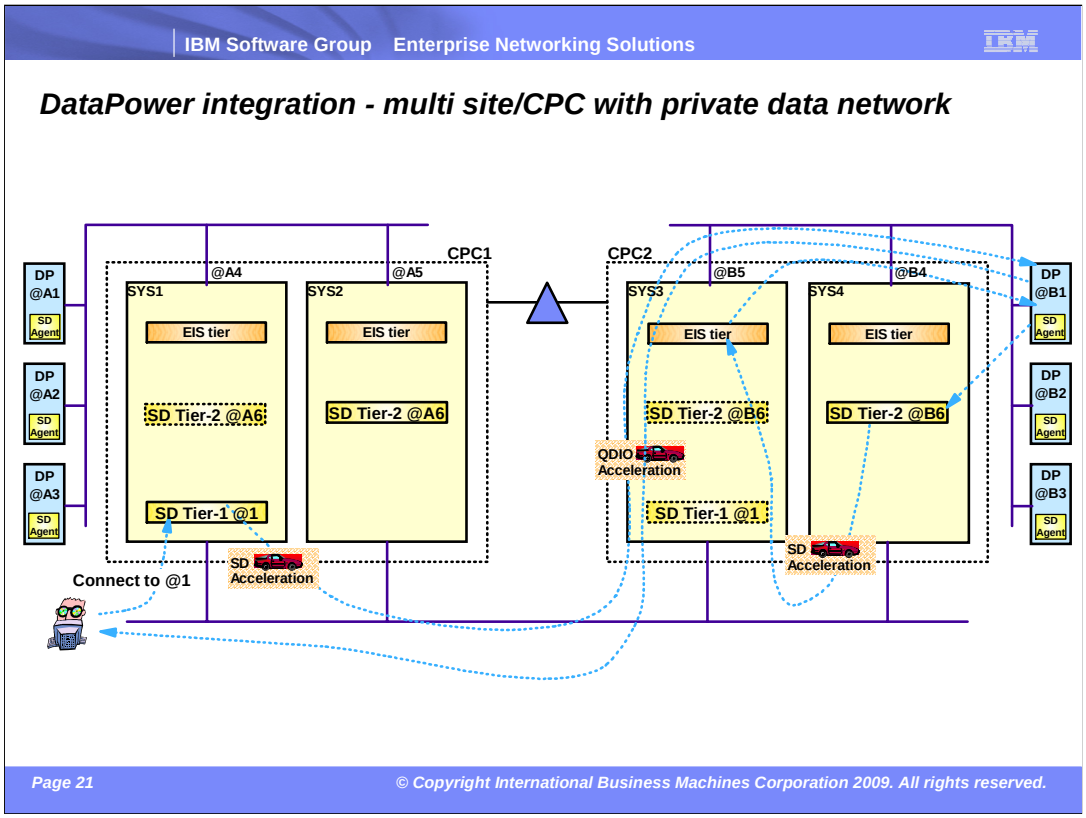
DataPower integration with sysplex distributor



z/OS V1R11 Communications Server continues to focus on logical integration between DataPower and z/OS by supporting a DataPower feedback technology to z/OS sysplex distributor.

This support allows sysplex distributor to include DataPower appliance availability, health metrics, and performance metrics when load-balancing connections to a set of DataPower appliances.

This logical integration allows sysplex distributor to make much higher quality load balancing decisions than any other existing load balancing technology, when load balancing connections to DataPower appliances.



This support applies to installations with a multi-site (multi-CPC) sysplex with DataPower appliances and tier-2 z/OS applications in both sites. In such cases, sysplex distributor is able to tie DataPower targets as tier-1 servers together with z/OS tier-2 servers. This can be done in such a way that cross-site communication for tier-2 connections can be eliminated.

When sysplex distributor selects a DataPower appliance in one of the two sites, it does so based on the availability, health, and performance of both tiers, the tier-1 DataPower appliances and the tier-2 z/OS servers.

If the network between the z/OS systems and DataPower appliances is implemented as a secured network, the overhead of doing connection-based encryption/decryption can be eliminated. This can further enhance the overall performance of the solution.

Trademarks, copyrights, and disclaimers

IBM, the IBM logo, ibm.com, and the following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

CICS	DataPower	DB2	DRDA	HiperSockets	OS/390	System z
WebSphere	z/OS	z/VM				

If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of other IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

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

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

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements or changes in the products or programs described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (for example, IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products.

IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance 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 or performance improvements equivalent to the ratios stated here.

Note to U.S. Government Users - Documentation related to restricted rights-Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract and IBM Corp.