IDUG The Worldwide DB2 User Community

EMEA Conference 2010

DB2 & System z Synergy

Christopher J. Crone
IBM DB2 for z/OS Development

Session: A11

November 10th, 2010 • 13:00 – 14:00
Platform: DB2 for z/OS

Abstract

DB2 for z/OS has a long history of exploiting z/OS and System z enhancements. Examples of these include ESA and 64 bit Architectures, and Parallel Sysplex (DB2 Datasharing). That's the past - this session focuses on the present and the future. In this session, we will discuss how DB2 is taking advantage of the latest changes in System z HW and SW to bring you a more robust, scalable, and affordable computing platform today, and in the future.

DB2 and System z Synergy

Most SW is designed to be platform neutral. DB2 is not only designed **NOT** to be platform neutral, it is designed to take advantage of the platform to provide industry leading capabilities that no other database SW can match.

DB2 Development works closely with the System z HW and SW teams to not only take advantage of enhancements on the platform, we drive many of the enhancements that you see on the platform.

This presentation will explore some of the more recent enhancements that continue to make System z the Gold Standard for mission critical – bet your business applications.

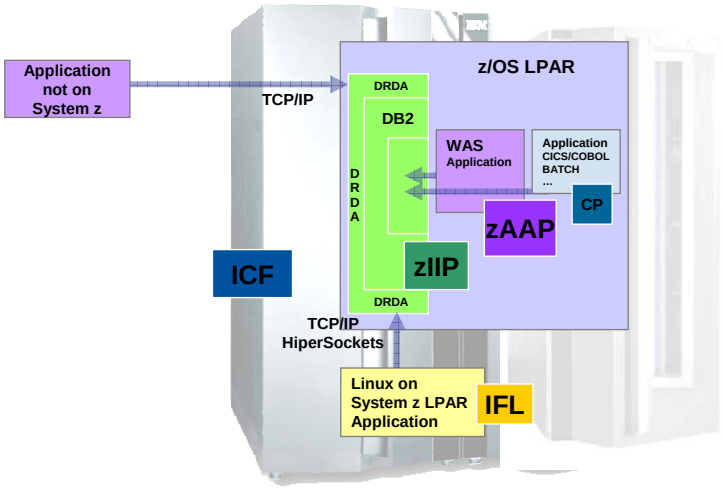
Agenda

- Overview of the System z Environment
- Building on a Solid Base- System z Architecture
- z/OS, WLM, and DB2
- Communications
- A peek into the Future

Overview of the System z Environment

System z:

The Premier Platform for Workload Consolidation



Challenges

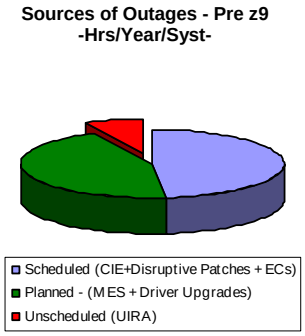
- Remove Constraints to growth
- Improve Reliability and Availability
- Continue to improve TCO
- Adapt to changing computing paradigms
- Improve Performance across the board

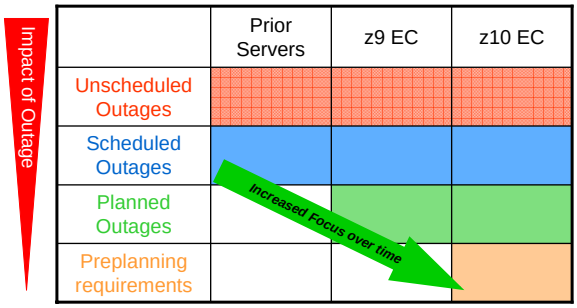
Building on a Solid Base- System z Architecture

Problem

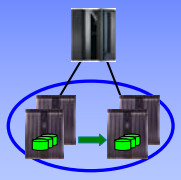
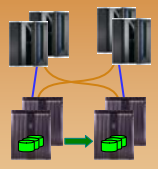
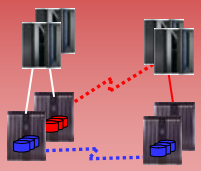
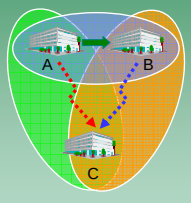
- Planned Outages are making it hard to meet my SLAs

Solution: System availability is central to our design





Solution: Continuous Data Availability

Continuous Availability of Data within a Data Center	Continuous Availability / Disaster Recovery within a Metropolitan Region	Disaster Recovery at Extended Distance	Continuous Availability Regionally and Disaster Recovery Extended Distance
Single Data Center Applications remain active Continuous access to data in the event of a storage subsystem outage  GDPS/HyperSwap Mgr RPO=0 & RTO=0	Two Data Centers Systems remain active Multi-site workloads can withstand site and/or storage failures  GDPS/PPRC RPO=0 & RTO<1 hr	Two Data Centers Rapid Systems Disaster Recovery with "seconds" of Data Loss Disaster recovery for out of region interruptions  GDPS/GM & GDPS/XRC RPO secs & RTO <1 hr	Three Data Centers High availability for site disasters Disaster recovery for regional disasters  GDPS/MGM & GDPS/MzGM

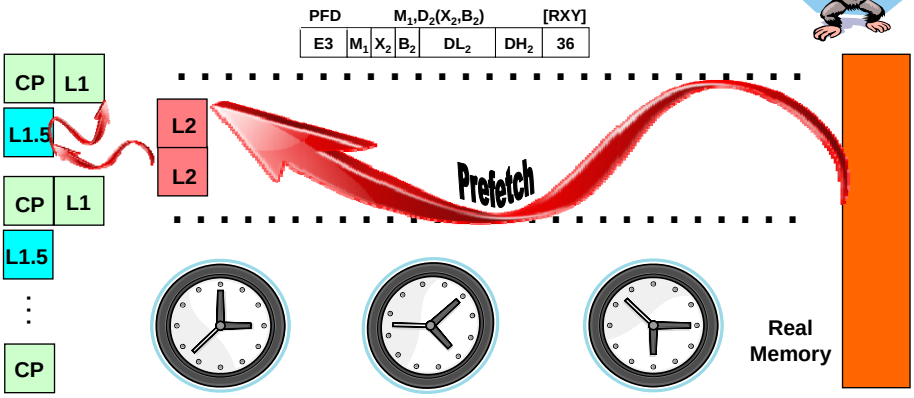
Problem

- Memory Speed Not Keeping Pace With Processor Speed

Problem: Memory and Cache Latencies

Cache and memory latency on a hypothetical modern server

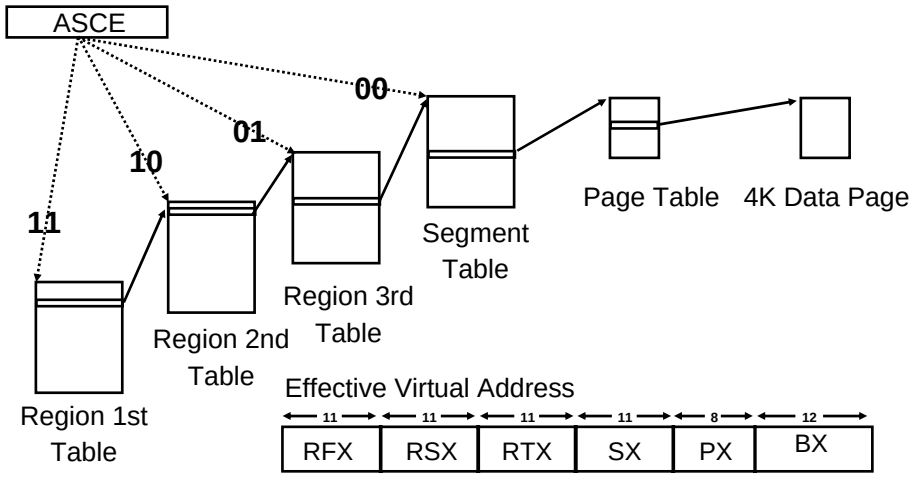
- **L1 Cache – 1 machine cycle**
- **L1.5 Cache – 4 machine cycles**
- **L2 Cache – variable, 10's of machine cycles**
- **Real memory – ~ 600 machine cycles**



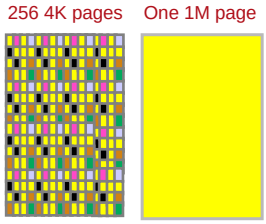
Problem

- Accessing Larger Memory Creates Inefficiencies

Problem: Address Translation for 4K Pages

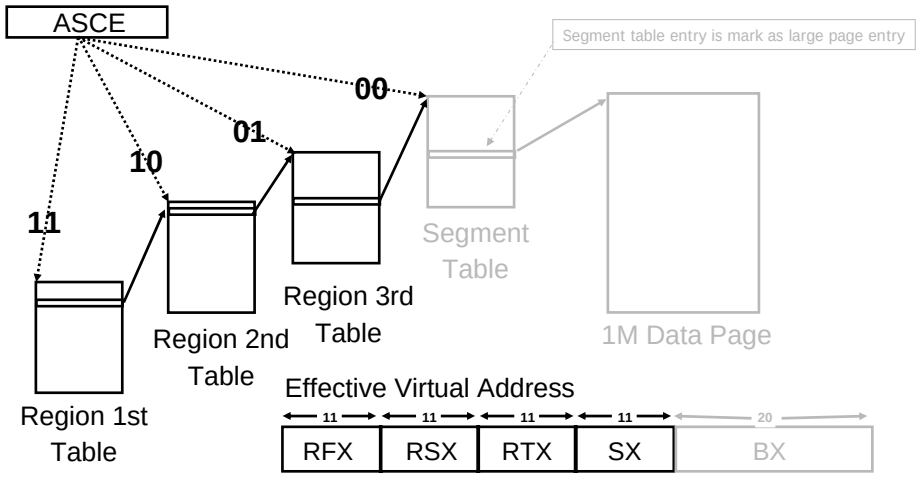


Solution: Large Pages



- Problem: Translation Lookaside Buffer (TLB) Coverage shrinking as % of memory size
  - Over the past few years application memory sizes have dramatically increased due to support for 64-bit addressing in both physical and virtual memory
  - TLB sizes have remained relatively small due to low access time requirements and hardware space limitations
  - Therefore TLB coverage today represents a much smaller fraction of an application's working set size leading to a larger number of TLB misses
  - Applications can suffer a significant performance penalty resulting from an increased number of TLB misses as well as the increased cost of each TLB miss
- Solution: Increase TLB coverage without proportionally enlarging the TLB size by using large pages
  - Large Pages allow for a single TLB entry to fulfill many more address translations
  - Large Pages will provide exploiters with better TLB coverage
- Benefit:
  - Better performance by decreasing the number of TLB misses that an application incurs
  - DB2 10 uses Large Pages to back PGFIX(YES) Bufferpools

Address Translation for 1M Pages



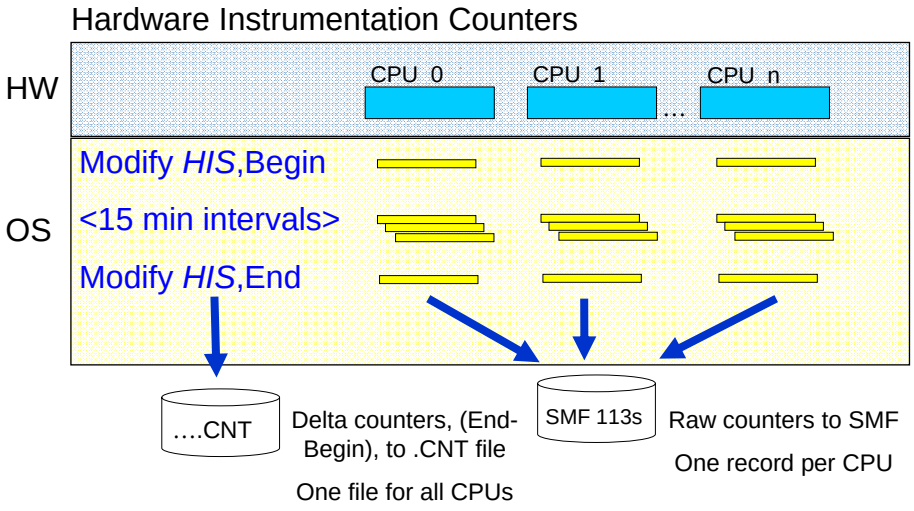
Problem

- Debugging and Tuning Applications and Middleware is difficult because traces and other tools like APA are not practical in a production environment

Solution: The z10 CPU Measurement Facility

- New hardware instrumentation facility “CPU Measurement Facility” (CPU MF)
  - Available on System z10 EC GA2 and z10 BC
  - Supported by a new z/OS component (Instrumentation), Hardware Instrumentation Services (HIS)
- Potential Future Uses – for this new “cool” virtualization technology
  - CPU MF provides support built into the processor hardware
    - So exploiting mechanism allows the observation of performance behavior with nearly no impact to the system being observed
  - Potential Uses
    - Future workload characterization
    - ISV product improvement
    - Application Tuning

How it works


Hardware Instrumentation Counters
HW
CPU 0
CPU 1
...
CPU n
OS
Modify *HIS*,Begin
<15 min intervals>
Modify *HIS*,End
….CNT
Delta counters, (End-Begin), to .CNT file
One file for all CPUs
SMF 113s
Raw counters to SMF
One record per CPU

Problem

- Floating Point numbers are approximations
- Where a calculation is executed (COBOL, JAVA, DB2 for z/OS, DB2 LUW…) in the stack determines the result
- Packed Decimal fixed point calculations rules are complex and limiting

Solution: Decimal Floating Point

- New datatype DECFLOAT
  - –Well suited to typical financial calculations
  - –Similar to “calculator” mathematics
    - •Eliminates rounding errors by using base 10 math
    - •Has up to 34 digits of precision
    - •Floating point convenience with fixed point precision!!!
  - –Hardware support will be provided in the next System z processor generation (new IEEE standard) – z9 millicode, z10 HW support
    - •Software emulation provided for other models
- Implemented across the DB2 Family

Problem

- Compute intensive applications such as BI don't coexist well in OLTP environments.
- Schema design and Tuning in a BI environment are labor intensive

Solution: DB2 & zIIP

Portions of the following DB2 for z/OS V8 and DB2 9 workloads may benefit from zIIP (DB2 9 in blue)*:

1 - **ERP, CRM, Business Intelligence or other enterprise applications**

- **Via DRDA over a TCP/IP connection (enclave SRBs)**
- **DB2 9 for z/OS Remote native SQL procedures**
- **DB2 9 XML parsing via DRDA to fully utilize zIIP**

New Specialty Engine

2 - Data warehousing applications

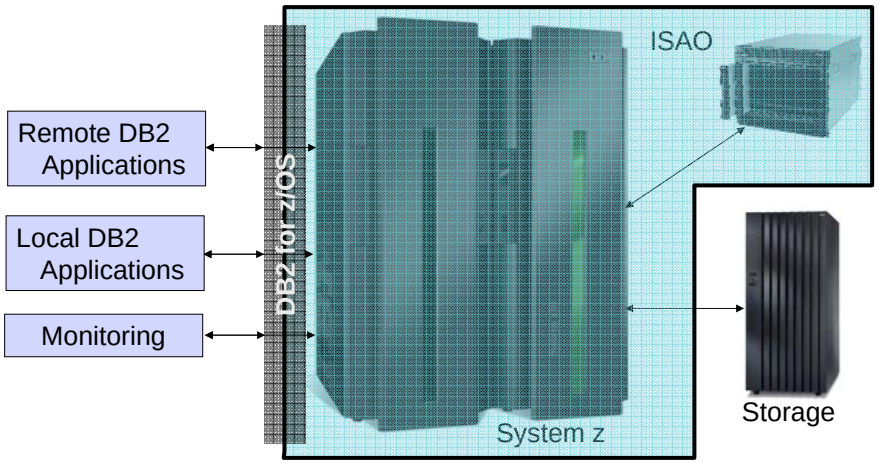
- **Requests that use parallel queries, including star schema**
- **DB2 9 Dynamic Index ANDing, higher percentage of parallel queries zIIP eligible**

3 - DB2 Utilities LOAD, REORG & REBUILD

- **DB2 utility functions used to maintain index structure**

*** zIIP allows a program working with z/OS to have all or a portion of its enclave Service Request Block (SRB) work directed to zIIP. Above types of DB2 work are those running in enclave SRBs, of which portions can be sent to zIIP.**

Solution: InfoSphere Smart Analytics Optimizer



Note – more on ISAO in Session A17

Defining, which data to accelerate

- A MART is a logical collection of tables which are related to each other. For example all tables of a single star schema would belong to the same MART.
- The administrator uses a rich client interface to define the tables which belong to a MART together with the information about their relationships.
- DB2 for z/OS creates definitions for these MARTs in it's own catalog. The related data is read from the DB2 tables and transferred to the Data Warehouse Accelerator.
- The InfoSphere Warehouse Accelerator transforms the data into a highly compressed, scan optimized format which is kept locally (in memory) on the Accelerator

Explaining the flow from

Defining the mart within the GUI (shown in more detail on the following slide)

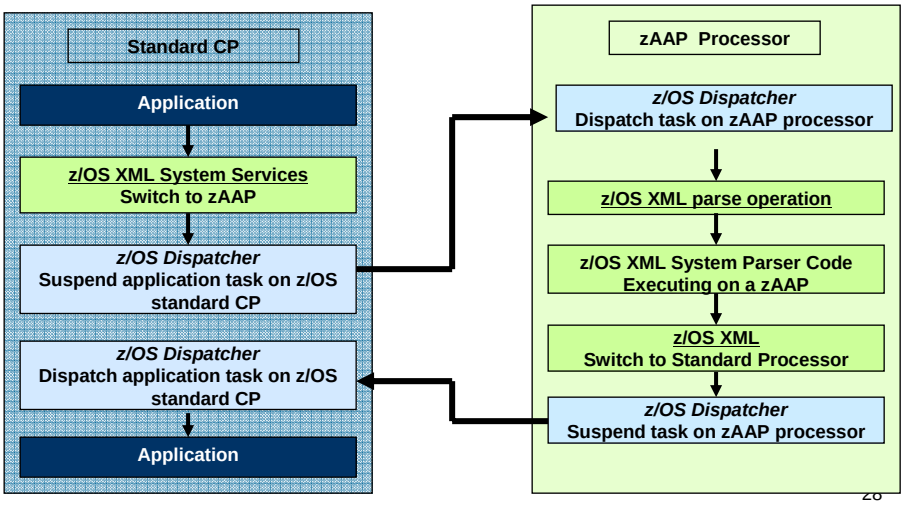
Deploying the definitions to DB2

Starting the LOAD and transformation of the data into DWA

Problem

- XML applications require complex and CPU intensive processing (encryption/decryption, parsing, formatting…)

Solution: z/OS XML System Services processing on zIIP/zAAP



Solution: DataPower Enhancements for DB2

- DataPower acts as a transformation engine into SQL
  - DataPower communicates with DB2 via DRDA
- Potential use cases:
  - High performance Web Services processing and mapping to SQL/CALL statement as well as XSLT transformation
  - Acceleration of insert of XML data into DB2
- DB2 and DataPower tooling support

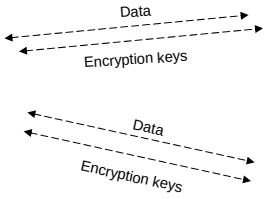
Problem

- Encryption of data at rest is needed to prevent unauthorized access to sensitive data, and to ensure that data shipped offsite is safe from prying eyes

Solution: Extending Encryption to IBM TotalStorage

- Encryption of “data at rest”
- Continuous, real-time encryption of individual drives and tape
- Expected to:
  - Have no performance impact
  - Require no application changes
- Uses Tivoli Key Lifecycle Manager
  - Key management via ICSF and RACF®
  - Audit via SMF

Enterprise-wide Key Management



Tape TS1120

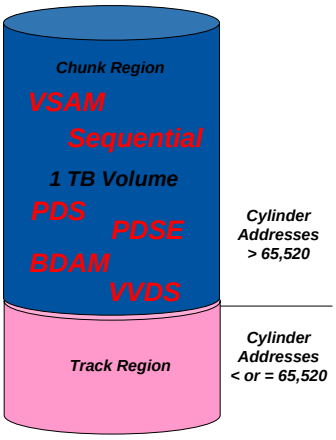
Problem

- Data Volume is growing exponentially
- I/O Rates are also increasing exponentially
- I/O Configuration is Labor Intensive

Solution: Data and Index Compression

	Data	Index
Level	Row	Page (1)
Comp on disk	Yes	Yes
Comp in Buffer Pool	Yes	No
Comp in Log	Yes	No
Comp Dictionary	Yes	No (2)
Average Comp Ratio	10% - 90%	25% - 75% (3)

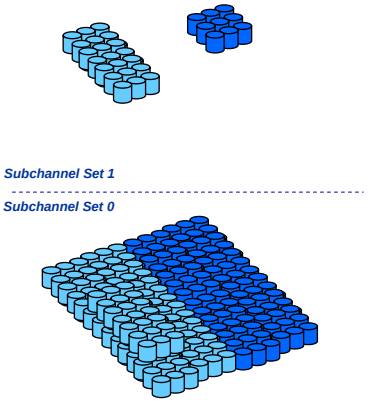
Solution: Extended Addressability Volumes (EAV)



3390-9 EAV Volume

- Continue Exploitation (z/OS 1.11 – z/OS 1.12)
  - Non-VSAM Extended Format Datasets
  - Sequential Datasets
  - PDS
  - PDSE
  - BDAM
  - BCS/VVDS
- XRC Secondary's in Alternate Subchannel Set
- Larger Volumes
  - 1 TB Volume Sizes
- Dynamic Volume Expansion
  - With Copy Service Intact
  - Automatic VTOC and Index Rebuild

Solution: Using fewer UCBs



- Subchannel Sets
  - 50% reduction in device numbers
- HyperPAV
  - 10x reduction in PAV devices by assign PAV aliases on demand as the work load requires
  - PAV-aliases virtualized across operating system images for more efficient use on addressing constructs
- Extended Addressability Volumes
  - 223 GB Volume Sizes
  - Reduce system resources and overhead with managing fewer resources
- Dynamic Volume Expansion
  - Without copy services intact
- MIDAWs
  - Performance enhancement for using larger datasets
- TDMF and zOSDMC (aka LDMF)
  - Non-disruptively consolidate data on to a single larger volume and consolidate device number ranges

Solution: High Performance FICON (zHPF)

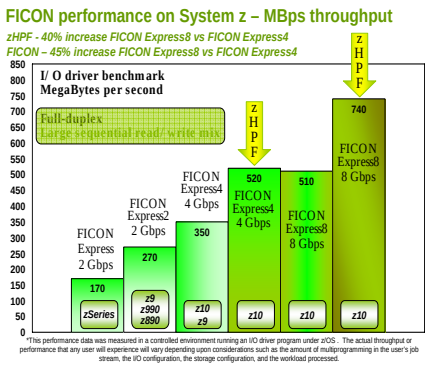
- Improve FICON Scale, Efficiency and RAS
  - As the data density behind a CU and device increase, scale I/O rates and bandwidth to grow with the data
    - Leverages HBA hardware optimizations done while preserving System z QOS
    - Significant improvements in I/O rates (2-3x) for small block transfer
    - Improved I/O bandwidth (ability to fill the link at 8 Gbs and beyond)
    - New ECKD commands for improved efficiency
  - Improved first failure data capture
  - Additional channel and CU diagnostics for MIH conditions
- Value
  - Reduce the number of channels, switch ports, control unit ports and optical cables required to balance CPU MIPS with I/O capacity
  - Enhanced resilience for work load spikes and failing components
  - Reduced job elapsed times
  - Improve First Failure Data Capture
  - Improved workload management

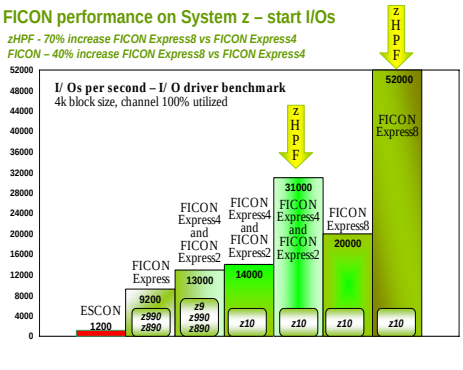
This is new technology that you want to deploy

z10 High Performance FICON for System z (zHPF)

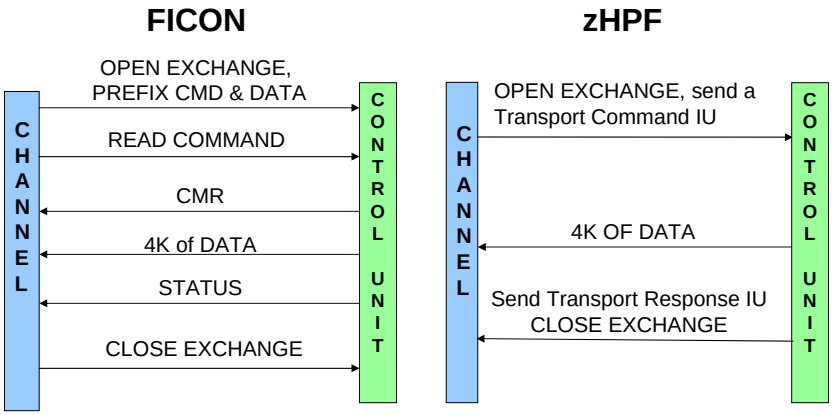
Optimization of storage area network (SAN) traffic using zHPF to improve performance, especially with FICON Express8 and SSD

- Maximum number of I/Os per second can be increased by up to 100%*
- For OLTP workloads (DB2, VSAM, PDSE, and zFS) that transfer small blocks of fixed size data (4K blocks)



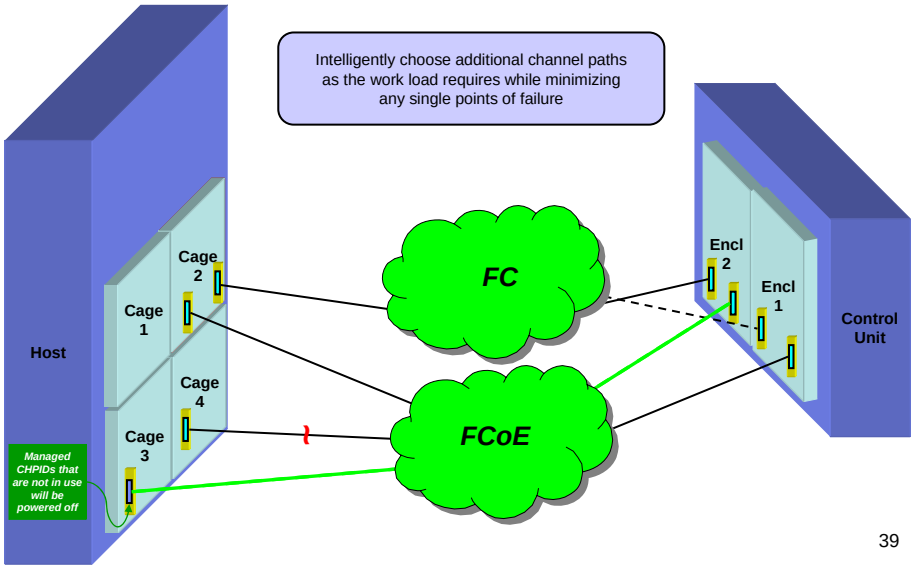


Link Protocol Comparison (4KB READ)



zHPF provides a much simpler link protocol than FICON

Solution: FICON Dynamic CHPID Management

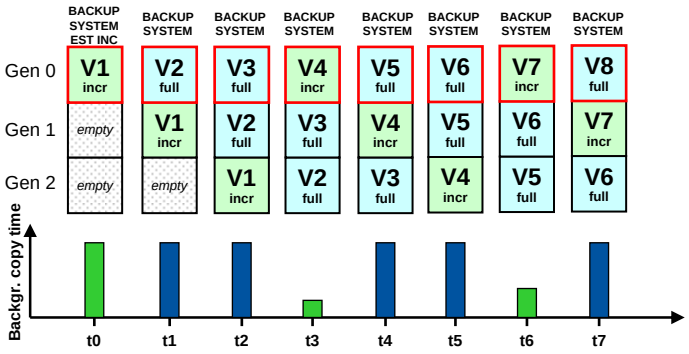


Problem

- Data Volume is growing exponentially
- How do I develop and maintain a backup and recovery strategy

Solution: Flashcopy

- Incremental Flashcopy with three FC versions:

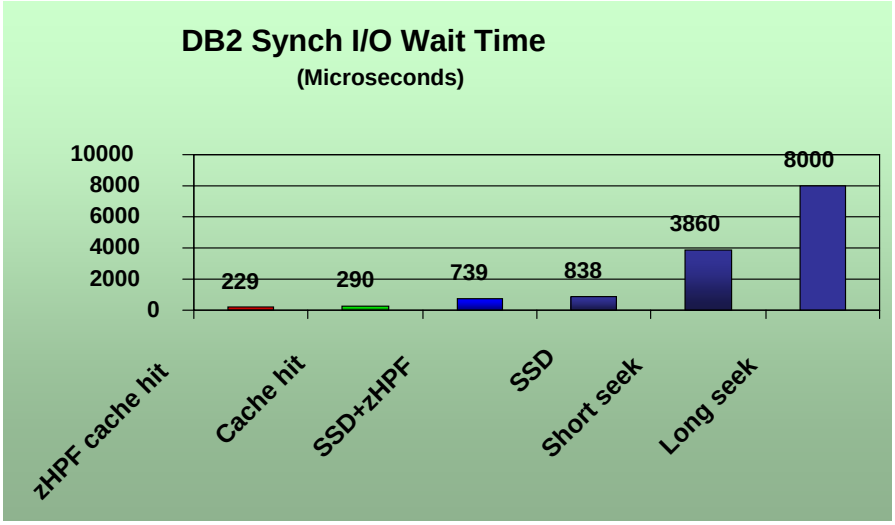


- Data set level FlashCopy option
- FlashCopy backups with consistency and no application outage
- FlashCopy backups as input to:
  - RECOVER (fast restore phase)
  - UNLOAD
  - COPYTOCOPY, DSN1COPY1
- FlashCopy Migrate/Restore from Tape

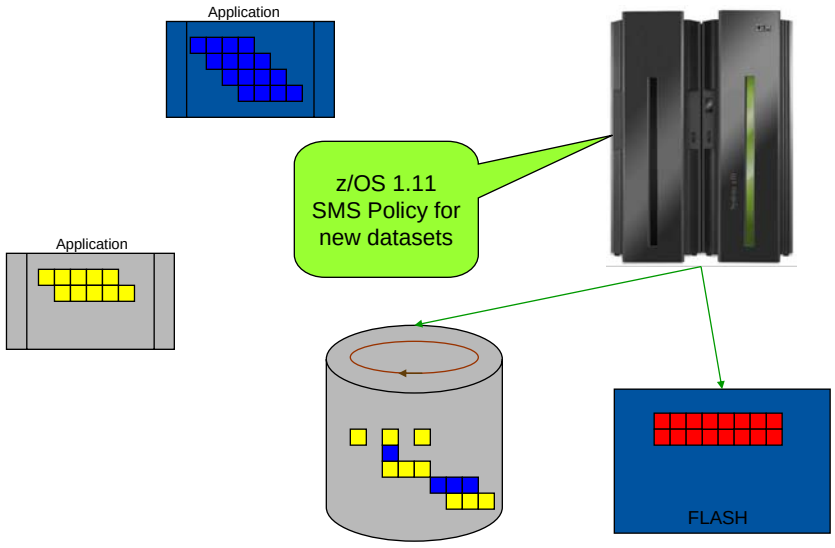
Problem

- I/O performance and I/O inefficiency are adversely affecting my DB2 system
- How can I take advantage of SSD

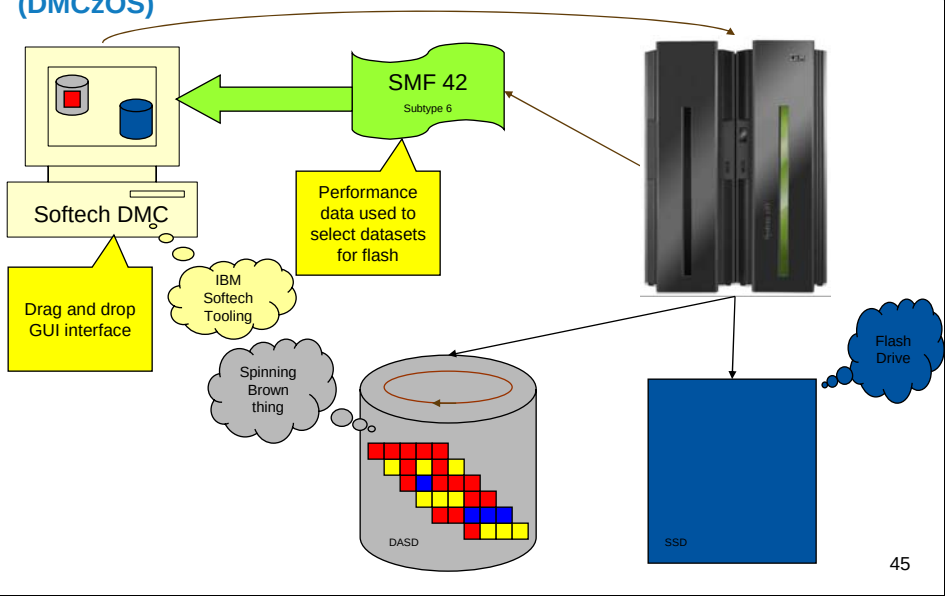
Solid state disk in storage pyramid



Solution: Policy Based Storage Management



Solution: Flash Memory and IBM Softech Data Mobility Console (DMCzOS)

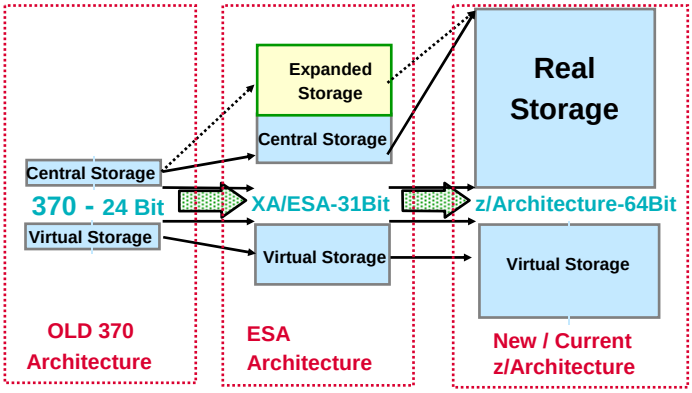


z/OS, WLM, and DB2

Problem

- Storage Constraints inhibit growth, add management complexity, and increase costs

Solution: Over Forty years of continuous (and compatible) evolution

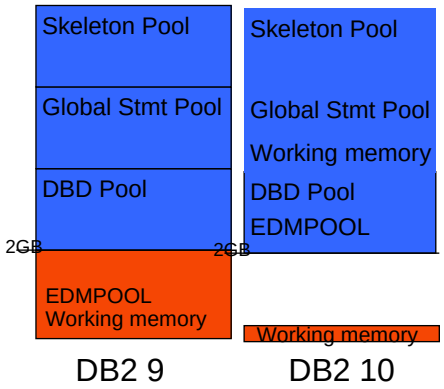


- S/390 V2R10 provided initial z/Architecture real addressing support:
  - Supporting 24-bit, 31-bit, and 64-bit applications
  - 1.5 TB (current max on z10) of central storage
  - 64-bit registers, etc...

Solution: DB2 10: 64 bit Evolution (Virtual Storage Relief)

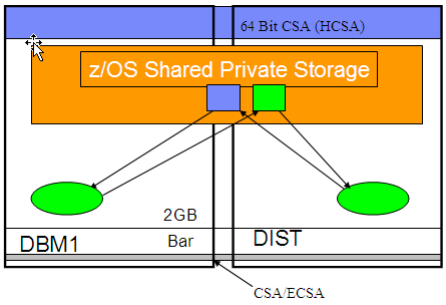
Scalability: Virtual storage constraint is still the number 1 requirement for many DB2 customers.

- DB2 9 helped (~ 10% – 15%)
- DB2 10 expects to move 90%
  - More concurrent work
  - Reduce need to monitor
  - Consolidate LPARs
  - Reduced cost
  - Easier to manage
  - Easier to grow



Solution: Enhanced Shared Storage

- In DB2 9, the DDF address space runs in 64-bit addressing mode
  - Shared 64-bit memory object avoids xmem moves between DBM1 and DDF – this improves performance and provides VSCR
- In DB2 10, some control structures will move from CSA/ECSA to z/OS 1.10 64 bit CSA. This provides VSCR

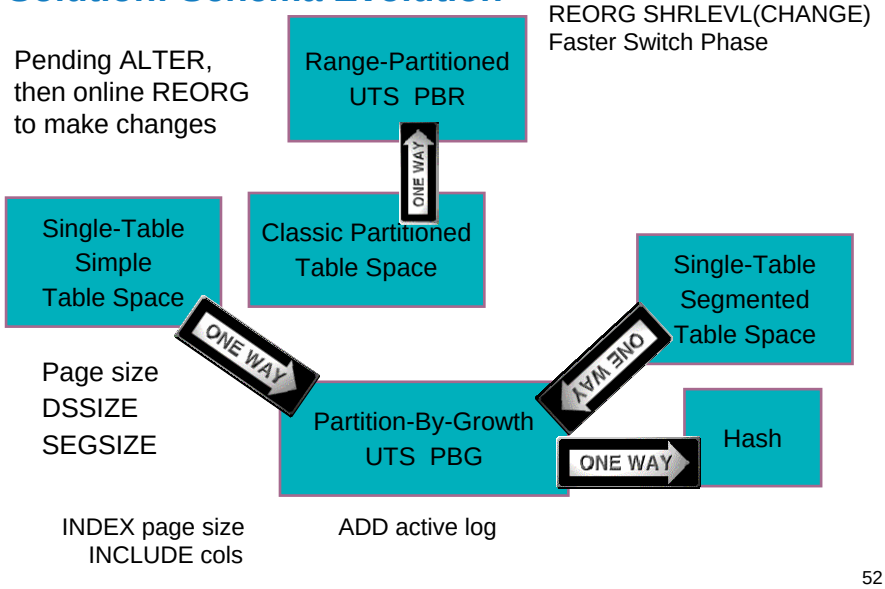


- Shared memory: new virtual storage type allowing multiple address spaces to share storage.
- Similar to ECSA – always addressable, avoids AR and XM moves.
- Different from ECSA – only available to those address spaces registering with z/OS to share this storage.
- Reduces data formatting and data movement
- Reduces virtual storage
  - It exists once, instead of in each address space
- 64 Bit CSA introduced in z/OS 1.10

Problem

- Planned Outages are making it hard to meet my SLAs

Solution: Schema Evolution



Problem

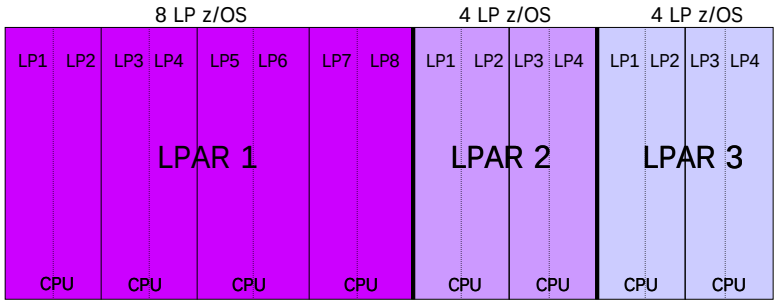
- UNI scaling by increasing GHZ is coming to an end
- Parallel Processing will be the new GHZ

Solution: HiperDispatch

- Hardware cache can be optimized when a given unit of work is consistently dispatched on the same physical CPU (or related set of CPUs)
  - In the past, System z hardware, firmware, and software have remained relatively independent of each other
  - But, the realities of modern processor and memory designs now make a change appropriate.
    - Different CPUs in the complex have different distances to the various sections of memory and cache (here, “distance” is measured in CPU cycles.)
    - Memory access times can vary from less than 10 cycles to several hundred cycles depending upon cache level and whether the access is local or remote.

Horizontal CPU Management

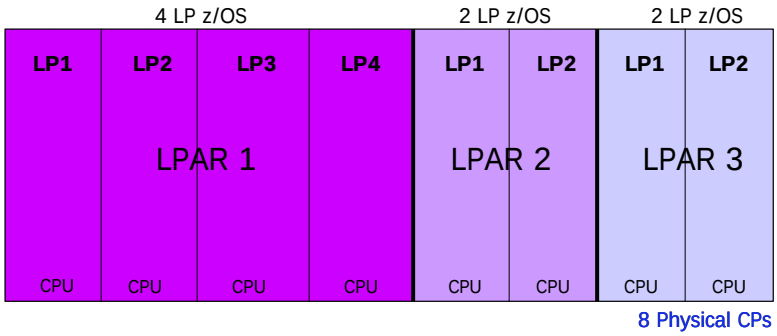
Hiperdispatch **NO**



- Typical PR/SM 2-to-1 Logical Processor to physical CP overcommitment
- High competition for physical processors
- z/OS must use all LPs to consume full LPAR weight

Vertical CPU Management

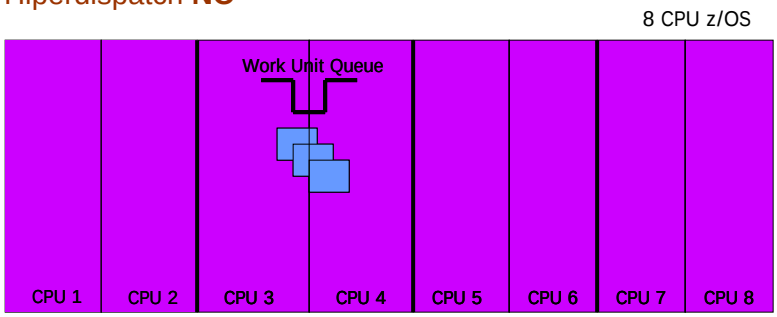
Vertical CPU Mgmt **OFF**



- PR/SM changes to allow z/OS to consume full weight with fewer LPs
- Vertical CPU management - LP to CP mapping is relatively static

z/OS Dispatcher Affinity

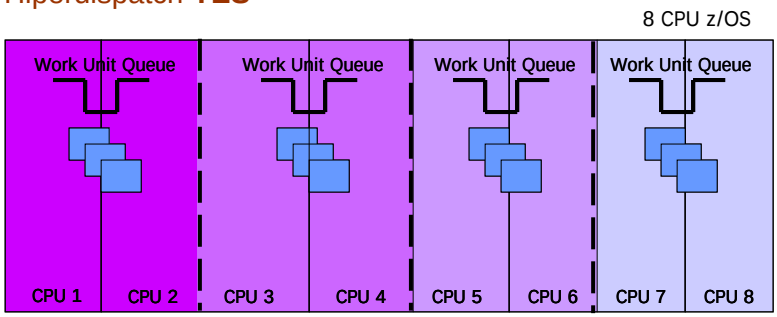
Hiperdispatch **NO**



- All work units have access to all processors of the proper type.
- Cache optimization is minimal.

z/OS Dispatcher Affinity

Hiperdispatch **YES**



- Work units normally have access to only a subset of processor.
- Work imbalances are address by a “needs help” algorithm.
- Cache optimization is improved.

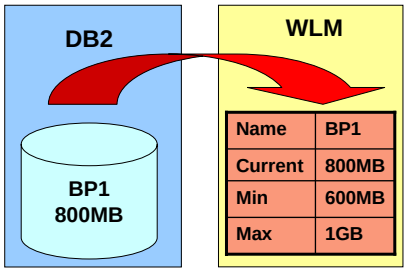
Note that this depiction is simplified for illustrative purposes.
Typically an actual processor subset will have at least 4 processors.

Problem

- Allocating Bufferpools for disparate workloads wastes resources

Solution: WLM Buffer Pool Management

- WLM-assisted buffer pool management
  - -ALTER BUFFERPOOL () AUTOSIZE(YES)
  - z/OS 1.8
  - DB2 registers BP to WLM and reports synch read I/O delays to WLM
  - DB2 periodically reports BP hit stats to WLM
  - WLM projects effect of adjusting BP size on workload performance goals
    - Takes into account overall system storage usage
  - WLM drives DB2 exit to adjust size if appropriate
    - V9 restricts to +/- 25%

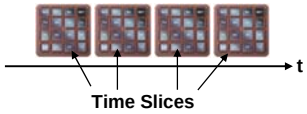


DSNB555I WLM RECOMMENDATION TO ADJUST SIZE FOR BUFFER POOL bpname HAS COMPLETED
OLD SIZE = csize BUFFERS
NEW SIZE = nsize BUFFERS

Problem

- WLM does a great job of prioritizing my OLTP over my batch – but now my batch work is suspended and causing problems in my OLTP work

Solution: Blocked Workload Support



- Usually work uses the CPU for a predefined amount of time
  - Named a time slice
    - After that period the work gets re-dispatched
  - z/OS uses a major and a minor time slice
- Approach
  - Convert the capacity of the LPAR into multiple of major time slices

$$\text{LPAR WCap} = \frac{\text{LPAR Weight}}{\text{Total Weight} + \text{Phantom Weight}} \bullet \text{Shared Regular CPs}$$

$$\text{LPAR Capacity} = \text{MIN}(\text{LPAR WCap}; \text{Logical Regular CPs of LPAR})$$

  - Define a percentage value of how many time slices can be used for blocked workloads (as trickles)
  - Parameter: BLWLTRPCT
    - Values: 0.. 200
    - 200 = 20%
    - Default: 5 = 0.5% of the LPAR capacity

Problem

- Management of workload distribution in a n-tier environment is complex

Solution: Workload distribution algorithm enhancements by Sysplex Distributor and Workload Manager in z/OS V1R11

- New workload at Importance level 2
- Which LPAR is best?
  - They both have 500 service units of displaceable workload
  - Before R11, they would be equal
- z/OS V1R11 takes importance level of displaceable workload into consideration
  - LPAR2 will be preferred

New workload at IL=2 (can displace IL=3 to IL=7 workload)

IL 0: High
IL 7: Low

	LPAR1		*LPAR2*	
	IL	**SUs**	**IL**	**SUs**
	0	0	0	0
	1	0	1	0
	2	0	2	0
	3	500	3	0
	4	0	4	0
	5	0	5	500
	6	0	6	0
	7	0	7	0

- New workload at Importance level 2
- Which LPAR is best?
  - They both have equal amount of displaceable service units
  - Before R11, they would be equal
- z/OS V1R11 takes amount of crossover to CP of displaceable workload into consideration
  - LPAR2 will be preferred since it has the least amount of crossover

New workload designed to use 90% zAAP and 10% CP

New workload at IL=2 (can displace IL=3 to IL=7 workload)

	LPAR1			*LPAR2*		
	IL	**CP SUs**	**zAAP SUs**	**IL**	**CP SUs**	**zAAP SUs**
	0	0	0	0	0	0
	1	0	0	1	0	0
	2	0	0	2	0	0
	3	900	100	3	100	900
	4	0	0	4	0	0
	5	0	0	5	0	0
	6	0	0	6	0	0
	7	0	0	7	0	0

Problem

- multi-tier environments require complex cross-domain security
- I can't tell who did what

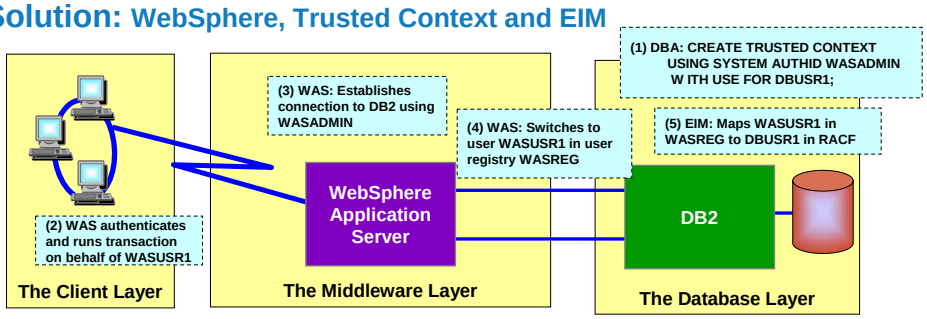
Solution: RACF access control

- Policy and people implications
  - Roles will change
  - Authorities will change
    - Use RACF facilities more, e.g. groups & patterns
    - Not a completely compatible change
  - Need both DB2 & RACF knowledge for implementation and for administration
  - Mix of RACF and DB2 Authorization
- Security group should define authorization
- Centralized Security Control Point
- Use patterns, not individual access authority

Solution: End-to-End Auditing for Remote Applications

- Applications accessing DB2 without RACF user IDs
- Inbound IDs are different across systems
- Exploits z/OS Security Server user mapping SAF plug-in service
  - Default implementation is RACF's Enterprise Identity Mapping Feature (based on LDAP)
  - Retrieves RACF auth ID (used as DB2 primary authorization ID) from the remote user ID (non-RACF or distributed ID)
  - Provides many to one mapping
- Remote user IDs and DB2 auth IDs are included in both DB2 and RACF audit records

Solution: WebSphere, Trusted Context and EIM



- Configure WebSphere Application Server to create a trusted connection and send the WAS user registry name to the DB2 server
- DB2 maps the WAS User ID to obtain RACF User ID
- DB2 establishes DB2 Auth ID form RACF User ID
- DB2 checks if the DB2 Auth ID is allowed to use the trusted connection
- WAS User ID, DB2 Auth ID, RACF User ID and WAS registry name are recorded in RACF and DB2 audit records

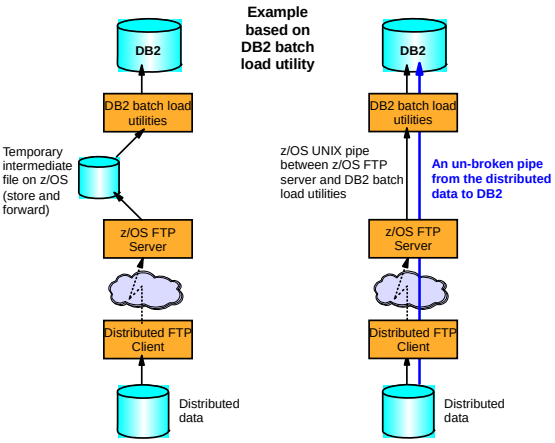
Communications

Problem

- How do I convert my SAP system to Unicode within the window I have to do the work?
- How do I efficiently load data from my distributed systems into DB2 for z/OS

FTP access to UNIX named pipes

- Helps save total processing time when pre/post processing is needed for files transferred from/into z/OS
- Support available in both the z/OS FTP client and server
- FTP can be either the reading end or the writing end of the pipe
- PTFs for prior releases
  - APAR PK71213 provides z/OS FTP server support
  - z/OS V1R8, V1R9, and V1R10
- The SAP on DB2 for z/OS Unicode FASTLOAD conversion utility exploits named pipes



Problem

- I'm running out of IP addresses

Solution: DDF V9 IPv6 Support

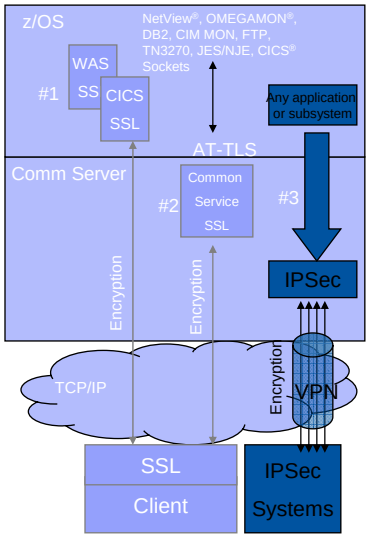
- Requires TCP/IP dual-mode stack.
  - Allows transition to IPv6.
  - DDF uses bilingual socket calls, 32bit or 128bit, depending on presence of dual-mode TCP/IP stack.
- DB2 always displays addresses in IPv6 form.
  - So don't panic when you see something like
    - ::FFFF:1.2.3.4
- Communications Database (CDB) changes.
  - SYSIBM.IPLIST and SYSIBM.IPNAMES
    - IPv4 Dotted Decimal address (max 15 chars).
    - IPv6 Colon Hexadecimal address (max 39 chars).
    - Domain Name (max 254 chars).
      - Resolves to IPv4 or IPv6 address.

Problem

- My Auditors say I have to Secure my DB2 connections

Solution: Network encryption

- The z/OS Communications Server (z/OS CS) encrypts application data
  a number of ways:
  - #1 Application layer encryption (per session)
    - Application is coded with encryption
  - #2 Application is encrypted in network layer (also per session)
    - "common service" AT-TLS (z/OS 1.7)
    - Transparent to the application
  - #3 "Platform to platform" encryption (Virtual Private Networks using IPSec)
    - All traffic may be encrypted – transparent to all applications
- When do you use one form of encryption versus another?
  - Depends on client, application, topology, performance requirements.....
  - IPSec can be used for some or all traffic – can create a VPN



A Peek into the Future

Problem

- The future of computing seems uncertain

Solution: DB2, z/OS and System z – An Unbeatable Combination

- DB2 10 removed the last real roadblock to scaling (from a VSTOR perspective)
- System storage will continue to grow
  - 1.5 TB today -> 3+ TB tomorrow
  - Flash Memory/SSD will create opportunities
- Number of cores will continue to increase with each generation of HW
- Specialty accelerators will continue to evolve
- Shift to parallel processing will accelerate
- Size of Data and Volume of transactions continues to grow
- System Complexity will increase

DB2, System z, and z/OS will continue to provide you the most reliable, scalable, and available platform for your data needs.

Christopher J Crone
cjc@us.ibm.com

IDUG The Worldwide DB2 User Community

EMEA Conference 2010

DB2 & System z Synergy

Christopher J. Crone
cjc@us.ibm.com
IBM DB2 for z/OS Development