The Modern Mainframe…
At the Heart of Your Business

A Mainframe Primer - Mainframe Clustering

Superior Qualities of Service

- How does the mainframe deliver superior qualities of service?
    - ▶ Unmatched scale-up
    - ▶ Continuous operation
    - ▶ Systematic disaster recovery

- Mainframe clustering technology hardware and software are optimized to provide these qualities of service
    - ▶ Unique Parallel Sysplex design is superior

Mainframe Clustering is Superior

System z
- Specialized hardware for clustering
- Dedicated high speed fiber interconnect
  - Low latency
- Integrated exploitation by operating system and all software subsystems

Distributed
- No special hardware
- No exploitation of special networking
  - Full software path length
- Each subsystem (database, application server) is designed to run on commodity servers

1. **Very low overhead in clusters yields ultimate scalability**
2. **Unrivaled high availability**

A Primer on Mainframe Clustering

- **Coupling Facility**
  - Dedicated processor with specialized microcode to coordinate shared resources
  - Large amounts of fast memory
  - High speed inter-connect to clustered systems
  - Hardware invalidation of local cache copies
  - Special machine instructions
  - Timing facilities to maintain logical execution-order across coupled systems
 - Fault-Tolerant

- **Parallel Sysplex**
  - Multiple z/OS images clustered using the coupling facility for coordination

This presentation will use the word "image" to refer to a node in a sysplex cluster, "LPAR" may also be used to describe this

Parallel Sysplex – What is it ?

- Hardware
  - Redundant coupling facilities providing processing and shared storage
  - Timing facilities
    - Sysplex timers (Hardware)
    - STP protocol (Software)
  - Dedicated high speed interconnections (up to 16 Gigabits/sec, up to 10 meters)
  - Fiber switch provides access to data

- Micro-code + Software
  - CFCC (coupling facility control code)
    - High throughput, low latency, micro code control program for the coupling facility

- Clustering service APIs within z/OS
  - XES APIs support program connectivity
  - XCF connectivity configuration

- Workload Management
  - WLM (workload manager within a z/OS instance)
  - IRD (intelligent resource director across LPAR's)
 - Both manage workload across the sysplex

Coupling Facility

Sysplex Timer

System z9

zSeries

ESCON/FICON

Shared data
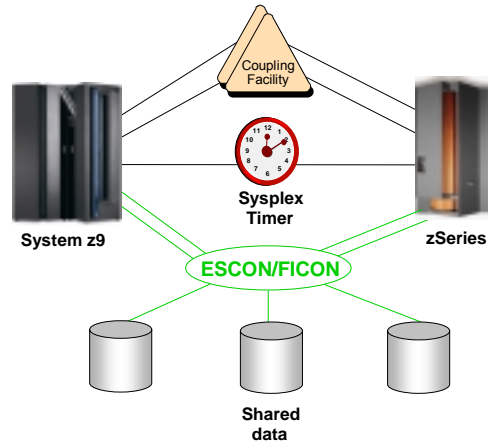
…A Key IBM Unique Differentiator in the IT Industry

04 - Mainframe Clustering v2.5.ppt   6

Implementation of Coupling Facility

Logical Relationship

ICF

A

B

ICF
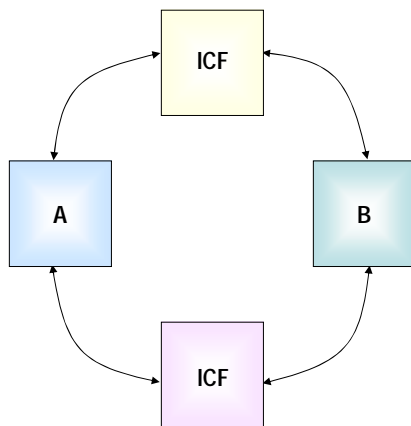
Physical Deployment
ICF within System z LPARs

A

B

ICF

ICF

No extra servers needed

04 - Mainframe Clustering v2.5.ppt   7
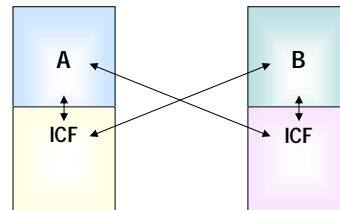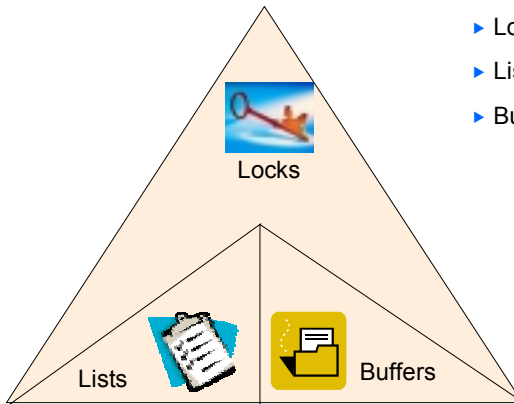
Coupling Facility is an Optimized Hardware Technology for Coordinating Clusters
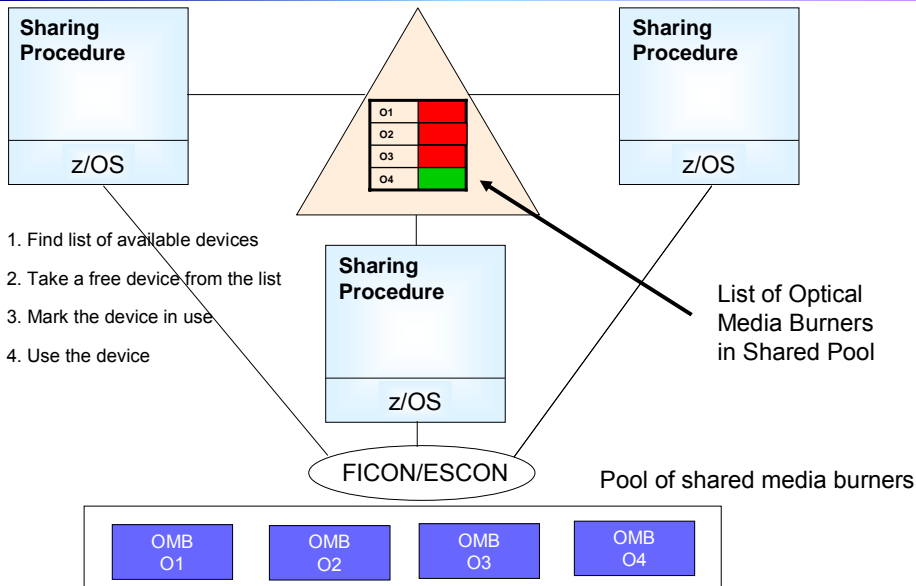
The Coupling Facility implements

- ▶ Locks for synchronizing data
- ▶ Lists for sharing data
- ▶ Buffers for database consistency

Locks

Lists   Buffers

Let's look at examples of how each of these are used in a cluster

---

Using the List Capability for Sharing Devices Getting a Device for Making a Backup Copy

Sharing Procedure

z/OS

Sharing Procedure

z/OS

O1	
O2	
O3	
O4	

1. Find list of available devices

2. Take a free device from the list

3. Mark the device in use

4. Use the device

Sharing Procedure

z/OS

List of Optical
Media Burners
in Shared Pool

FICON/ESCON   Pool of shared media burners

| OMB O1 | OMB O2 | OMB O3 | OMB O4 |

4

**Using the List Capability for Sharing Devices
Releasing a Device for Use by Others**

Sharing Procedure
z/OS

O1	
O2	
O3	
O4	

Sharing Procedure
z/OS

1. Find list of available devices

2. Mark the device as available

Sharing Procedure
z/OS

List of Optical
Media Burners
in Shared Pool

FICON/ESCON

Pool of shared media burners

OMB O1	OMB O2	OMB O3	OMB O4
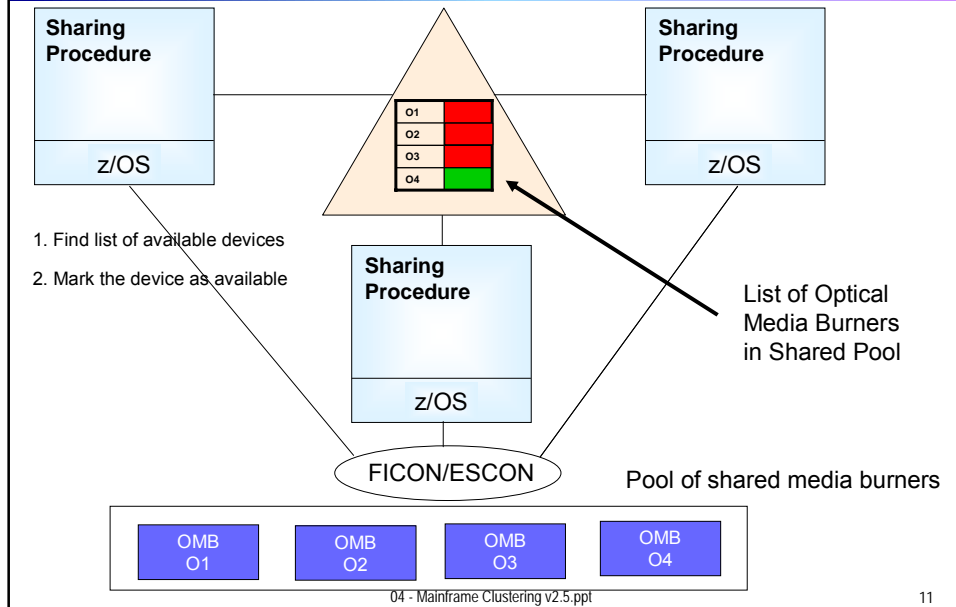
Other System Uses of Lists

- Shared Resources
  - ▶ Tapes
  - ▶ Files
  - ▶ Consoles
  - ▶ Etc

- Sysplex-wide information
  - ▶ Workload-balancing information
  - ▶ Status of each system in the sysplex

- Subsystem information
  - ▶ Logfiles for recovery
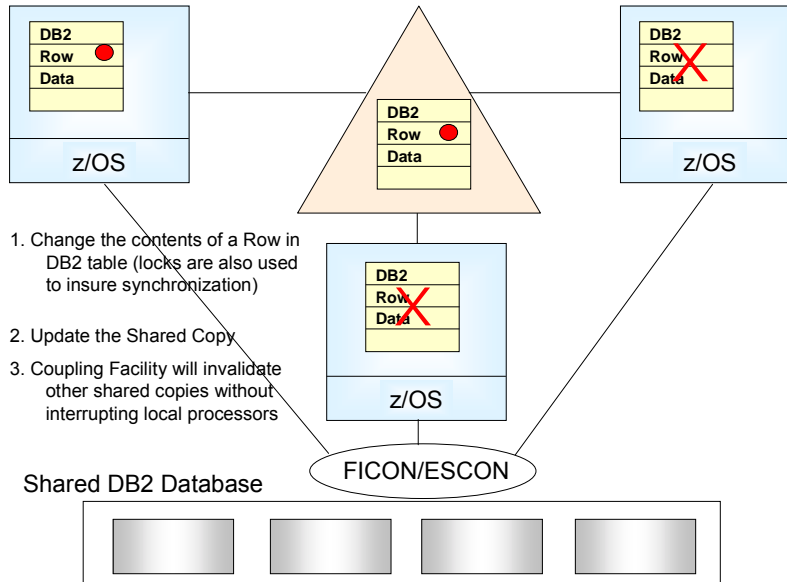 - ▶ Configuration and Restart Data

Use of the Lock Capability for Updating Information

Updating Procedure

z/OS

1. Get the lock for Disk 1

2. Update the directory for Disk 1

3. Release the lock for Disk 1

Updating Procedure

z/OS

Lock for Disk 1

Updating Procedure

z/OS

Disk Directory For Disk 1

FICON/ESCON
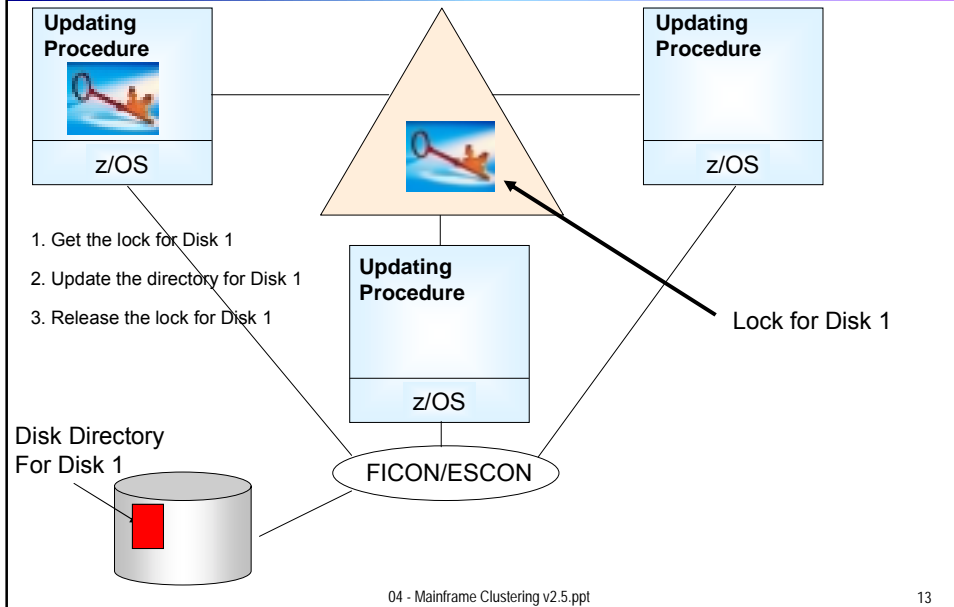
Other System Uses of Locks

- Any synchronization of shared information
  - ▶ Files
  - ▶ Databases
  - ▶ System-wide resources

Using the Buffer Capability for DB2 Data Consistency

DB2
Row ●
Data

z/OS

DB2
Row ●
Data

DB2
Row ✗
Data

z/OS

DB2
Row ✗
Data

z/OS

1. Change the contents of a Row in DB2 table (locks are also used to insure synchronization)

2. Update the Shared Copy

3. Coupling Facility will invalidate other shared copies without interrupting local processors

FICON/ESCON

Shared DB2 Database

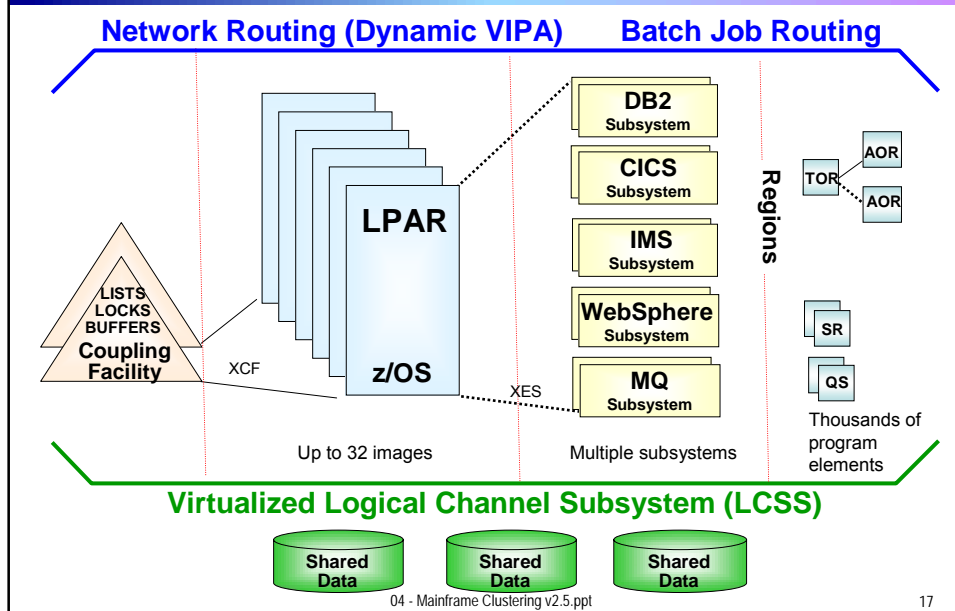## Other Uses of Buffers for Data Consistency

- DB2 for System z

- IMS

- VSAM

- Computer Associates IDMS

- Computer Associates Datacom

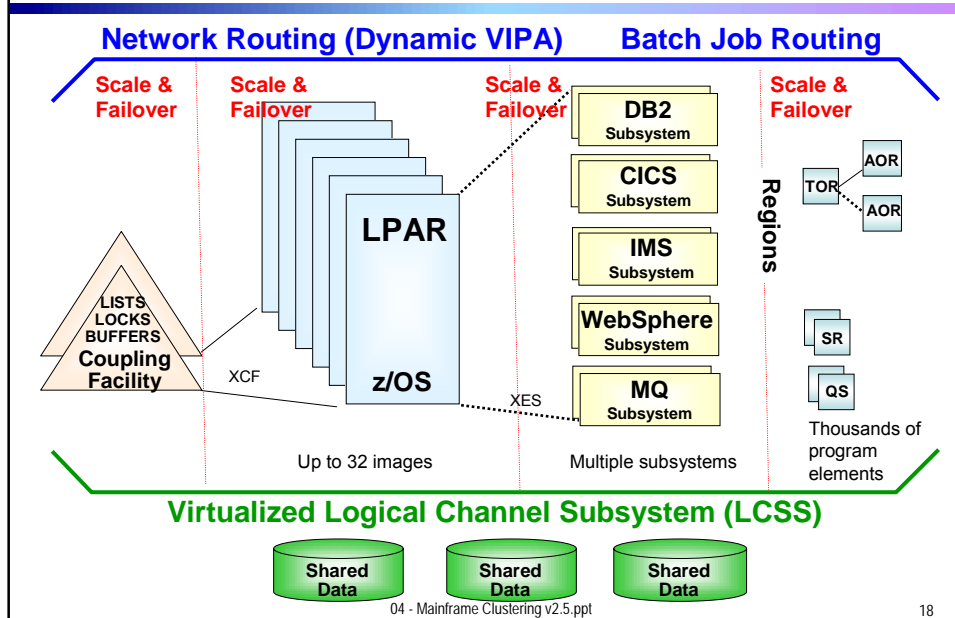## Thousands of Program Elements Can Be Coordinated in a Single System Image

Network Routing (Dynamic VIPA)   **Batch Job Routing**

LISTS LOCKS BUFFERS
Coupling Facility

XCF

LPAR

z/OS

XES

DB2 **Subsystem**
CICS **Subsystem**
IMS **Subsystem**
WebSphere **Subsystem**
MQ **Subsystem**

Regions

TOR
AOR
AOR

SR

QS

Thousands of program elements

Up to 32 images   Multiple subsystems

Virtualized Logical Channel Subsystem (LCSS)

Shared Data   Shared Data   Shared Data

04 - Mainframe Clustering v2.5.ppt   17

Scale and Availability Within Each Layer

Network Routing (Dynamic VIPA)   **Batch Job Routing**

Scale & Failover   **Scale & Failover**   **Scale & Failover**   **Scale & Failover**

LISTS LOCKS BUFFERS
Coupling Facility

XCF

LPAR

z/OS

XES

DB2 **Subsystem**
CICS **Subsystem**
IMS **Subsystem**
WebSphere **Subsystem**
MQ **Subsystem**

Regions

TOR
AOR
AOR

SR

QS

Thousands of program elements

Up to 32 images   Multiple subsystems

Virtualized Logical Channel Subsystem (LCSS)

Shared Data   Shared Data   Shared Data

04 - Mainframe Clustering v2.5.ppt   18
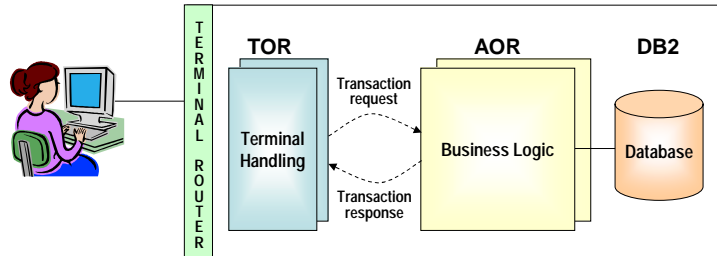

CICS - Regions

- CICS takes a transaction request from an end user, accesses a database, performs business logic and returns a response (similar to J2EE)
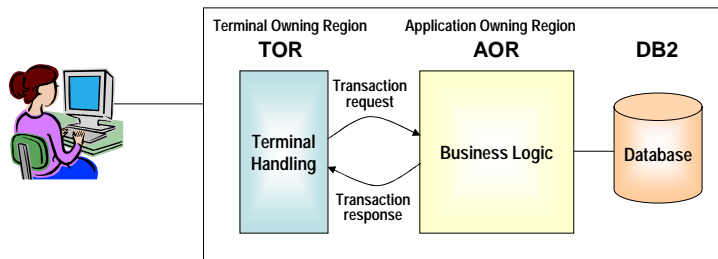
- Each CICS region (TOR and AOR) provides a single thread of execution for a program
- Regions provide transaction isolation

CICS – Multiple Regions in an Image

- Terminal router routes transaction to appropriate TOR

- Multiple TORs and AORs scale by adding system resources (threads, memory, etc)
- Multiple TORs and AORs provide availability
  - ▶ A software failure could bring down a region (e.g. programmer error)
  - ▶ Current in flight transactions are rolled back
  - ▶ New transactions are routed to other regions
  - ▶ CICS restarts failed region

# Throughput is Maintained in the Event of Software Failure in a Region

- If an AOR or TOR fails the resources it was consuming, processors and memory, etc. are released

- These resources are immediately available for remaining regions

- Throughput can be maintained

- This important capability is lacking in the distributed world

CICS – Multiple Images in a Sysplex

- Multiple regions on multiple machines in a parallel sysplex
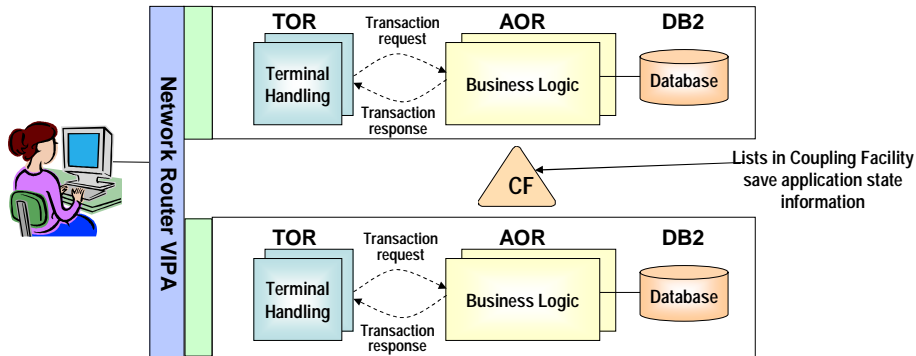
- Scalability is enhanced In that processing resources from up to 32 images in the sysplex can be utilized
- The work of a failed region can be taken over by any other region in the sysplex
- Protects against machine hardware failure or operating system failure

---

Parallel Sysplex Performance

- High performance interconnect and low latency in coupling facility causes minimal overhead.

- Typical overhead
  - Multisystem Management - 3%
  - Resource Sharing - 3%
  - Application data sharing - <10%
  - Incremental cost of adding an image - 1/2%

- Result
  - Near-linear scalability as more systems are added
  - Better efficiency than other clustering schemes

Mainframe Clustering Delivers Near-Linear Scalability
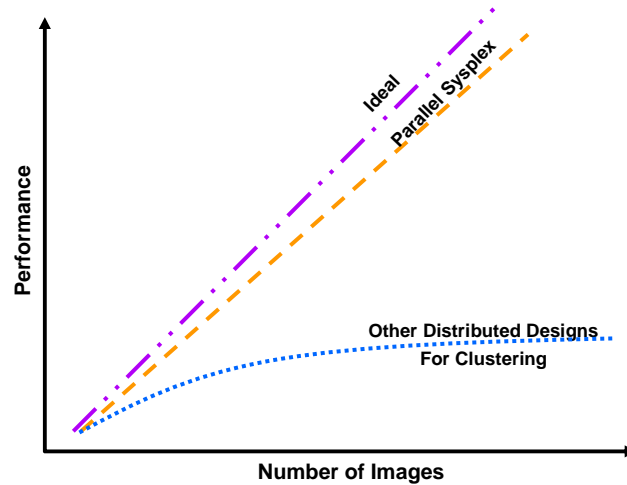
Imagine the Scale…

- A single LPAR image in a 54-way System z delivers 17,801 MIPs and huge I/O bandwidth

  - ▶ This is roughly 6 times the processing capacity of the largest HP Itanium Superdome with 768 processor cores*

- Up to 32 of these images can be clustered in a parallel sysplex, single system image

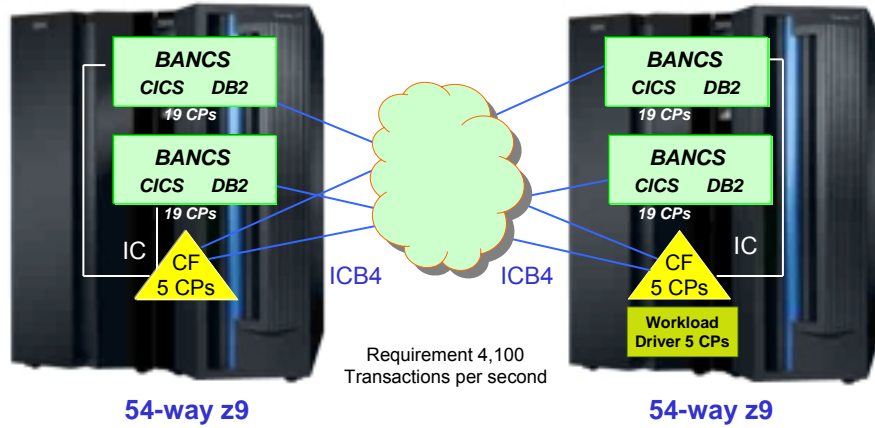* Based on equivalence factor of 1 MIP = 122 RPE's from HP presentations

Bank of China Parallel Sysplex Benchmark

Database
- 380 million accounts
- 52 TB Storage
- 4 DS8300

BANCS
CICS   DB2
19 CPs

BANCS
CICS   DB2
19 CPs

IC

CF
5 CPs

ICB4   ICB4

BANCS
CICS   DB2
19 CPs

BANCS
CICS   DB2
19 CPs

IC

CF
5 CPs

Workload
Driver 5 CPs

Requirement 4,100
Transactions per second

54-way z9   **54-way z9**

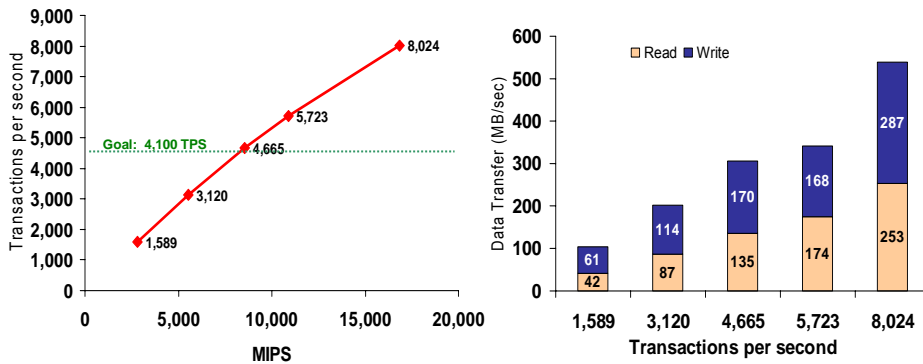
Bank of China Parallel Sysplex Benchmark

Near-Linear Scalability on a Parallel Sysplex running CICS and DB2 in a single system image with No Partitioning Required

Transactions per second vs MIPS:
- 1,589
- 3,120
- 4,665
- 5,723
- 8,024

Goal: 4,100 TPS

Data Transfer (MB/sec), Read / Write by Transactions per second:

Transactions per second	Read	Write
1,589	42	61
3,120	87	114
4,665	135	170
5,723	174	168
8,024	253	287

Huge scale up, requires huge I/O bandwidth capacity

Mainframe Parallel Sysplex Summary

- Layered approach enables thousands of program elements to cooperate in a single system image with very low overhead

- Ultimate scalability
  - ▶ Up to 32 hardware systems each with 54 processors

- Unrivaled high availability
  - ▶ Protection against hardware and software failures

- Foundation for a systematic disaster recovery capability