z/VM Live Guest Migration

MVMUA
July, 2008

Romney White
romneyw@us.ibm.com
IBM System z Software – Strategy and Design

IBM Systems

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

DB2*
DB2 Connect
DB2 Universal Database
e-business logo
GDPS*
Geographically Dispersed Parallel Sysplex
HyperSwap
IBM*
IBM eServer
IBM logo*
Parallel Sysplex*
System z
Tivoli*
VM/ESA*
WebSphere*
z/OS*
z/VM*
zSeries*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Intel is a registered trademark of the Intel Corporation in the United States, other countries or both.

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

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

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.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

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.

Agenda

- Motivation
- Alternatives
- Early Steps
- Technology
- Conceptual Migration Process
- Technology Demonstration
- Challenges
- Summary

Motivation

- z/VM is extremely reliable
  - Customers “complain” about having to IPL to/from Daylight Time
  - Still, z/VM is a single point of failure
  - More importantly, perhaps, it is a single point of service
    - Planned hardware and software outages predominate
- VMware, Xen, pHyp, and other hypervisors have found value in guest migration
  - Addressing a somewhat different set of problems than z/VM has
    - Reliability
    - Scalability limitations
  - A differentiating factor nevertheless
  - Caused us to reconsider its importance

Alternatives

- Concurrent patch
  - Firmware approach
  - Must be able to apply and remove patches
  - Number of combinations grows exponentially
    - Difficult to test
  - Could cause more problems than it solves
- Application migration
  - E.g., MetaCluster
  - Probably leaves virtual machine impotent
  - Knowledge at the wrong level
- Multi-system virtualization
  - “Single system image” including Live Guest Migration
  - Breadth of z/VM virtualization leads to large, complex challenge

Early Steps

- IBM Research interest in problem of z/VM Live Guest Migration
- Started prototype work in 2004
- Speed Team created in summer 2006
  - Cross-site (Poughkeepsie, Endicott) team with Research assistance
  - Brought prototype forward to z/VM 5.3 base – Endicott
  - Designed Migration Diagnose – Endicott/Poughkeepsie
  - Developed Migration Diagnose – Endicott
  - Developed service machine ("moving van") to orchestrate migration – Poughkeepsie
    - Based on CSE and ISFC

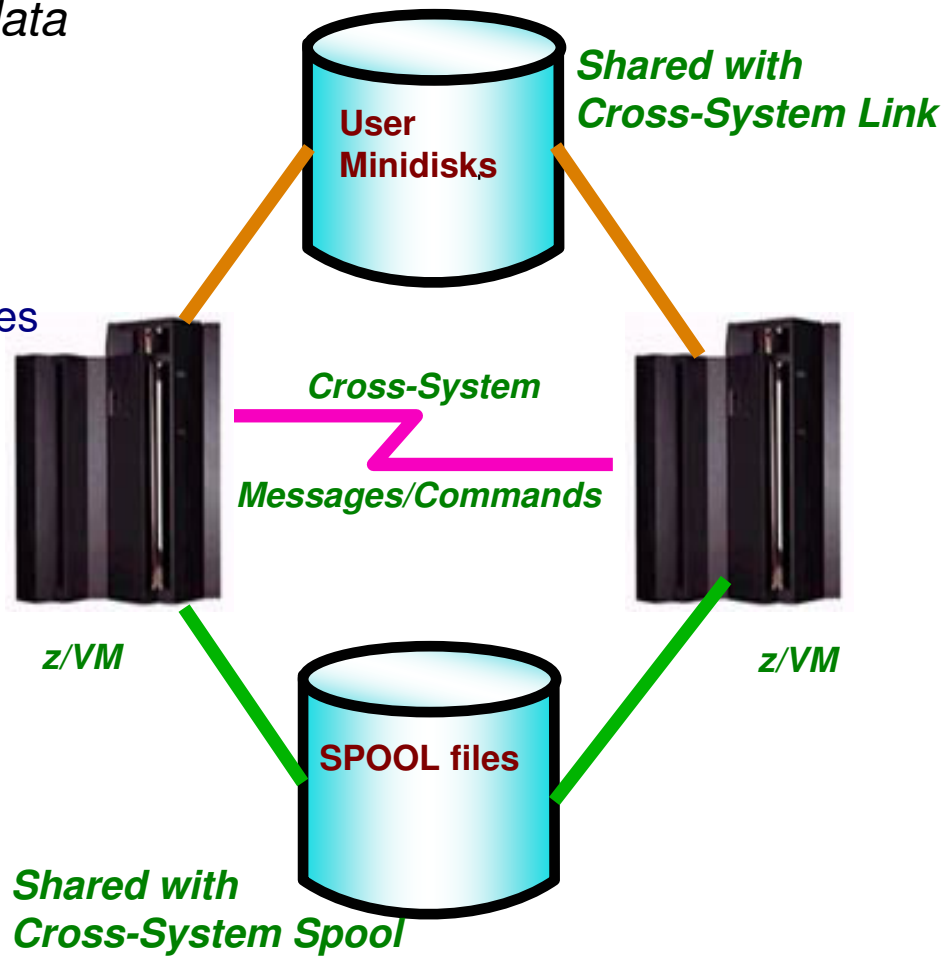
Technology

- **Cross-System Extensions (CSE)**
- **Inter-Systems Facility for Communications (ISFC)**
- **“TRACK” Diagnose**
- **Migration Diagnose**
- **Guest memory change tracking**

Cross-System Extensions (CSE)

Virtual Machines may access their data from any z/VM image in a cluster

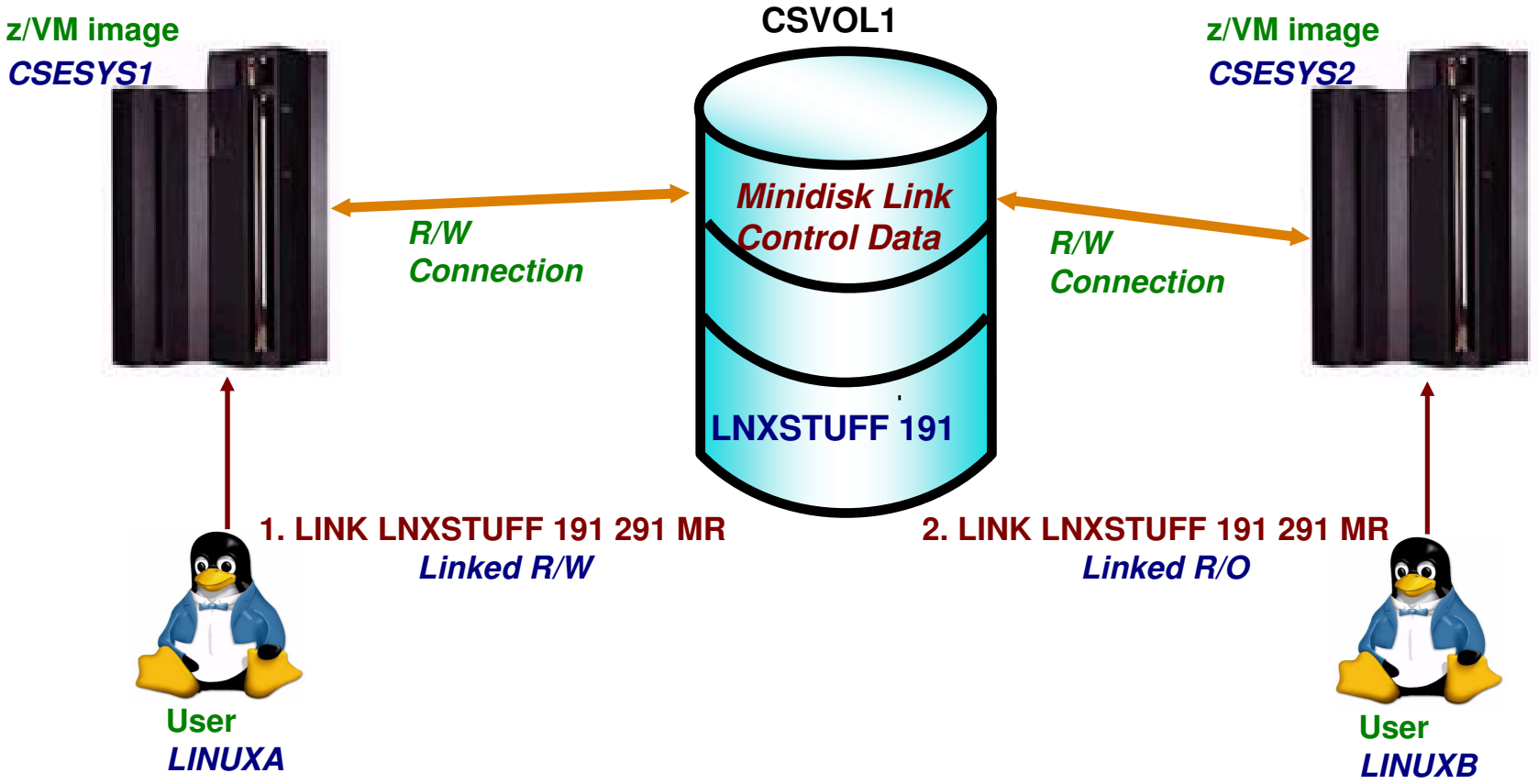
- Capability to share
  - Minidisks
  - Spool files
- Commands may be sent among images in the cluster
  - Messages
  - Query
  - Link
  - Spool File Commands



CSE Cross-System LINK

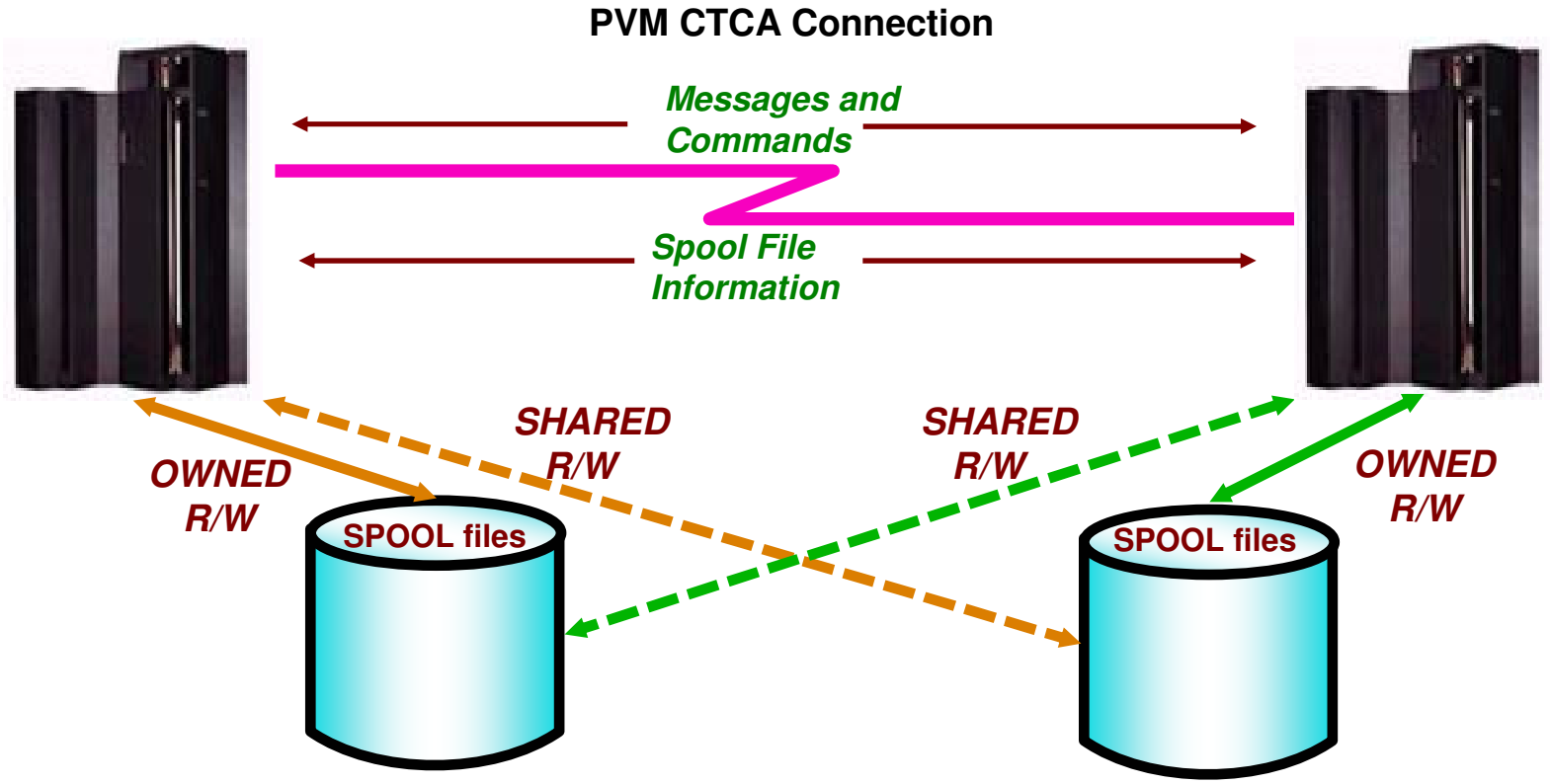
Shared Minidisk Volumes

- Link control information for all systems is kept on the volume

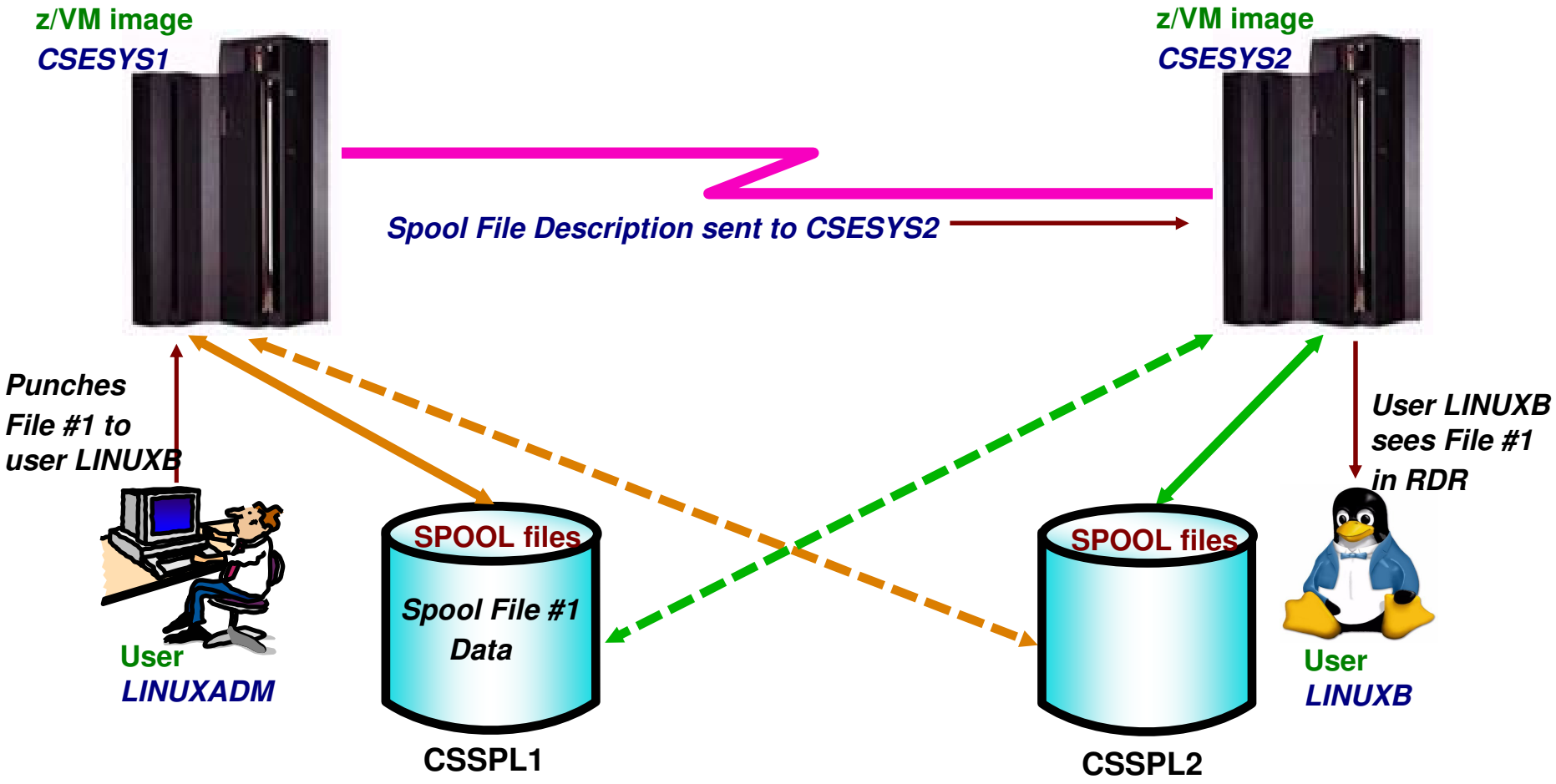


CSE Communication and Spool

Up to 4 z/VM Images can share spool files



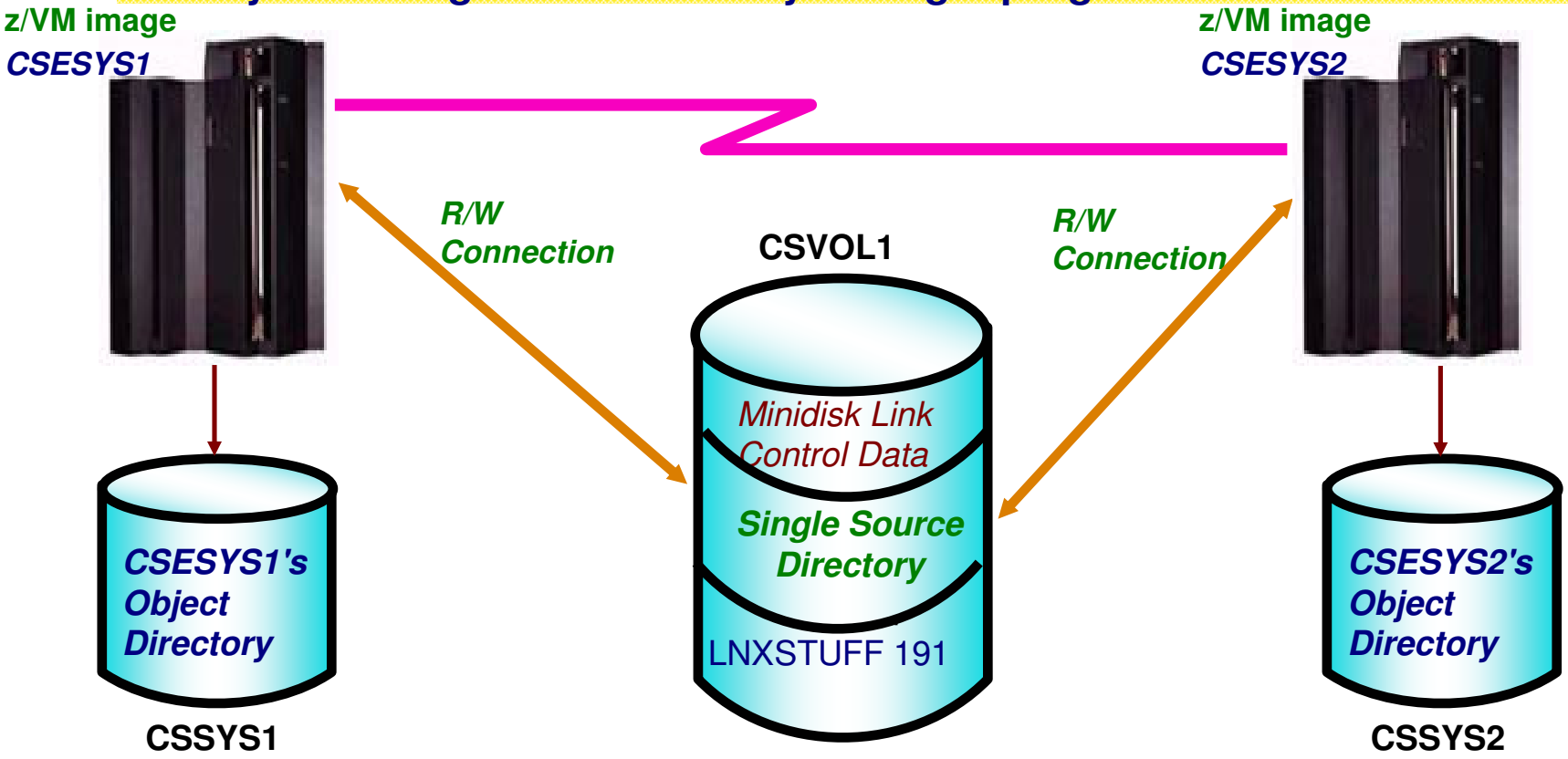
CSE Communication and Spool …



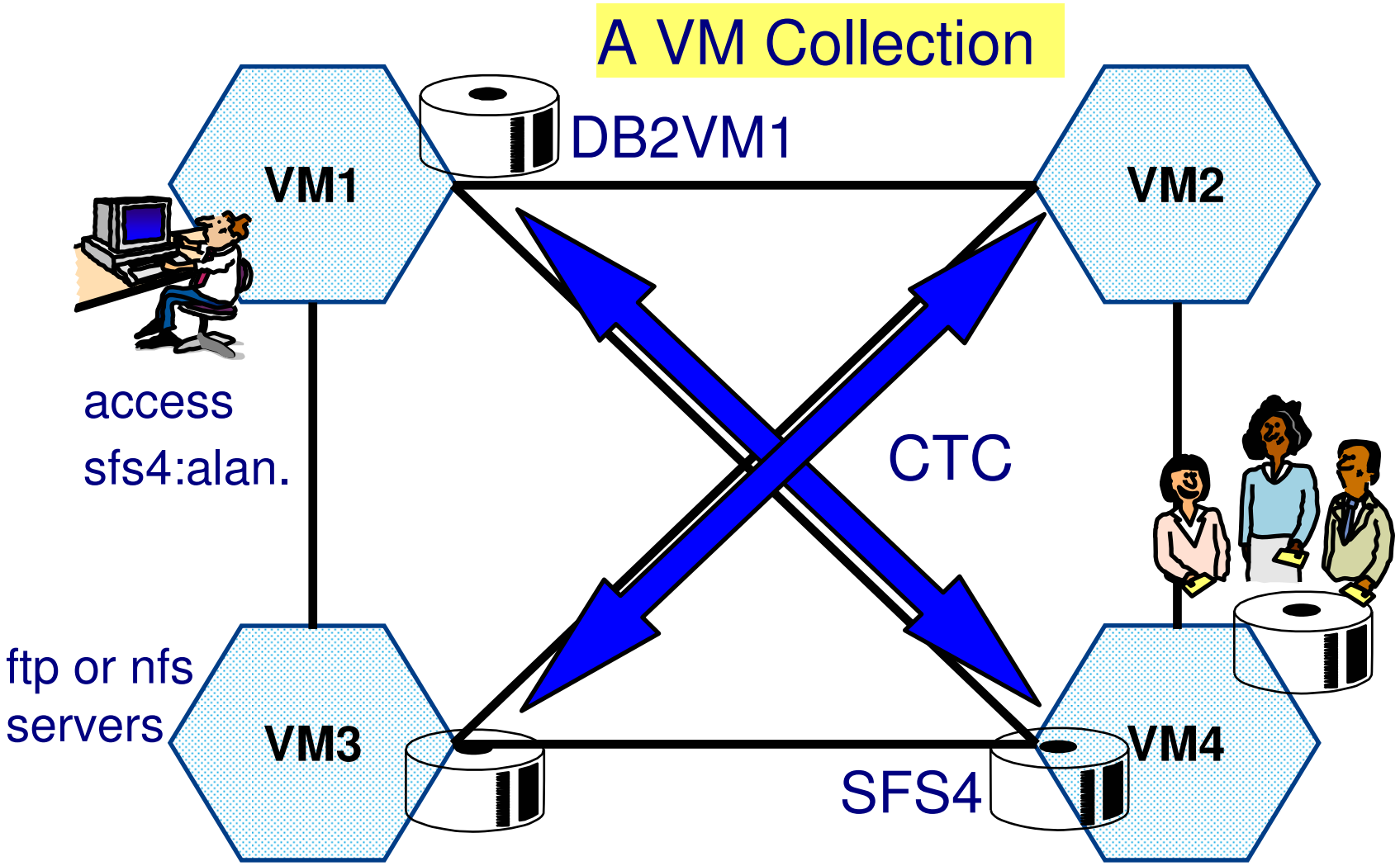
CSE Single Source Directory

All systems use the same source directory

- Each system has its own object directory
- May be managed with directory manager program such as DirMaint



Inter-System Facility for Communications



“TRACK” Diagnose

- TRACK tool originally from Princeton (Serge Goldstein) now maintained by Nationwide (Jim Vincent)
- z/VM 5.2 storage management changes provided motivation to dispense with TRACK’s use of LOCK, DISPLAY HOST, and Diagnose 4 (Examine Real Storage)
  - Proposed Diagnose interface to enable authorized guest to gain access to target’s base address space or System Execution Space as a data space
    - Natural use by exploiting Access Register mode
  - Code written but serialization issues never resolved => not released
  - Turned out to be useful for guest migration (with extension to allow read-only or read/write access to target’s address space)

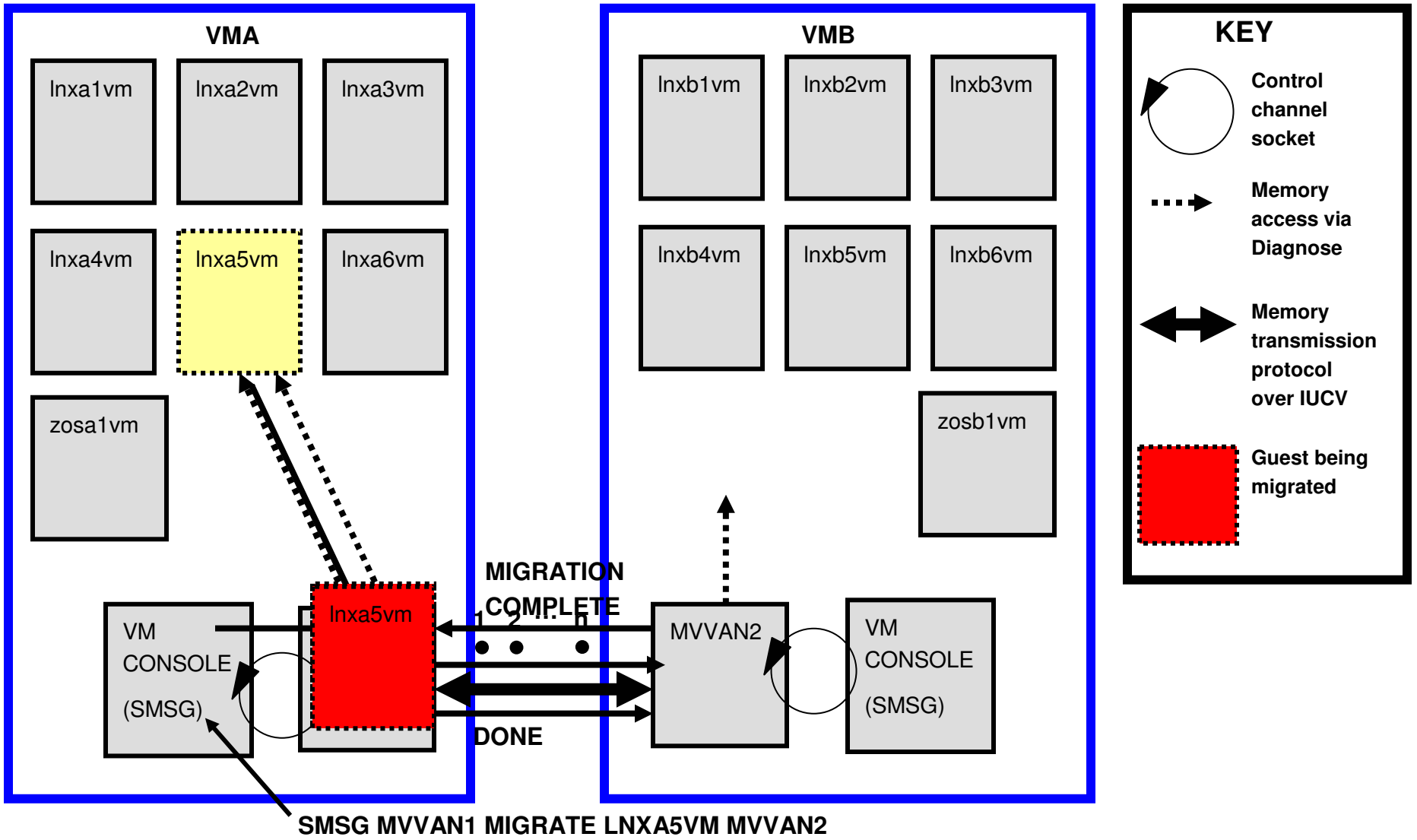
Migration Diagnose

- **Migrator interface to CP functions**
  - **Begin migration (outward or inward)**
  - **Get guest configuration**
  - **Set guest configuration**
  - **Retrieve migration change bits**
  - **Stun guest**
  - **Get guest state**
  - **Restore guest state**
  - **Abort migration**

Guest Memory Change Tracking

- **Initiated by Migration Diagnose "Begin outward migration" function**
  - **Causes target guest key operations to be intercepted**
  - **Keeps shadow copy of page change state for migration**
- **First invocation of "Retrieve migration change bits" returns a "1" bit for each non-zero target guest page and resets all migration change bits**
- **Subsequent invocations clear migration change bits and return a "1" bit for each page changed since last invocation**

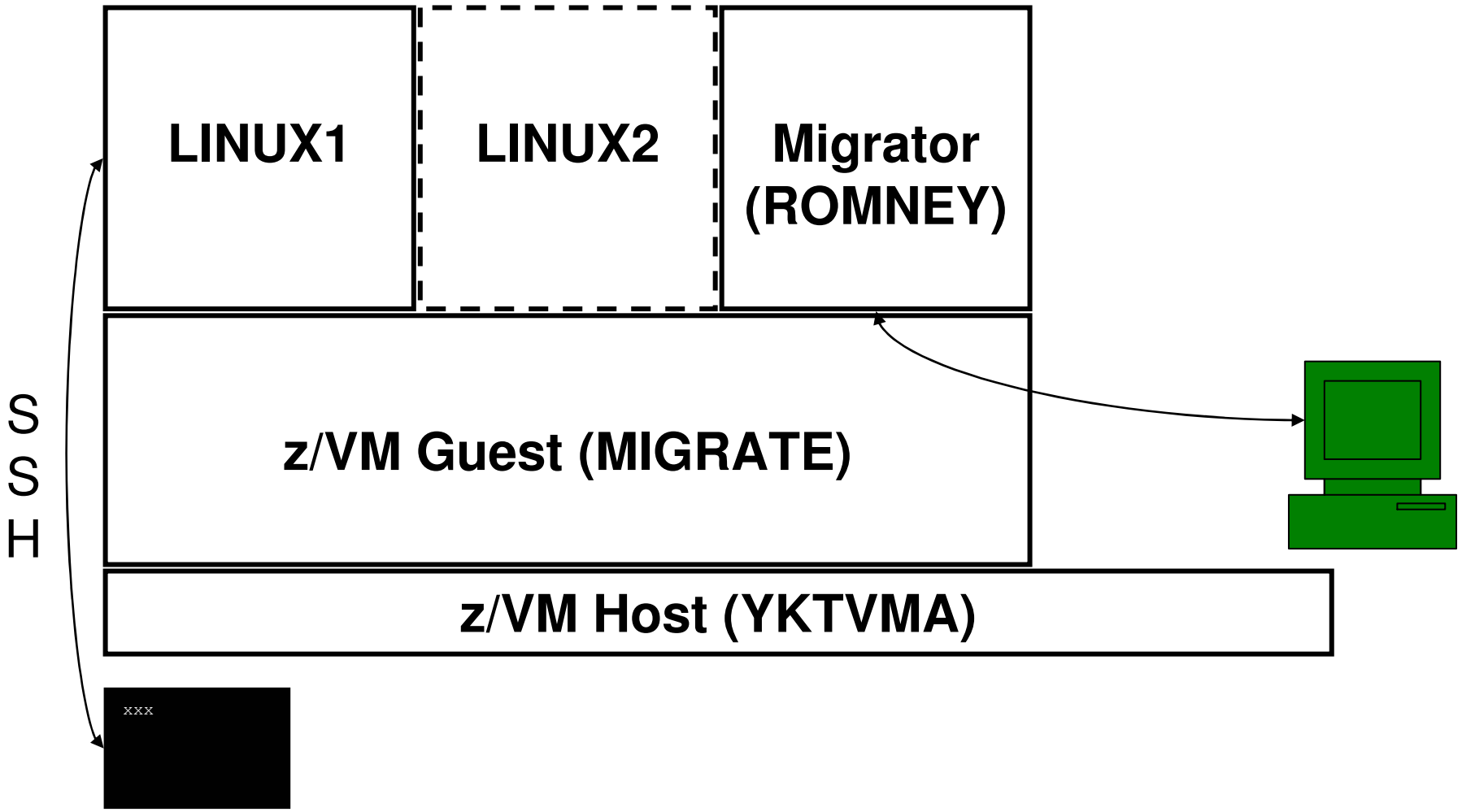
Conceptual Migration Process



Technology Demonstration

- **Configuration**
- **Caveats**
- **Problems**
- **Demo**

Technology Demonstration - Configuration



Technology Demonstration - Caveats

- This is a proof-of-concept
  - Same system still presents most challenges
  - Simpler to set up, control, and demonstrate
  - Guests are only 256MB
  - Not speed team moving van – REXX program orchestrates migration using functions that invoke Track and Migration Diagnoses
- Using a different user identifier is merely a convenience
  - Facilitates testing
  - Does not affect other aspects of migration
- Invocation via SMSG or as a CP command is well understood
  - Some additional considerations (e.g., serialization of requests) will have to be made

Technology Demonstration - Problems

- **Brief but inconvenient pause (PING) after migration completes and network interface reset**

Technology Demonstration

Challenges

- Release-to-release compatibility
- Existing CSE and ISFC customer environments
- Processor architecture and features
  - E.g., System z9 to z990
- CSE and ISFC duplication
  - Collection definition
  - Communication
- Distance
  - Shared I/O subsystem
- User name space
- Migration eligibility
  - Some current restrictions will disappear
  - Others will need to be removed

Summary

- **Multi-system virtualization on System z is feasible**
  - **Need to define objectives**
  - **Requires staged delivery plan**
- **We have a guest migration prototype**
  - **Work needed to make the function production-ready**


IBM
eserver