

Enhancing Live Migration Process for CPU and/or memory intensive VMs running Enterprise applications

Benoit Hudzia

CEC Belfast / SAP Research

08/2011

With the contribution of Aidan Shribman and Petter Svard



SAP RESEARCH

Agenda

- **Background: Enterprise Applications and Live Migration**
- **Warm Up**
- **Delta Compression**
- **Page Priority**
- **Future Works**



Background

Migrating Enterprise Class applications



Enterprise application and Live Migration

Issues

- **Enterprise class application:**

- Bigger than average resource requirement
- Average SAP ERP 16GB + per VM with 32 GB of swap more than common
- OLTP system such as ERP are very sensitive to time variation.
- Rely heavily on precise scheduling capabilities, triggers, timers and on the ACID compliance of the underlying

- **Challenge when migrating such application:**

- Disconnection of services:
 - Gigabit Ethernet timeout \approx 5 seconds (>500 MB memory left in stop and copy phase)
 - Downtime is workload dependent
- Disruption of services:
 - Migration progressively increasing the amount of resource dedicated to itself => gradually degrade performance of the coexisting systems / VMs.
- Difficulty to maintain consistency and transparency
- Unpredictability and rigidity



Warm Up for Live Migration

Increasing the flexibility of Live Migration



SAP RESEARCH

Warm Up

Increasing flexibility

Extended adaptive Pre-copy phase without triggering actual migration

Increased flexibility :

- “just in time“ triggering of live migration

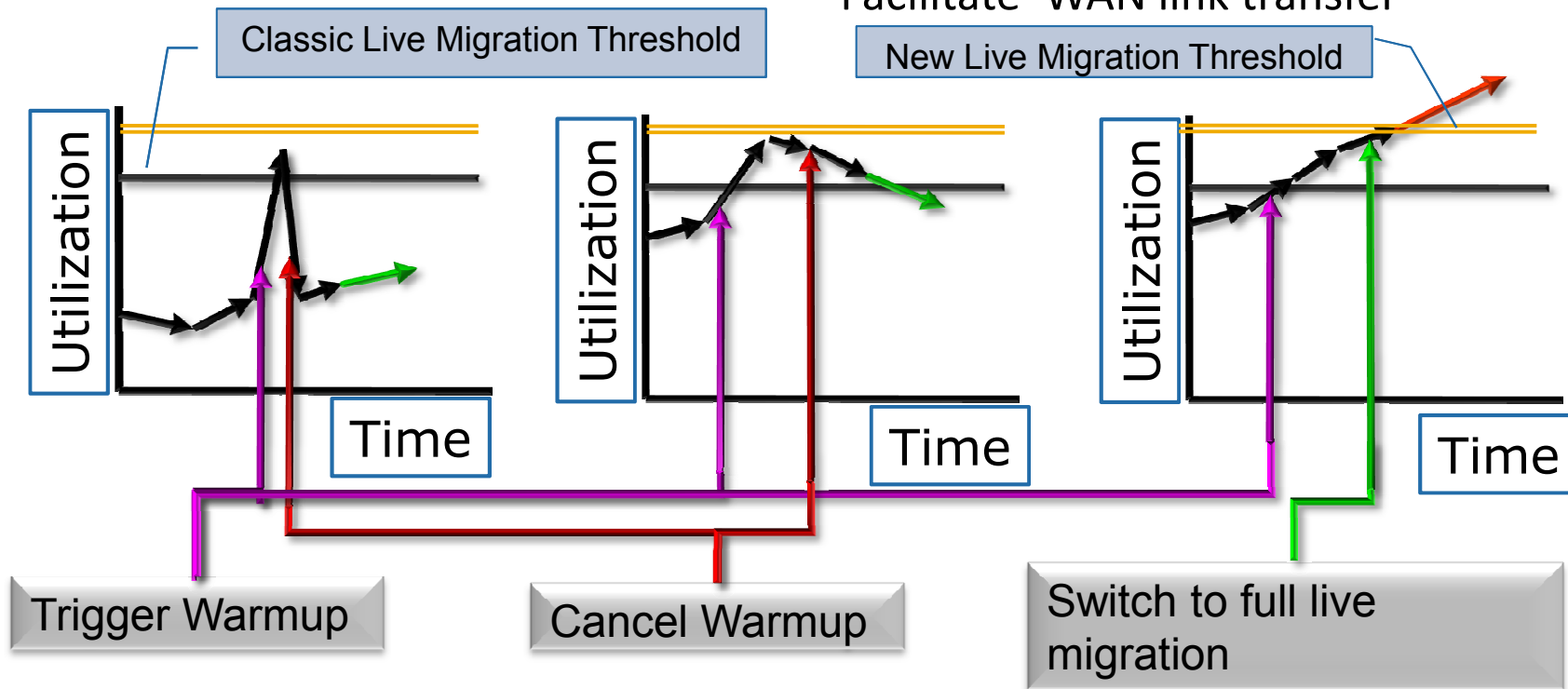
- Reduce down time

Dynamic adaptive bandwidth allocation

- Manual and automatic

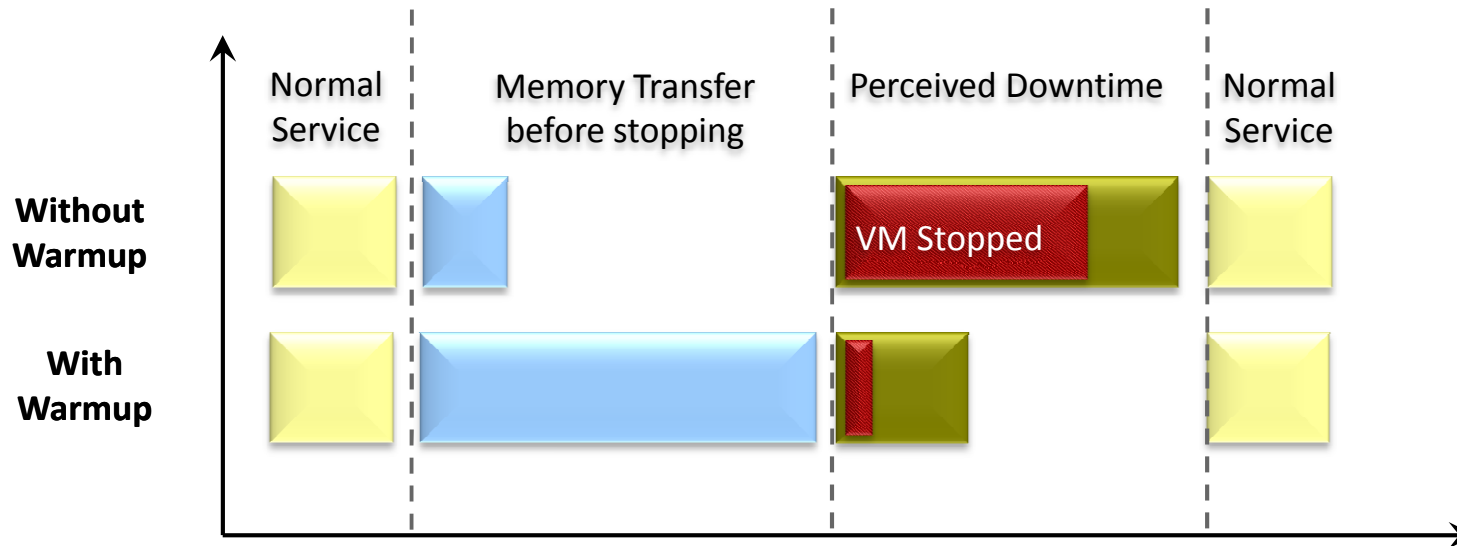
Allow “hot standby”

Facilitate WAN link transfer



Experimental Results: Warm-up Summary

SAP Sales and Distribution Benchmark



VM size : 4GB

SMP : 2 vCPU

Users : 150

Load ~= 80%

	CPU	Avg Response Time
Baseline	60%	2.18 sec
Warm-up	73%	2.16 sec

Downtime under load: <1 sec
Success ratio : ~99%



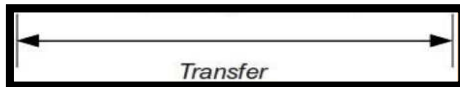
Delta Compression of Page

Limiting the impact of resending Page

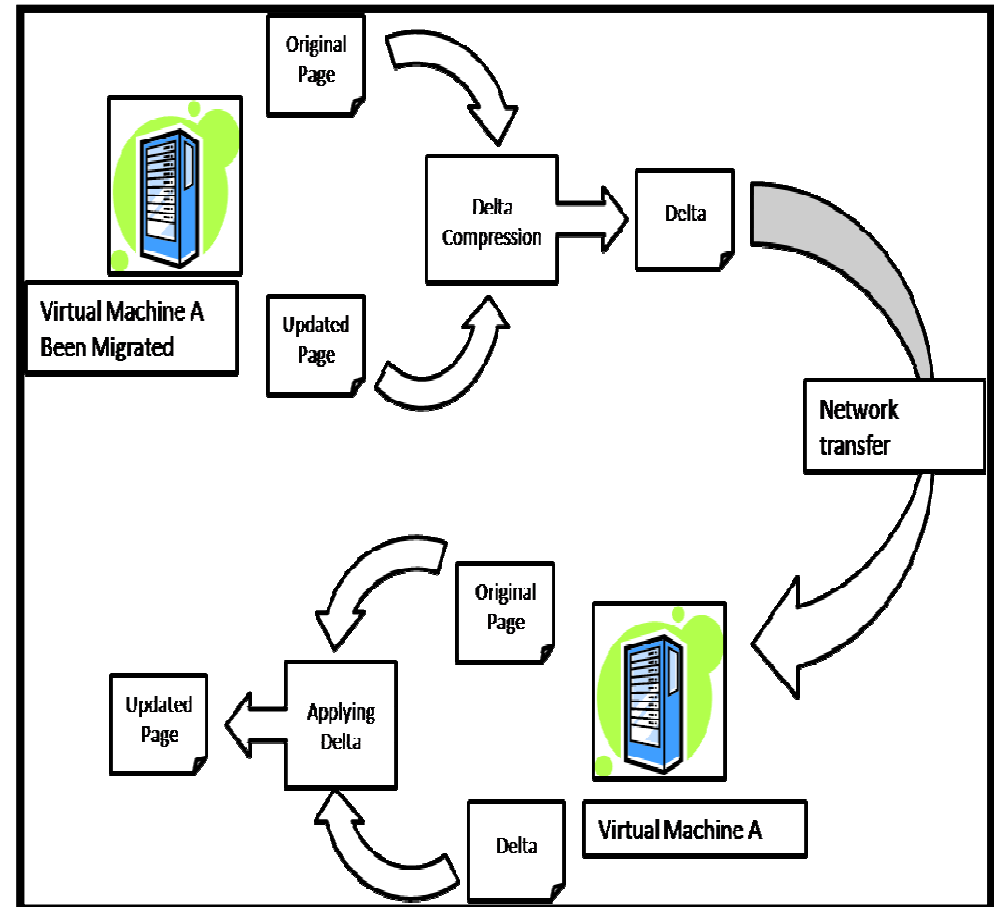
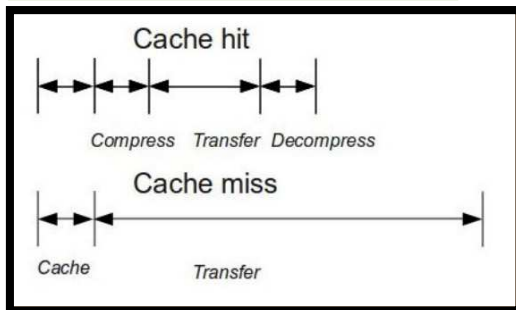
Dirty Page Delta Compression

- Cache page with highest dirtying rate during send operation
- Compression Algorithm:
 - XBRLE : XOR +binary run length encoding

Vanilla (no compr.)



Delta compression

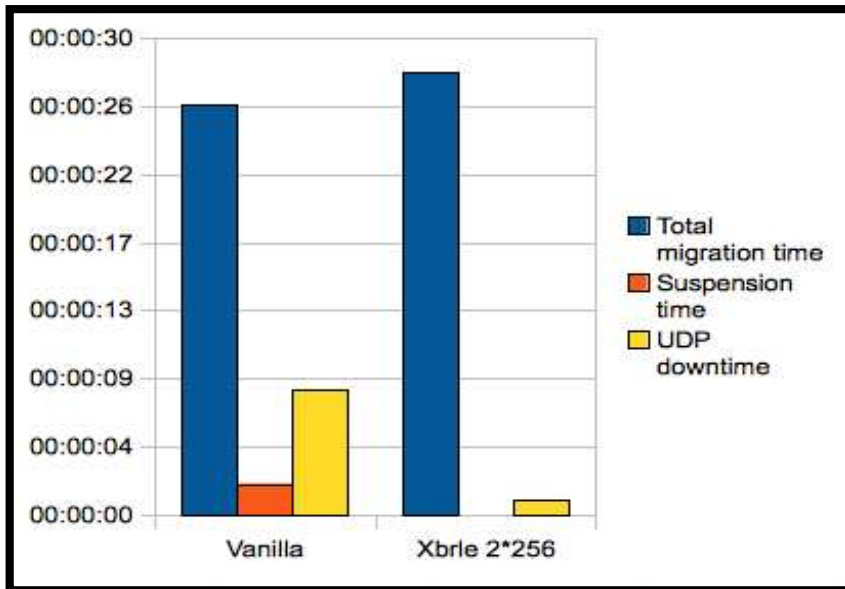


Evaluation

Benchmark

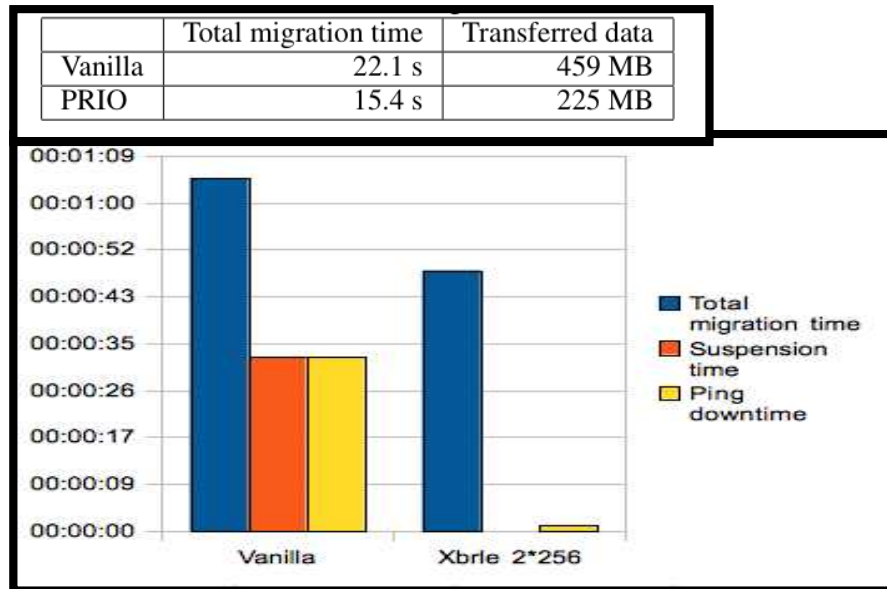
- **Memory write benchmark (Im_bench)**

- 1 GB RAM, 1 vcpu VM
- Near ideal case
- Downtime reduced by a factor of 100
- Throughput increased by 63 %



- **Transcoded HD Video (VLC)**

- 1 GB RAM, 1 vcpu VM
- Real-world, non-ideal case
- UDP downtime reduced from 8 s to 1
- Migration is transparent using XBRLE
- 31% faster, 51% less data sent

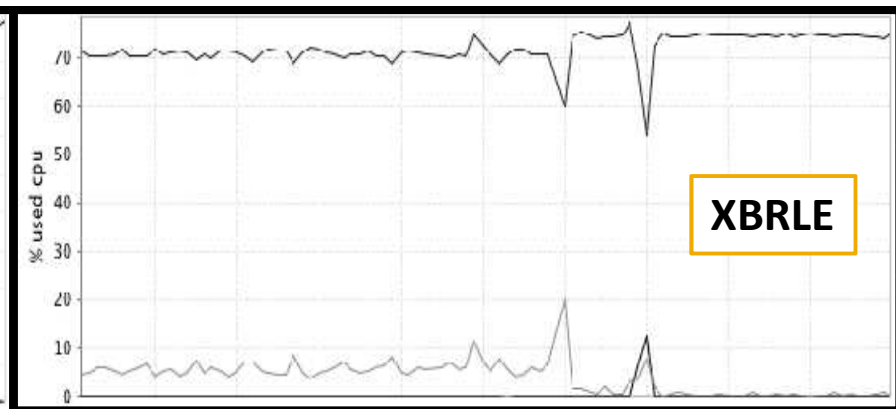
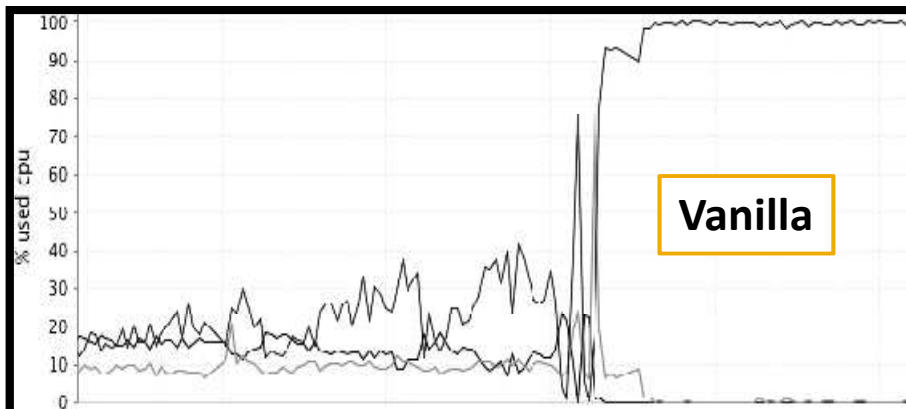


SAP RESEARCH

Evaluation- SAP ERP

Sales and Distribution benchmark, load 100%

- Non-responsive on resume with vanilla algorithm
- Survived using XBRLE
- >0.5s of downtime = risk of damaging the system
- Measured downtime was 0.2s for XBRLE and 2s for vanilla
- Live Migration Cpu usage directly impact (limit) the available resource for the ERP



HW:4x 3,0GHz Xeon dual-core 32GB RAM
16TB Raid 5, 6Gbits/s trunked NFS server
1000Mbit/s Network

VM:8 GB RAM, 4 vcpus VM
App: SAP ERP 7.0 / S&D Benchmark

SAP RESEARCH

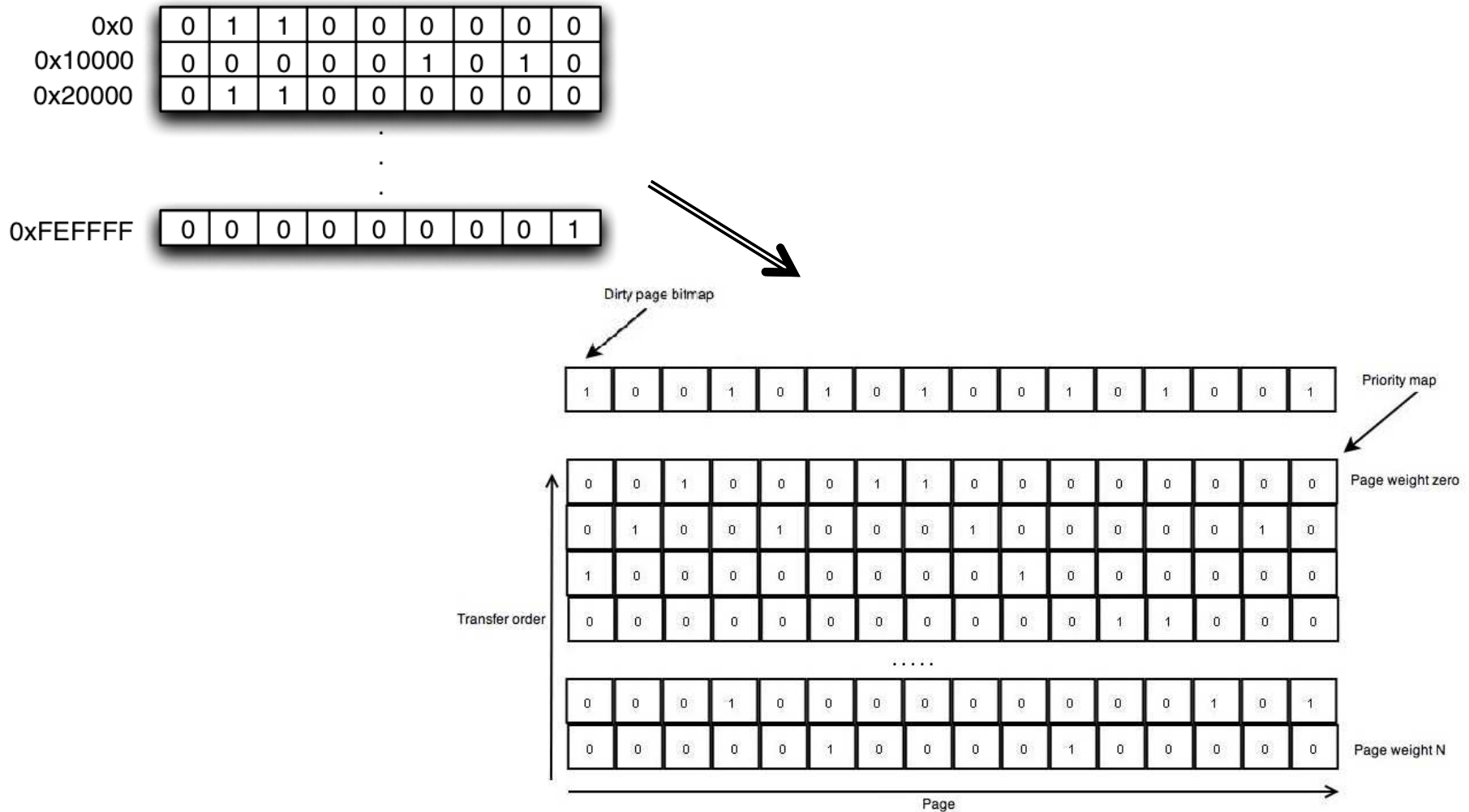


Page Prioritization

Dynamic page transfer reordering

Dynamic page transfer reordering

Prioritizing page sends (similar to writable working set concept in Xen)



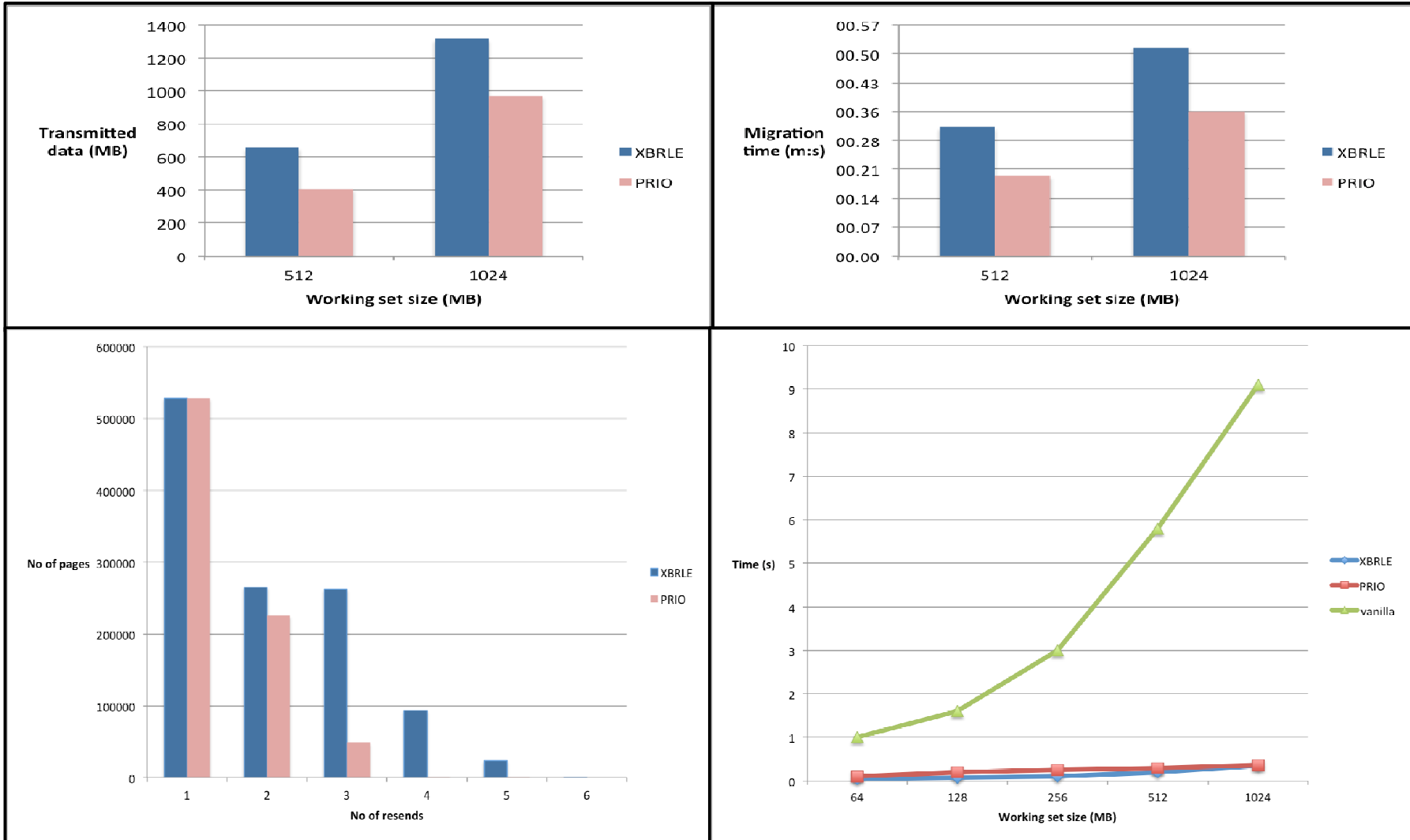
Dynamic page transfer reordering

Prioritizing page sends



Evaluation

Prio vs XBRLE : reveal Cache miss and compression efficiency Issue





Optimizing Compression

Making XBRL E more efficient



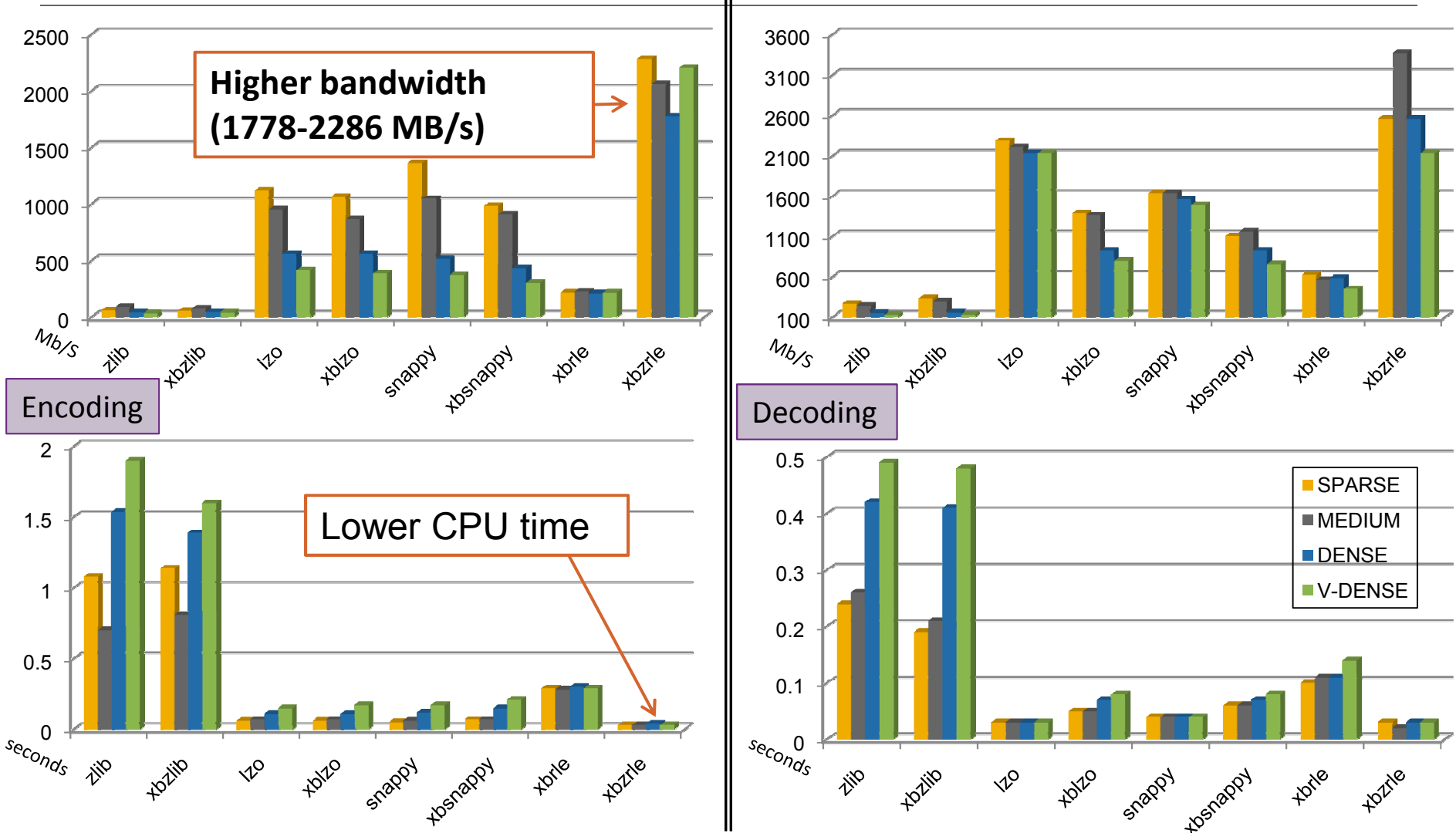
XBZRLE

Increase compression speed /efficiency

- Only compress unmodified data using word aligned encoding and only encodes runs of zeros
- For encoding page diffs XBZRLE is:
 - Compression :
 - 20% more efficient than XBRLE
 - 20% less efficient than LZ0/Snappy.
 - Speed:
 - Overall 2.5x-5x faster than XOR + LZ0/Snappy
 - 11x-9x faster than the original XBRLE
- Doesn't solve the impact of cache miss

Performance comparison

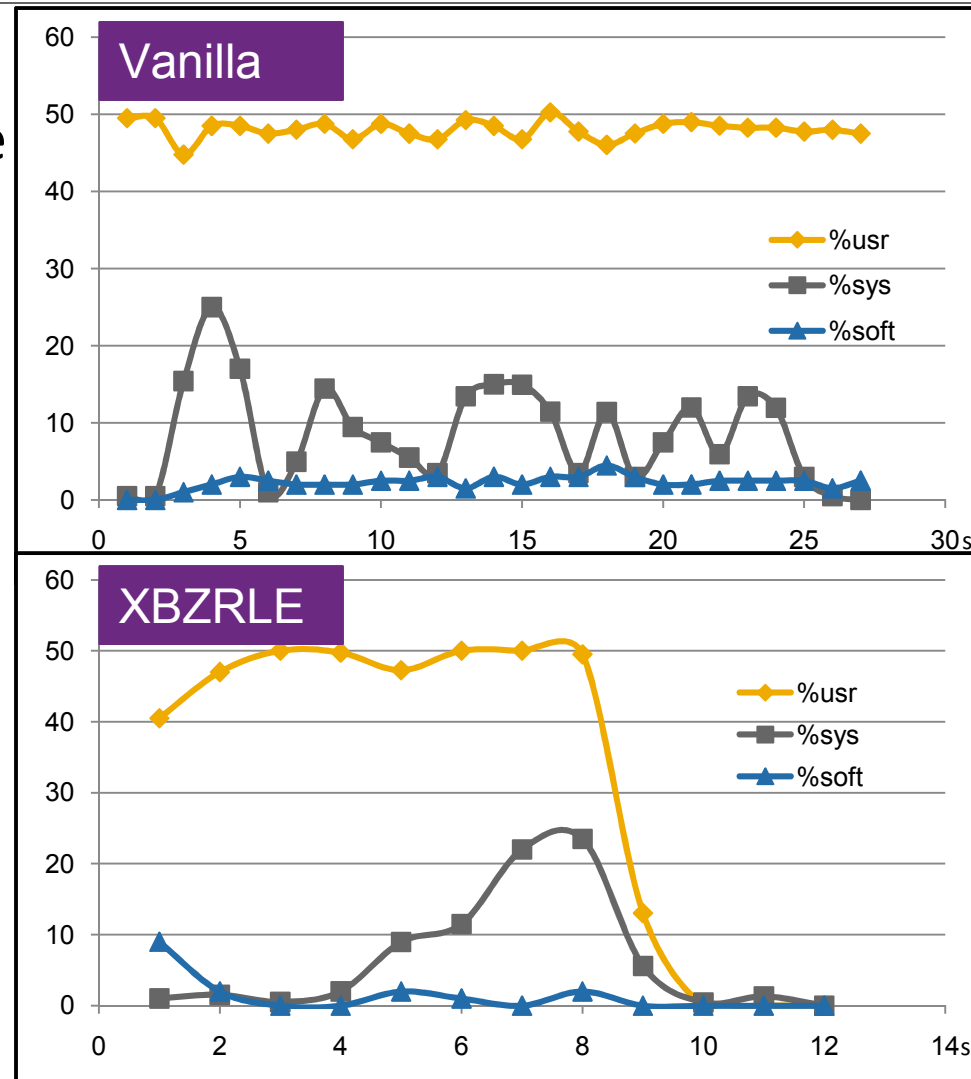
Synthetic benchmark representing enterprise workload



Performance comparison

Live Migration Benchmark

- Compute capacity used for live migration :
 - **xbzrle** : 50%
 - **vanilla**: between 30%-60%
- Live Migration:
 - **xbzrle** : terminate in seconds
 - **Vanilla** :not able to complete in the allocated time





Future Work



Future Works

- **Dynamically disable XBZRLE algorithm if the cache miss ratio is to important**
- **Combine Page priority algorithm and XBZRLE:**
 - Cache page with highest dirtying rate
 - Eliminate unnecessary cache check
 - Eliminate page compression with low potential return



Thank You!

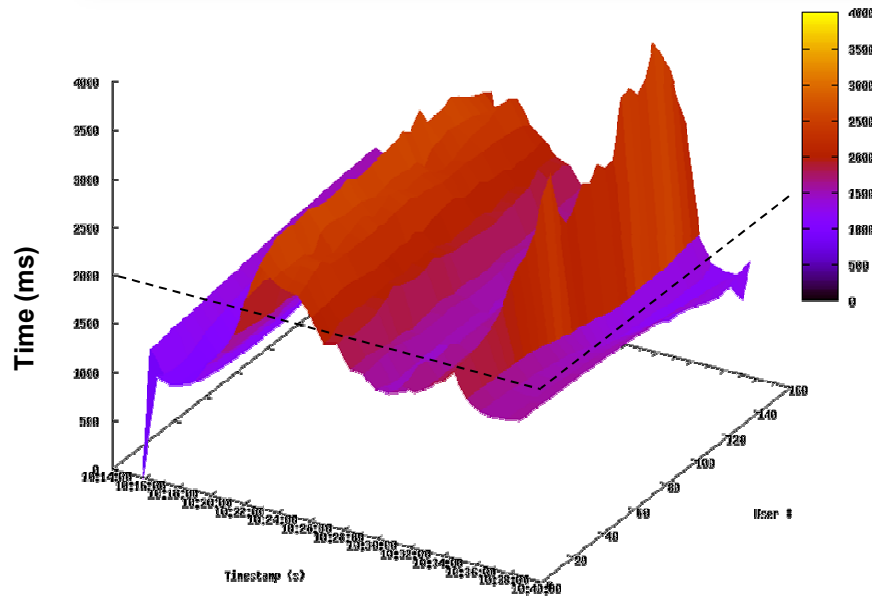
Contact information:

Dr. Benoit Hudzia
Senior Researcher
benoit.hudzia@sap.com

SAP RESEARCH

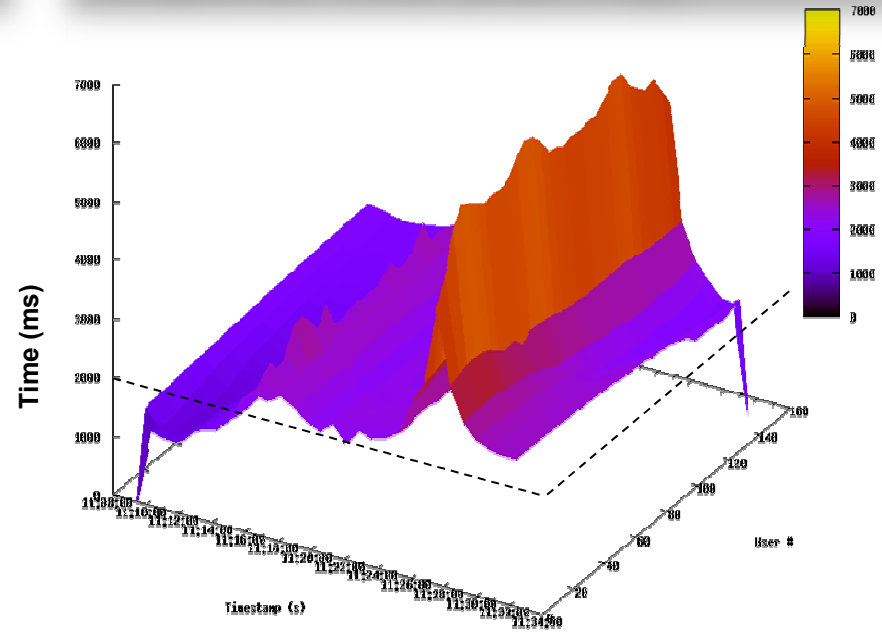
Experimentations Results: S&D Benchmark with/out warm-up

Response Time (baseline)



----- 2 s response time threshold

Response Time (warm-up)



----- 2 s response time threshold

VM size : 4GB

SMP : 2 vCPU

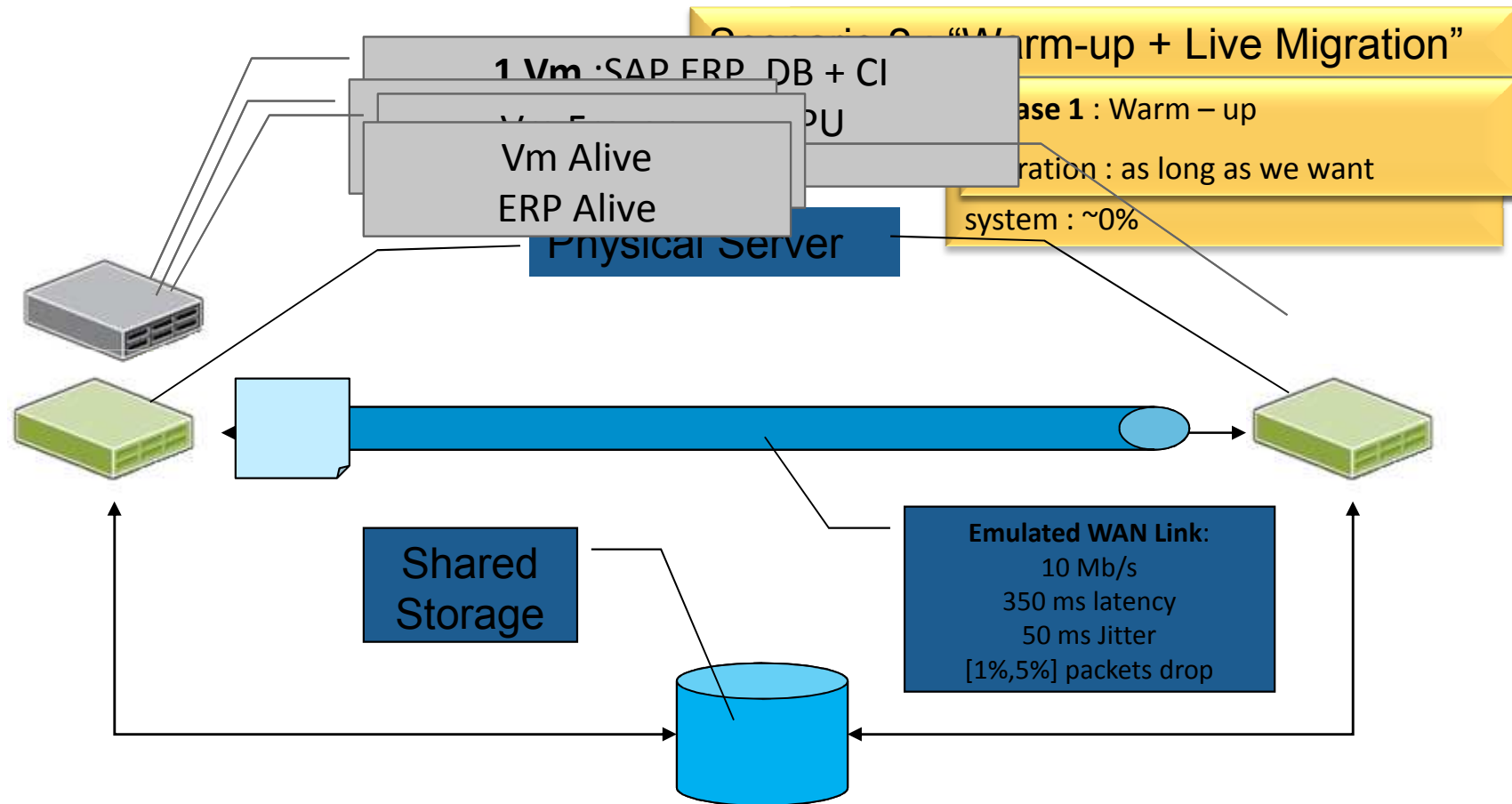
Users : 150

	CPU	Avg Response Time
Baseline	60%	2.18 sec
Warm-up	73%	2.16 sec

Downtime under load: <1 sec
Success ratio : ~99%

SAP RESEARCH

Live Migration over emulated WAN Link



No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

Microsoft, Windows, Excel, Outlook, and PowerPoint are registered trademarks of Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, System z9, z10, z9, iSeries, pSeries, xSeries, zSeries, eServer, z/VM, z/OS, i5/OS, S/390, OS/390, OS/400, AS/400, S/390 Parallel Enterprise Server, PowerVM, Power Architecture, POWER6+, POWER6, POWER5+, POWER5, POWER, OpenPower, PowerPC, BatchPipes, BladeCenter, System Storage, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, Parallel Sysplex, MVS/ESA, AIX, Intelligent Miner, WebSphere, Netfinity, Tivoli and Informix are trademarks or registered trademarks of IBM Corporation.

Linux is the registered trademark of Linus Torvalds in the U.S. and other countries.

Adobe, the Adobe logo, Acrobat, PostScript, and Reader are either trademarks or registered trademarks of Adobe Systems Incorporated in the United States and/or other countries.

Oracle is a registered trademark of Oracle Corporation.

UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems, Inc.

HTML, XML, XHTML and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Java is a registered trademark of Sun Microsystems, Inc.

JavaScript is a registered trademark of Sun Microsystems, Inc., used under license for technology invented and implemented by Netscape.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.

Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase, Inc. Sybase is an SAP company.

All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.

This document is a preliminary version and not subject to your license agreement or any other agreement with SAP. This document contains only intended strategies, developments, and functionalities of the SAP® product and is not intended to be binding upon SAP to any particular course of business, product strategy, and/or development. Please note that this document is subject to change and may be changed by SAP at any time without notice.

SAP assumes no responsibility for errors or omissions in this document. SAP does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

SAP shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The statutory liability for personal injury and defective products is not affected. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.