IBM and Novell:
The SLES 10 Success Story

White Paper

September, 2006

Introduction

Novell's SUSE Linux® Enterprise Server (SLES) 10 delivers new functionality, improved scalability and increased performance. To enhance IBM solutions on System x™, System p™, System z™, and BladeCenter® servers, SLES 10 includes over 180 features requested by IBM. SLES 10 features support new IBM hardware (devices and processors), cross platform functionality, and interoperability of IBM middleware and hardware. IBM implemented customer requirements, gained community acceptance and performed extensive testing of SLES 10 on IBM platforms. IBM's collaboration with Novell has continued to advance the SLES enterprise solution, with fast time to production, new technology and superior quality.

This paper primarily focuses on features with special value for IBM platforms and IBM contributions. For a complete overview of SLES 10 and SUSE Linux Enterprise Desktop (SLED) 10 content, please visit the Novell web pages http://www.novell.com/news/press/item.jsp?id=1031.

The following sections give an overview of key enhancements by major area and platform including the following areas:

- Virtualization
- Kernel
- Networking
- Reliability, Availability, Serviceability
- Security
- Toolchain

Virtualization

One of the key highlights of SLES 10 is the virtualization provided by Xen, supported on IBM's System x platforms. In addition to Xen, this chapter highlights SLES 10 specific enhancements for virtualization on System p, System z9™ and System z servers.

Virtualization on System x - Xen

Xen is a lightweight open source hypervisor. Xen on System x servers provides the ability to run multiple operating systems simultaneously in an isolated manner on a single server. This allows sparsely used machines running different operating systems and applications to be consolidated on to a single machine, helping to reduce server hardware and operating costs. Xen supports paravirtualized guest operating systems and unmodified guests in hardware-assisted full virtualization mode using Virtualization Technology designed by Intel® and AMD. In this release, SLES 10 is the only supported guest operating system. Support for other operating system releases are planned for future updates.

Xen provides isolated operating environments for applications by running on independent operating system instances. It reduces the impact of a failure of one application on another, without requiring the use of independent hardware.

Xen in the SLES 10 release provides the following features:

- Support for IBM's Secure Hypervisor Initiative components such as the Trusted Platform Modules, which provide hardware based security and trust, and support for other security mechanisms to provide isolation and access control between domains and system devices.
- Virtual devices are presented to guest operating systems (other SLES10 instances), which can protect them from or minimize the impact of failures in real devices.
- Support for up to 32-core SMP systems, and hot-plug re-provisioning of processors to dynamically adjust processor resources.
- Support for Physical Address Extensions, which allows 32-bit operating systems to address more than 4GB of memory.

- Checkpoint and restart functionality for the operating system, as well as certain limited support of the migration of an operating system instance to another physical server. This helps minimize downtime due to server upgrades, hardware failures, and maintenance.

Virtualization on System p servers

System p Advanced POWER™ Virtualization now scales from Linux images running on 1/10th of a processor up to 64 processors. Because processing power can be changed on the fly, this provides tremendous flexibility in managing multiple server images in a single Power Architecture™ technology-based system. Unique to POWER5™ servers, IBM has implemented Micro-Partitioning™ as an option in the server, bringing this function to a broader class of Linux environment clients and applications. The Linux operating system has been adapted and optimized for virtualization. The virtualization of physical processors in POWER5 systems (or Micro-Partitioning) introduces an abstraction layer that is implemented within the hardware microcode. From an operating system perspective, a virtual processor is the same as a physical processor. The key benefit of implementing partitioning in the hardware allows any operating system to run on POWER5 technology with little or no changes.

Virtualization on System z9 and System z (LPAR and z/VM) servers

System z servers provide both hardware (LPAR) and software virtualization (z/VM®). Each of the two virtualization schemes are proven to be robust and to satisfy the needs of mission-critical services in an enterprise. SLES 10 can run in either an LPAR or a z/VM guest and thus inherits the flexibility, scalability and manageability benefits of System z virtualization. In particular, z/VM V5.2 improves the scalability to host servers with large memory requirements and high network and storage traffic. Using SLES 10, you can benefit form various improvements related to accurate accounting, performance and RAS as described in the System z section later in this article.

Kernel Enhancements

SLES 10 provides core enhancements to the Linux kernel to help improve stability, scalability and improved features and functionality. Some of the key kernel enhancements include the following:

- Virtual File System (VFS) enhancements to support shared subtrees of name spaces. These enhancements allow full support for Rational® ClearCase® and Multi-Version File System (MVFS) software available from IBM.
- Madvise (MADV_REMOVE) and lazy paging support for shared memory allocations, which enable autonomic management of DB2® software.
- ext3 file system enhancements allowing users to add on new space to an existing partition without unmounting the file system first (online resize).
- Robust Virtual Machine design to handle memory pressure handling.
- Flexible resource management (CPUSETS, NUMA enhancements, Pluggable I/O schedulers and fork/exit event notification).
- Delay accounting statistics for enabling enterprise workload manager (eWLM).
- Enhanced device driver support.
- Hotplug Memory Add Support, which allows dynamic increases of memory in the system (in particular for virtual environments).

Networking

SLES 10 provides new networking functionality and IBM's contributions in these areas include the following:

- Network File System version 4 (NFSv4) is a standards-based state-oriented network file system protocol with built-in security support. NFSv4 support available through the SLES 10 kernel has been further stabilized with patch contributions from IBM.
- IPv6 Advanced Sockets API enables "advanced" IPv6 applications to access features such as interface identification options and IPv6 extension headers that are not addressed in the basic sockets API (RFC 2133). The IPv6 Advanced Sockets API also enables hop-by-hop options, destination options and other features making it compliant with RFC 3542.

- The Stream Control Transport Protocol (SCTP) connectx() API in SLES 10 provides a faster way to set up an association with a multi-homed server when the client is aware of the multiple IP addresses of the server. SCTP allows the caller to specify multiple addresses at which a peer can be reached and let the kernel do the retries.
- Other SCTP enhancements include Multiple SCTP bug-fixes, performance enhancements and updates to the API to bring it closer to the latest SCTP sockets API draft.

Reliability, Availability, Serviceability

With SLES 10, numerous RAS enhancements are part of IBM solutions. SLES 10 has a new crash dump capability named Kdump. Kdump is a First Failure Data capture (FFDC) mechanism and it is supported across multiple architectures. Reliability of the Kdump comes from the fact that it doesn't depend on the dying kernel to capture the dump, unlike previous solutions. SLES 10 includes the enhanced analysis tool, Crash, for advanced dump analysis capabilities.

IBM brand hardware capabilities, error recovery, remote management and service capability, are enabled in SLES 10. IBM solutions with SLES 10 deliver improved error handling at the hardware level and memory parity error checks. IBM further improved the enhanced error handling (EEH) capabilities on System z and on System p servers. On PowerPC® processor-based hardware, EEH helps to detect and recover from a wide assortment of PCI device and bus errors. System z servers now have better instrumentation to analyze problems with SCSI/FCP and networking in guest LANs. IBM platforms also have the ability to manage and service systems remotely with the help of remote management cards in the system.

SLES 10 provides advanced performance analysis and a debugging tool in the form of SystemTap. SystemTap provides a safe and easy scripting language for system administrators to analyze performance problems all the way from application programs to device drivers. The tool is also sophisticated enough for experienced system programmers to use in their debugging with its on-demand probing. The tool has zero overhead when not in use and very low overhead when in use, this makes it ideal for doing real-time performance analysis of production systems.

Security

SLES 10 delivers security enhancements to provide advanced cryptographic performance, hardware-enabled security, workload isolation, data confidentiality and regulatory compliance. IBM has aided in this effort by providing:

- Enhancements to openCryptoki, a PKCS#11 implementation, to exploit the new embedded cryptographic System z instructions providing improved cryptographic performance.
- A new dynamic OpenSSL engine to exploit the new embedded cryptographic System z instructions providing improved cryptographic performance within OpenSSL.

Support for the Trusted Computing Group (TCG) version 1.1 Trusted Platform Module (TPM). The TPM device enables smart card-like support to provide secure key storage to reduce software attack vectors.

Toolchain

SLES 10 brings a number of toolchain-related (gcc, binutils, glibc) security and performance enhancements.

For 32-bit PowerPC, SLES 10 provides the -msecure-plt option and has made it the default. The -msecure-plt option allows applications to be divided into pure read-only executable code and read-write no-execute data areas. This is important to preventing, so-called buffer overrun attacks. With SLES 10 there is also improved signal handling for 32- and 64-bit PowerPC so that the run-time stack does not need to be executable. Finally the gcc compiler added the -fstack-protector feature that detects buffer overflows on the stack.

The gcc-4.1.1 compiler provides numerous code generation improvements over previous gcc-3.3 or gcc-3.4 compilers. PowerPC gains are estimated at 10% improvement for 32-bit and 13% for 64-bit using SPECint2000 tests, when comparing SLES 9 with gcc 3.3 and SLES 10 with gcc 4.1 on a 4-core Power5

system. SLES 10 also offers improved VMX/Altivec support for the JS20 Blade with autovectorization features of gcc 4.1.

Key Enhancements by Platform

This chapter presents key enhancements for IBM System p, System x, System z9 and System z hardware platforms.

System p (Power Architecture)

Linux on POWER processor-based servers provides supported applications with high performance, high scaling, high capacity within a full family of servers, high reliability, leadership virtualization capabilities, and low operating costs compared to alternative server architectures. IBM is committed to leveraging Linux on POWER to provide a proven, open, and powerful computing platform to drive innovation which can enable competitive advantages and provide value to our client's company, customers and shareholders. Specifically, SLES 10 on Power Architecture provides high levels of hardware reliability and serviceability, a very efficient native virtualization environment, and high levels of system scalability (number of processors, amount of physical memory, and I/O bandwidth).

SLES 10 delivers the following features for POWER:

- 64-core SMP allows execution of single large OS images that requires large amounts of computing power operating on a single large shared memory. This is useful in certain high performance computing applications.
- NUMA optimization improves the performance of workloads running across large (greater than 4-8 processor cores) systems, by keeping memory accesses local to the processors doing the access (which has best performance).
- The iSCSI software initiator allows access to remote iSCSI storage devices across standard (Ethernet) network connections, providing a low cost remote storage solution.
- Post-Link Optimizations for Linux on POWER allow performance improvements to applications by reorganizing the compiled application code based on a certain workload to reduce page faults, cache misses, and so on. For more information about Post-Link Optimizations, refer to alphaWorks® at http://www.alphaworks.ibm.com/tech/fdprpro.
- Many portions of the operating system (kernel, libraries, tools) have been tuned to provide better performance on POWER systems.

VMX exploitation provides performance improvements to applications that require "vector operations" by providing a single instruction that operates on a number of data items at once. This capability is available in systems built around 64-bit PowerPC 970FX processors - like the IBM BladeCenter® JS20 and JS21 systems.

- Large page support improves performance by using a smaller number of translations that the processor has to deal with (and can cache in the processor) for a given large contiguous range of virtual memory.
- Device Driver error log analysis provides a mechanism for analyzing the errors generated by a device and recommending replacement actions.
- iSCSI TCP/IP offload adapter support provides a high performance mechanism for communicating with iSCSI storage devices by offloading the work of the TCP/IP stack to the adapter rather than the operating system.
- The ltrace, itrace and oProfile tools allow tracing and performance analysis of the operating system and applications to determine where performance problems are occurring so that they can be addressed. The itrace tool provides input to the IBM Performance Simulator for Linux on POWER, which is also available from alphaWorks at http://www.alphaworks.ibm.com/tech/simppc
- Physical memory add allows more physical memory to be added to a Linux partition without rebooting the partition. This allows some adjustment to the size of a partition to accommodate larger workloads. Physical memory cannot be removed from a Linux partition without rebooting the partition. Processors can be added and removed from a Linux partition without rebooting it.
- Serial Attach SCSI are a new type of SCSI storage device that uses serial rather than parallel communication paths. SCSI storage devices are transitioning from parallel to serial interface types.

- PCI-Express support allows BladeCenter JS21 systems to use a faster, higher bandwidth path to the I/O devices, providing better I/O performance and scalability.
- Trusted Computing Specification Support provides a mechanism for applications to validate that they are running in an approved environment for processing digital media and other protected content.
- Non-executable stack and heap prevents applications from executing code in their stack or heap areas, preventing the common security problem of buffer overflow attacks, where a security attack causes an application to load executable code by overflowing a buffer that is either in the stack or the heap, overwriting code for the application.
- High performance time system calls allow timestamps to be generated more quickly, improving the performance of transactional systems like databases.
- 64-bit tracing tools allow tracing of 64-bit applications, easing the debugging and tuning of 64-bit applications.
- Libraries tuned for each processor type allow improved performance by using code that is tuned differently for each processor type. Different processor types have different instructions, different instruction scheduling mechanisms, and different performance for different instructions and thus different code will run fastest on each processor type.

System x

SLES 10 offers improved functionality and increased scalability on System x servers. System x servers utilize Intel and AMD processors, which represent the majority of Linux installments worldwide. The System x platform differentiates itself from other Intel/AMD systems by providing world-class customer satisfaction and assurance with respect to robustness, reliability, availability, and stability. System x includes features that are typically only found in higher-end mainframe environments. SLES 10 leverages these high-end features to provide a superior hardware/software environment. IBM has contributed to the following SLES 10 features, some of which are available for the first time, to further expand System x solutions:

- Dual core support (two processor cores per socket) for both Intel and AMD processor packages.
- IBM Extended X-Architecture™ (EXA) chipset support enables support for large SMP NUMA-based IBM servers, primarily with 64-bit kernels. The EXA chipset provides the I/O bridge controller and the cache/memory and scalability controller for high-end NUMA platforms.
- Global timesource to synchronize timekeeping for NUMA systems. Support for both the High Precision Event Timer (HPET) and the ACPI PM timer are used to consistently track time in environments with multiple clock sources.
- Intelligent Platform Management (IPMI) support using the OpenIPMI driver and user level interfaces (for example, ipmitool), provides the ability to manage and monitor platform hardware in-band or out-of-band. It includes IPMI 2.0 compliance, extensions for SoL, VLAN and DHCP firewall support.
- Serial Attached SCSI (SAS) support, SCSI protocol with a serial interface vs. older parallel bus based interface, provides improved I/O bandwidth and performance.
- Advanced Configuration and Power Interface (ACPI) extensions for system reset, PM timer, PCI hotplug, power management and processor throttling,
- Support for legacy-free System x blades.
- New System x device support: ServeRAID 8i/8k driver (aacraid) support and synchronization, Broadcom 5706/5708 (bnx2) NIC support, Broadcom 5714/5715 (tg3) NIC support and nVidia, ATI graphics accelerators.
- Execute Disable (XD) – Intel processor feature that protects against execution of malicious software at the hardware level.
- Active PCI – IBM hot plug features that extend existing PCI hot plug capabilities – PCI slot/device identification and notification handler extensions.
- iSCSI allows for remote access to storage devices over standard ethernet connections.
- PCI Express - offers extended PCI configuration space, improved bandwidth and faster speeds to improve overall performance and scalability.
- System x 4-node x460 with support for up to 64 logical processors and 256GB memory.

- NUMA support, which enables the kernel and applications, (through export APIs), to make intelligent decisions with respect to memory placement and process scheduling on systems that don't have uniform memory characteristics (for example, local and remote memory latencies differ).
- Improved Power Management capabilities that provide power cost savings by throttling processor frequency on unused and under-used processors. Full support for Demand Based Switching (Intel) and PowerNow! (AMD).
- Support for the Trusted Computing Group (TCG) version 1.1 Trusted Platform Module (TPM) device to enable hardware-enabled security.

System z9 and System z

SLES 10 on System z servers provides a cost efficient and reliable scale-up and scale-out application-hosting environment specifically meant to augment z/OS® enterprise deployments. To serve this goal, SLES 10 exploits the latest features of System z9 – the newest System z architecture level.

A new Fibre Channel Protocol (FCP), SCSI over Fibre Channel, Host Bus Adapter Virtualization Technology (N_Port ID Virtualization: logical WWPNs) on System z9 now allows sharing of FCP adapters in a fully SCSI standards-compliant way. This comprises SAN access right management and disk sharing through virtual HBAs.

SLES 10 network adapters have been enhanced to enable the Communication Controller for Linux products on System z9 for 374x NCP virtualization. To provide optimal performance in a security-critical environment, SLES 10 supports the new cryptographic adapters of System z servers and the new crypto instructions for AES and SHA-256 of the System z9 processors, the latter being not only accessible to applications, but also to the kernel.

Further enhancements in SLES 10 that relate also to pre-System z9 machines comprise miscellaneous virtual server enhancements. Linux images can now use the clock comparator for virtual elapsed time (rather than a wall clock) which is much more hypervisor-friendly. It is now possible to analyze network traffic on virtual network connections (Guest LANs) and there is support for concurrent I/O through multiple paths for ESCON/FICON (channel)-attached disk storage subsystems using the System z Parallel Access Volume technology.

With SLES 10 execute in place, support is fully integrated in the ext2 file system. This technology can be leveraged by customers running multiple similar servers in the System z z/VM Hypervisor. It allows for efficiently shared memory for executables and library code across multiple Linux guests running in the same z/VM system.

Further SLES 10 improvements are enhancements for z/OS disaster recovery integration (for planned outages) both for Linux in LPARs and Linux operating with FCP I/O attachments and support of IBM Director 5.1 to provision and manage Linux images in z/VM.

© IBM Corporation 2006

IBM Corporation
Marketing Communications
Systems Group
Route 100
Somers, New York 10589

Produced in the United States of America
September 2006
All Rights Reserved

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM's future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only. Any statements about support or other commitments may be changed or cancelled at any time without notice. This Information is provided "AS IS" without warranty of any kind.

IBM, the IBM logo, alphaWorks, BladeCenter, ClearCase, Micro-Partitioning, Power Architecture, POWER, POWER5, PowerPC, Rational, System p, System x, System z, System z9, X-Architecture, Z/OS, z/VM are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both. A full list of U.S. trademarks owned by IBM may be found at http://www.ibm.com/legal/copytrade.shtml.

Intel is a registered trademark of Intel Corporation in the United States and/or other countries.

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

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

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

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

SLES 10 is a non-IBM product and is not warranted by IBM. Information concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

LXW03001-USEN-00

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven, ClusterProven or BladeCenter Interoperability Program products. Support for these third-party (non-IBM) products is provided by non-IBM Manufacturers.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM.

The IBM home page on the Internet can be found at http://www.ibm.com.

The System p home page on the Internet can be found at http://www.ibm.com/systems/p