



Linux Performance

UT Austin - CS378 Presentation

Duc J. Vianney, Ph.D.

dvianney@us.ibm.com

Linux Performance

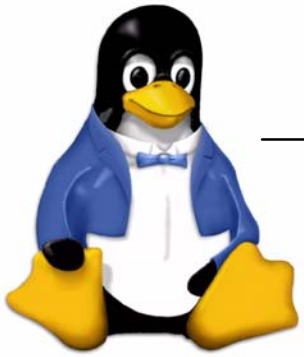
IBM Linux Technology Center

April 28, 2003

The IBM logo is located in the bottom left corner. It consists of the letters "IBM" in a bold, sans-serif font, with each letter made of horizontal stripes.

Linux Technology
Center

Legal Statement



- This work represents the views of the author(s) and does not necessarily reflect the views of IBM Corporation.
- The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries: IBM (logo), e-business (logo), developerWorks, DB2, iSeries, OS/2, pSeries, WebSphere, xSeries, zSeries. A full list of U.S. trademarks owned by IBM may be found at <http://www.ibm.com/legal/copytrade.shtml>.
- Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
- Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
- Linux is a registered trademark of Linus Torvalds.
- Other company, product, and service names may be trademarks or service marks of others.

The IBM logo, consisting of the letters 'IBM' in a stylized, striped font, is located in the bottom left corner of the slide.

Linux Technology
Center

Outline



- Overview of Linux Technology Center
- Linux® Performance Team
- Benchmarks and Workloads
- Linux Performance Tools
- Linux 2.5 Performance Line Items
- Base Kernel
- Web Serving
- Database
- File Serving
- File Systems
- Java™-based Benchmarks
- Hyperthreading
- Linux Performance Team Members

The IBM logo, consisting of the letters 'IBM' in a stylized, striped font.

Linux Technology
Center

IBM Linux Technology Center



Mission

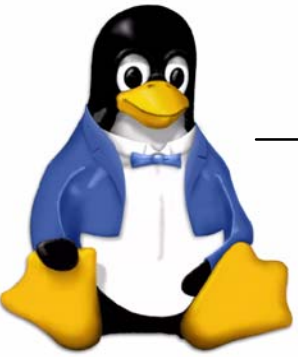
Work towards enterprise enablement of the Linux OS through the development and contribution of technology, utilities, tools and code

**250+ Developers Across
22 Worldwide Locations**

- Adelaide
- Austin
- Bangalore
- Beaverton
- Boeblingen
- Boulder
- Canberra
- Chicago
- Denver
- Haifa
- Hawthorne
- Hursley
- Mount Laurel
- Portland
- Poughkeepsie
- Raleigh
- Rochester
- San Francisco
- Seattle
- Somers
- Yamato
- Yorktown

The IBM logo is located in the bottom left corner of the slide.

Linux Technology
Center



The Role . . .

Linux Community - Core Components

Linux Technology Center

- **Mission: "Strengthen Linux"**
 - Actively accelerate the growth of Linux as an enterprise operating system
 - Working as a trusted, valued member of the Linux community
 - Providing Linux expertise to IBM's technical community
- 70+ active projects
- <http://www.ibm.com/linux/ltc>

Linux Distributions

xSeries™ - pSeries™
iSeries™ - zSeries™

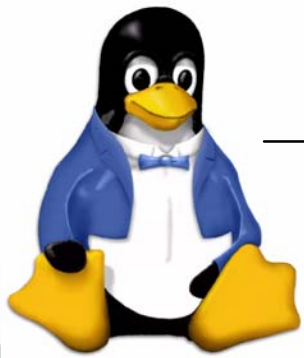
Storage

Software

IBM

Linux Technology
Center

Enterprise Focus Areas . . .

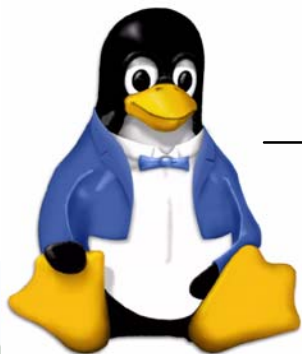


- Kernel Scalability
- POSIX Threading
- I/O & Storage Improvements
- Serviceability
- File/Print Enhancements
- Directory
- Security
- HPC Clusters
- Level 3 Support Team
- Accessibility
- Internationalization
- Enterprise Volume Management System
- Journalled File System
- Networking
- Embedded
- Systems Management
- High Availability
- Standards
- Linux Test Project
- Linux Documentation Project
- Graphics Workstation Support

IBM

Linux Technology
Center

Enterprise Focus Areas *(cont'd)*



■ Performance Benchmarking/Analysis

- ▶ Database Workloads
- ▶ VolanoMark
- ▶ Specweb99 / SPECweb99ssl
- ▶ Netperf
- ▶ tiobench
- ▶ Netbench
- ▶ SPECsfs
- ▶ SPECjbb2000
- ▶ SPECjAppServer
- ▶ Trade2 / Trade3

■ Performance Instrumentation / Tools

- ▶ Readprofile
- ▶ Kernprof
- ▶ Lockmeter
- ▶ Resource Monitoring



Linux Technology
Center

LTC Linux Performance Team



■ Mission

- ▶ Make Linux better by improving Linux kernel performance, with special emphasis on SMP scalability

■ Methodology

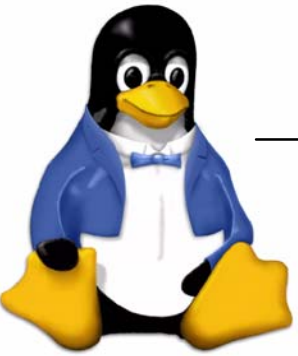
- ▶ Measure, analyze and improve the performance and scalability of the Linux kernel
- ▶ Focus on platform-independent issues
- ▶ Benchmarks that provide coverage for data center, carrier space, security and web server workloads
- ▶ Migration to newer kernels will occur as needed

■ Plan Assumptions

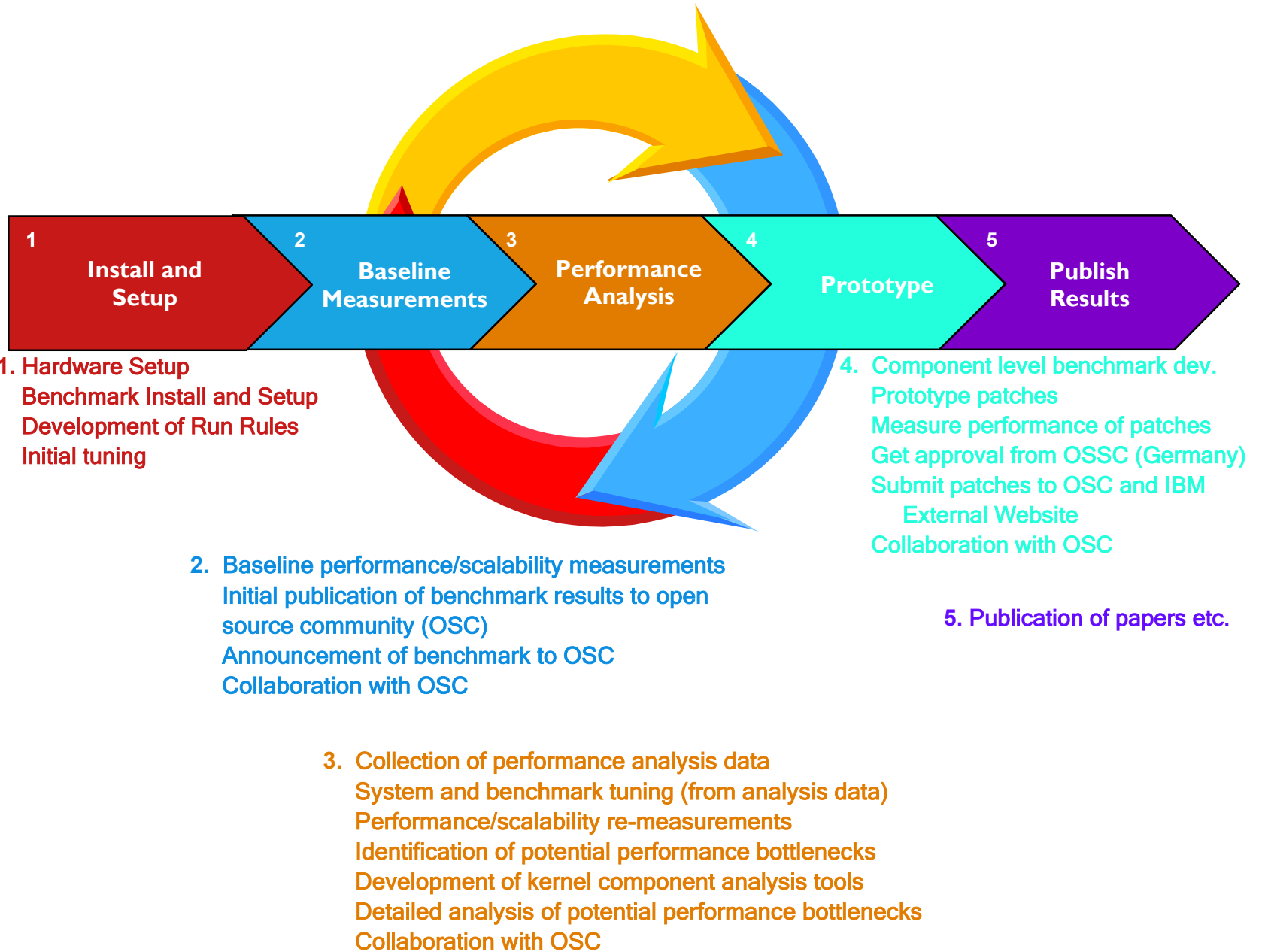
- ▶ Work items may change based on IBM strategy and acceptance from the open source community
- ▶ Work items may change as measurement results unfold and/or hardware requirements increase
- ▶ Baseline measurements currently on Linux 2.4 and 2.5 kernel.org

IBM

Linux Technology
Center



Benchmark Activities



Performance Tools



■ Readprofile

- ▶ Build into the kernel; kernel functions and idle; profile=2 in boot manager

■ Kernprof (SGI)

- ▶ patch needed; kernel functions, idle, user (time only); annotated call graph

■ Lockmeter (SGI)

- ▶ patch needed; examine lock contention, identify hot locks

■ Resource and performance monitoring

- ▶ vmstat, iostat, sar, top, /proc/meminfo, /proc/slabinfo, ...

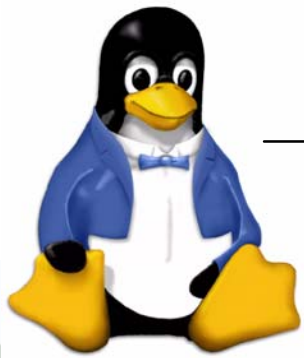
■ Disk statistics

- ▶ modify query plans or redesign the database layout

The IBM logo, consisting of the letters 'IBM' in a stylized, striped font.

Linux Technology
Center

Linux Kernel Focus



- Base kernel
 - ▶ Scheduler
 - ▶ Memory management (large memory support, large page support, page cache)
- File systems
 - ▶ Local file systems: EXT2, EXT3, JFS, ReiserFS
 - ▶ Network file systems: NFS, SMB
- I/O subsystems
 - ▶ Async I/O, Direct I/O
 - ▶ Device drivers
- Network manager
 - ▶ 100 Mbps and 1000 Mbps Ethernet
 - ▶ TCP/IP
- Non-Uniform Memory Access (NUMA)
- Performance focus on 16-way systems

The IBM logo, consisting of the letters 'IBM' in a stylized, striped font, is located in the bottom left corner of the slide.

IBM

Linux Technology
Center

Linux 2.5 Performance Line Items



■ Kernel

- ▶ O(1) scheduler for improving scalability
- ▶ Preemptible kernel support
- ▶ Pagetables in highmem support
- ▶ Large page table support
- ▶ New VM with reverse mappings
- ▶ Syscall interface for CPU task affinity
- ▶ Fastwalk dcache
- ▶ Futexes – Fast Lightweight Userspace Semaphore
- ▶ Read-Copy Update (RCU) mutual exclusion
- ▶ Remove long-held locks for low scheduling latency
- ▶ POSIX threading support for signals
- ▶ Faster system calls

IBM

Linux Technology
Center



Linux 2.5 Performance Line Items *(cont'd)*

■ I/O

- ▶ Rewrite of the block IO layer
- ▶ New IO scheduler
- ▶ Asynchronous IO (aio) support
- ▶ Better IO performance with epoll
- ▶ Zerocopy NFS
- ▶ Wider range of major/minor device numbers to support more disk devices

■ NUMA

- ▶ NUMA aware scheduler extensions
- ▶ Discontigmem support
- ▶ Parallelizing page replacement

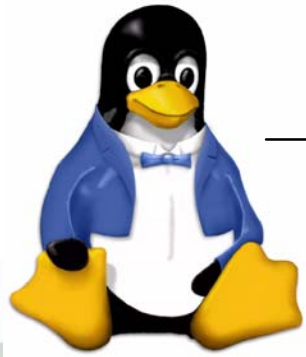


Base Kernel

- 2.5 is equal or faster in an SMP test environment
- SUT: 4-way SMP 500 MHz, 2.5 GB RAM,
- Stock kernels, LMBench 3.0 benchmark, time in microseconds

Linux Kernels	2.2.23	2.4.20	2.5.59
Simple syscall	0.69	0.71	0.69
Simple read	1.13	1.36	1.37
Simple write	1.01	1.14	1.18
Simple stat	6.11	8.35	7.74
Simple fstat	1.31	1.53	1.70
Simple open/close	7.64	10.55	9.92
Pipe latency	9.88	14.36	13.46
Process fork+exit	469.67	555.13	586.03
Process fork+execve	1824.67	1957.00	2146.56
Process fork+/bin/sh -c	67206.67	8859.67	9175.00
UDP latency using localhost	63.52	50.40	47.96
TCP latency using localhost	87.87	64.90	67.77
RPC/udp latency using localhost	104.58	91.86	93.59
RPC/tcp latency using localhost	130.31	116.03	119.48
TCP/IP connection cost to localhost	182.83	129.92	207.86

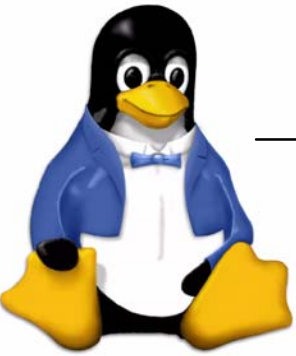
Web Serving



- Linux is traditionally strong in this area
- Typically 4-way SMP, horizontal scaling (clustering)
- TUX and Zeus perform better than Apache in general
 - ▶ Apache is most widely used web servers
- More improvements are required for larger SMP machines
- Benchmarks: SPECweb99, netperf3
- Performance Tips
 - ▶ Tuning of the web servers ; need large memory, many file handles
 - ▶ Tuning of the network stack
 - ▶ Take advantage of the network card capability
 - TSO (TCP Segmentation Offload)
 - Rx and Tx interrupt delay

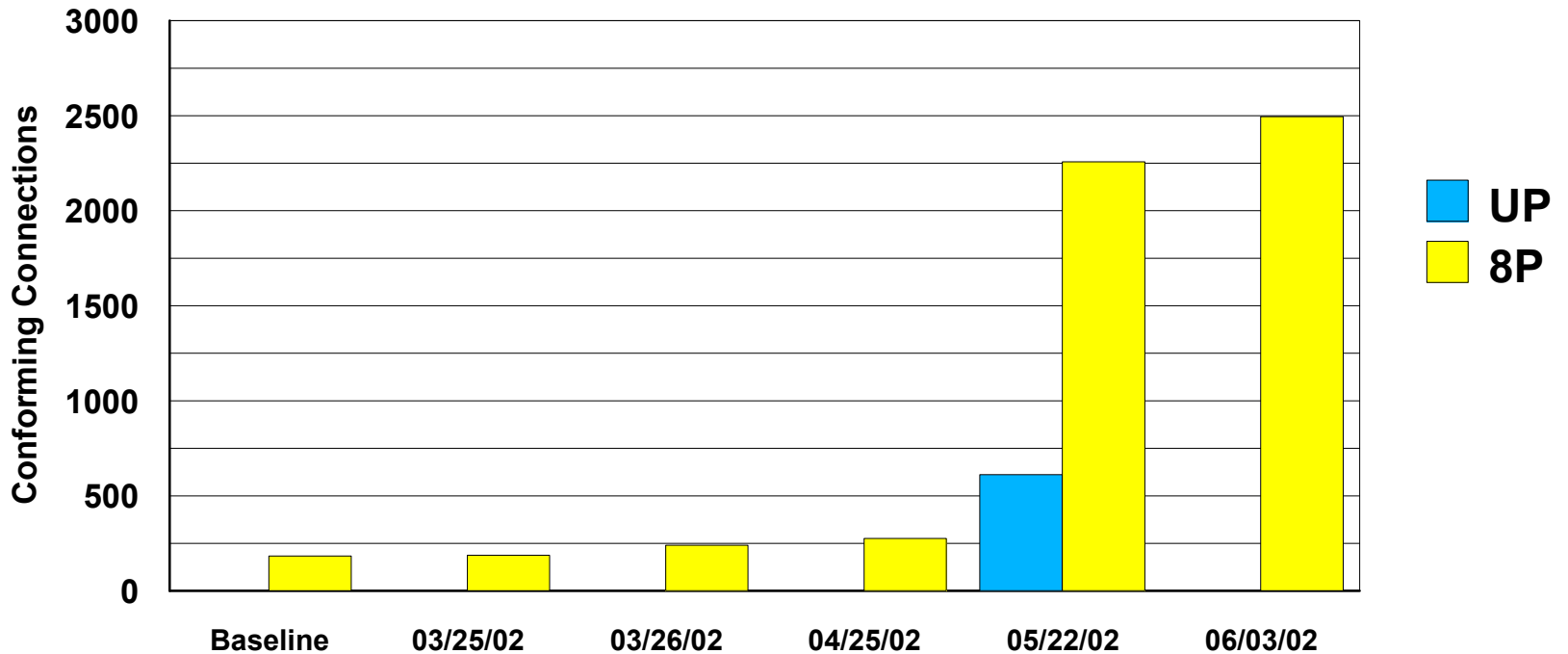
The IBM logo, consisting of the letters 'IBM' in a stylized, striped font.

Linux Technology
Center



SPECweb99 Results **

SPECweb99 8-way Performance with Apache Dynamic and Static Content



Intel 8-Way 900 MHz Pentium III with 2MB L2 Cache, 32 GB RAM, (4) Gb Ethernet
Red Hat 7.1, Linux Kernel 2.4.17, Apache 2.0.36, SPECweb99 1.02

** SPEC™ and the benchmark name SPECweb are registered trademarks of the Standard Performance Evaluation Corporation. The benchmarking was conducted for research purposes only and were non-compliant with the following deviations from the rules:

1. Executed on hardware that does not meet the SPEC availability-to-the public criteria.
2. access_log wasn't kept for full accounting. It was written, but deleted every 200 seconds.

For the latest SPECweb99 benchmark results, visit <http://www.spec.org>.

IBM

Linux Technology
Center

Database

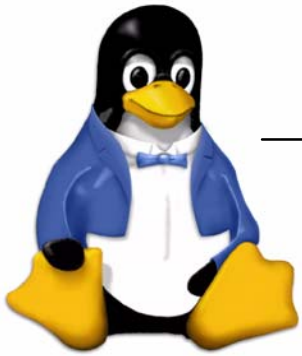
- DB2®/Oracle/Sybase
- Typically 8-way SMP or higher, vertically scaled
- Decision support systems and online transaction processing
- Block I/O has been improved significantly
 - ▶ Bounce buffer elimination, per-queue lock, large I/O block size, efficient vector I/O, lightweight kiobufs; direct I/O on files
- Async I/O (file and network)
- Performance tips
 - ▶ Tuning of the database systems, designing of the database layout
 - ▶ Avoid using bounce buffers
 - **With 4GB memory, make sure the controller and driver can do 32-bit DMA**
 - **With > 4GB memory, make sure the controller and driver can do 64-bit DMA**
 - ▶ Per-queue lock for each device
 - ▶ Large I/O block size



IBM

Linux Technology
Center

File Serving



■ Samba

- ▶ Very popular on Linux, competitive replacement for Windows & OS/2® file & print servers
- ▶ Currently scales to 4 way systems
- ▶ Performance tips
 - Use Samba 2.2.7 or later, including sendfile() API
 - Use lots of system memory for file caching
 - Gigabit Ethernet tuning including interrupt coalescing
 - Tune with these tools: [dbench](#), [tbench](#), and [smbtorture](#)
<http://samba.org/ftp/unpacked/dbench/README>
 - P4 hyperthreading: 25% throughput improvement with 1 CPU, and 5% improvement with 4 CPUs

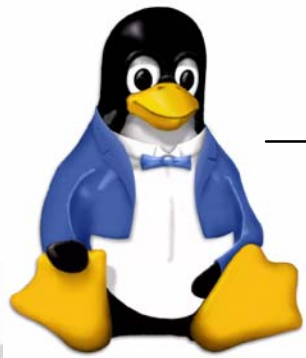
■ NFS - performance analysis underway

- ▶ Current resource for Linux NFS: <http://nfs.sf.net>

The IBM logo is located in the bottom left corner. It consists of the letters 'IBM' in a bold, sans-serif font, with horizontal lines through the letters.

Linux Technology
Center

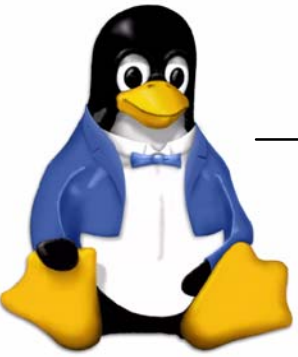
File Systems



- EXT2, EXT3, JFS, ReiserFS
- Focus on JFS performance
- In general, JFS performance is as good as ReiserFS and EXT3
 - ▶ Patch 1: allocation fragmentation problems on multiple I/Os
 - ▶ Patch 2: pjpw (partial journal page write) - don't write to the disk until the transaction buffer is full
- Benchmarks: tiobench, iozone, mongo, postmark
- Async I/O, direct I/O
- Performance tips
 - ▶ Read performance under all journaling file systems are about the same
 - ▶ JFS performs best for large files, 4KB block size, e.g., 100 MB (write)
 - ▶ ReiserFS performs best for small files, e.g., less than 4 KB (write)
 - ▶ Short directory searches resulted in better performance

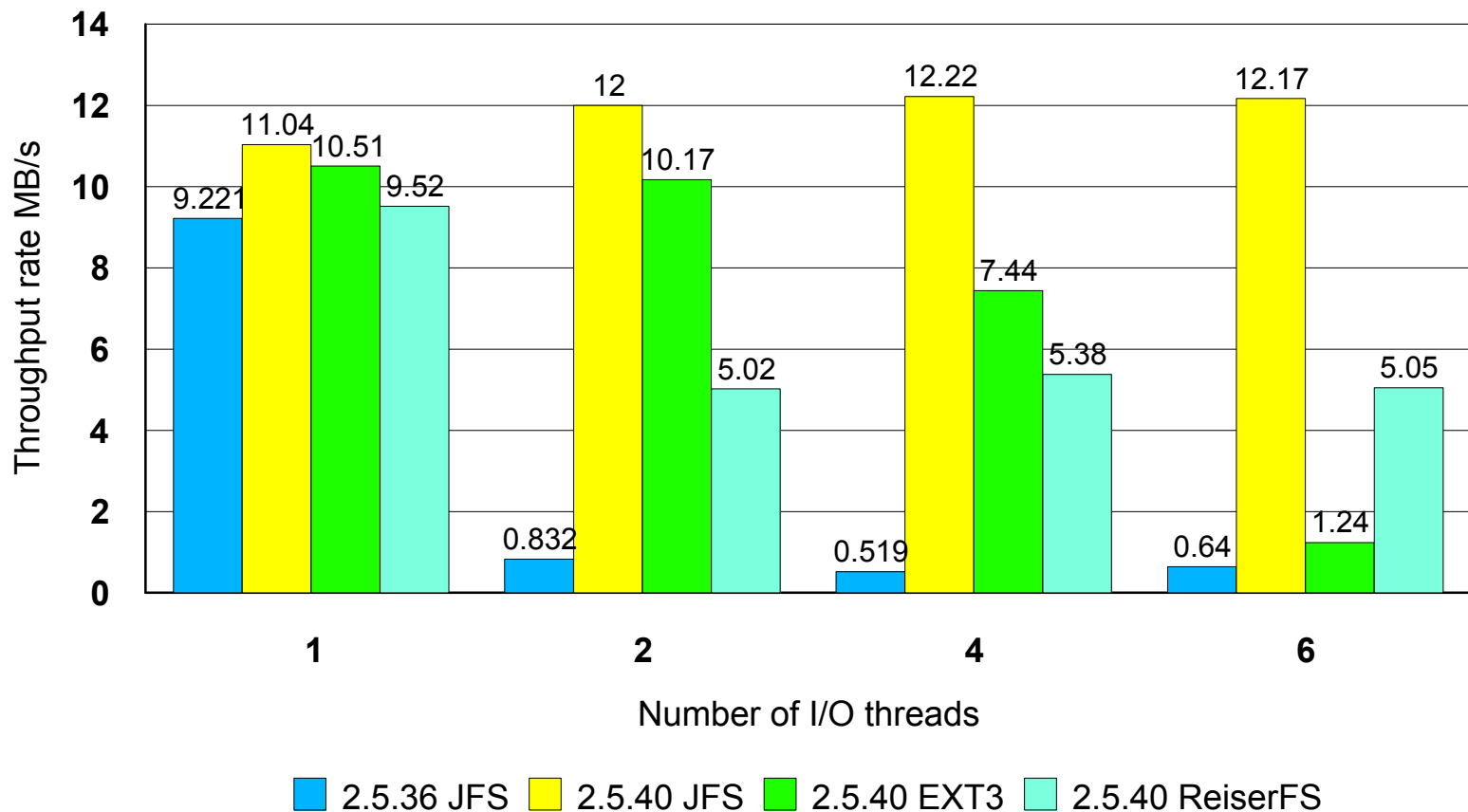
The IBM logo, consisting of the letters 'IBM' in a stylized, striped font.

Linux Technology
Center



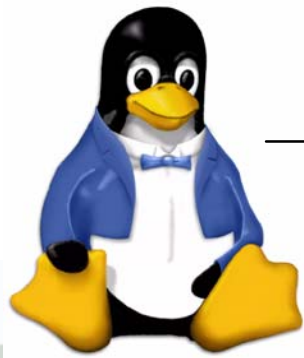
File System Performance

File System Performance - Sequential Write tiobench benchmark - 120 MB file, 4KB block 4-way 500MHz 2.5GB RAM stock kernels 2.5.40 and 2.5.36



2.5.40JFS has dbAlloc patch integrated
2.5.36 is the base kernel, before dbAlloc patch

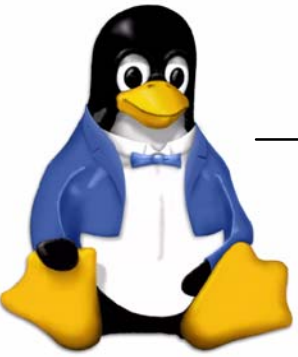
Java-based Applications



- Benchmarks
 - ▶ VolanoMark (loopback and network modes)
 - It has been used to expose scheduler limitation
 - Help drive O(1) scheduler to the kernel
 - ▶ SPECjbb2000
 - 95+% spent on user level code, no network, no I/O
 - ▶ Trade2/Trade3, SPECjAppServer
 - DB2, WebSphere® Application Server
- Compare different JVMs: IBM JVM, BEA JRockit JVM, Sun JVM
- Evaluate different thread models
 - ▶ linuxthreads
 - ▶ NGPT (Next Generation POSIX Threading)
 - ▶ NPTL (Native POSIX Thread Library)
- Performance tip
 - ▶ Use large heap size to reduce frequency of garbage collection

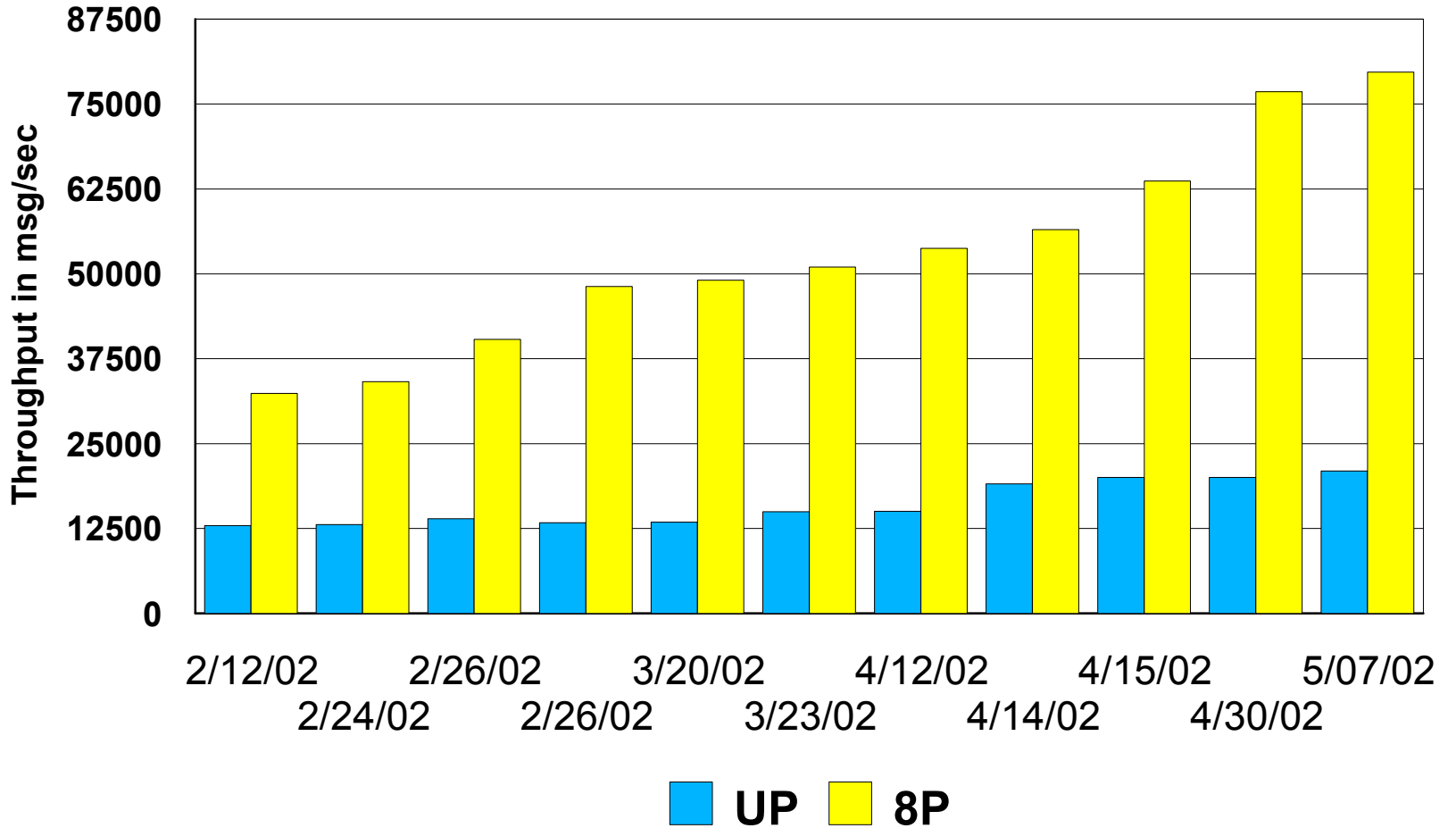
The IBM logo, consisting of the letters 'IBM' in a stylized, bold font.

Linux Technology
Center

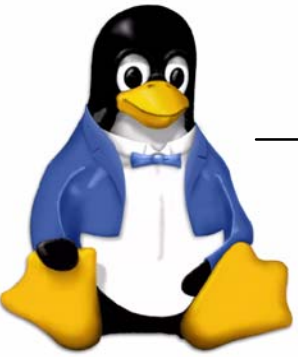


VolanoMark Results

VolanoMark Loopback

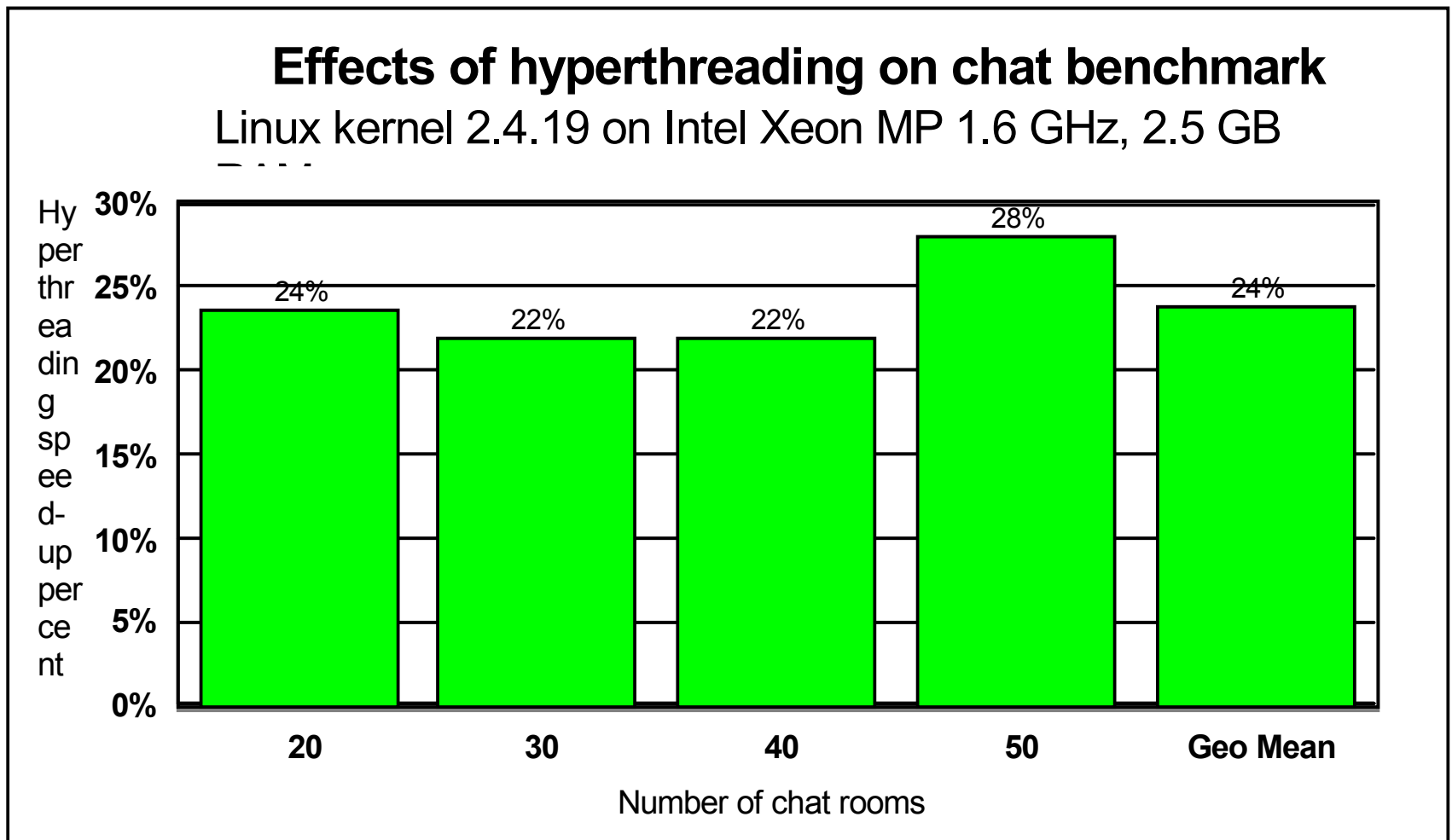


Netfinity 8500R, 8-Way 700 MHz Pentium III with 2MB L2 Cache, 4 GB RAM
Red Hat 7.1, Linux Kernel 2.4.17 + Patches + Tuning

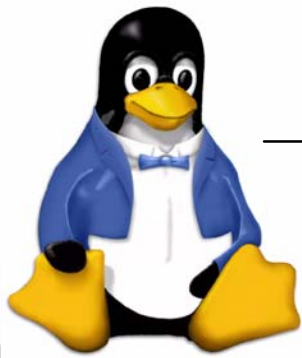


Hyperthreading

- No performance impact on base kernel functions
- Great performance impact on multithreaded workloads



LTC Performance Team Contact



- **Manager:** Ray Venditti - @us.ibm.com
- **Tech Lead:** Bill Hartner - bhartner@us.ibm.com
- **Database/Transaction Processing/Decision Support:** Peter Wong - wpeter@us.ibm.com, Ruth Forester - rsf@us.ibm.com, Mike Sullivan - mksully@us.ibm.com
- **Networking Performance:** Barry Arndt - barndt@us.ibm.com
- **Web Serving:** Troy Wilson - wilson@us.ibm.com
- **NFS / File Serving:** Duc Vianney - dvianney@us.ibm.com
- **File systems / I/O:** Steve Pratt - splratt@us.ibm.com
- **I/O Benchmarks:** Ben Rafanello - benr@us.ibm.com
- **Virtual Memory Performance** - Mark Peloquin - peloquin@us.ibm.com
- **NUMA:** Andrew Theurer - atheurer@us.ibm.com
- **Java Benchmarks:** Mala Anand - manand@us.ibm.com, Partha Narayanan - partha@us.ibm.com
- <http://oss.software.ibm.com/developerworks/opensource/linuxperf>
- <http://oss.software.ibm.com/developerworks/projects/linuxperf>
- ★ **Performance Architect:** Sandra K Johnson - sandraja@us.ibm.com