Linux for z/Series Performance
Overview

Klaus Bergmann
IBM Linux for z/Series Lab, Boeblingen, Germany
Klaus_Bergmann@de.ibm.com
3/4/2002
Agenda

- Relative system capacity
- z900 Hardware
- Scalability
- DASD I/O
- Networking
- Crypto
- Linux under VM
Relative System Capacity

- Components of Capacity
  - Processor
  - Memory Hierarchy
  - I/O Structure

- Processor, Memory and I/O time vary greatly with application

- Processor designs vary greatly in the balance of capacity across components which results in a wide range of relative capacity.
Relative System Capacity

- Data Intense work such as BI, Very Large Data Base, Classical OLTP or "cache killer" workloads (Object oriented code or context switching) will potentially run much better on S/390.

- CPU Intense work such as SPECint, Deep Computing, Graphic Rendering, will perform relatively poorly on S/390.
Relative System Capacity

- S/390 is not for "Deep Computing"
- Most commercial applications range between CPU and Data intensity
- z/ Series is very good at "data intense" work for many users.
z900 Systemstructure: Optimized for maximum external bandwidth

Memory (up to 2 x 16 GB)

Level 2 Cache (16 MB)

Tape, DASD, Network:
- 16 Port ESCON
- 2 Port FICON (ESS!)
- 2 Port GbE

6 x 1 GB/s
From 2064-10x (z900) to 2066-yyy (z800)

- 12 PU Chips @ 1.3 ns
- 2 SAP's, 1 spare
- Up to 9 CP's
- Up to 8(9) ICF's/IFL's

- 5 PU Chips @ 1.6 ns
- 1 SAP
- Up to 4 CP's
- Up to 4 ICF's/IFL's
  - Linux Only version (OLF), CF
  - 16 IO-Slots (z900 Cards)
- New Pricing-Models
CMOS 8SE and 8S Technologie
- Internal Copper-wiring

10 Chips
- 5 CMOS 8SE
- 3 CMOS 8S
- 2 CMOS 7S

MCM Packaging
- 71 mm x 71 mm
- 5 Processor Units (PUs)
  - 17.9 mm x 9.9 mm
  - 44M transistors
  - L1 cache/CP
    - 256 KB I-cache
    - 256 KB D-cache
  - 1.6 ns Cycle Time
- 2 System Data (SD) Cache Chips
  - L2 Cache
    - 234M transistors
    - 4 MB/Chip, 8MB/Module
- 1 Storage Control (SC) Chip
- 1 Memory Bus Adapter (MBA) Chip
- 1 Clock (CLK) chip (7S)
- Glass-Ceramics (42 layers), Thin Film
z800 Model 0LF (announced 01/29/2002)

1 - 4 z800 Integrated Facility for Linux (IFL) Processors

- 4 Standard-Configurations:

<table>
<thead>
<tr>
<th>Feature Nr.</th>
<th>CP's (IFL's)</th>
<th>SAP's</th>
<th>Spare</th>
<th>Memory (GB)</th>
<th>ESCON Channels</th>
<th>OSA-E Channels</th>
<th>FICON Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>3605</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>28</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3606</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>28</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3607</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>28</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3608</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>28</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

- 16 I/O Slots
  - 16-Port ESCON channels, max. 240 ports
  - FICON-Express channels, max. 32 ports
  - OSA-Express, max. 24 ports
  - PCICA, max. 12 ports

- 3 Years Hardware Support
  - 1 year warranty + 2 years maintenance

z/VM Version 4
- 3 years support
Relative Performance: G6 vs z900

- **G6 Turbo**
- **z900 2-bus (2064-10x)**
- **z900 4-bus (2064-1Cx,11x)**
z800 Performance

- Rx5 (G4)
- RA..D6 (G5)
- z800-0C1
- z800-0B1
- z800-0C1/A2
- Xx7 (G6)
- z800-00x

CP's: 1, 2, 3, 4
Our Hardware for measurements

2064-116 (z900)
- 1.3ns (770MHz)
- 2 * 16 MB L2 Cache (shared)
- 64 GB
- LPAR
- ESCON
- FICON
- HiperSockets
- OSA Express GbE

2105-F20 (Shark)
- 384 MB NVS
- 16 GB Cache
- 128 * 36 GB disks
- 7200 RPM
- 4 FCP (1 port)
- 6 ESCON (2 port)
- 4 FICON (1 port)

8681-7RY (8-way Netfinity)
- Pentium III, 700 MHz
- 8 * 1 MB L2 Cache (private)
- 4.5 GB
- 2 * 37 GB SCSI
But isn't Linux SMP scalability limited to a 4 way??

....well maybe on some platforms but not from what we are seeing on S/390

WebSphere

Business Intelligence
Scalability: file system benchmark

netfinity 8-way, ext2, kernel 2.4.14

z900, 16-way LPAR, ext2, 31 bit, kernel 2.4.14

z900, 16-way LPAR, ext2, kernel 2.2.16

Scalability

Throughput [MB/sec]

# processes

Throughput [MB/sec]

# processes

Scaling factor

# CPUs

netfinity, kernel 2.4.14
z900, kernel 2.2.16
z900, kernel 2.4.14
Scalability, Webbased benchmark

- Number of CPUs: 1, 2, 4, 8, 12, 16
- Connected Clients: better

- 31 Bit Kernel 2.4.3
- 31 Bit Kernel 2.4.7
- 64 Bit Kernel 2.4.7
context switching

Netfinity 8-Way

z900

kernel 2.4

# processes

microseconds

0K

4K

8K

16K

32K

64K

0

50

100

150

200

250

microseconds

0

50

100

150

200

250

microseconds
DASD I/O

- ESCON
- FICON (Express)
ESCON, FICON (Express)

**ESCON**
- 17 MB half duplex
- 3 km w/o repeaters
- 43 km w/ repeaters
- data rate droop at 9km
- separate CTC function

**FICON (Express)**
- (> 100 MB full duplex
- 20 km w/o repeaters
- 100 km w/ repeaters
- no data rate droop
- integrated CTC function
- consolidation of 4-8 ESCON channels to 1 FICON channel

FICON and FICON Express Channel Performance
White Paper GM13-0120-00
ESCON vs. FICON

Single DASD

<table>
<thead>
<tr>
<th></th>
<th>Write</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCON</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>FICON</td>
<td>15</td>
<td>25</td>
</tr>
</tbody>
</table>

MB/sec
Sequential DASD I/O

4 FICON

MB/sec

# of disks

Write
Read
DASD access translations
today:

Linux buffer cache

DASD driver

FICON

ESS

DASD

color coding:

block format

translation block <-> ECKD

ECKD format
DASD access w/o translations

today

- Linux buffer cache
- DASD driver
- FICON
- ESS
- DASD

we are working on:

- Linux buffer cache
- SCSI driver
- OpenFCP
- ESS
- DASD
Networking

- OSA Express
- HiperSockets
Features for z800, z900, and S/390 G5, G6
- Gigabit Ethernet LX (long wavelength) full duplex, 50/62.5 micron multimode (with mode conditioning patch cables) or 9 micron single mode
- Gigabit Ethernet SX (short wavelength) full duplex, 50/62.5 micron multimode
- Fast Ethernet (10 or 100 Mbps full duplex)
- 155 ATM Single Mode (9 micron)
- 155 ATM Multimode (62.5 micron)

Feature for z800, z900
- Token Ring (4, 16, or 100 Mbps full duplex)

Up to 12 OSA-Express Features per System
- S/390 features have one port, z800, z900 features have 2 ports
- independent of OSA-2

Direct attach to Self-Timed Interconnect (STI) bus
- S/390 STI bus supports up to 333 megabytes/second
- z900 STI bus supports up to 1 gigabyte/second

Port attaches to:
- LAN (full-duplex support if connected to a switch) or ATM network
- Direct connected workstation (or server)
- Multiple LPARs (port sharing)
HiperSockets Hardware Elements
('Network in a box')

- Synchronous data movement between LPARs and virtual servers within a zSeries server
  - Provides up to 4 "internal LANs". HiperSockets accessible by all LPARs and virtual servers
  - Up to 1024 devices (TCP/IP stacks) across all 4 HiperSockets
  - Up to 4000 IP addresses
  - Similar to cross-address-space memory move using memory bus
- Extends OSA-Express QDIO support
  - LAN media and IP layer functionality (internal QDIO = iQDIO)
  - Enhanced Signal Adapter (SIGA) instruction
    - New "thin interrupt" without use of System Assist Processor (SAP)
    - Optional dispatcher polling mechanism
- HiperSockets Hardware I/O Configuration with new CHPID type = IQD
  - Controlled like regular CHPID
  - Each CHPID has configurable Maximum Frame Size
- Works with both standard and IFL CPs
- No physical media constraint, no physical cabling, no priority queueing
- Secure connections
- Both 31 bit and 64 bit operating systems supported
- Pre-req: IBM eServer zSeries 900 Licensed Internal Code (LIC)
  Update
HiperSockets Performance (Interactive Transactions)
Linux to Linux

RR 1/1 Response Time Summary

RR Transaction Rate
200 Bytes out; 1000 Bytes in

RedHat 7.2, kernel 2.4.9, 31-bit
OCO-modules from developerworks (1/2002)
Two LPARs, each w/ 4 non-dedicated CPUs and 2 GB
OSA-Express GbE card shared between LPARs
Default TCP Send/Receive buffer size (64 KB)
MTU: 9000 (GbE), 32K (HiperSockets)
HiperSockets Performance (Bulk Data Xfer)
Linux to Linux

GbE / HiperSockets
Stream (Put) throughput summary

RedHat 7.2, kernel 2.4.9, 31-bit
OCO-modules from developerworks (1/2002)
Two LPARs, each w/ 4 non-dedicated CPUs and 2 GB
OSA-Express GbE card shared between LPARs
20 MB out, 20 Bytes in
Default TCP Send/Receive buffer size (64 KB)
MTU: 9000 (GbE), 32K (HiperSockets)
Crypto performance

- Linux SSL-RC4
- VM Linux SSL-RC4
ITR (ETR + idle) for LINUX SSL-RC4 MD5 US
Non-Cached, 1024 bit keys, 2*2048 bytes
2064-116 by # processors and PCICAs
ITR (ETR + idle) for VM LINUX SSL-RC4 MD5 US
Non-Cached, 1024 bit keys, 2*2048 bytes
2064-104, 1 PCICA by # of LINUX Guest

Native
LINUX GbE
VMTCIP
Linux 'jiffies'\

- 100 Hz timer 'wakes up' kernel
- Linux uses Clock Comparator
- timer count stored in variable named 'jiffies'
- to VM, the guest is always busy
- affects VM's paging
- overhead for idle Linux guest: 0.3% of one G5 CPU
The Timer Patch

- Recommended Patch on DeveloperWorks
- Reduces overhead for idle guest close to zero
- Interrupts only when necessary
- Clock Comparator for absolute timer events
- CPU Timer for process related events
- 'jiffies' is checked on every system entry