IBM eServer zSeries

IBM z/VSE 3.1 und SCSI Performance Update

Ingo Franzki (ifranzki@de.ibm.com) 4/25/2005

Trademarks

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

CICS*
DB2*
DB2 Connect
DB2 Universal Database
e-business logo*
Enterprise Storage Server
HiperSockets
IBM*
IBM logo*
IMS
Intelligent Miner
Multiprise*
MQSeries*
OS/390*
S/390*
SNAP/SHOT*
Virtual Image Facility
VM/ESA*
VSE/ESA
VisualAge*
VTAM*
WebSphere*
xSeries
z/Architecture
z/VM
z/VSE
zSeries

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

LINUX is a registered trademark of Linus Torvalds

Tivoli is a trademark of Tivoli Systems Inc.

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.

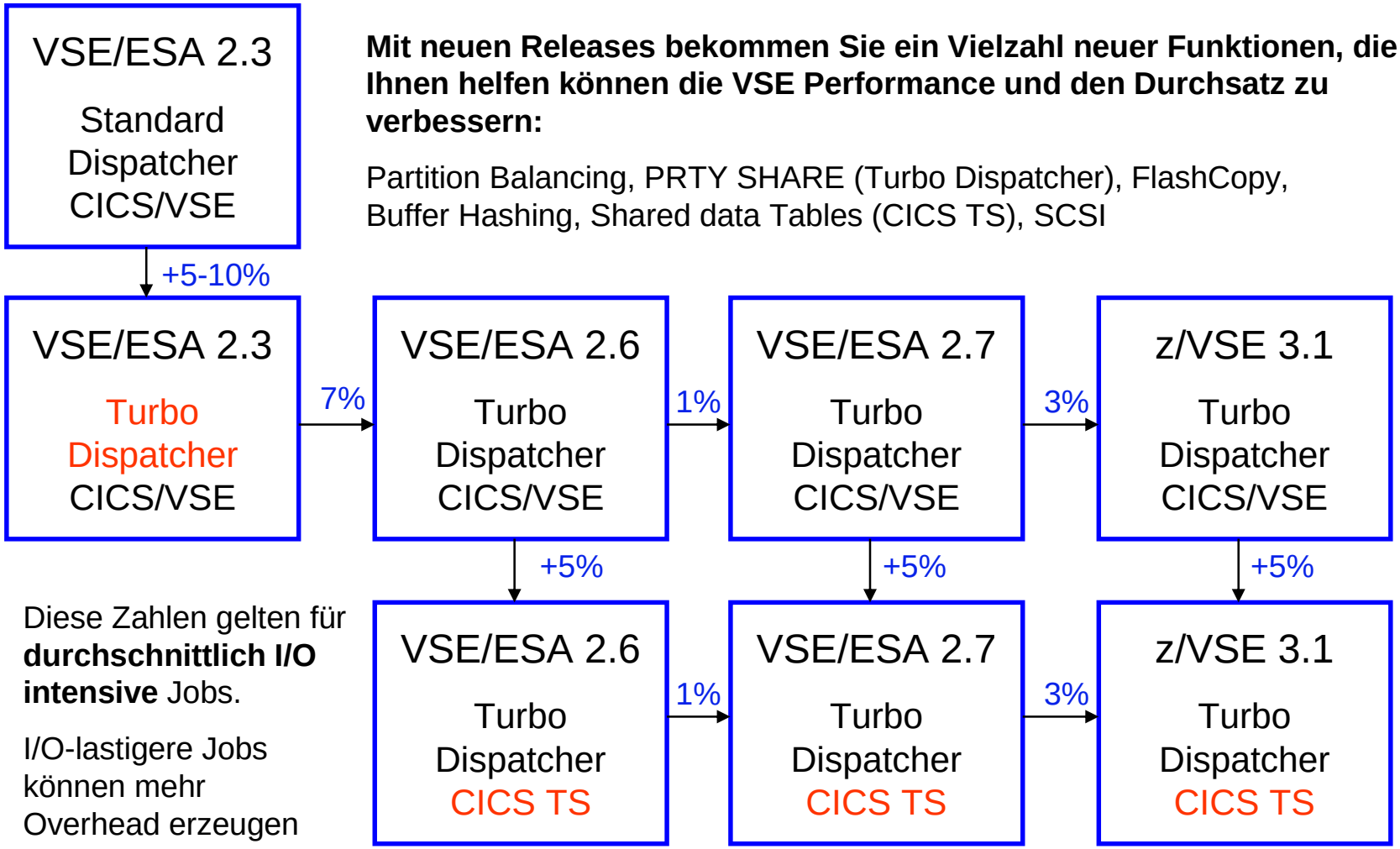
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

Intel is a registered trademark of Intel Corporation.

Agenda

- **VSE Release Overhead**
- **SCSI mit z/VSE**
- **z/VSE Hardware Unterstützung**
- **z890 und z990**
- **z/VM 5.1**
- **Hints und Tipps**

Overhead Deltas für VSE Releases



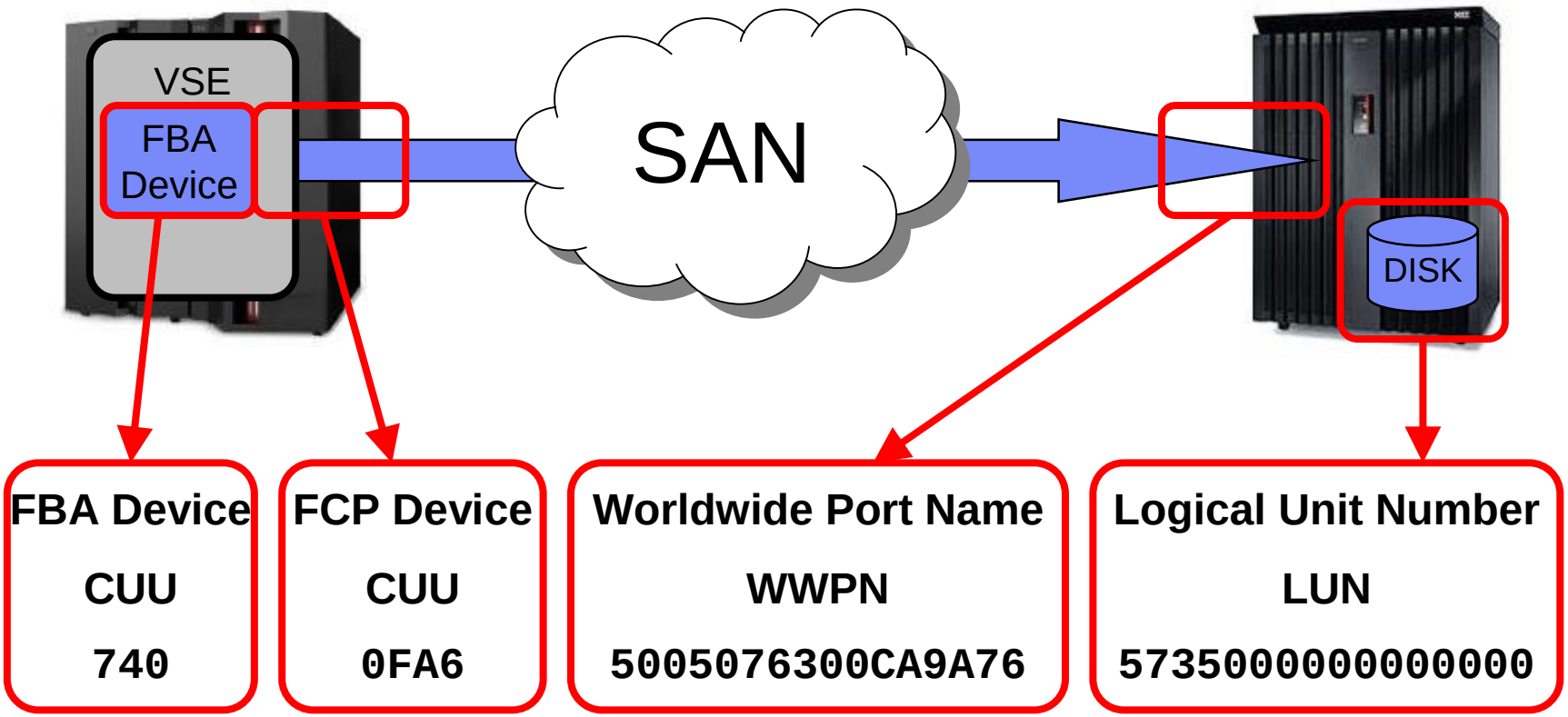
Mit neuen Releases bekommen Sie ein Vielzahl neuer Funktionen, die Ihnen helfen können die VSE Performance und den Durchsatz zu verbessern:

Partition Balancing, PRTY SHARE (Turbo Dispatcher), FlashCopy, Buffer Hashing, Shared data Tables (CICS TS), SCSI

Diese Zahlen gelten für **durchschnittlich I/O intensive** Jobs.

I/O-lastigere Jobs können mehr Overhead erzeugen

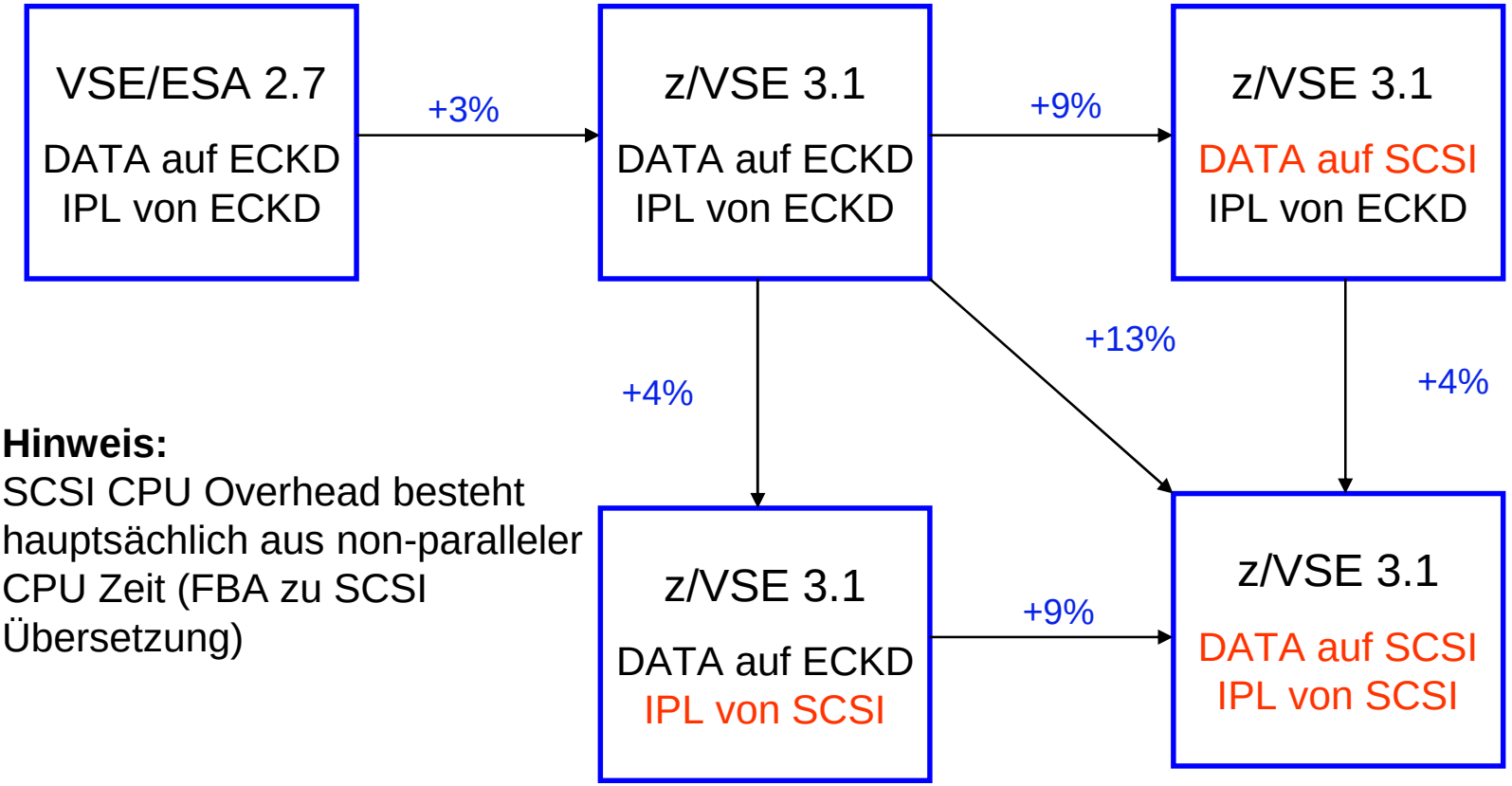
SCSI Adressierung im VSE



SCSI Setup für z/VSE (Beispiel)

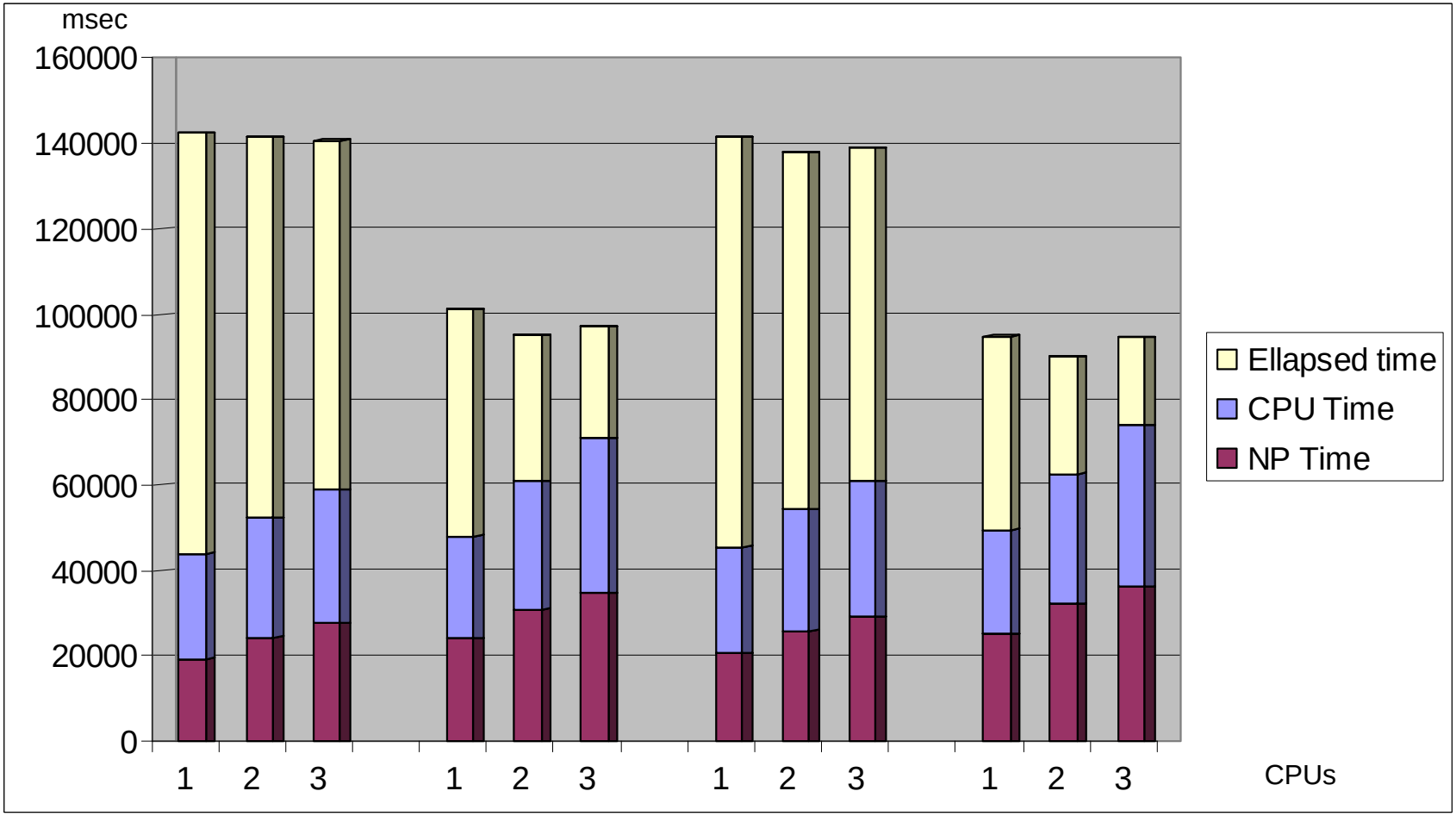
- **FCP Devices:**
  - ADD 4A7:4A9,FCP
- **FBA Devices:**
  - ADD 608:61B,FBA
- **Define SCSI:**
  - DEF SCSI,FBA=608,FCP=4A7,WWPN=5005076300CA9A76,LUN=5710
  - DEF SCSI,FBA=609,FCP=4A7,WWPN=5005076300CA9A76,LUN=5711
  - ...
- **IPL von SCSI (VM)**
  - Minimum 32M Memory
  - SET LOADDEV PORT 50050763 00CE9A76 LUN 57350000 00000000
  - IPL 4B8 (IPL von FCP device)

Overhead Deltas für SCSI



Hinweis:
SCSI CPU Overhead besteht hauptsächlich aus non-paralleler CPU Zeit (FBA zu SCSI Übersetzung)

SCSI Overhead



SCSI I/O Counts

- **Für PACEX16 Workload**
  - ECKD:
    - 18000 ECKD I/Os pro Disk
  - SCSI:
    - 20000 FBA I/Os pro Disk
  - ECKD → FBA:
    - 11% höhere I/O Counts
- **Generell wird 1 FBA I/O in 1 SCSI I/O übersetzt**
- **Ausnahmen:**
  - Lange CCW-Ketten
  - Überlappende Adressen (z.B. PHASE laden)

SCSI Multipathing

- **Ein oder mehrere alternative Pfade zur selben SCSI Disk**
  - Erhöht die Verfügbarkeit
  - Nicht: Workload-Balancing
- **Jeder Pfad muss über einen anderen FCP Adapter definiert sein**
  - Eine FCP Karte kann mehrere FCP Adapter (CHPID) haben
  - Um die Ausfallsicherheit zu erhöhen, sollte man FCP Adapter auf verschiedenen physikalischen FCP Karten verwenden
- **Am besten auch über unterschiedliche Switches und/oder Ports**
- **Beispiel:**

```
DEF SCSI,FBA=DA1,FCP=FA0,WWPN=5005076300CA9A76,LUN=5600
DEF SCSI,FBA=DA1,FCP=FB0,WWPN=5005076300C29A76,LUN=5600
```

- **QUERY SCSI**

```
AR 0015 FBA-CUU FCP-CUU WORLDWIDE PORTNAME LOGICAL UNIT NUMBER
AR 0015 DA1     FA0     5005076300CA9A76   5600000000000000
AR 0015 DA1     FB0     5005076300C29A76   5600000000000000
```

  - Der zuerst gezeigte Pfad wird derzeit verwendet um auf die SCSI Platte zuzugreifen

Vergleich: ESCON – FICON/FCP

	ESCON	FICON/FCP
Max # of channels	256	4 x 256
Max # device addresses per link	4096	65536
Max # logical CU-paths per port	64	256
Device addresses per channel	1024	16384
Link rate	20 MB/sec	200 MB/sec
Max achievable transfer rate	17 MB/sec	170 MB/sec
Full duplex	No	Yes
Concurrent I/O operations	1	Up to 32
CCW execution	Synchrony	Asynchrony (FICON)

Wann sollte man SCSI (nicht) benutzen?

- **Wann sollte man SCSI benutzen?**
  - Wenn genug CPU Power verfügbar ist, um den zusätzlichen SCSI Overhead abzuarbeiten
  - SCSI Overhead ist fast nur non-paralleler Code.
- **Wann sollte man SCSI NICHT benutzen?**
  - Wenn man schon bei 80-100 % CPU ist
    - Wenn Sie heute schon bei 80 % CPU sind, würde der SCSI Overhead das System auf 100 % CPU bringen

Migration von ECKD nach SCSI

- **FSU von ECKD zu SCSI nicht möglich**
  - Base Installation nötig
- **Datei-Allokationen müssen angepasst werden**
  - Umrechnen von Tracks/Cylinder zu Blocks (für 3390):
    - 1 Track = ca. 112 Blocks
    - 1 Cylinder = ca. 1680 Blocks
  - VSAM Space
    - Tipp: Cluster-Größe in RECORDS angeben, nicht in Tracks
  - Sequentielle Dateien
  - VSE Libraries
    - Tipp: 1 LIBR Block = 1024 Bytes = 2 SCSI Blocks
- **Programme müssen mit FBA Disks umgehen können**
  - Am besten Device-unabhängig implementiert

z/VSE 3.1 Hardware Unterstützung

- **z/VSE 3.1 und VSE/ESA 2.7 läuft auf den folgenden Systemen:**
  - zSeries: z800, z900, z990, z890
  - 9672 Parallel Enterprise Server (G5/G6)
  - Multiprice 3000 (7060)
  - Gleichwertige Emulatoren (Flex-ES)
- **z/VSE 3.1 und VSE/ESA 2.7 basieren auf dem Hardware Instruction-Set beschrieben im Manual 'ESA/390 Principles of Operation' (SA22-7201).**
  - Es wird davon ausgegangen, dass alle diese ESA/390 Instruktionen und Funktionen benutzt werden können.

z/VSE 3.1 Hardware Unterstützung (2)

- **z/VSE 3.1 unterstützt:**
  - IBM eServer zSeries 890 and 990
  - SCSI Disks verbunden über zSeries FCP Kanäle
  - OSA-Express2 und FICON Express2 Adapter
  - Crypto Express2 und CP Assist for Cryptographic Function (CPACF)
  - IBM TotalStorage 3494 Virtual Tape Server
  - IBM 3494 Tape Library
  - IBM TotalStorage DS8000 and DS6000 series Storage Servers
  - IBM TotalStorage Enterprise Storage Server (ESS)

Unterstützte VSE Releases

VSE Release	Verfügbar	End of Marketing	End of Service
z/VSE 3.1	04. 03. 2005		
VSE/ESA 2.7	14. 03. 2003	geplant 3Q2005	
VSE/ESA 2.6	14. 12. 2001	14. 03. 2003 (nicht mehr bestellbar)	31. 03. 2006
VSE/ESA 2.5	29. 09. 2000	14. 12. /2001	31. 12. 2003 (out of service)
VSE/ESA 2.4	25. 06. 1999	29. 09. 2000	30. 06. 2002 (out of service)
VSE/ESA 2.3	12. 07. 1997	30. 06. 2000	31. 12. 2001 (out of service)

VSE Server Unterstützung

IBM zSeries eServer	z/VSE 3.1	VSE/ESA 2.7	VSE/ESA 2.6	VSE/ESA 2.5	VSE/ESA 2.4/2.3
zSeries 890, 990	Ja	Ja	Ja (PTF required)	Ja (PTF required)	Nein
zSeries 800, 900	Ja	Ja	Ja	Ja	Ja
S/390 Parallel Enterprise Server G5/G6	Ja	Ja	Ja	Ja	Ja
S/390 Multiprise 3000	Ja	Ja	Ja	Ja	Ja
S/390 Parallel Enterprise Server G3/G4	Nein	Nein	Ja	Ja	Ja
S/390 Multiprise 2000	Nein	Nein	Ja	Ja	Ja
S/390 Integrated Server	Nein	Nein	Ja	Ja	Ja
S/390 Parallel Enterprise Server G2 / G1 (out of Service)	Nein	Nein	Ja	Ja	Ja
ES/9000 – 9221, 9121, 9021 (out of Service)	Nein	Nein	Ja	Ja	Ja
P/390 and R/390 (out of Service)	Nein	Nein	Ja	Ja	Ja

VSE Hardware Unterstützung

VSE Release	HiperSockets	OSA Express (QDIO mode)	Hardware Crypto
z/VSE 3.1	Ja	Ja	Ja (PCICA, CEX2C, CPACF)
VSE/ESA 2.7	Ja	Ja	Ja (PCICA, CPACF)
VSE/ESA 2.6	Nein	Ja	Nein
VSE/ESA 2.5	Nein	Nein	Nein
VSE/ESA 2.4	Nein	Nein	Nein
VSE/ESA 2.3	Nein	Nein	Nein

PCICA:	PCI Cryptographic Accelerator	verfügbar für z800, z900, z890, z990
CEX2C:	Crypto Express2	verfügbar für z890, z990
CPACF:	CP Assist for Cryptographic Function	verfügbar für z890, z990

zSeries Hinweise

- **Vor-zSeries-Systeme (z.B. G6) hatten den selben Cache für Daten und Instruktionen**
- **zSeries-Systeme haben je einen eigenen Cache für Daten und Instruktionen**
- **Performance Implikationen:**
  - Wenn Programm-Variablen und Code welcher diese Variablen updaten in der selben Cache-Line sind (256 Bytes):
    - Update der Programm-Variable invalidiert auch den Instruktions-Cache
    - Performance Verschlechterung wenn das in einer Schleife sehr oft passiert
    - Siehe APAR PQ66981 für FORTRAN Compiler

zSeries Hinweise - Beispiel

Kein Problem:

```
         LA    R1,PHASNAME   POINT AT PHASE NAME
         CDDELETE (1)
+*       SUPERVISOR - CDDELETE - 5686-032-06
+        CNOP  0,4
+        BAL   15,*+8
+        DC    A(B'00010010')
+        L     15,0(,15)
+        SVC   65            ISSUE SVC FOR CDDELETE
         DS    0H
```

CDDELETE benutzt ein Inline-Flag-Byte, aber es modifiziert es nicht

Möglicherweise ein Problem:

```
         WTO TEXT=DATA
+        CNOP  0,
+        BAL   1,IHB0003A   BRANCH AROUND MESSAGE
+        DC    AL2(8)                TEXT LENGTH
+        DC    B'0000000000010000'   MCSFLAGS
+        DC    AL4(0)              MESSAGE TEXT ADDR
         ...
+IHB0003A DS   0H
+        LR    14,1          FIRST BYTE OF PARM LIST
+        SR    15,15         CLEAR REGISTER 15
+        AH    15,0(1,0)     ADD LENGTH OF TEXT + 4
+        AR    14,15         FIRST BYTE AFTER TEXT
+        LA    15,DATA       LOAD TEXT VALUE
+        ST    15,4(0,1)     STORE ADDR INTO PLIST
+*       SUPERVISOR - SIMSVC - 5686-032
         ...
+        SVC   35            ISSUE SVC 35
 @GE00016 DS   0H
```

WTO benutzt eine Inline-Parameter-Liste, und modifiziert diese

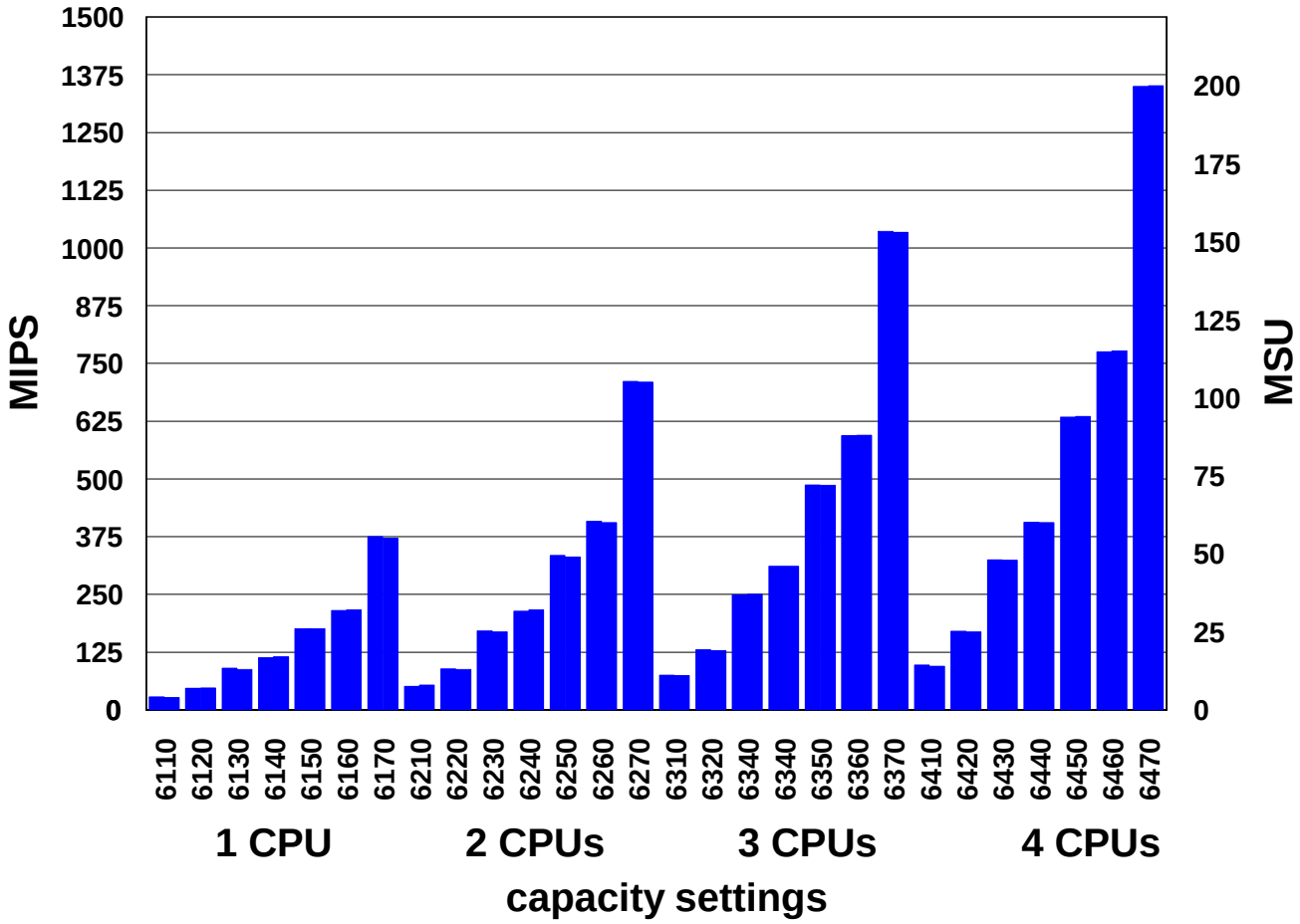
Hinweis: WTO kann so codiert werden, dass es eine externe Parameter-Liste benutzt:

WTO …,MF=(E,addr)

z890 und z990 Betrachtungen

- **Die z890 und z990 Systeme sind LPAR-only Maschinen**
  - Kein Basic-Mode mehr
  - Selbst wenn Sie nur ein VSE System betreiben, läuft es in einer LPAR
  - VSE Systeme unter z/VM bedeutet:
    - VSE in z/VM in einer LPAR
  - Kein I/O Assist in LPARs
    - Nur verfügbar, wenn z/VM im Basic-Mode laufen würde, aber kein Basic-Mode mehr verfügbar auf der z890, z990

IBM eServer zSeries 890



z890 consists of one Model (A04) and 28 capacity settings

z/VM 5.1 Betrachtungen

- **z/VM 5.1 unterstützt keine V=R und V=F Gäste mehr**
- **z/VM 5.1 unterstützt kein I/O Assist mehr**
  - Wenn Sie heute mit Preferred Guests laufen, können sie von einer erhöhten CPU Belastung ausgehen, da diese Gäste dann zu V=V Gästen werden.
  - Siehe auch „Preferred Guest Migration Considerations“ unter http://www.vm.ibm.com/perf/tips/z890.html
- **Vorhersage der CPU Belastung**
  - **Fehlendes I/O Assist:** Lassen Sie Ihre Workload mit CP SET IOASSIST OFF laufen und messen Sie die CPU Belastung
  - **Fehlendes V=R/F:** Lassen Sie Ihre Workload mit V=V laufen, und benutzen Sie den CP Monitor um die CPU Auslastung zu beobachten.
- **Tuning**
  - **Dedicated Processors:** CP SET SHARE ABSOLUTE
  - **Dedicated Memory:** CP SET RESERVED
  - **I/O Assist:** Verwenden Sie Minidisks, setzten Sie Minidisk-Caching an (MDC)

Mögliches Performance-Problem mit PPRC

- **Problem tritt auf wenn**
  - PPRC benutzt wird
  - VSE im Native-Mode oder im LPAR betrieben wird
  - Nicht alle Devices, die im IOCP definiert sind, dem VSE per ADD Statement bekannt gemacht wurden
- **Wenn ein PPRC State Change auftritt, werden Interrupts an alle LPARs geschickt, bei denen das Device im IOCP definiert ist.**
  - Wenn das Device dem VSE bekannt ist (ADD), tritt kein Problem auf: VSE behandelt den Interrupt.
  - Wenn das Device dem VSE NICHT bekannt ist, wird der Interrupt vom VSE nicht behandelt, und der Interrupt wird dem LPAR sehr schnell erneut zugestellt
    - Das resultiert in sehr hoher Channel-Aktivität (bis zu 100%)
- **Lösung**
  - Definieren Sie ALLE Devices per VSE ADD die auch im IOCP definiert sind

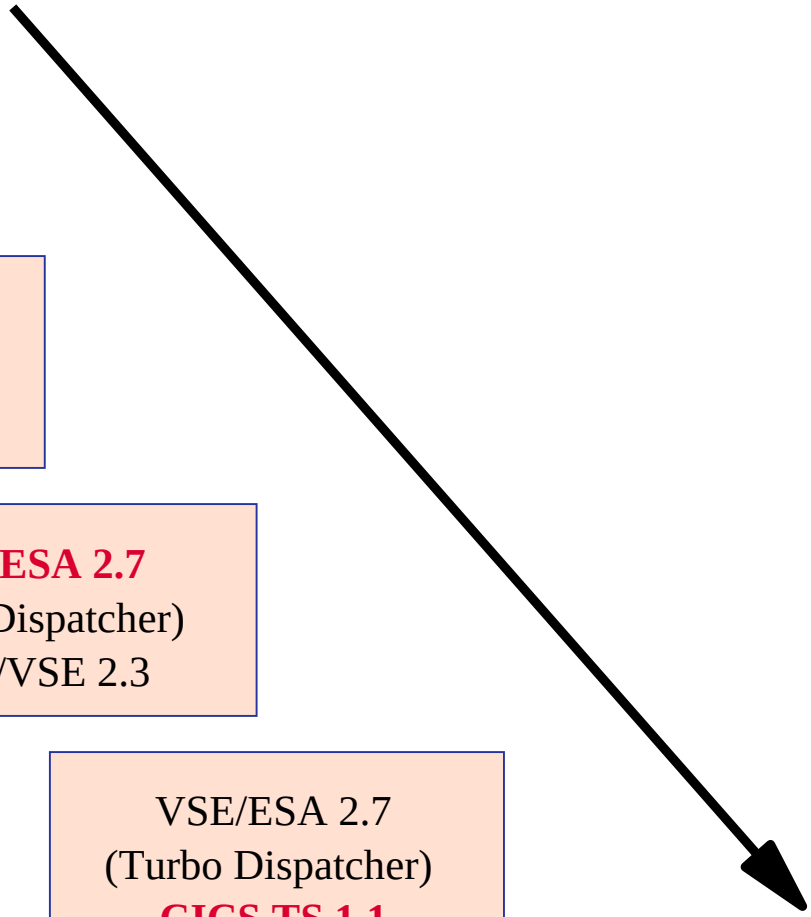
Migration

VSE/ESA 2.3
Standard Dispatcher
CICS/VSE 2.3

VSE/ESA 2.3
Turbo Dispatcher
CICS/VSE 2.3

VSE/ESA 2.7
(Turbo Dispatcher)
CICS/VSE 2.3

VSE/ESA 2.7
(Turbo Dispatcher)
CICS TS 1.1



Jeweils nur eine
Änderung pro Schritt !

Dann kann man nachvollziehen
welcher Schritt ein Problem
gebracht hat.

Performance Tipps

- **Eine Partition kann jeweils nur eine CPU ausnutzen**
  - 2 CPUs bringen nichts für eine CICS Partition
  - 2 oder 3 CPUs können nur mit einer bestimmen Anzahl von Partitionen ausgenutzt werden
- **Aktivieren Sie nur so viele CPUs wie sie wirklich benötigen**
  - zusätzliche CPUs bringen nur Overhead, aber keinen Vorteil
- **Partitions Aufteilung**
  - nutzen Sie mehrere parallele Batch und/oder (unabhängige) CICS Partitionen
  - Trennen Sie CICS Produktions Partitionen in mehrere Partitionen

Performance Tipps (2)

- **Eine CPU muss in der Lage sein die gesamte non-parallel Workload zu verkraften**
- **Non-paralell Code kann die Ausnutzung von mehreren CPUs verhindern**
  - siehe QUERY TD: NP/TOT = NPS
  - NPS vor der Migration messen
  - max CPUs = 0.9 / NPS

NPS	Anzahl der CPUs	NPS	Anzahl der CPUs
0.20	4.5 (4)	0.40	2.2 (2)
0.25	3.6 (3)	0.45	2.0 (2)
0.30	3.0 (3)	0.50	1.8 (1)
0.35	2.6 (2)	0.55	1.6 (1)

Performance Tipps (3)

- **Non-paralleler Code begrenzt den Durchsatz**
  - verhindert die Ausnutzung von mehreren CPUs
- **System Code (Key 0) erhöht den non-parallel Anteil**
  - Vendor Code kann dabei auch einen grossen Anteil haben
- **Der System-Overhead steigt, wenn non-paralleler Code den Durchsatz limitiert**
- **Data In Memory (DIM) verkleinert non-parallel Code**
  - da weniger System Calls (I/Os) gemacht werden
  - kann damit den Durchsatz erhöhen
- **Generell ist es besser EINE schnellere CPU zu haben als mehrere langsamere**
  - Auch wenn rechnerisch mehrere CPUs insgesamt eine höhere Leistung hätten

Performance Tipps (4)

Der schnellste Uni-Prozessor ist (fast immer) der beste Prozessor !

Dokumentation

- **z/VSE Homepage:**
  - http://www.ibm.com/servers/eserver/zseries/zvse/
- **VSE Performance:**
  - http://www.ibm.com/servers/eserver/zseries/zvse/documentation/performance.html
- **z/VM Homepage:**
  - http://www.ibm.com/vm
- **z/VM 5.1 Preferred Guest Migration Considerations**
  - http://www.vm.ibm.com/perf/tips/z890.html
- **IBM eServer zSeries 890 and 990:**
  - http://www.ibm.com/servers/eserver/zseries/z890/
  - http://www.ibm.com/servers/eserver/zseries/z990/
- **IBM TotalStorage DS8000 and DS 6000:**
  - http://www.ibm.com/servers/storage/disk/ds8000/index.html
  - http://www.ibm.com/servers/storage/disk/ds6000/index.html
- **IBM TotalStorage 3494 Virtual Tape Server:**
  - http://www.ibm.com/servers/storage/tape/3494vts/index.html
- **IBM 3494 Tape Library:**
  - http://www.ibm.com/servers/storage/tape/3494/index.html

Fragen ?