IBM Information On Demand 2007

IBM

# DB2 9 for z/OS System Performance

*Akira Shibamiya, IBM,*
*shibamiy@us.ibm.com*

*Session: 1228, Track: Data Servers – System z – DB2 and Tools*

Act.Right.Now.

IBM INFORMATION ON DEMAND 2007
October 14 – 19, 2007
Mandalay Bay
Las Vegas, Nevada

# Abstract


- **This presentation provides a look at the performance impact of DB2 9 for z/OS from the system performance viewpoint, including catalog migration, synergy with new hardware, virtual and real storage, utility, index compression, and general transaction CPU usage trend.**

# Acknowledgment and Disclaimer

- Measurement data included in this presentation are obtained by the members of the DB2 performance groups at the IBM Silicon Valley and Poughkeepsie Laboratory.
- Powerpoint notes are provided by Roger Miller.
- The materials in this presentation are subject to
  - enhancements at some future date,
  - a new release of DB2, or
  - a Programming Temporary Fix
- The information contained in this presentation has not been submitted to any formal IBM review and is distributed on an "As Is" basis without any warranty either expressed or implied. The use of this information is a customer responsibility.

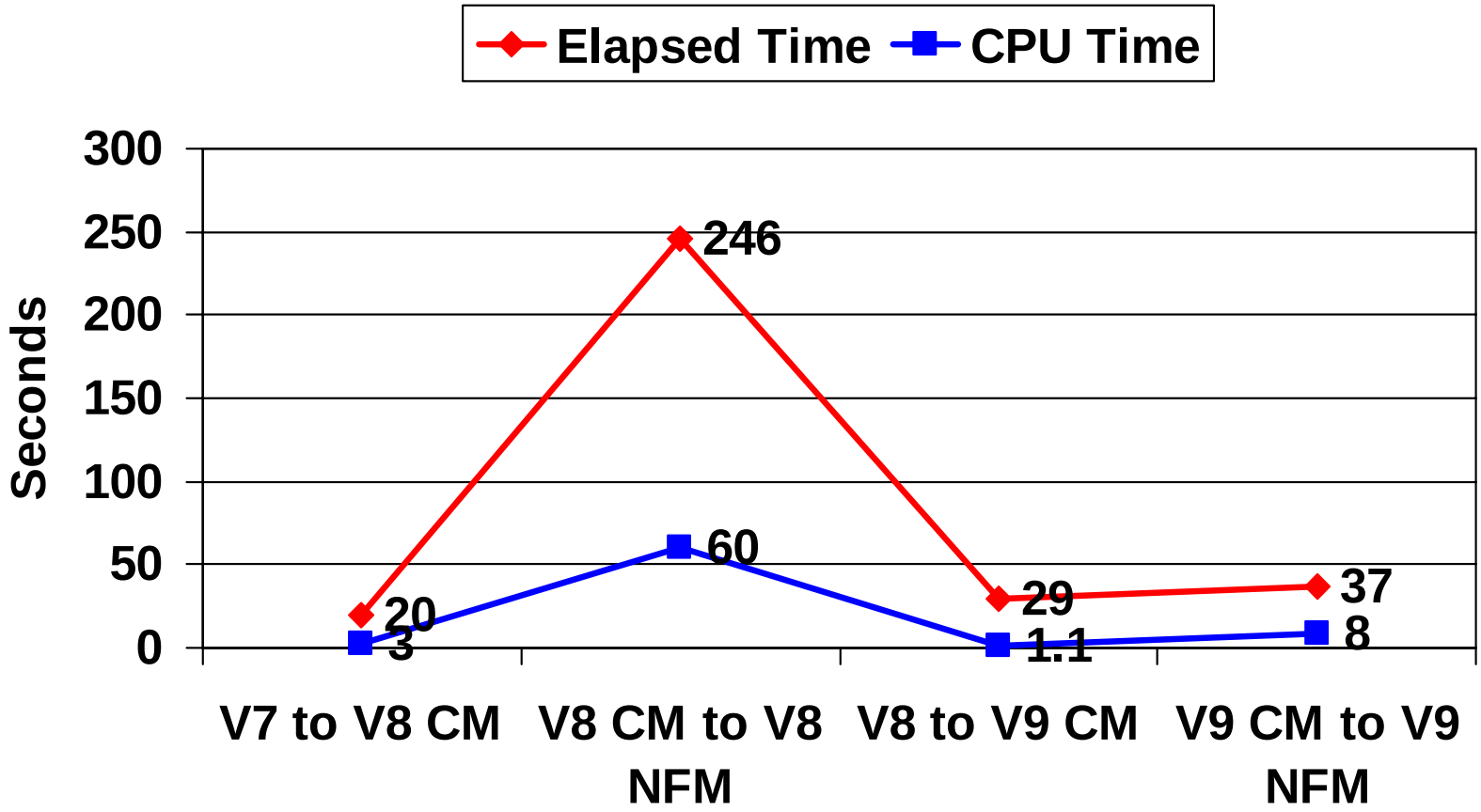
# Agenda

1. Catalog Migration
2. Synergy with New Hardware
3. Virtual/Real Storage
4. Utility
5. Index Compression
6. Miscellaneous

- For each new V9 performance feature, (CM) or (NFM) is shown to indicate if it is supported in Compatibility Mode or New Function Mode.



# V7-V8-V9 Catalog Migration – 700MB catalog



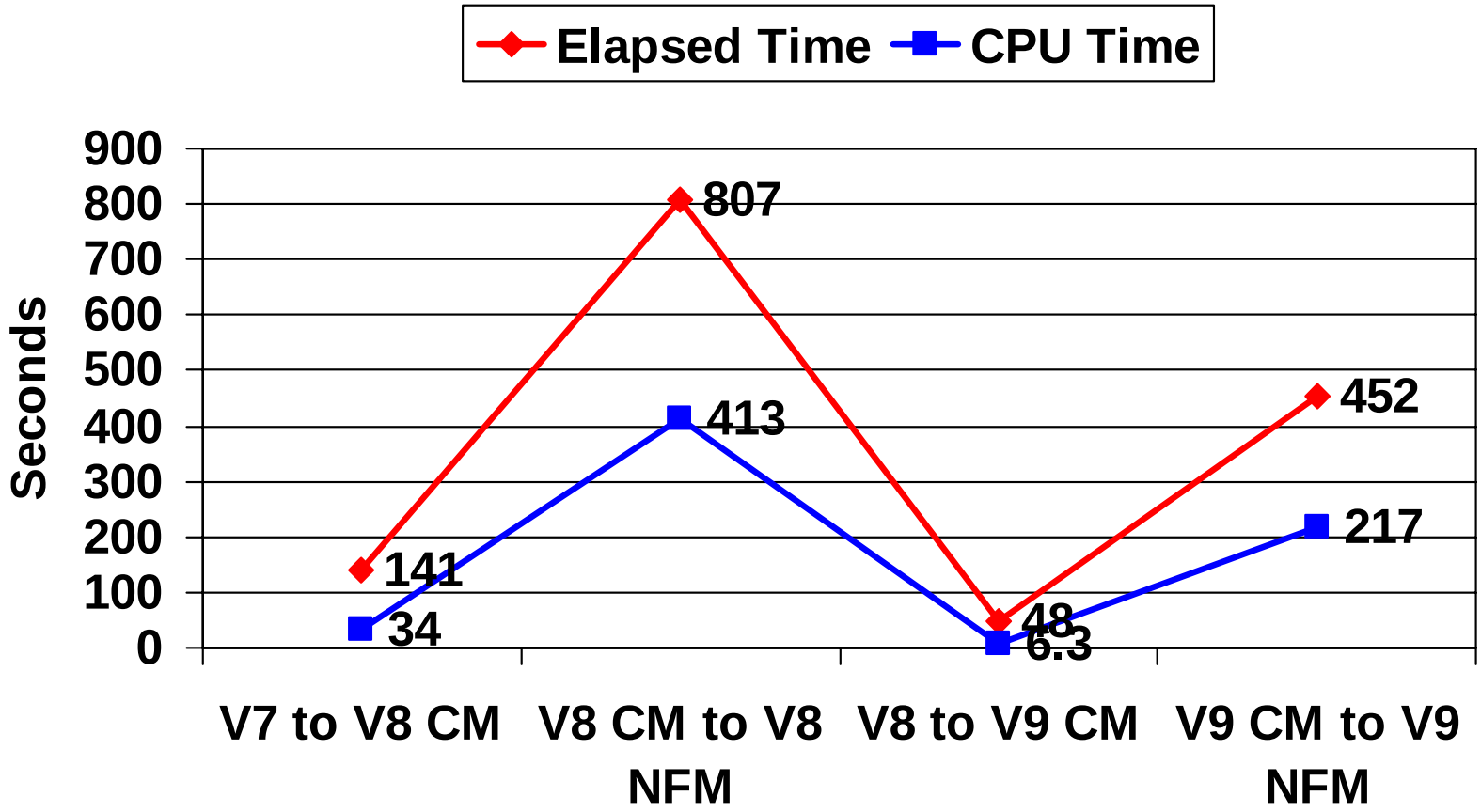
# NOTES

- Elapsed time and CPU time of CATMAINT/CATENFM job is shown for four migration steps
  - V7 to V8 CM
  - V8 CM to V8 NFM
  - V8 NFM to V9 CM
  - V9 CM to V9 NFM

  for different size of catalogs from 3 different customers
  - 700MB
  - 15GB
  - 28GB



# V7-V8-V9 Catalog Migration – 15GB catalog

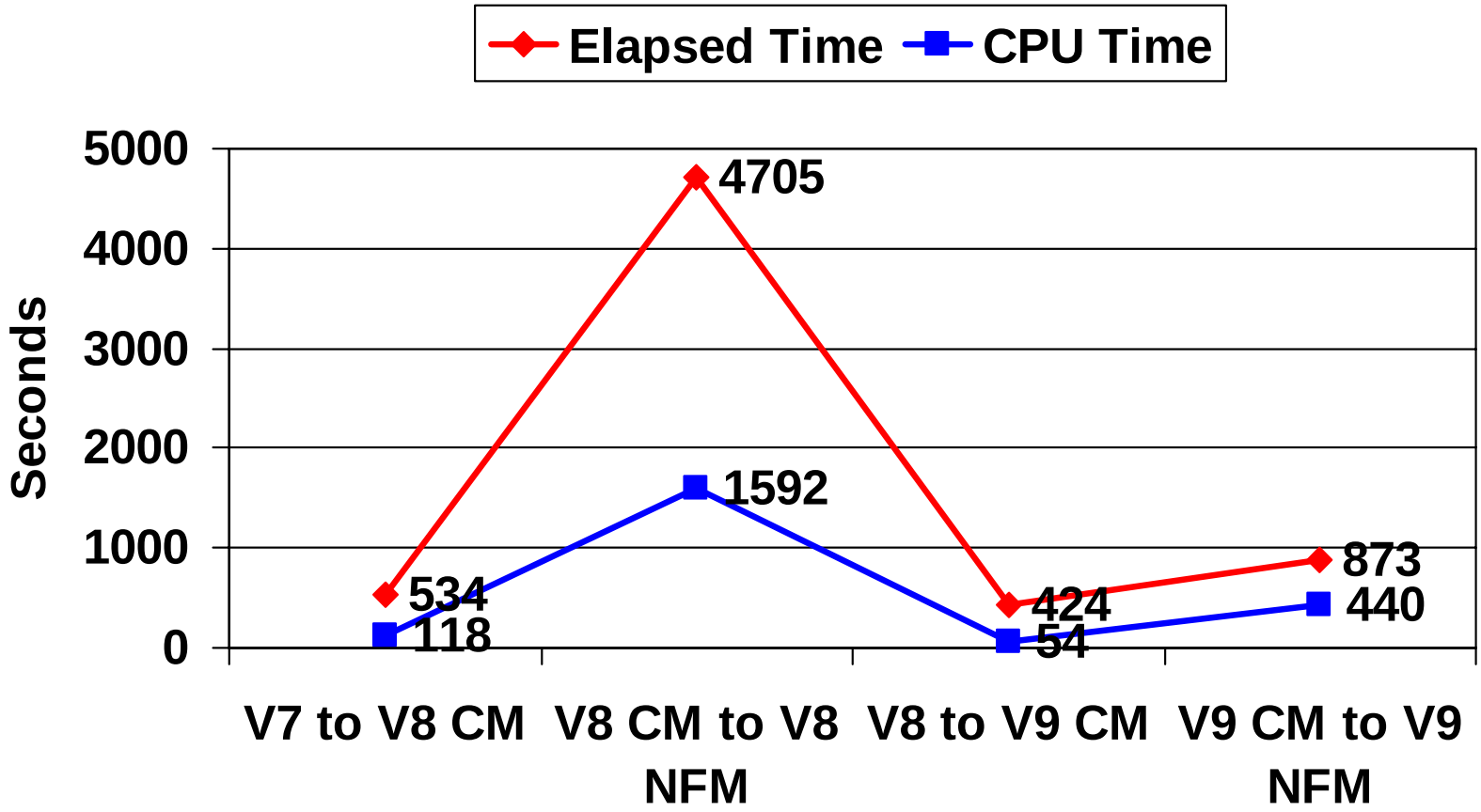


# NOTES

- **In all cases, V8 CM (Compatibility Mode) to V8 NFM (New Function Mode) takes the most time because of Online Reorg Sharelevel Reference of SPT01 and 17 catalog tablespaces such as SYSPKAGE and SYSDBASE.**

- **V9 CM to V9 NFM takes the next highest time, although much faster than V8 CM to V8 NFM, as online reorg of 2 catalog table spaces SYSPKAGE and SYSOBJ takes place.**

# V7-V8-V9 Catalog Migration – 28GB catalog



# NOTES

- **How much time for catalog migration primarily depends on the size of table spaces reorg'd and I/O hardware used.**

- **Configuration used**
  - **4way Z9 (2094)**
  - **ESS mod 800**
  - **z/OS 1.7**



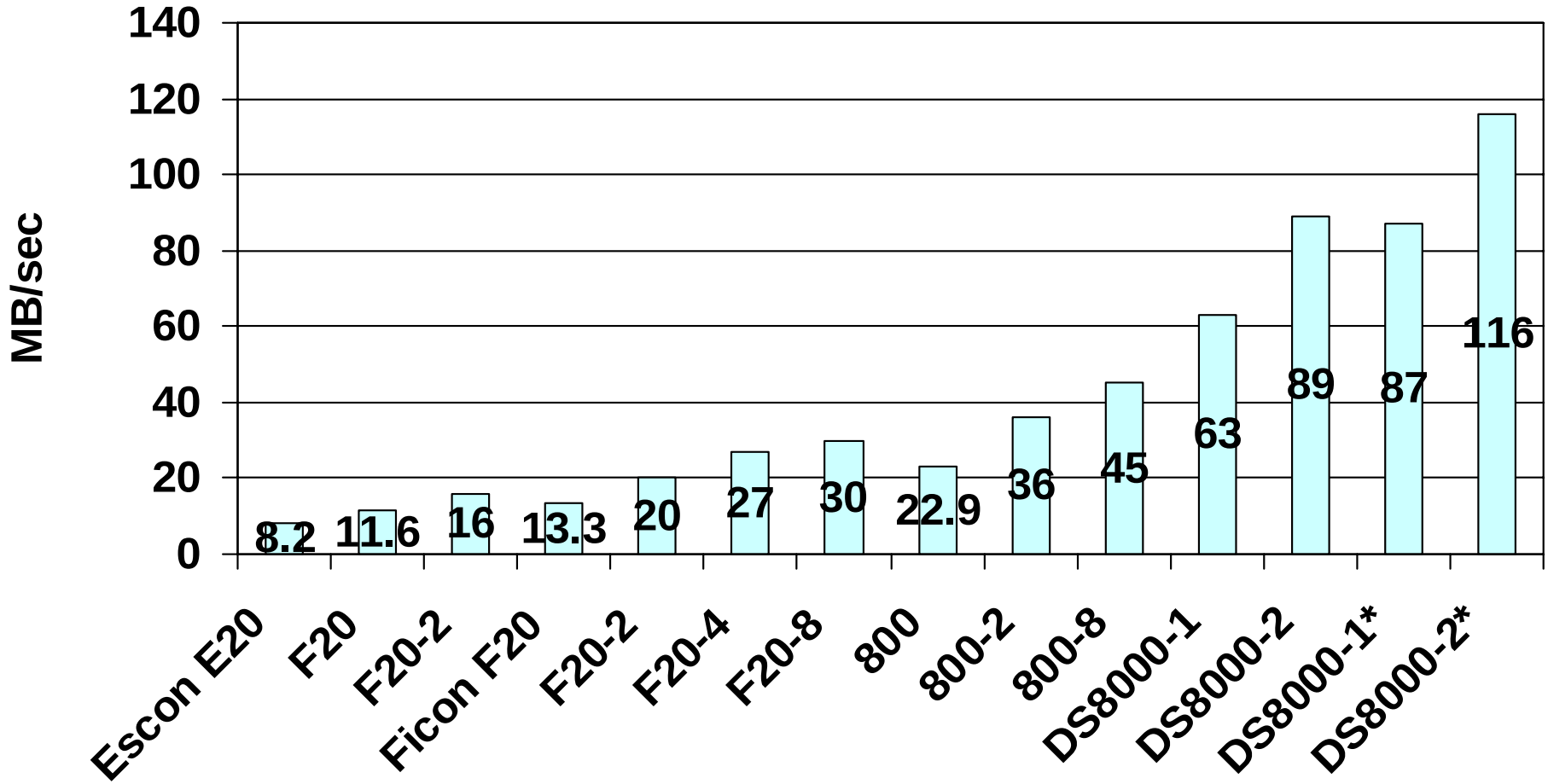
# Synergy with new I/O hardware

- **DS8000 with Ficon Express and MIDAW (Modified Indirect Data Address Word)**
  - **MIDAW requires z9 (2094) and z/OS1.6 OA10984 8/05, 13324/13384 9/05**
  - **Sequential read throughput from cache**
    - **40MB/sec on ESS 800**
    - **69MB/sec with DS8000**
    - **109MB/sec with DS8000 and MIDAW**
    - **138MB/sec with 2 stripes**
  - **Bigger read, write, preformat quantity in DB2 9**
    - **183MB/sec in sequential read with 2 stripes**
  - **Similarly for write**
  - **Performance gap between EF(Extended Format) and nonEF datasets practically gone**

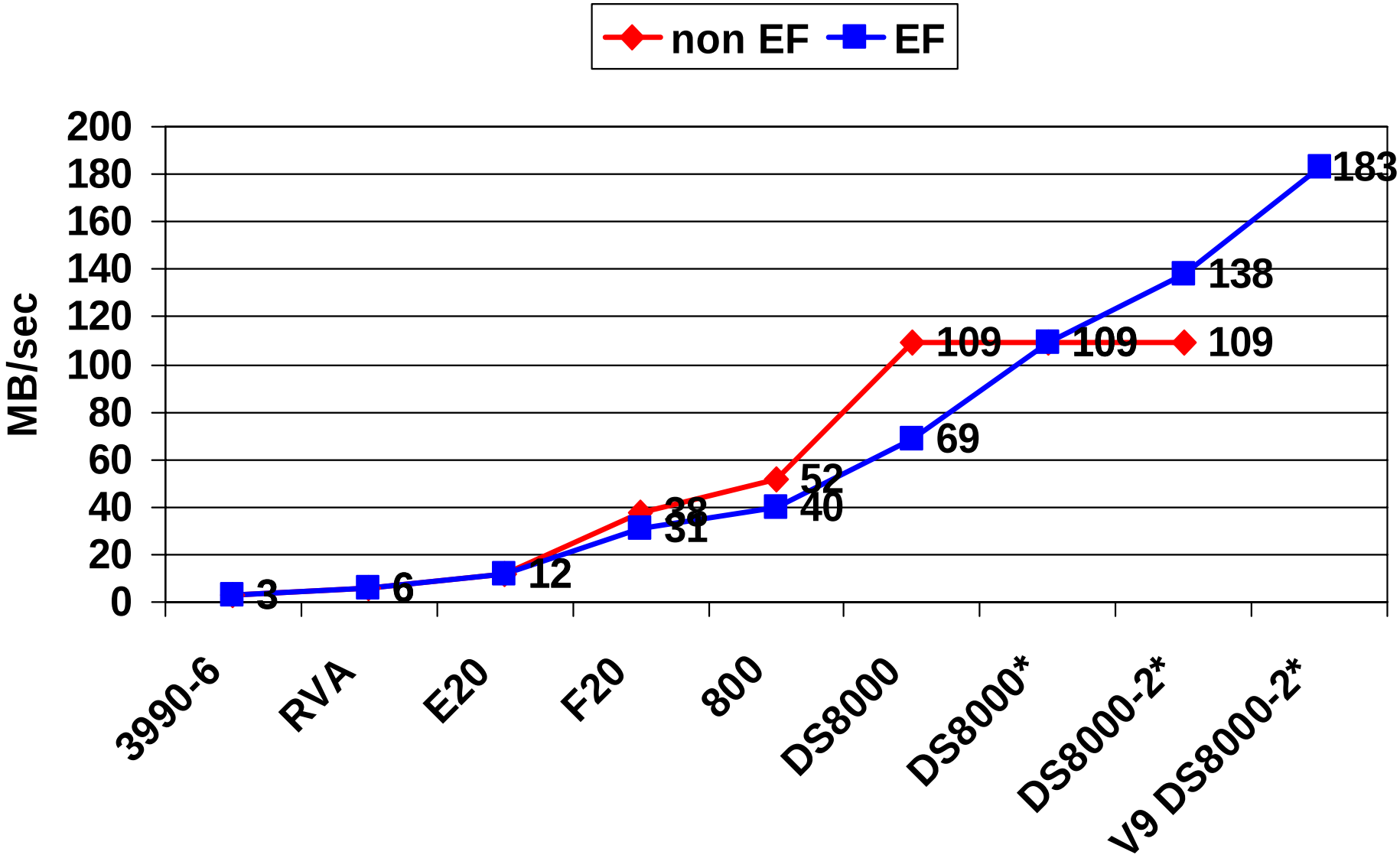


# Maximum observed rate of active log write

- First 3 use Escon channel, the rest is Ficon.
- -N indicates N i/o stripes; * MIDAW



# MB/sec in sequential prefetch from cache (*MIDAW)



# NOTES

- **10 to 20% more throughput with DS8300 turbo**



# Buffer Manager Performance

- **Efficient large BP (>5GB) support**
  - **Fix RSM UIC update for LRU in z/OS1.8**
  - **More evenly balanced buffer hash chain**
    - **Also in V8 PK29626 8/06**

- **Bigger prefetch and deferred write quantity for bigger buffer pool (CM)**
  - **Max of 128 V8 ->256KB V9 in SQL tablespace scan**
  - **256 V8 ->512KB V9 in utility**
  - **+36% MB/sec in non striped prefetch**
  - **+47% in 2-striped prefetch -> more effective striping**



# NOTES

- RSM = Real Storage Manager
- UIC = Unreferenced Interval Count
- LRU = Least Recently Used

- “Bigger buffer pool”
  - For sequential prefetch, if VPSEQT*VPSIZE> 160MB for SQL, 320MB for utility
  - For deferred write, if VPSIZE> 160MB for SQL, 320MB for utility

# Synergy with New CPU Hardware

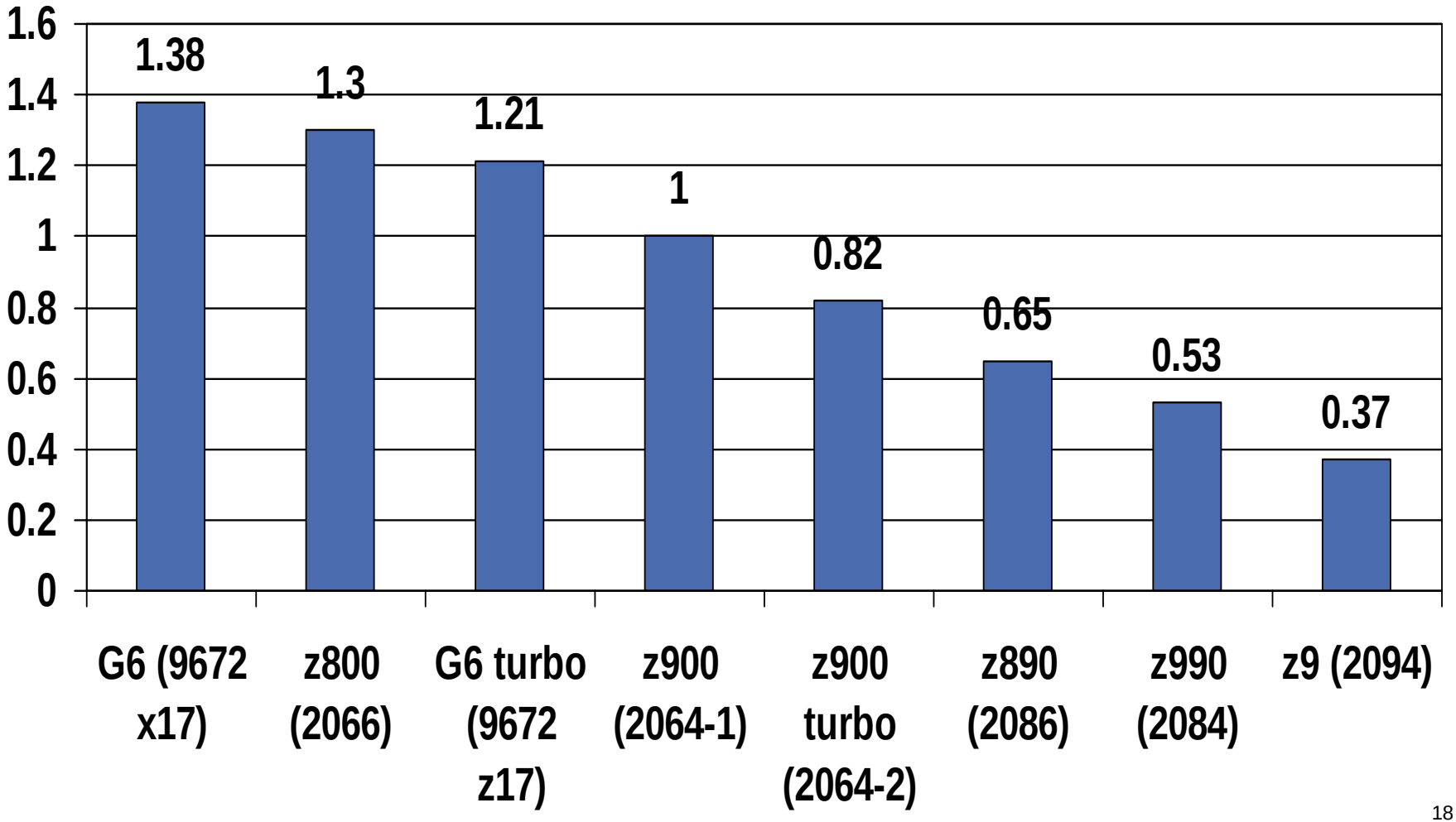
- **Data compression**
  - **Z900 (2064-1) up to 5 times faster than G6 turbo (9672), instead of normal 1.15 to 1.3 times, in compression and decompression**
  - **Z990(2084) 1.4 times additional speed up compared to z900 turbo in decompression**
    - **Z990 1.5 times faster than z900 turbo on average**
    - **But decompression is 1.5x1.4=2.1x faster**
- **Faster Unicode conversion with z900, and more with z990**



# Synergy with new CPU hardware - continued

- **Z990 (2084)**
  - **More than 2 times faster row-level encryption**
  - **V9 long displacement instruction hardware support, simulated by microcode on z900**
    - **Most impact on input/output column processing**
    - **V9 cpu vs V8 on z990 or later: -5 to -15% if column-intensive application**
    - **V9 cpu vs V8 on z900: +5 to 15%, more if many columns**
- **Z9 (2094)**
  - **MIDAW to improve I/O performance**
  - **zIIP offload to reduce total cost of ownership**

# CPU Time Multiplier for various processor models



# NOTES

- **In addition to the raw speed improvement per engine, there are more engines (up to 54 for z9) and special performance improvement tied to a given hardware**

# Z9 Integrated Information Processor (zIIP)

- **ZIIP intended to reduce the total cost of ownership**
- **Prereuisites: DB2 for z/OS V8 (CM), z/OS 1.6, z9 processor**
- **SYS1.PARMLIB(IEAOPTxx) PROJECTCPU=YES for projection without zIIP**
- **Off-loadable enclave SRBs in 3 areas**
  - **DRDA over TCP/IP**
  - **Load, Reorg, Rebuild Utility**
  - **Parallel query**



# DRDA over TCP/IP - PK18454 6/06

- External stored procedure, user defined function, and SNA are not zIIP-eligible
  - However, stored procedure call, result set, and commit processing that run under enclave SRB are eligible for zIIP redirect
- V9 native SQL procedure is off-loadable under DRDA
  - Runs in DBM1, not Workload Manager, address space under enclave SRB



# Load, Reorg, Rebuild Utility

- PK19920 6/06, PK27712 8/06, PK30087 9/06

- Example of effective offloaded CPU time with 4 CPs and 2 zIIPs
  - 5 to 20% Rebuild Index
  - 10 to 20% Load/Reorg partition with one index or entire tablespace
  - 40% Rebuild Index logical partition of NPI
  - 40 to 50% Reorg Index
  - 30 to 60% Load/Reorg partition with more than one index

- Higher percentage redirect as the ratio of #zIIPs to #CPs goes up



# NOTES

- **Variations in percentage redirect for various utilities are primarily determined by the percentage of CPU time consumed by build index processing, which is redirected, to the total CPU time for a given utility.**
  - **E.g. Partition Load/Reorg spends most of the CPU time in build index phase and consequently is in a position to gain the biggest redirect percentage, especially with more indexes.**

- **Less percentage redirect in Rebuild Index with more indexes because of added cost of sort. This also explains smaller percentage redirect than Reorg Index which does no sort.**



# Parallel Query – PK27578 7/06

- For relatively long-running queries, not short
  - E.g. seconds rather than milliseconds of z9 CPU time
  - A portion of the child task processing redirected after certain CPU usage threshold is reached for each parallel group
- More query parallelism leading to more zIIP offload in V9
  - Optimized access path under parallelism separate from sequential access
  - #CPs and #zIIPs both considered in parallel degree



# NOTES

- For more details on zIIP off-load, please refer to Gopal Krishnan's "Leveraging zIIP with DB2 for z/OS" session 1782 in 2007 IOD conference

- zAAP (System z Application Assist Processor) offload of XML System Services in TCB (zIIP offload if DRDA SRB)
  - Prerequisites: DB2 9 NFM, z9, z/OS 1.9 or 1.7/1.8 with PTF
  - XML Insert/Update/Load offload percentage depends on document size, complexity, number of indexes, etc.
  - XML zAAP whitepaper on http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101088



# DBM1 Virtual Storage below 2GB (CM)

- 3 major storage areas still below 2GB in V8
  - EDM pool containing SKCT/SKPT and CT/PT
  - Local dynamic statement cache
  - Thread and stack storage

- EDM pool
  - SKCT/SKPT moved above 2GB
  - A portion of CT/PT moved above 2GB
  - Average estimated reduction of 60% but there is a wide fluctuation from 20 to 90%
  - Rough estimation of V9 EDM pool below 2GB
    = 0x[V8 SKCT/SKPT pages used]
    + 70%x[V8 CT/PT pages used]
    + [free pages based on maximum CT/PT usage]



# NOTES

- Needs V9 rebind to get EDM pool below 2GB relief
- Potential reduction in DBM1 virtual storage below 2GB can range from 0 to 300MB depending on how much thread/stack storage usage
- V8 PK20800 8/07 Display Thread(*) Service(Storage) for agent-local virtual storage, real storage, and auxiliary storage used by DB2

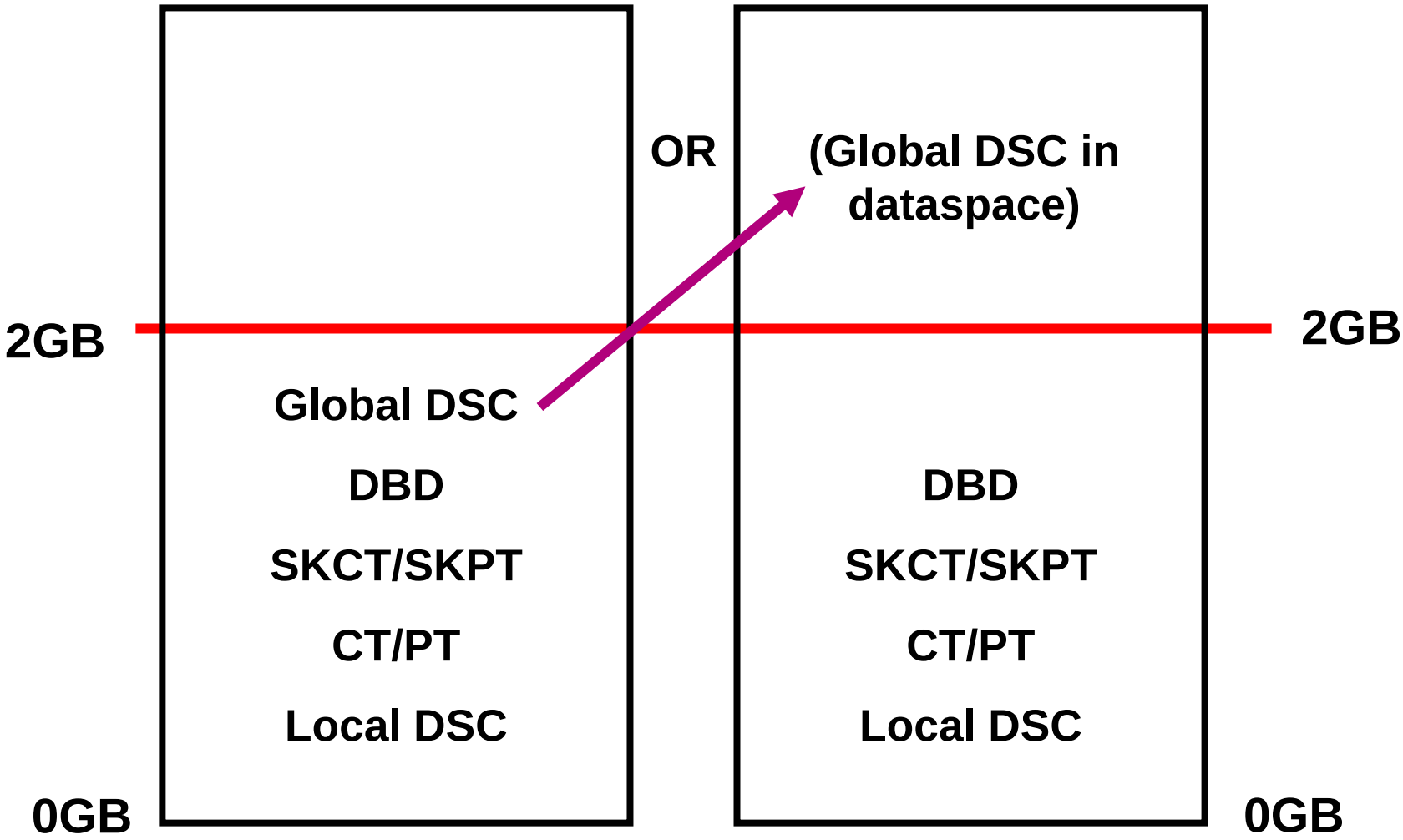


# DBM1 Virtual Storage - continued

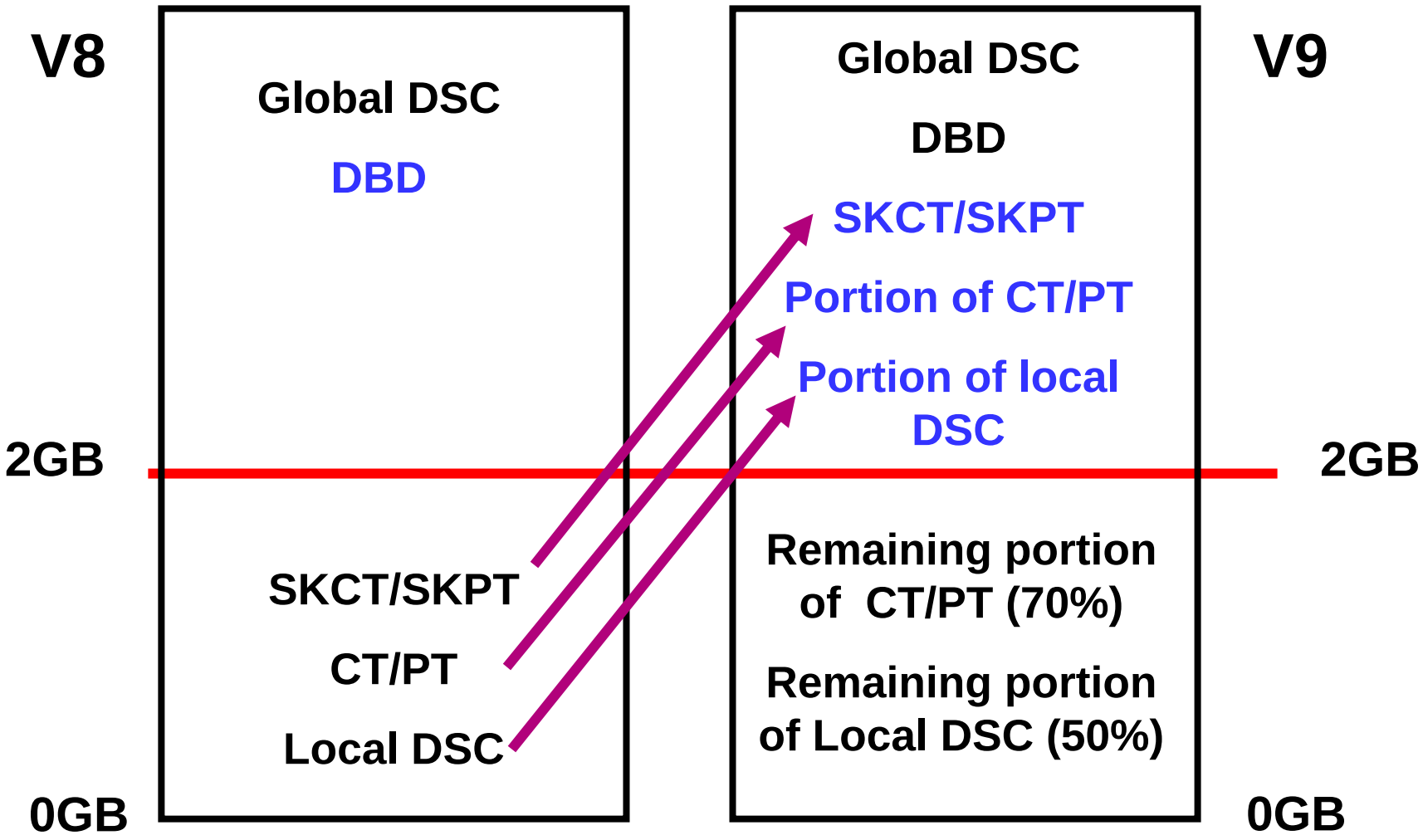
- **Local dynamic statement cache**
  - 60% reduction in one measurement, 40% in another
  - Rough estimation of V9 local dynamic statement cache = 50% of V8 local dynamic statement cache
- **User thread storage, System thread storage, Stack storage**
  - Current expectation of less than 10% difference overall



# Summary of EDM-Related Storage in V7



# Summary of EDM-Related Storage in V8 and V9



# Virtual and Real Storage - continued

- **Real storage – If everything under user control such as buffer/storage pool size, #concurrent threads, etc. is kept constant,**
  - **5 to 25% increase in overall real storage from V7 to V8, primarily depending on active buffer pool size**
  - **Less than 10% from V8 to V9**

- **DDF virtual storage  below 2GB**
  - **15 to 40% reduction in V9**
    - **via shared storage between DDF and DBM1 above 2GB (CM)**



# One measurement example

| Virtual storage below 2GB from DB2 Stats | V8 | V9 |
|---|---|---|
| DBM1 below 2GB used | 1091MB | 819MB |
| Local dynamic statement cache | 466 | 172 |
| Thread/stack storage | 500 | 528 |
| EDM | 110 | 110 |
| DBM1 real storage | 1935 | 2203 +14% |

| Real storage frames from RMF | V8 | V9 | %delta |
|---|---|---|---|
| DBM1 | 497K | 564K | +14% |
| DDF | 171K | 115K | -33% |
| MSTR | 17K | 16K | |
| IRLM | 4K | 4K | |
| TOTAL | 690K | 699K | +1% |



# NOTES

- **In this measurement comparing V8 and V9 virtual and real storage usage,**
  - **DBM1 virtual storage usage below 2GB is reduced by 272MB primarily from local dynamic statement cache**
  - **DBM1 read storage reported in DB2 Statistics indicates 14% increase.**
  - **Real storage frames from RMF indicates 14% increase for DBM1 also. However, because of 33% reduction in real storage usage by DDF, the overall real storage usage increased 1%.**
  - **1 frame represents 4KB, so that 497K and 564K frames reported in RMF match well with DBM1 real storage usage reported in DB2 Statistics Report.**



# Utility CPU time reduction (CM) – primarily from index processing

- 5 to 20% in Recover index, Rebuild Index, Reorg Tablespace/Partition
- 5 to 30% in Load
- 20 to 60% in Check Index
- 35% in Load Partition
- 30 to 50% in Runstats Index
- 40 to 50% in Reorg Index
- Up to 70% in Load Replace Partition with NPIs and dummy input



# NOTES

- **Biggest utility CPU time reduction in DB2 history**
- **Percentage improvement depends on the percentage of index processing in a given utility**
- **Less %improvement in Load, Reorg, and Rebuild if zIIP redirect in both V8 and 9**
  - **Also less % zIIP offload since index maintenance cost has been dramatically reduced**



# Reorg Utility Performance (CM)

- **Parallel unload/reload by partition**
  - **10 to 40% faster in one measurement**

- **Eliminate Build2 phase in Online Reorg Partition with NPI (Non Partitioning Index) for better availability**
  - **But higher CPU time and elapsed time when few out of many partitions, especially with more NPIs, are Reorg'd as entire NPIs copied to shadow dataset**
    - **Additional temporary DASD space needed**
    - **NPIs are automatically Reorg'd also**

- **Fast Log Apply in Online Reorg**



# Miscellaneous Utility Enhancement (CM)

- **Tablespace Image Copy with Checkpage option always**
  - **Added overhead for Checkpage practically eliminated**
  - **LRU (Least Recently Used)->MRU (Most Recently Used) buffer steal**
    - **Protects contents of buffer pool**
  - **15% less CPU than V8 with Checkpage option, same as V8 without Checkpage option in one measurement**

- **Check Index parallelism for sharelevel reference**
  - **Up to -30% elapsed time with less than +5% cpu**



# Miscellaneous Utility Enhancement NOTES

- V9 PK44026 6/07 Restore dynamic prefetch on index scan in Load, Reorg, Runstats Index
- Large Block Interface (System Determined Tape Blocksize) for tapes
  - Tablespace Copy: -10% cpu, -27% elapsed time
  - Restore: -3% cpu, -35% elapsed time
- Active log read buffers per Start IO increased from 15 to 120 for up to +70% recovery throughput



# Other Online Utility Improvement (CM) NOTES

- **Online Check Index V8 PQ96956 10/05**
- **Online Rebuild Index**
  - **Read/Write instead of Read-only access allowed**
  - **For unique index, update of non indexed columns and all deletes, but not insert, allowed**
- **Online Check LOB**
- **Online Check Data**

# Index Compression (NFM)

## Difference between data and index compression

| | Data | Index |
|---|---|---|
| Level of compression | Row | Page (1) |
| CPU overhead | In Acctg | In Acctg and/or DBM1 SRB |
| Comp in DASD | Yes | Yes |
| Comp in BP and Log | Yes | No |
| Comp Dictionary | Yes | No (2) |
| 'Typical' Comp Ratio CR | 10 to 90% | 25 to 75% (3) |



# NOTES

- No compression or decompression in each Insert or Fetch; instead at I/O time → CPU overhead critically sensitive to index BP hit ratio
  - Bigger index BP strongly recommended for index compression
  - No change in acctg CPU time if index pages are brought in by prefetch
- Load or Reorg not required for compression
- Based on a limited survey thus far
  - Higher for relatively unique indexes with long keys
  - Maximum CR limited by index page size: 50% with 8K, 75% with 16K, 87.5% with 32K page



# Index Compression - continued

- **DSN1COMP utility to simulate compression ratio without real index compression**
- **Work area for compressed index I/O is long term page fixed.**
  - **So do not page fix compressed index buffer pool.**
  - **If a mix of compressed and non compressed indexes, use a separate buffer pool.**

# NOTES

- **For some customers, especially in Data Warehouse/Business Intelligence, indexes take up more DASD space than data**
  - **Index compression can be very valuable in such an environment.**
  - **The cost of index compression is under user control.**



# Dataset Open/Close Performance

- **Buffer Manager Open/Close service tasks increased from 20 to 40 to speed up massive dataset open/close (CM)**
  - **Also only one close at a time while up to 20 concurrent open prior to V8 PK28008 8/06 and PK33496 1/07 which avoids the problem of deferred close not being able to keep up with dataset open resulting in the number of open data sets much greater than DSMAX.**
    - **Also V8 PK42106 5/07 to limit the number of deferred closes scheduled which causes performance degradation.**
  - **Up to 40 concurrent dataset open or close in V9**



# NOTES

- **Deferred close**
  - **V7: When the number of open datasets reaches 99% of DSMAX, 3% of DSMAX of Least-Recently-Used datasets, with CLOSE YES datasets picked first, are closed.**
  - **V8/V9: When the number of available open dataset slots reaches a smaller of 100 or 1% of DSMAX, a smaller of 300 or 3% DSMAX of LRU datasets, with CLOSE YES datasets picked first, are closed.**
- **ACCESS DATABASE(…) SPACENAM(…)**
  - **MODE(OPEN) to physically open dataset on the local member (CM)**
  - **MODE(NGBPDEP) to convert to non GBP dependency (CM)**



# Instrumentation CPU Time Reduction

- **Minimize phantom or orphaned trace records (CM)**
  - **Example from customer's DB2 V8 statistics report in IFC records per commit**
  - **(1) Phantom or orphaned trace because monitoring (eg vendor tool) stopped but not DB2 trace. The same CPU overhead as real trace.**
  - **V9 tries to eliminate orphaned trace records**

| IFC DEST | Written | Others (1) |
|---|---|---|
| SMF | 2 | 0 |
| OP5 | 0 | 4 |
| OP6 | 0 | 4 |
| OP7 | 0 | 4 |
| OP8 | 2 | 0 |
| Others | 0 | 0 |

# Instrumentation NOTES

- **Capture missing wait time in class3 accounting to reduce NOT ACCOUNTED time (CM)**
  - **Active log read**
  - **TCP/IP to transmit the LOB**
- **Package-level trace filter in Trace Command**



# General Transaction CPU Usage Trend

- 5 to 10% increase on average from V7 to V8
- No change on average from V8 to V9
  - 0 to 5% improvement for column-intensive transaction, especially with varchar (NFM). CPU reduction also from
    - Long displacement instructions (CM)
    - DDF/DBM1 shared storage (CM)
    - Partition declaim (CM)
    - Index access (CM)
    - DSNXECW (CM)



# NOTES

- Long displacement instructions on z990 and higher
  - To avoid module split and promote more efficient register-to-register instructions
  - Bigger benefit with many input host variables and/or output columns
  - Up to 15% CPU reduction in many column Fetch/Insert
  - Simulated by microcode on z900 and can result in 10% or higher cpu time for column-intensive applications
- Table space, index, partition declaim cpu time reduction by only accessing partitions claimed
- DSNXECW cpu time reduction by cleaning up unused statement section at commit for JDBC/CLI



# Reference

- Redbooks at www.redbooks.ibm.com
  - DB2 9 for z/OS Technical Overview SG24-7330
  - DB2 9 for z/OS Performance Topics SG24-7473

- DB2 for z/OS home page at www.ibm.com/software/db2zos
  - E-support (presentations and papers) at www.ibm.com/software/db2zos/support.html