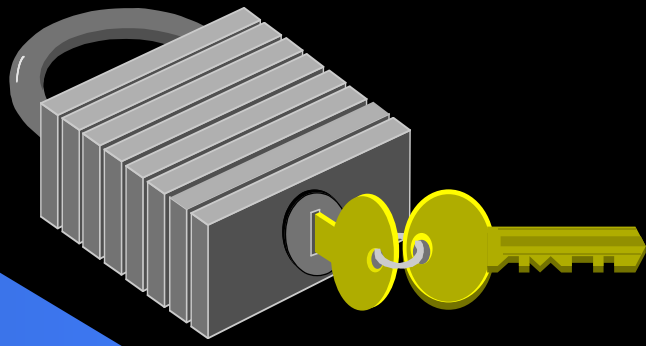
OpenVMS Locking Concepts



Christian Moser
OpenVMS Engineering
Finland Development Center
cmos@hp.com

Norman Lastovica
Oracle Rdb Engineering
New England Development Center
norman.lastovica@oracle.com

Revised: June 20, 2005

Agenda

- Synchronization techniques
- VMS Distributed Lock Manager
- Rdb's uses of locking
- Tools and tricks

Synchronization Techniques

- Atomic Updates
  - Compiler builtins
  - LDx_L/STx_C (Alpha)
  - fetchadd, xchg, cmpxchg (IA64)
- Spinlocks
  - Static (SCHED, IOLOCK8, MMG)
  - Dynamic (PCB, Mailbox, TCPIP)
- Mutexes & Semaphores
  - Posix Threads
- Locks & Resources
  - Distributed Lock Manager
  - Used by VMS, XQP, RMS, Rdb, etc.

Distributed Lock Manager

- Cooperating processes use lock manager to synchronize access to shared resources
- Locks are used to control access to resources
- Resource may be just about anything
  - Device, File, Record, Bucket, Database, Page
- Works across all nodes in a cluster environment

Distributed Lock Manager (cont'd)

- Lock modes: NL, CR, CW, PR, PW, EX
- Locks "requested" by thread/process and "granted" by lock manager
  - Thread/process put in wait state until lock granted
- Many locks may be taken out on single resource

Compatibility of Lock Modes

Requested Mode	Mode of Currently Granted Lock: NL	CR	CW	PR	PW	EX
NL	Yes	Yes	Yes	Yes	Yes	Yes
CR	Yes	Yes	Yes	Yes	Yes	No
CW	Yes	Yes	Yes	No	No	No
PR	Yes	Yes	No	Yes	No	No
PW	Yes	Yes	No	No	No	No
EX	Yes	No	No	No	No	No

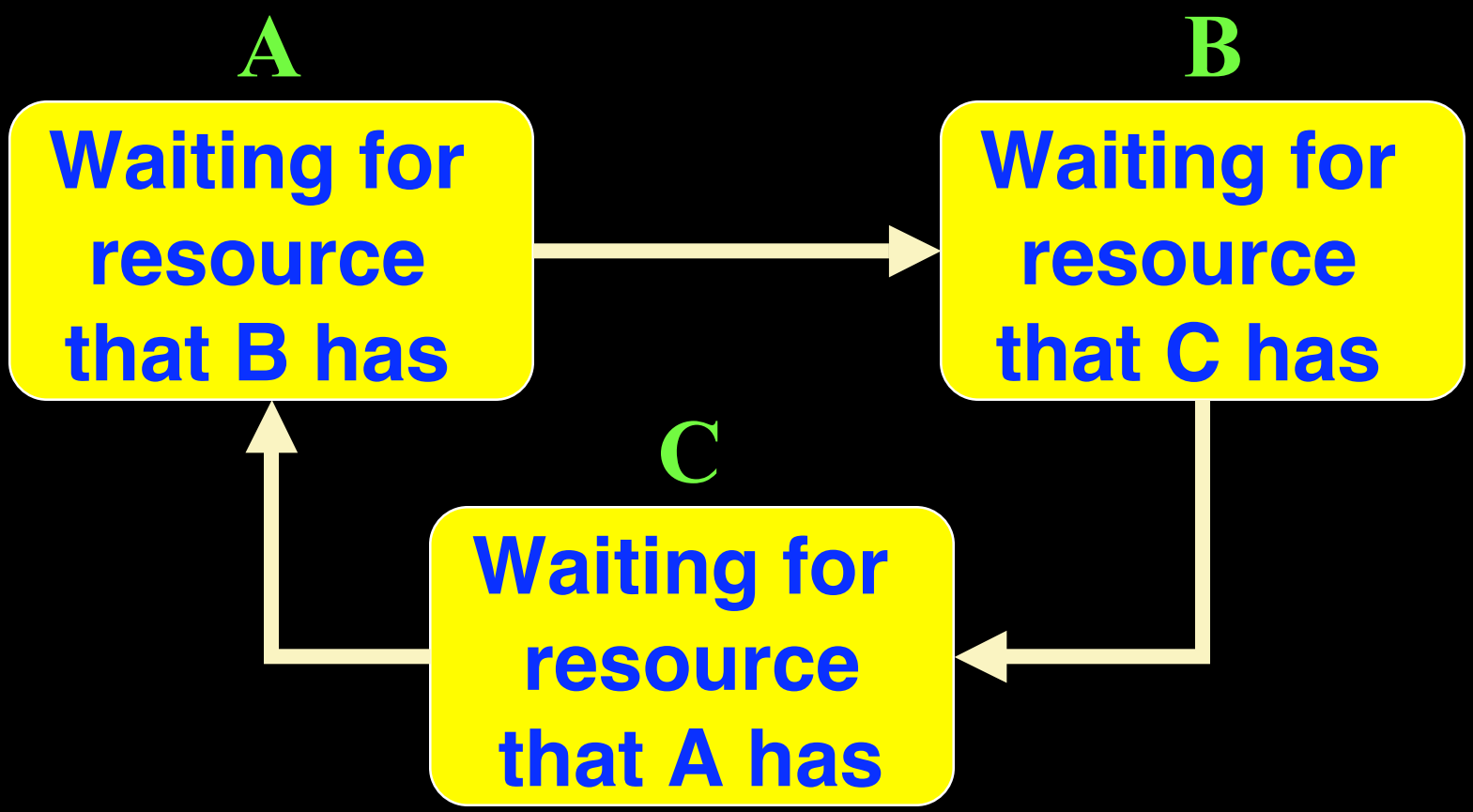
Distributed Lock Manager (cont'd)

- Lock conversion
  - "UP" to more restrictive mode
  - "DONE" to less restrictive mode
  - Convert to NL mode & keep lock is faster than enqueue/dequeue new locks
- Lock queues
  - Granted, Waiting & Conversion
- Synchronous lock request $ENQW or lock completion AST with $ENQ

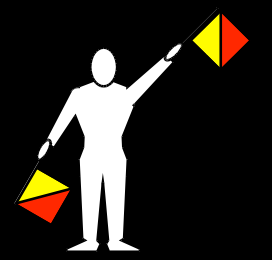
Deadlocks

- Occur when any group of locks are waiting for each other in a circular fashion
  - There can be 2 or more locks involved in deadlock
- VMS checks waiting locks once per second
  - DEADLOCK_WAIT to disable or control frequency of searches
- Deadlock search is complex & quite costly if nothing found
  - Use LCK$M_NODLCKWT & LCK$M_NODLCKBLK to avoid searches on ”doorbell” locks
- VMS chooses victim to break deadlock
  - Returns SS$_DEADLOCK
  - Process needs to demote all locks & restart

3-Member Deadlock



Lock Blocking AST



- When lock is requested, blocking AST routine may be specified
- When incompatible lock is requested by another process, blocking AST routine is called
  - Notifies program that it is blocking some other lock request
  - Also called ‘BLAST’ (BLockingAST)
- Cheaper to hold on to lock with BLKAST than to convert up & down at high rate
  - RMS bucket and record lock
  - Rdb Page lock
- Used also for event notification

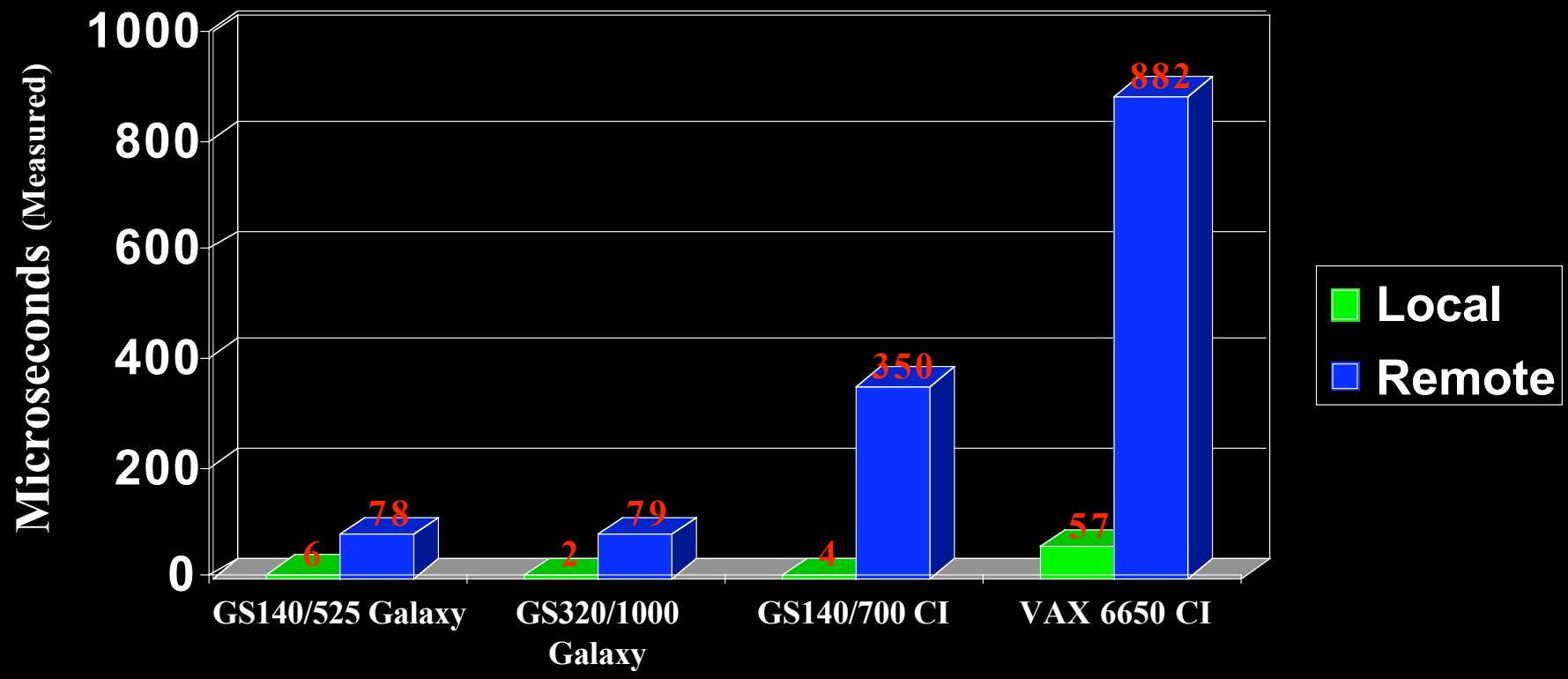
Lock Value Block

- Optional information stored with resource
  - 16 or 64 bytes
  - Content written from LKSB into RSB when convert down or dequeue from EX or PW
  - Content read from RSB into LKSB when lock granted or convert to equal or higher mode
- Used to pass information between processes
  - Resonably fast way to pass small amount of data between multiple nodes in a cluster
  - Often used for cache coherency
- Volatile: content becomes invalid if node/process fails

Lock (Re)Mastering

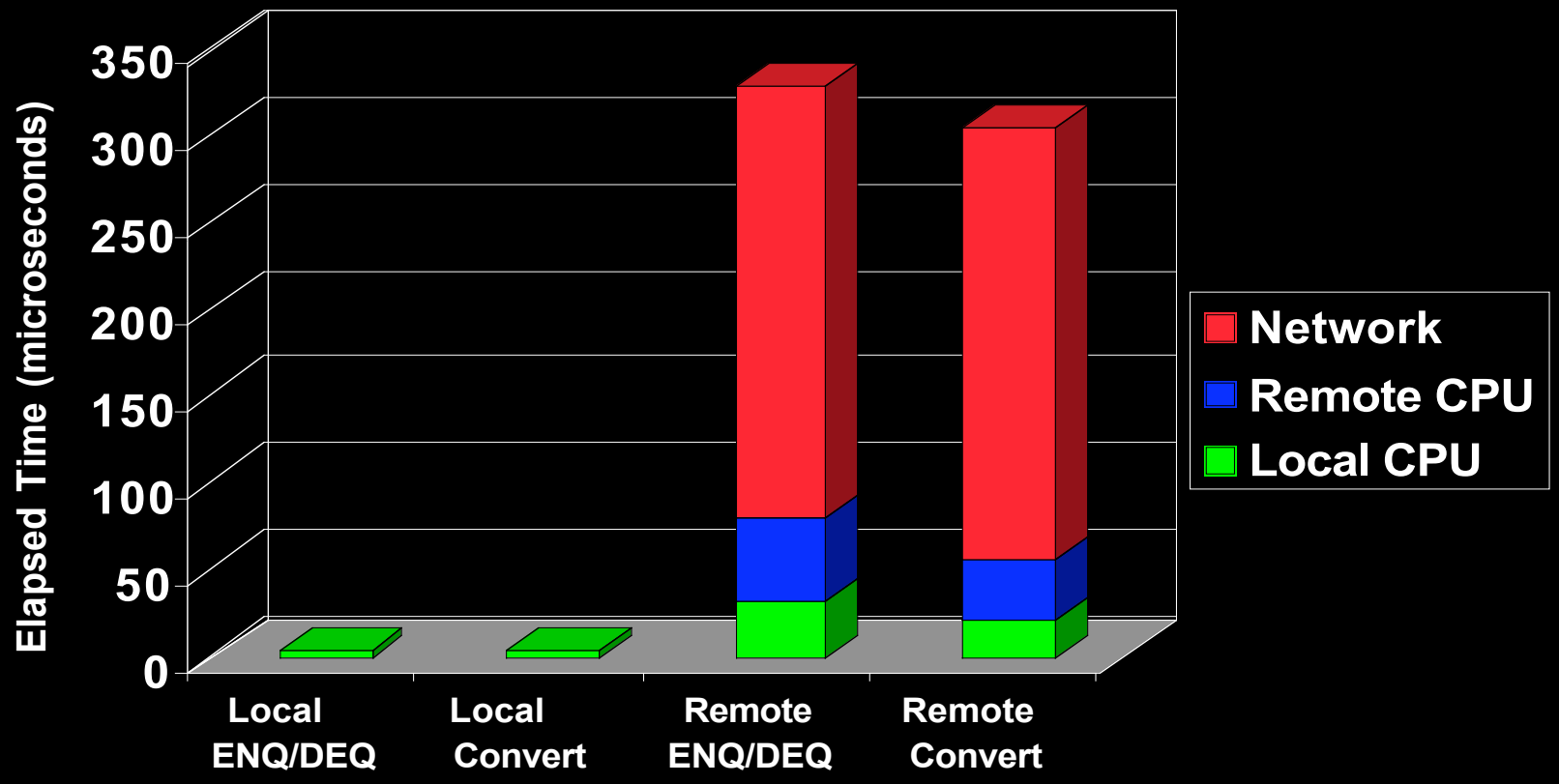
- Resources mastered on one node at a time
  - Entire lock tree (lock & sublocks) mastered on one node
- Master is responsible to coordinate access to locks
- New resource mastered on local node
  - Moves to nodes with non-zero LOCKDIRWT if more than one node has locks on resource
- Lock tree remastered if node removed from cluster
- Dynamic remastering based on activity
- Local lock operation is fast
- Remote lock operation is slow (orders of magnitude)

Distributed Locking Response Times



Distributed Locking Costs

GS140 - CI Interconnect (estimated)



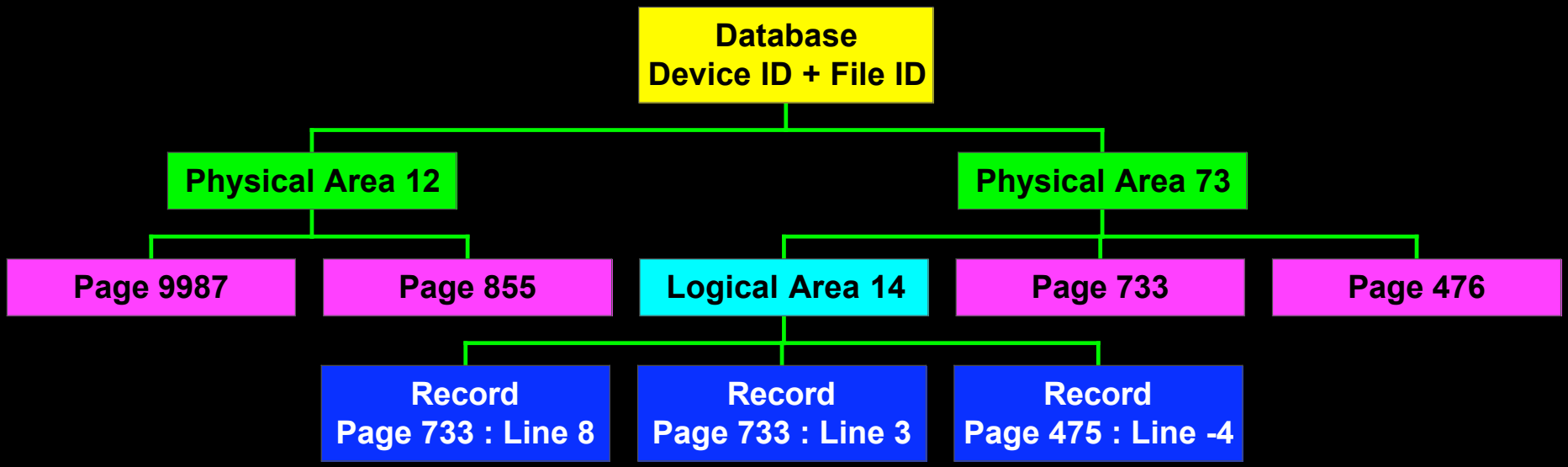
Dedicated CPU Lock Manager

- "Dedicate" a CPU for all local lock management operations
  - Spins waiting for queued work requests
  - Very low latency
  - Avoids contention for lock management spin lock
  - Keep CPU caches "hot"
  - Virtually eliminates MPSYNC time for some applications
  - Remote lock operations do not use
  - Avoid device I/O & interrupts on this CPU

Locking Parameters

- **ENQLM process quota**
  - 32K in SYSUAF means unlimited
- **LOCKIDTBL initial size of lock id table**
  - Can grow on the fly
- **RESHASHTBL size of resource hash table**
  - If too small can result in long hash chain walks
- **DEADLOCK_WAIT**
  - Control deadlock searches, overhead if too small
- **LOCKDIRWT**
  - Controls portion of lock directory for this node
- **PE1 controls dynamic remastering**
- **LCKMGR_MODE enables dedicated lckmgr**
- **LCKMGR_CPU controls CPU affinity of dedicated lckmgr**

Rdb Lock Tree



Rdb Lock Tools

- $ RMU /SHOW STATISTICS
- $ RMU /SHOW LOCKS
  - [/MODE=BLOCKING]
  - [/MODE=WAITING]
  - [/MODE=CULPRIT]

Top 8 Interesting Rdb Resource Types

Type	Name
B	Logical Area
C	Snap Area Cursor
G	TSNBLK
K	Database key scope
L	DBKEY (page/line)
P	Page
R	SEQBLK
U	Client (subtypes: DDL, PSN, DDLctr)

Example Rdb Client Resource Name

- `Resource: client '....7...C1  '`
  `20203143000000370000000400000055`
  - Client Lock
  - Lock Type
    - ☒Relation/View = 00000004
    - ☒Modules = 00000015
    - ☒Routines = 00000016
  - Object number
  - Additional information (usually 4 byte start of ASCII name)
- Enhanced formatting in Rdb 7.2

Example Uses of Lock Value Block

Resource Name: TSN block 2
Lock Value Block: 00002171 00000000 00000FE2 03000001

- 00002171 00000000 - oldest TSN in block (8561)
- 00000FE2 - TSNBLK sequence (4066)
- 0001 - There is a WIP TSN in this block
- 00 - Filler byte – not used
- 03 - VALBLK_VALID + SYNCH

Resource Name: channel 4
Lock Value Block: 1A19EBC5 00290000 112407A3 03010000

- 1A19EBC50000 - Device 'type' in RAD50 (1DGA)
- 0029 - Unit number (41)
- 112407A30000 - File ID (1955,4388,0)
- 03 - VALBLK_VALID + SYNCH

Tools & Utilities

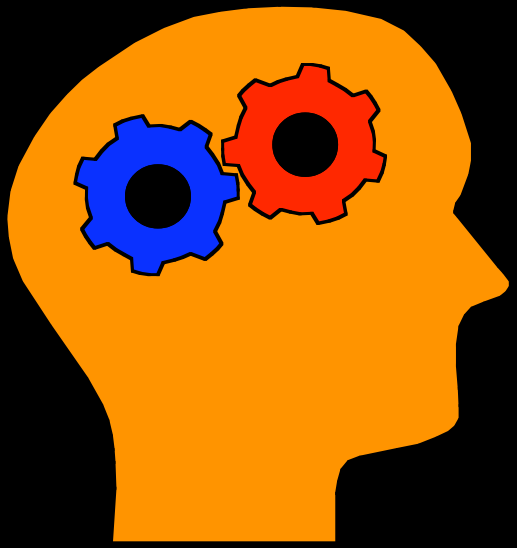
- `$ MONITOR LOCK`
- `$ MONITOR DLOCK`

- `SDA> SHOW PROC /LOCK [/BRIEF]`
- `SDA> SHOW LOCK [/SUMMARY]`
- `SDA> SHOW RESOURCE [/LOCK=n]`
- `SDA> SHOW RESOURCE [/CONTENTION]`
- **Examples:**
  - **sda_lock_summary.txt**
  - **rdb_active.txt**

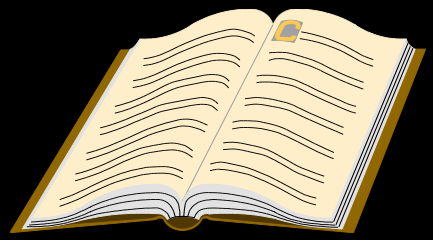
More Tools & Utilities

- `SDA> LCK STATISTIC`
- `SDA> LCK SHOW ACTIVE`
- `SDA> LCK SHOW CONTENTION /INTER=0.1`
- `SDA> LCK SHOW LCKMGR /INT=10 /REP=5`
- Examples
  - lck_active.txt
  - lck_statistic.txt
  - lck_lckmgr.txt
  - lck_process.txt

Anything Else...



- OpenVMS New Features and Release Notes
- OpenVMS Programming Concepts
- OpenVMS System Services Reference
- OpenVMS Internals and Data Structures



Questions?

Comments?