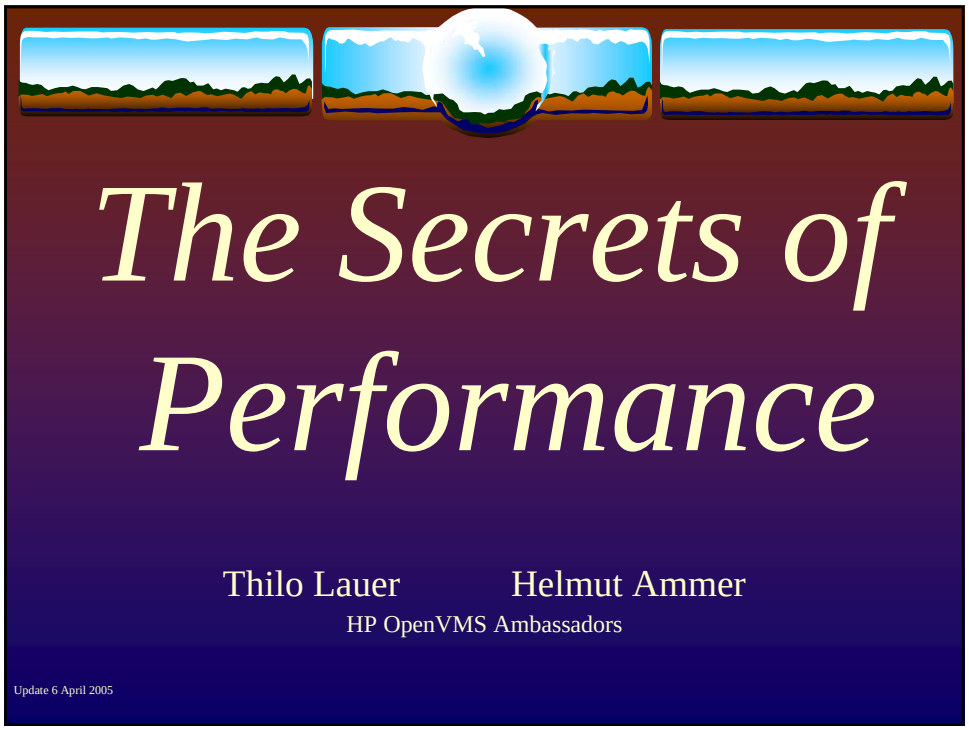
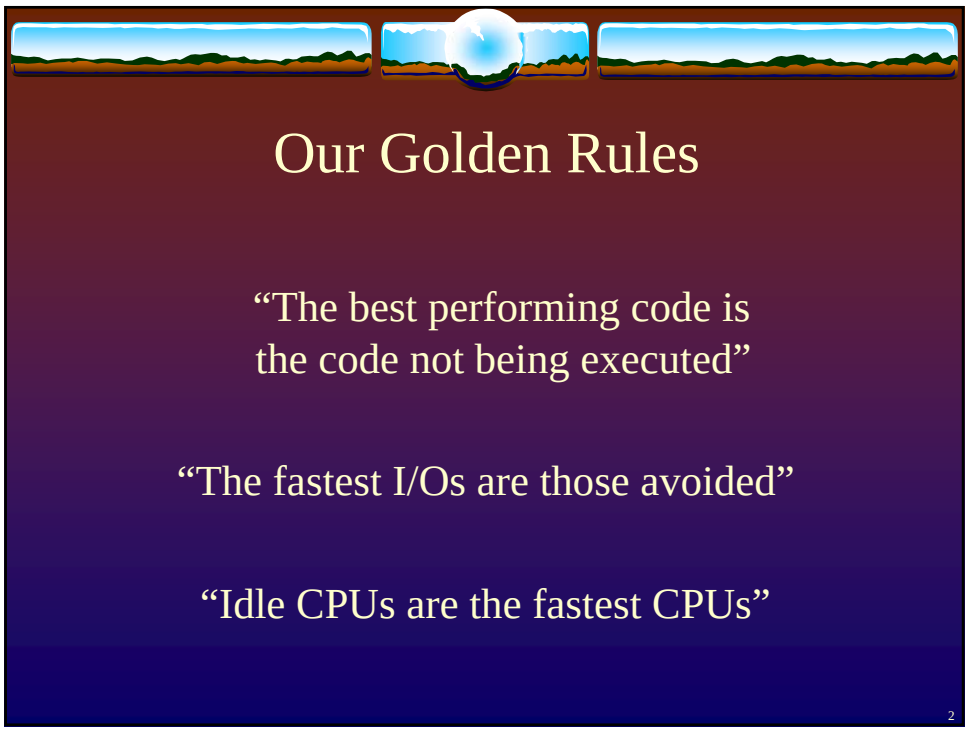
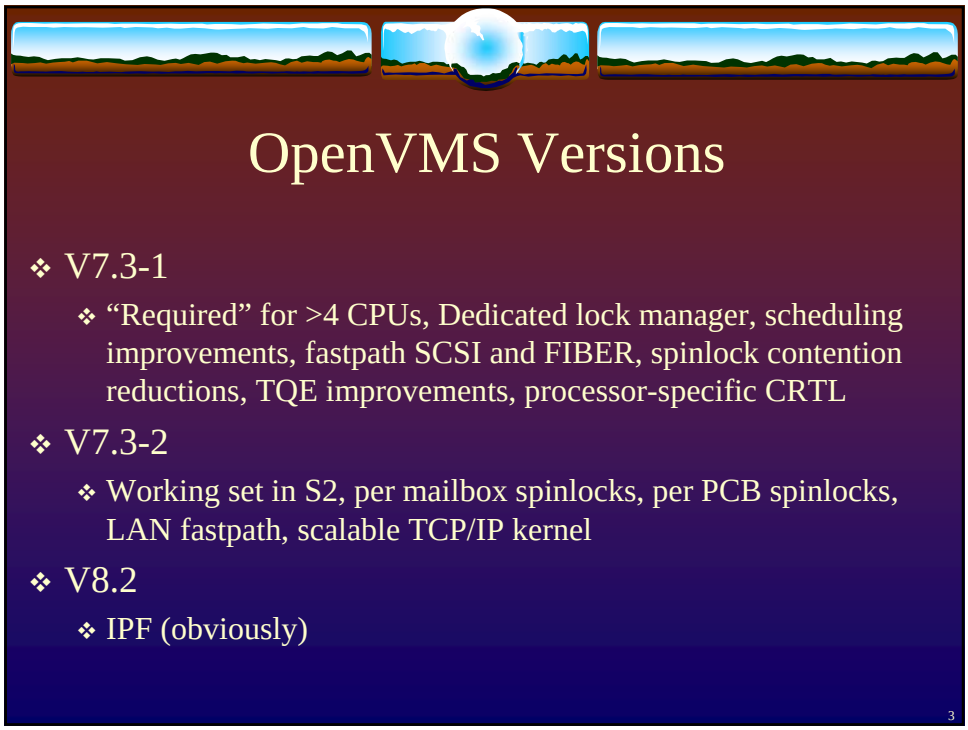
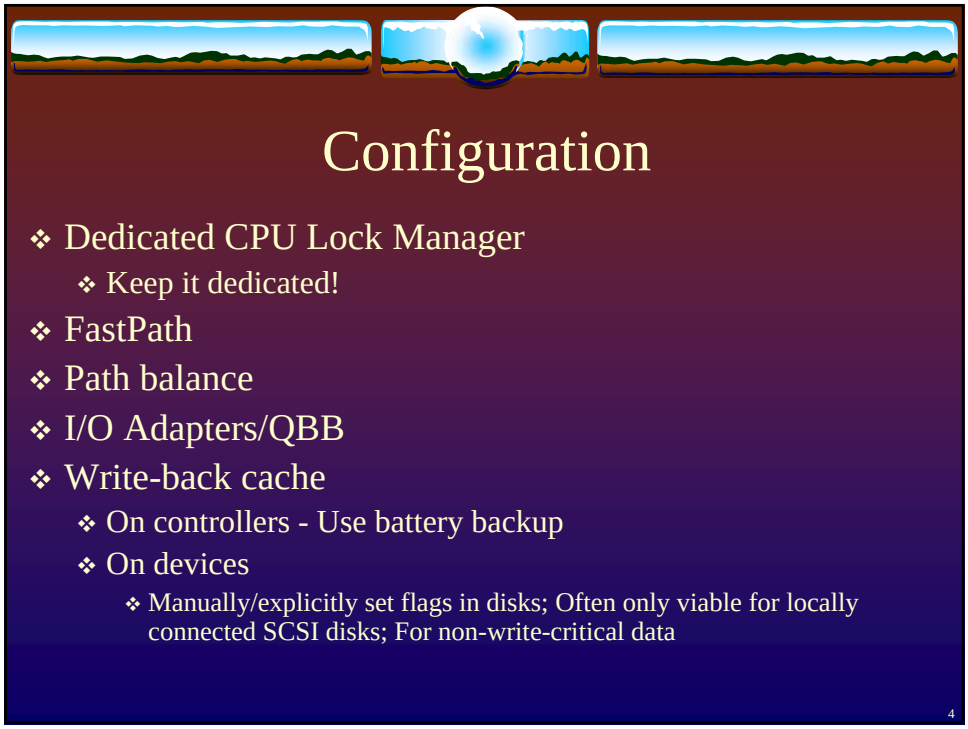
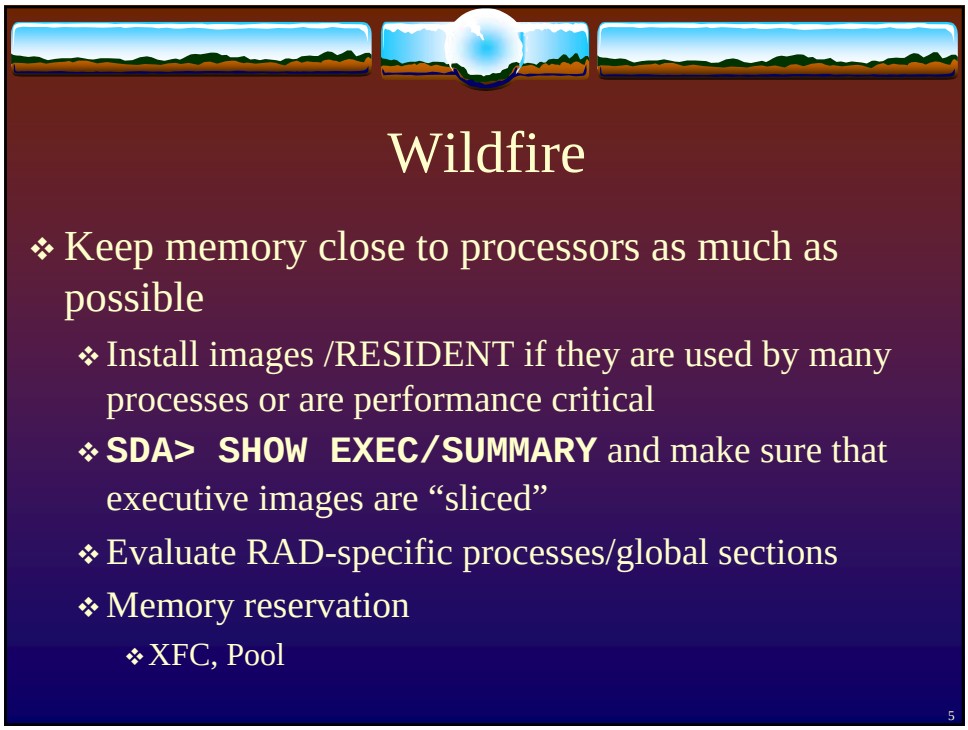
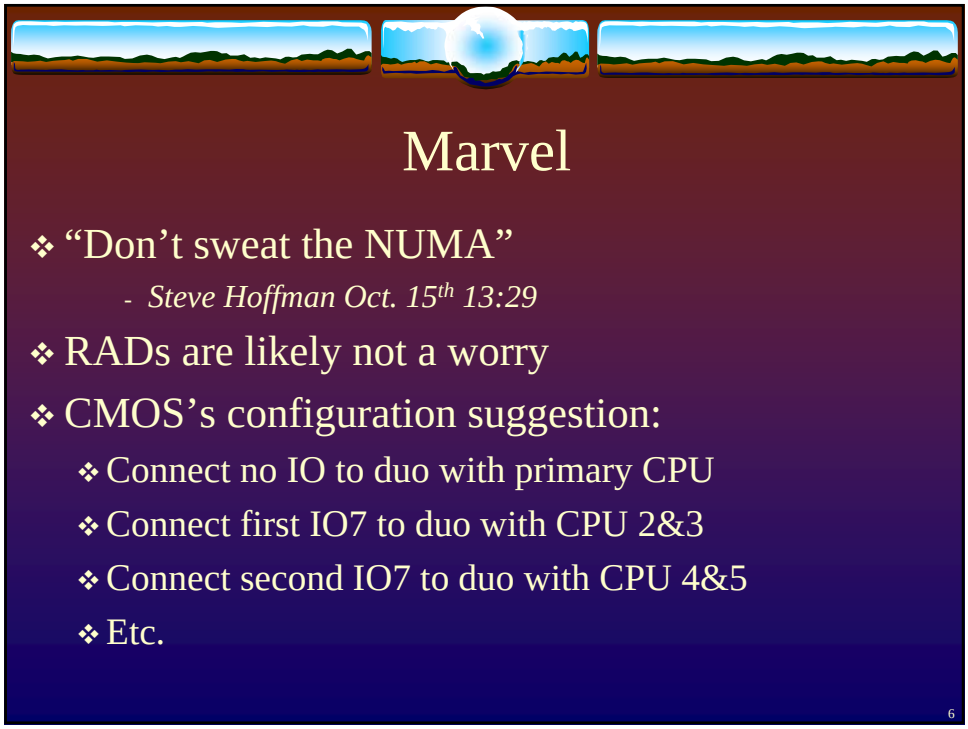

The Secrets of Performance
Thilo Lauer
Helmut Ammer
HP OpenVMS Ambassadors
Update 6 April 2005


Our Golden Rules
"The best performing code is the code not being executed"
"The fastest I/Os are those avoided"
"Idle CPUs are the fastest CPUs"
2


OpenVMS Versions
V7.3-1
"Required" for >4 CPUs, Dedicated lock manager, scheduling improvements, fastpath SCSI and FIBER, spinlock contention reductions, TQE improvements, processor-specific CRTL
V7.3-2
Working set in S2, per mailbox spinlocks, per PCB spinlocks, LAN fastpath, scalable TCP/IP kernel
V8.2
IPF (obviously)
3


Configuration
Dedicated CPU Lock Manager
Keep it dedicated!
FastPath
Path balance
I/O Adapters/QBB
Write-back cache
On controllers - Use battery backup
On devices
Manually/explicitly set flags in disks; Often only viable for locally connected SCSI disks; For non-write-critical data
4


Wildfire
Keep memory close to processors as much as possible
Install images /RESIDENT if they are used by many processes or are performance critical
SDA> SHOW EXEC/SUMMARY and make sure that executive images are “sliced”
Evaluate RAD-specific processes/global sections
Memory reservation
XFC, Pool
5

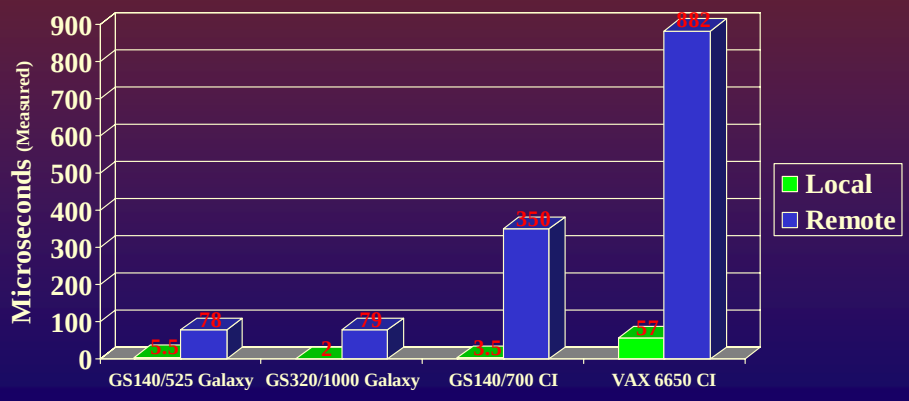

Marvel
“Don’t sweat the NUMA”
- Steve Hoffman Oct. 15th 13:29
RADs are likely not a worry
CMOS’s configuration suggestion:
Connect no IO to duo with primary CPU
Connect first IO7 to duo with CPU 2&3
Connect second IO7 to duo with CPU 4&5
Etc.
6

Locking

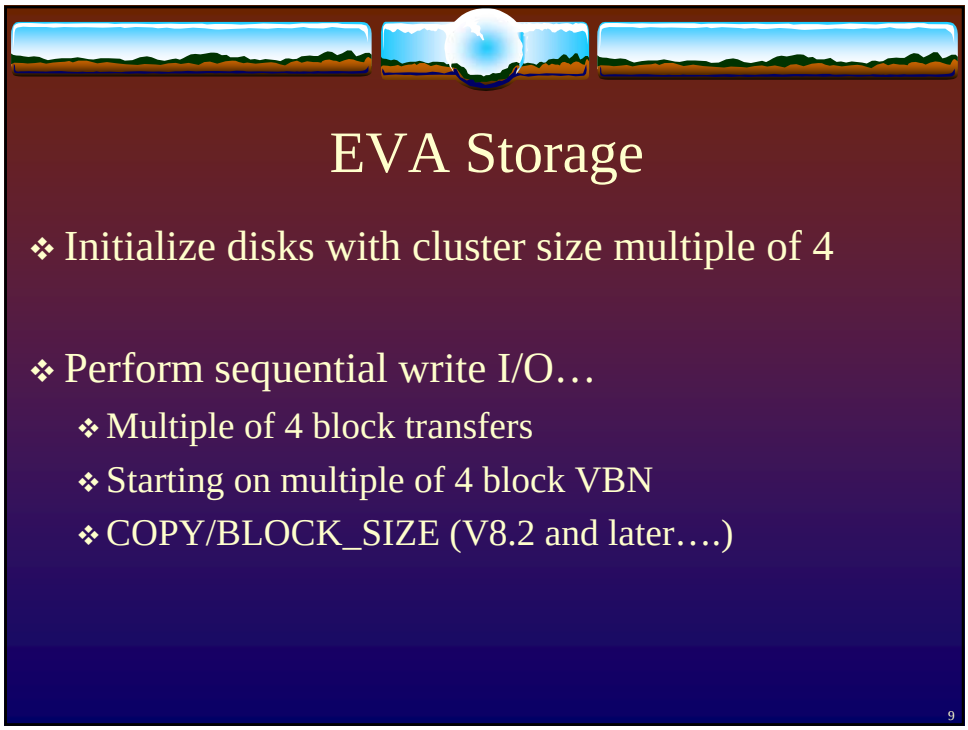
- Balance LOCKDIRW based on CPU power
  - Consider GS1280 clustered with VAX 6000-400
- **MIN_CLUSTER_CREDITS=128**
  - For big/fast machines
- **DEADLOCK_WAIT=1**
  - It ain't 1982 any longer

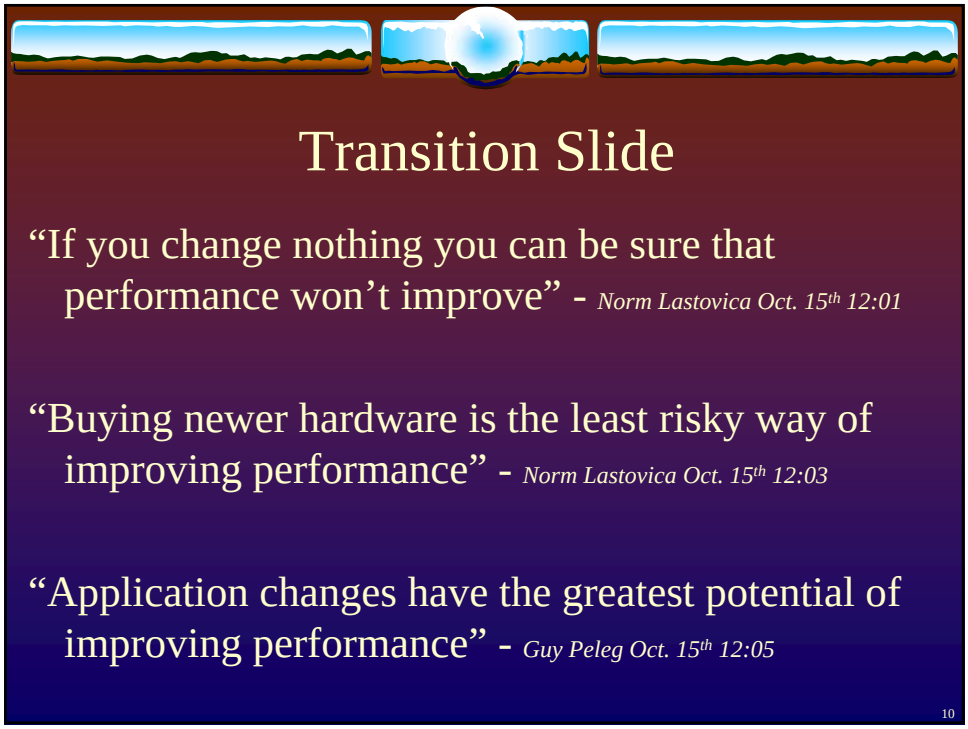
7

Local Locks are Faster Locks

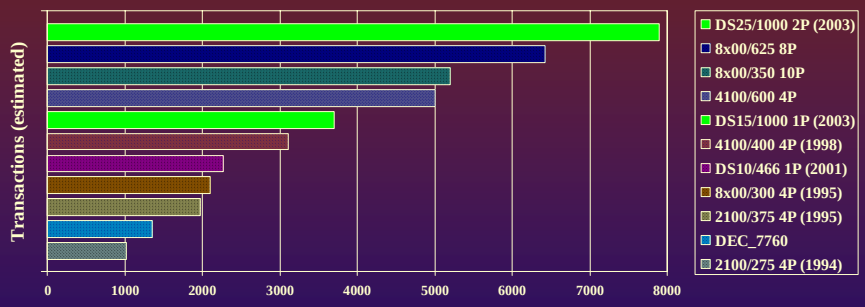


8


EVA Storage
Initialize disks with cluster size multiple of 4
Perform sequential write I/O…
Multiple of 4 block transfers
Starting on multiple of 4 block VBN
COPY/BLOCK_SIZE (V8.2 and later….)
9


Transition Slide
"If you change nothing you can be sure that performance won't improve" - Norm Lastovica Oct. 15th 12:01
"Buying newer hardware is the least risky way of improving performance" - Norm Lastovica Oct. 15th 12:03
"Application changes have the greatest potential of improving performance" - Guy Peleg Oct. 15th 12:05
10

Slowest Alpha today is faster than the fastest was 'back then'



- ***Significant*** I/O improvements - bus, disk, controller, network
- Often a good consideration
  - Low risk – cost savings?
  - Likely no application changes, no recompiles and no relinks required

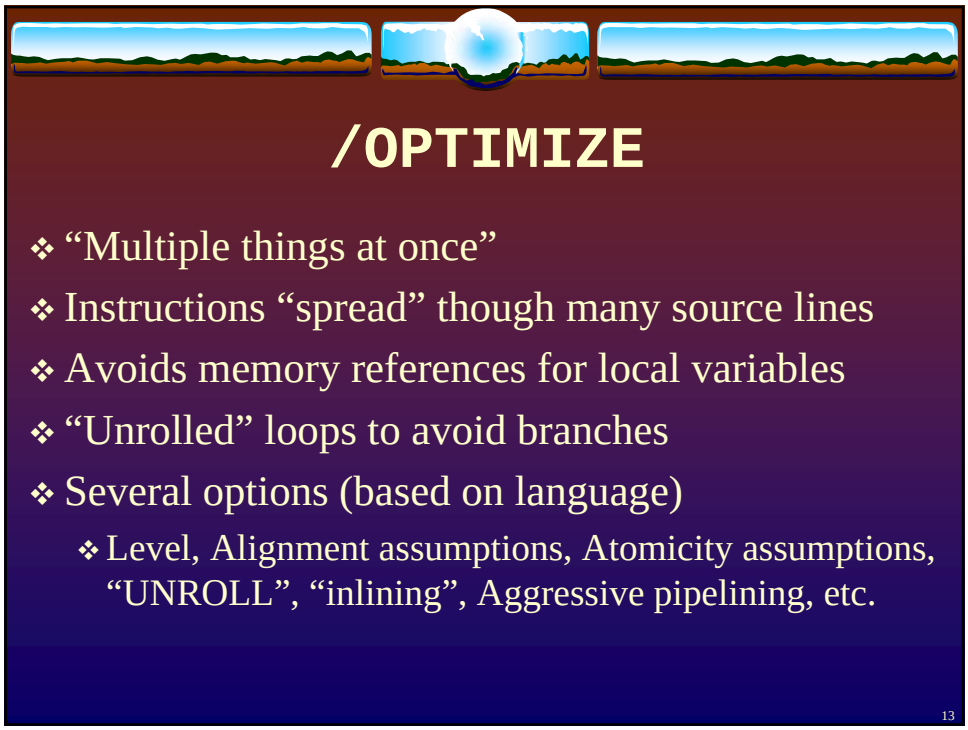
DS10L ~2" tall, uses 1.9A
DS25 ~9" tall, holds 16Gb

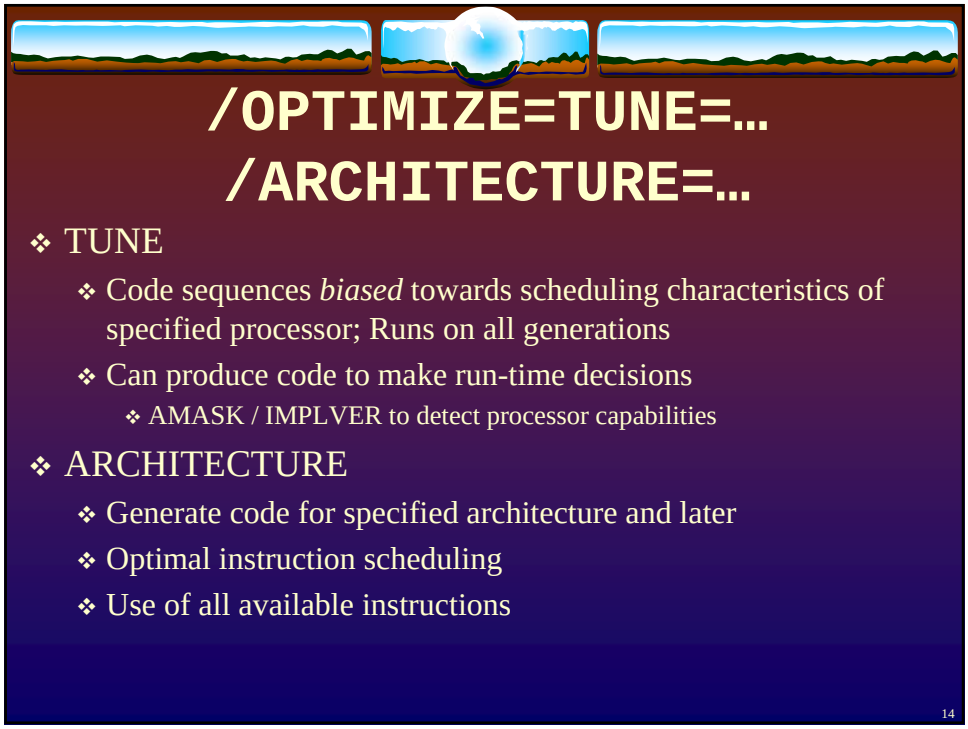
11

/NOOPTIMIZE

- Typically for debugging
- Many more memory references for local variables
- Longer instruction stream - "One thing at a time"
- Sometimes used to work around program bugs

12


/OPTIMIZE
❖ “Multiple things at once”
❖ Instructions “spread” though many source lines
❖ Avoids memory references for local variables
❖ “Unrolled” loops to avoid branches
❖ Several options (based on language)
❖ Level, Alignment assumptions, Atomicity assumptions, “UNROLL”, “inlining”, Aggressive pipelining, etc.
13


/OPTIMIZE=TUNE=…
/ARCHITECTURE=…
❖ TUNE
❖ Code sequences *biased* towards scheduling characteristics of specified processor; Runs on all generations
❖ Can produce code to make run-time decisions
❖ AMASK / IMPLVER to detect processor capabilities
❖ ARCHITECTURE
❖ Generate code for specified architecture and later
❖ Optimal instruction scheduling
❖ Use of all available instructions
14

Examples of …TUNE & /ARCHITECURE

- **/OPTIMIZE=TUNE=EV56**
  - Execute on all Alpha generations
  - Biased towards EV56
- **/OPTIMIZE=TUNE=EV6 /ARCHITECTURE=EV56**
  - Execute on EV56 and later (Byte/Word instructions)
  - Biased for EV6 (quad issue)
- **/ARCHITECTURE=EV6**
  - Execute on EV6 and later (Integer-Floating conversion, Byte/Word & Quad-issue scheduling)
- **/ARCHITECTURE=HOST**
  - Code intended to run on processors the same type as host computer
  - Eexecute on that processor type and higher

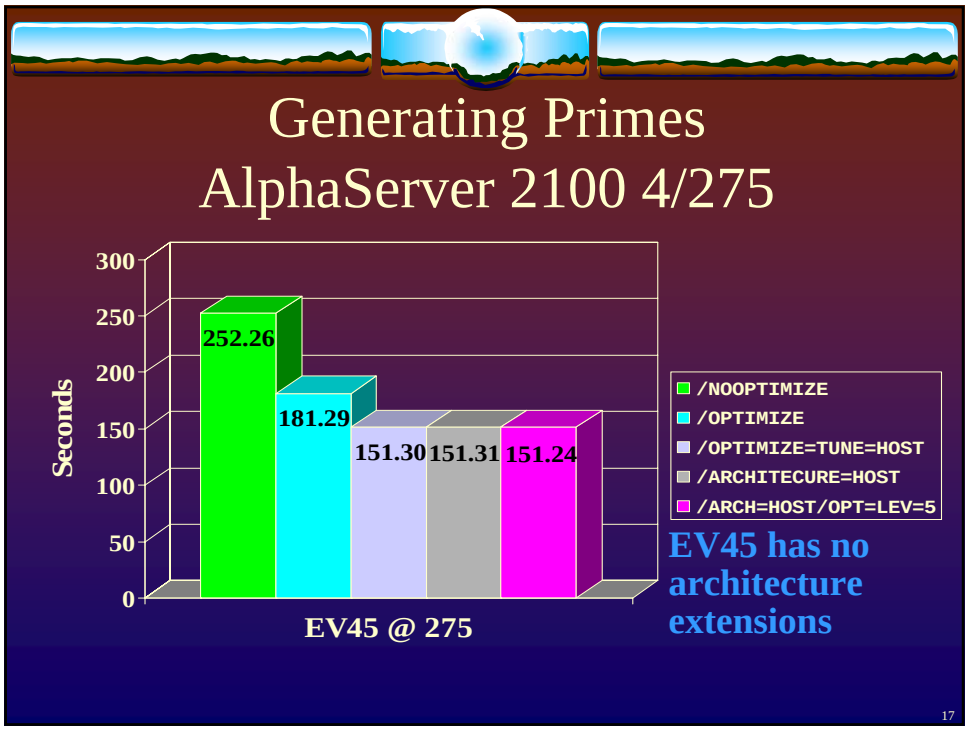
15

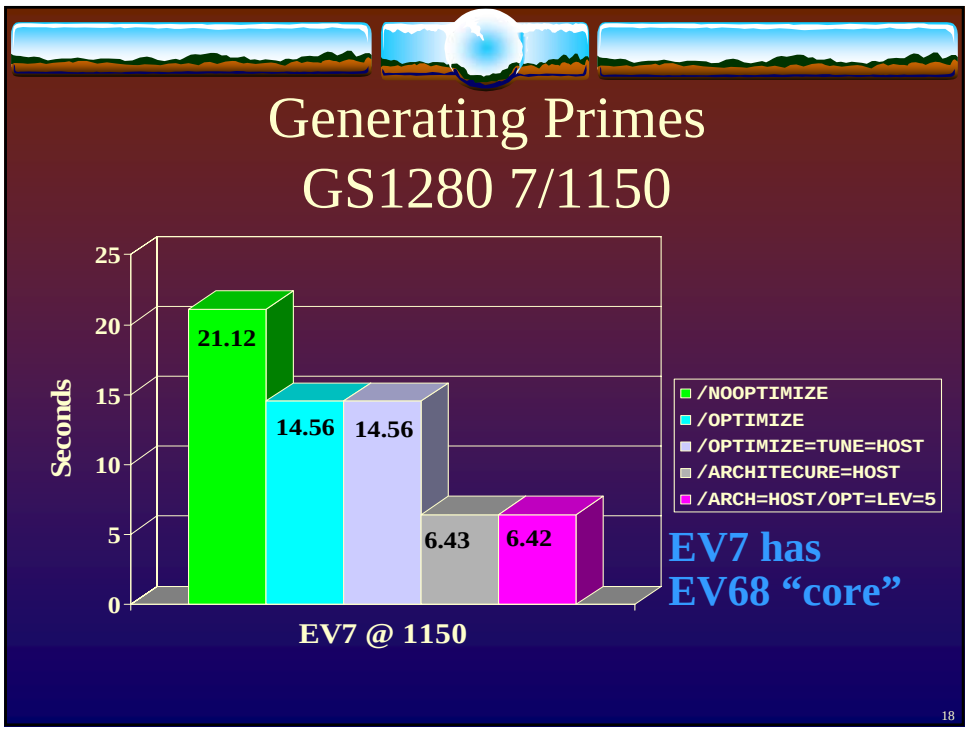
Prime Numbers Test

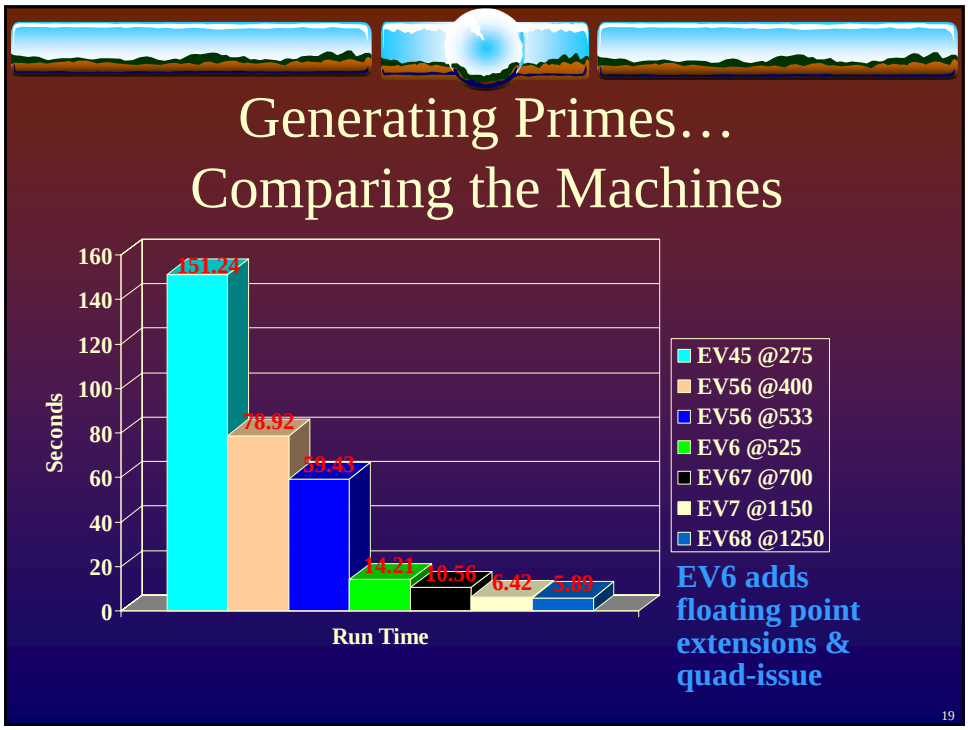
- Find first 1,000,000 prime numbers

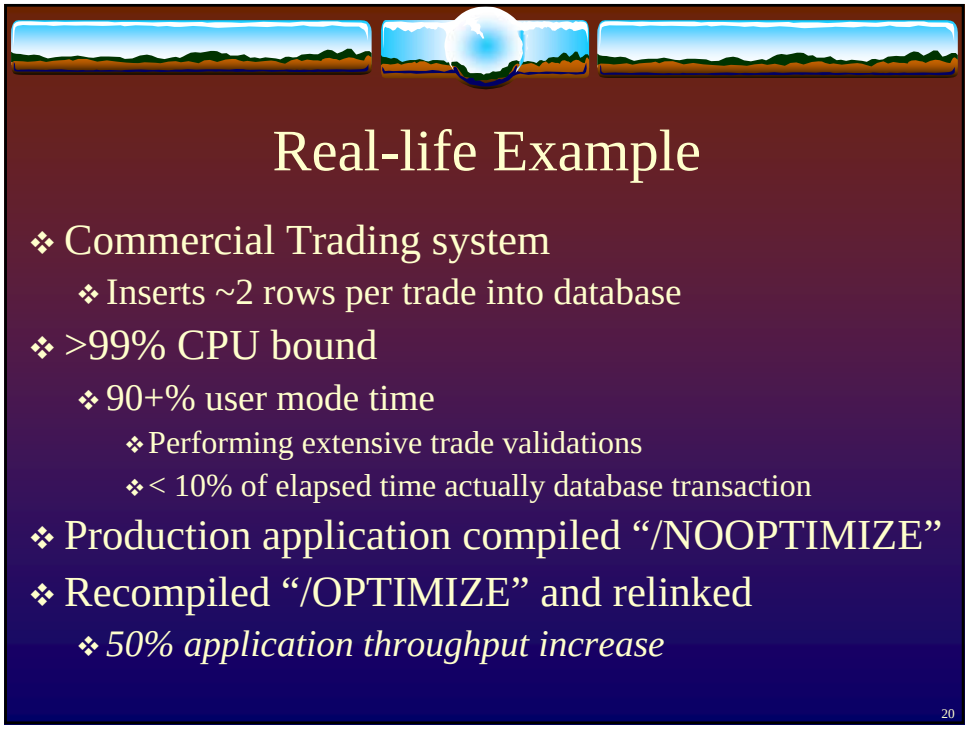
```
        primes(1) = 3
        hi_prime = 3
        hi_prime_index = 1
        hi_prime_divisor_index = 1
        do 100 i = 5,2000000000,2
        if (primes(hi_prime_divisor_index)**2 .lt. i)
            hi_prime_divisor_index = hi_prime_divisor_index + 1
        do 20 j = 1, hi_prime_divisor_index
        if (mod(i, primes(j)) .eq. 0) go to 100
20      continue
        hi_prime_index = hi_prime_index + 1
        primes(hi_prime_index) = i
        hi_prime = i
        if (hi_prime_index .eq. n_primes) go to 200
100     continue
200     …
```

16


Generating Primes
AlphaServer 2100 4/275
Seconds
300
250
200
150
100
50
0
252.26
181.29
151.30
151.31
151.24
EV45 @ 275
/NOOPTIMIZE
/OPTIMIZE
/OPTIMIZE=TUNE=HOST
/ARCHITECURE=HOST
/ARCH=HOST/OPT=LEV=5
EV45 has no architecture extensions
17


Generating Primes
GS1280 7/1150
Seconds
25
20
15
10
5
0
21.12
14.56
14.56
6.43
6.42
EV7 @ 1150
/NOOPTIMIZE
/OPTIMIZE
/OPTIMIZE=TUNE=HOST
/ARCHITECURE=HOST
/ARCH=HOST/OPT=LEV=5
EV7 has EV68 “core”
18


Generating Primes…
Comparing the Machines
160
140
120
100
80
60
40
20
0
Seconds
151.24
78.92
59.43
14.21
10.56
6.42
5.89
Run Time
EV45 @275
EV56 @400
EV56 @533
EV6 @525
EV67 @700
EV7 @1150
EV68 @1250
EV6 adds floating point extensions & quad-issue
19


Real-life Example
Commercial Trading system
Inserts ~2 rows per trade into database
>99% CPU bound
90+% user mode time
Performing extensive trade validations
< 10% of elapsed time actually database transaction
Production application compiled "/NOOPTIMIZE"
Recompiled "/OPTIMIZE" and relinked
50% application throughput increase
20

Linker Hints

- **LINK /VAX**
  - VAX 6650 - 153 seconds
  - GS1280 – 6 seconds

21

Images

- **$ PIPE SHOW DEV/FILE/NOSYS SYS$SYSDEVICE: | -**
  SEARCH SYS$INPUT: .EXE;
- Look for many copies of the same .EXE
- **$ INSTALL ADD …**
  - **/OPEN /SHARE /HEADER [/RESIDENT]**

22

RMS

- **SYSGEN SET RMS_SEQFILE_WBH 1**
- **SET FILE /STATISTICS & MONITOR RMS**
- Use larger buffers & more of them
- Specify FAB/RAB parameters:
  RAH, WBH, DFW, SQO, NOSHR, ALQ, DEQ, MBC, MBF
- RMS After Image Journaling
  - Data protection
  - RMSJNLSNAP freeware tool

23

Copying 800MB file disk to disk

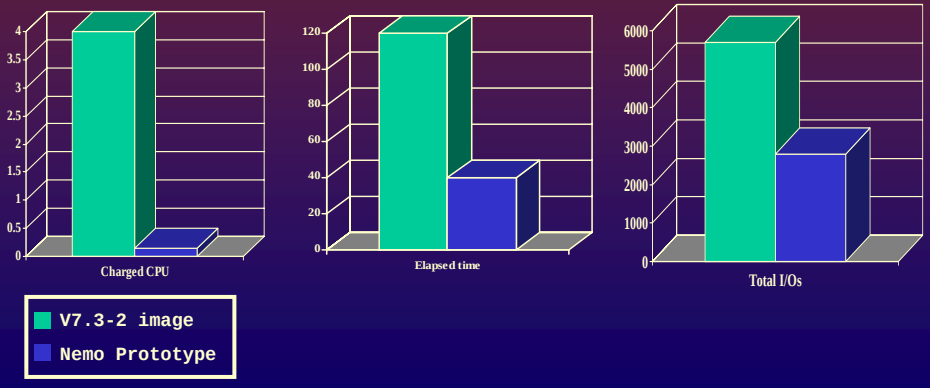
```
Accounting information:    ! VMS V7.3-1
  Buffered I/O count:              61     Peak working set size:        2352
  Direct I/O count:             51758     Peak virtual size:          168672
  Page faults:                    206     Mounted volumes:                 0
  Charged CPU time:     0 00:00:11.22     Elapsed time:     0 00:03:23.67

Accounting information:    ! VMS V7.3-2
  Buffered I/O count:              61     Peak working set size:        2480
  Direct I/O count:             26115     Peak virtual size:          168672
  Page faults:                    217     Mounted volumes:                 0
  Charged CPU time:     0 00:00:07.69     Elapsed time:     0 00:02:12.82
```

One line change – **RAB$B_MBC=127**

24

Sometimes….redesign is the answer

Copy 170MB file disk to disk



25

Indexed Files

- **ANALYZE /RMS /FDL**
- Evaluate larger bucket sizes
- “Null Key” can help you
- Long duplicate chains can kill performance
- Global buffers are a good thing

ask Thilo in the VMS clinic or anywhere else…

26

DECram

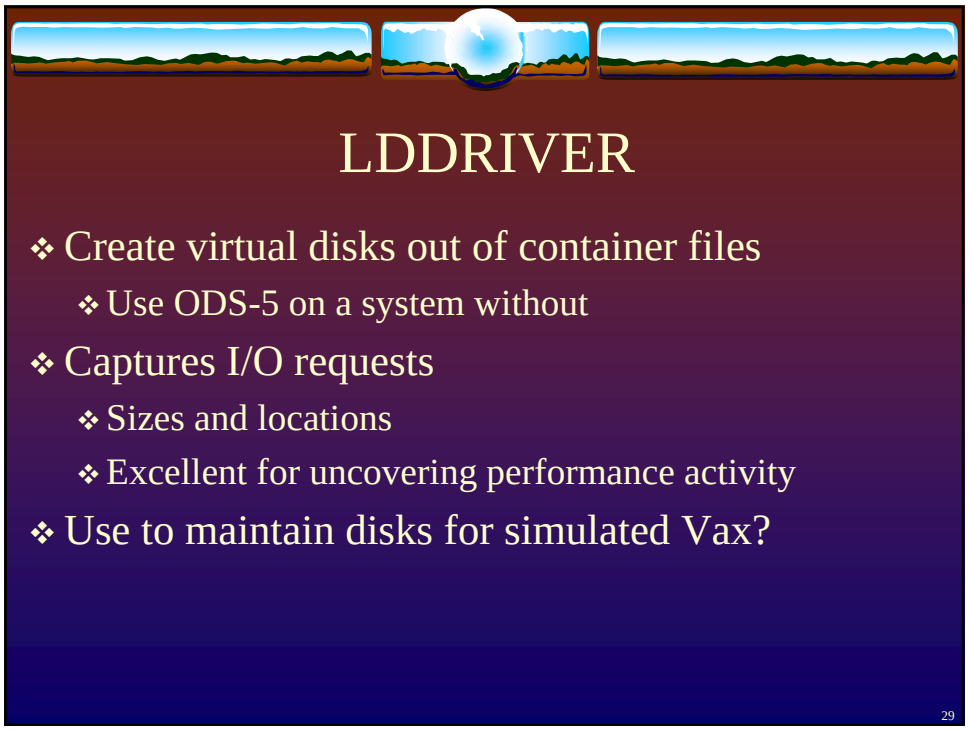
- Create virtual disk from system memory
- When temp/work files can not be avoided
- Integrated into OpenVMS V8.2
- May be shadowed with a physical disk
  - Shadowing smart enough to read from memory

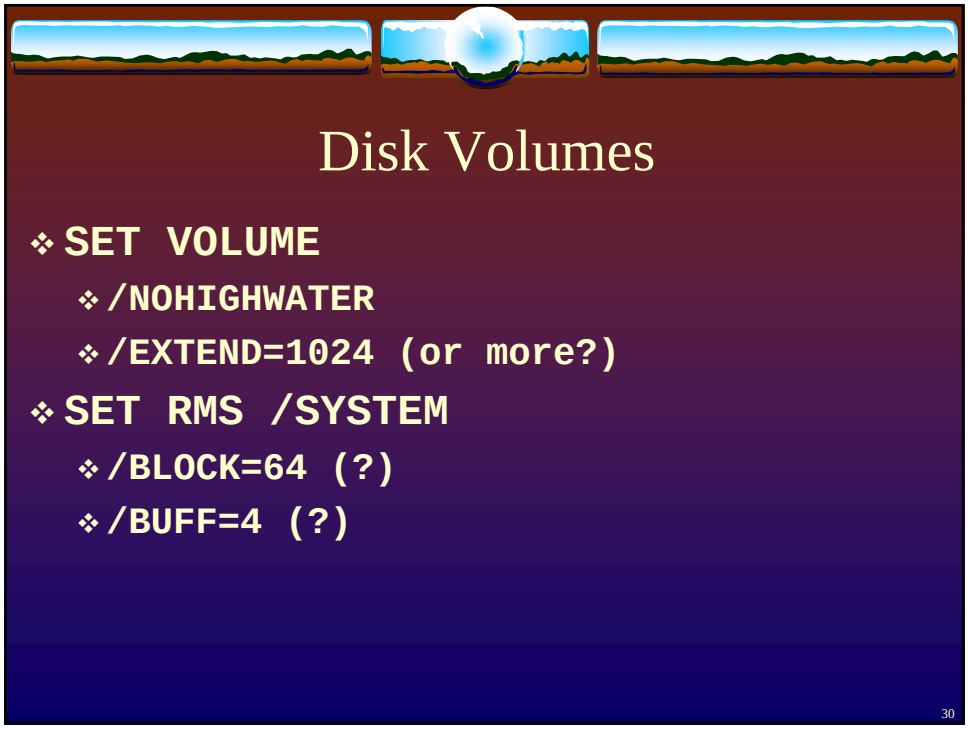
27

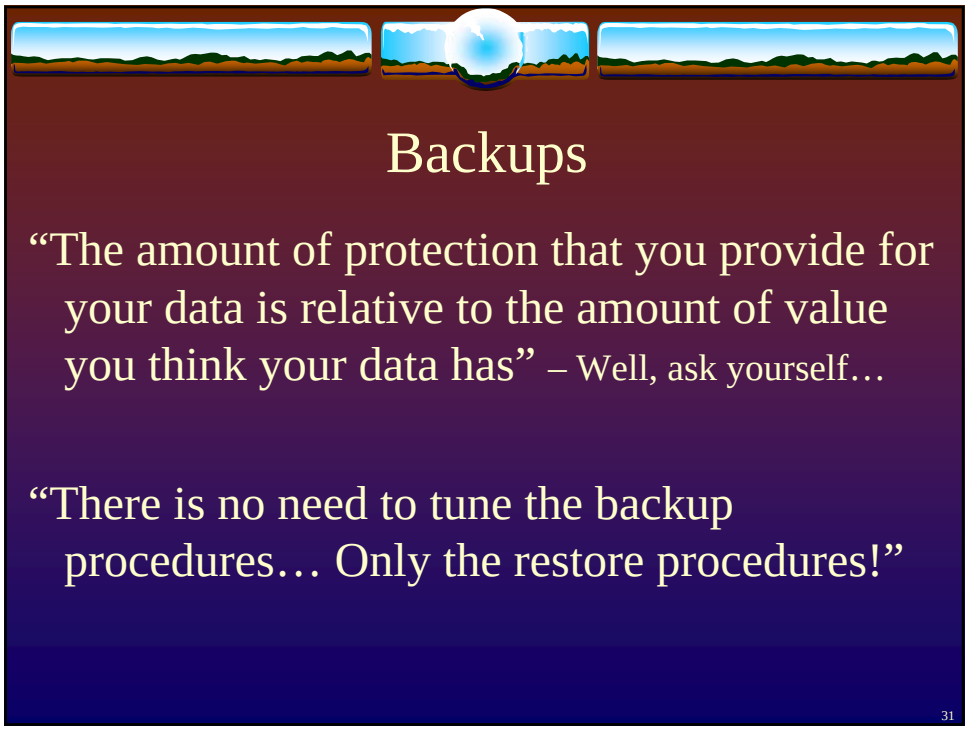
Software RAID

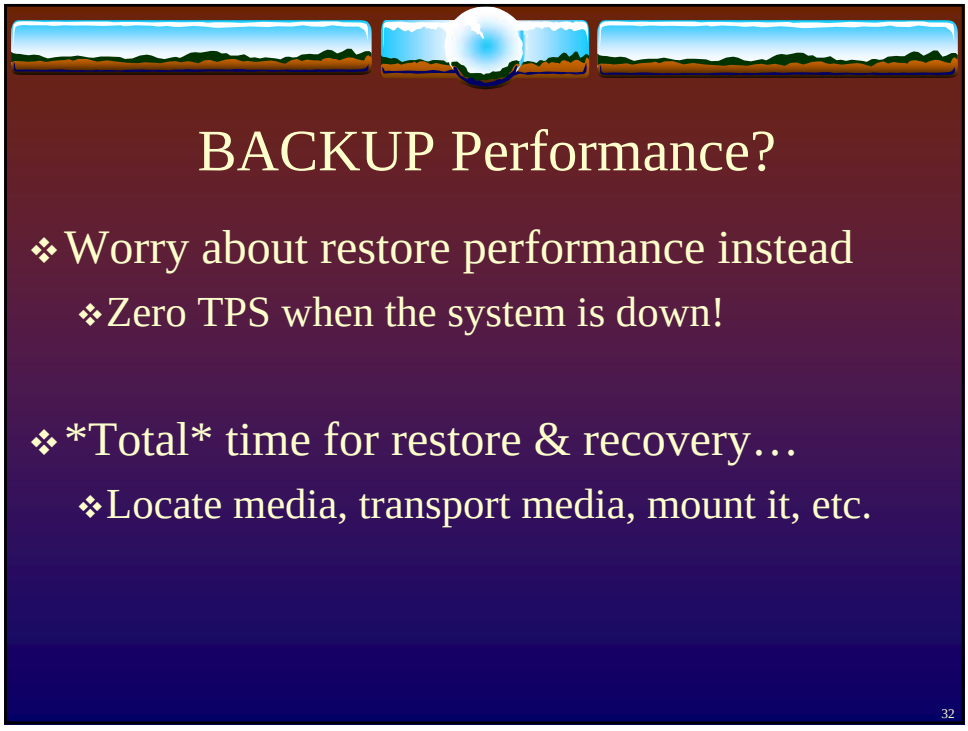
- Bind local disks into RAID (0 or 5) sets
- “Magically” distribute I/O load among spindles
- Partition RAID arrays into logical units
- Small CPU overhead vs. I/O distribution
- Or….Use hardware controllers

28


LDDRIVER
Create virtual disks out of container files
Use ODS-5 on a system without
Captures I/O requests
Sizes and locations
Excellent for uncovering performance activity
Use to maintain disks for simulated Vax?
29


Disk Volumes
SET VOLUME
/NOHIGHWATER
/EXTEND=1024 (or more?)
SET RMS /SYSTEM
/BLOCK=64 (?)
/BUFF=4 (?)
30


Backups
"The amount of protection that you provide for your data is relative to the amount of value you think your data has" – Well, ask yourself…
"There is no need to tune the backup procedures… Only the restore procedures!"
31


BACKUP Performance?
Worry about restore performance instead
Zero TPS when the system is down!
Total time for restore & recovery…
Locate media, transport media, mount it, etc.
32

BACKUP qualifiers

- /**CRC** /**VERIFY** – end-to-end protection
- /**JOURNAL** – so you can find files more easily
- /**TAPE_EXPIRATION** – avoid mistakes
- /**BLOCK_SIZE**=<large> for modern tapes
- /**MEDIA_FORMAT=COMPACTION** where possible
- /**GROUP=100**
  - Perhaps for tape drives that do additional data protection or for disk-based savesets

33

BACKUP /PHYSICAL

- Starting with V8.2 no longer requires identical sized input & output volumes

34

Online Indexed File Backup

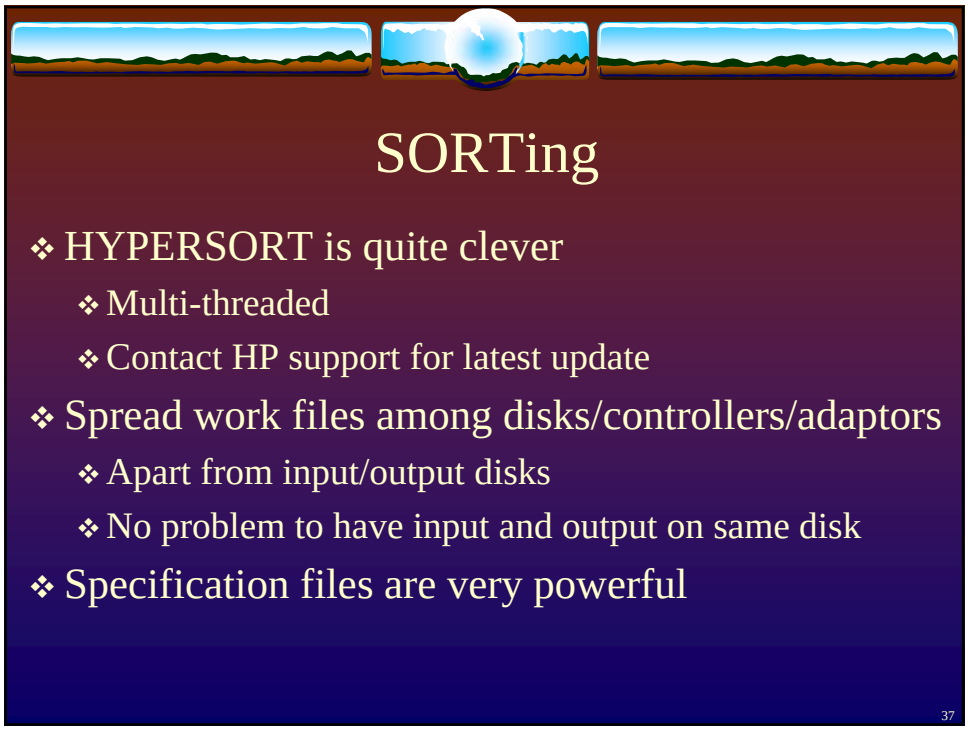
- **`CONVERT /SHARE`**
  - Record copy of an indexed file
  - Uncorrupted output file
- Run prior to online VMS backup for things like SYSUAF, NETUAF, RIGHTSLIST, etc.
- Discoordinated updates between files an issue

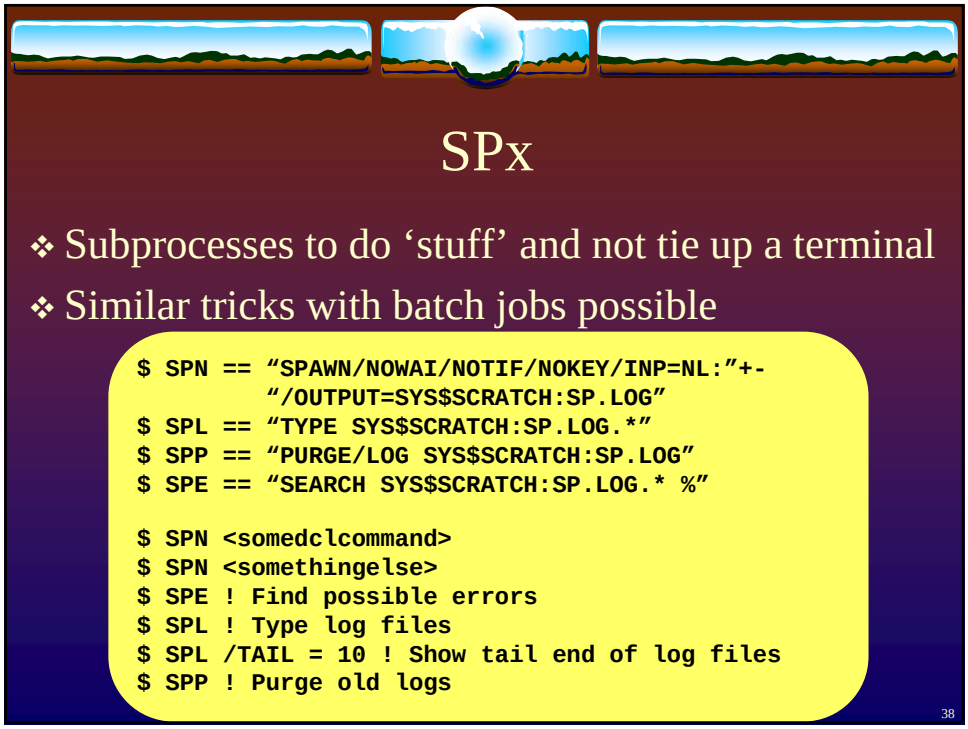
35

DELETE

- DELETE /LOG requires that files be opened prior to being deleted!
  - Can dramatically increase I/O

36


SORTing
❖ HYPERSORT is quite clever
❖ Multi-threaded
❖ Contact HP support for latest update
❖ Spread work files among disks/controllers/adaptors
❖ Apart from input/output disks
❖ No problem to have input and output on same disk
❖ Specification files are very powerful
37


SPx
❖ Subprocesses to do ‘stuff’ and not tie up a terminal
❖ Similar tricks with batch jobs possible
$ SPN == “SPAWN/NOWAI/NOTIF/NOKEY/INP=NL:”+-
“/OUTPUT=SYS$SCRATCH:SP.LOG”
$ SPL == “TYPE SYS$SCRATCH:SP.LOG.*”
$ SPP == “PURGE/LOG SYS$SCRATCH:SP.LOG”
$ SPE == “SEARCH SYS$SCRATCH:SP.LOG.* %”
$ SPN <somedclcommand>
$ SPN <somethingelse>
$ SPE ! Find possible errors
$ SPL ! Type log files
$ SPL /TAIL = 10 ! Show tail end of log files
$ SPP ! Purge old logs
38

Handy SDA commands

- Timer activities
  - **TQE LOAD**
  - **TQE START TRACE**
  - **TQE SHOW TRACE [/SUMMARY]**
- Locking activities
  - **LCK SHOW ACTIVE**
  - **LCK SHO LCK /INT=10/REP=10**

39

Handy SDA commands

Logical Name Translation

```
SDA> LNM LOAD
SDA> LNM START TRACE
SDA> LNM START COLL /LOGICAL
SDA> LNM SHO COLL
       Count    Logical Name
-----------    --------------------------
         324    TZ
         218    SYS$SYSROOT
         130    SYS$SHARE
         118    SYS$COMMON
          70    COSI_SRC
          68    SYS$DISK
          60    COSI$CMS
          56    SYS$SPECIFIC
          49    SYS$SYSTEM
          42    TCPIP$INET_DOMAIN
          31    PDEV$COSI
          30    GBL$INS$B3B500D0
SDA> LNM SHO TRACE ...
```

40

Handy SDA commands

FLT Alignment Fault Tracing

- Ideal is no alignment faults at all!
- Poor code and unaligned data structures do exist

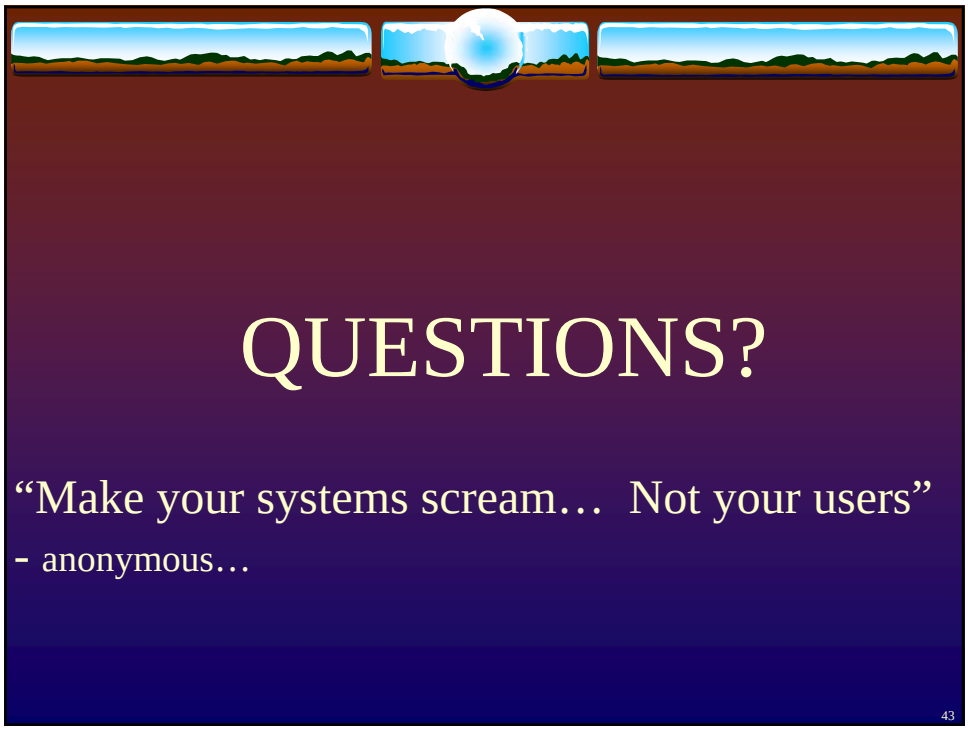
- Alignment fault summary…
  - `SDA> FLT START TRACE`
  - `SDA> FLT SHOW TRACE /SUMMARY`
  - `flt_summary.txt`
- Alignment fault trace...
  - `SDA> FLT START TRACE`
  - `SDA> FLT SHOW TRACE`
  - `flt_trace.txt`

41

Tools & FreeWare
Don't Leave Home Without…

- GREP
- AWK
- TECO
- RZDISK
- ICALCV
- MBU
- ZIP & UNZIP
- DFU
- AlphaPatch (or VMS 8.2)
- RMS_TOOLS
- Ethereal
  `(http://www.ethereal.com/)`

42


QUESTIONS?
“Make your systems scream… Not your users”
- anonymous…
43