DBMS SCALABILITY
THE MYTHS

David McGoveran
Alternative Technologies
13150 Highway 9, Suite 123
Boulder Creek, CA 95006
Telephone: 408/338-4621
www.AlternativeTech.com
mcgoveran@AlternativeTech.com

THE DBMS SCALABILTY PROJECT
AN ON-GOING STUDY

This Presentation Is Available From Our Website

www.AlternativeTech.com - see "Publications"

DBMS Scalability Report (*Overview* and *Full Report*) available too!

A Variety of Sources

- **CASE STUDIES, MARKET REQUIREMENTS, ANALYST REPORTS, USER SURVEYS BY OTHER GROUPS, 20 YEARS OF CLIENTS**

Purpose

- **EXPOSE NUMEROUS MYTHS, FALLACIES, AND FLIM-FLAM**
- **PROVIDE UNBIASED INFORMATION ABOUT SCALABILITY**

Primary Method

- **DETAILED SITE AUDITS *(NOT SURVEY AVERAGES!)***
- **IDENTIFICATION OF MYTHS BY COUNTER-EXAMPLES**

THE DBMS SCALABILTY PROJECT
AN ON-GOING STUDY

RDBMS Scalability Myths

"BELIEFS OR PERCEPTIONS WHICH, ALTHOUGH WIDELY HELD TO BE TRUE, ARE ACTUALLY MISSTATEMENTS OF THE FACTS."

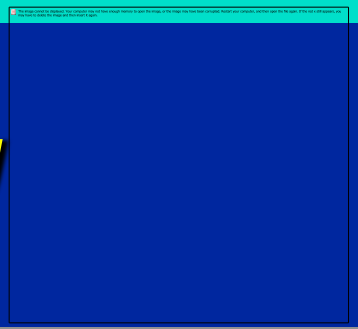
Case study sites

- PUSH SOME OR ALL RDBMS PRODUCT LIMITS
- ORACLE AND SYBASE DETAILED CASE STUDIES TO-DATE
- PRELIMINARY STUDIES OF INFORMIX, DB2, AND OTHERS
- ALL VENDORS WILL HOPEFULLY PARTICIPATE
- MANY PLATFORMS AND ARCHITECTURES (SMP, CLUSTER, "MPP")
- TP MONITORS NOT ADDRESSED *PER SE* (YET!)

Status

- INITIAL FOCUS ON SINGLE NODE NON-MAINFRAME RDBMS
- NOW EXPANDING TO MULTI-NODE

WE WANT YOU



- **Interested in being a participant in the DBMS Scalability Project?**
  - **VERY LARGE DATABASE (>500 GB)?**
  - **HIGH TRANSACTION RATES OR COMPLEX TRANSACTIONS?**
  - **LARGE USER POPULATION (>500 CONCURRENT USERS)?**
  - **DOES YOUR APPLICATION HAVE *ANY* OF THE ABOVE?**
  - **WANT A FREE 3-DAY AUDIT WITH RECOMMENDATIONS?**
    - » **If selected for the study, up to 3-days of onsite audit are free!**
- **Contact Alternative Technologies for more information**
  - **SEND E-MAIL TO: mcgoveran@AlternativeTech.com**
  - **TELEPHONE 408/338-4621**

MARKET REQUIREMENTS DEMAND OPEN-ENDED SCALABILITY

Four marketing requirements for open-ended scalability

1. **TENS OF THOUSANDS OF USERS ONLINE**
2. **VERY LARGE DATABASES**
3. **VERY HIGH TRANSACTION RATES**
4. **ELECTRONIC COMMERCE BUSINESS TRANSACTIONS**

($327 Billion by 2002 -- FORRESTER RESEARCH, INC.)

GOTCHA!

- Couldn’t build indexes or took too long
- Poor incremental CPU or node usage
- Too many users used up server memory
- Performance at 100GB was great, but when size tripled…
- Small amounts of wasted space added up
- 300 GB took over a terabyte of storage
- Transaction rates were way lower than TPC numbers…
- Our “scalable, parallel” DBMS didn’t scale
- We even went 3-tier and clustered, but...

THE TOP TWENTY MYTHS

MYTHS - STATED IN THIS AREA

REALITY - PRESENTED IN THIS AREA

SLIDES ARE CODED AS TO THE "REALITY SOURCE" IN THE LOWER LEFT. CODES FOR PRIMARY SOURCES WILL BE UNDERLINED, FOR SECONDARY SOURCES WILL BE *ITALICIZED*:

- **A - THE DATA IS BASED ON AUDITED SITES**
- **B - THE DATA IS BASED ON PUBLISHED SOURCES ABOUT SITES**
- **C - THE RELATIONSHIPS AMONG VARIABLES ARE THEORETICAL LIMITATIONS AND CANNOT BE CIRCUMVENTED**
- **D - THE DATA IS BASED ON VENDOR RECOMMENDED COMPUTATIONS**
- **E - THE DATA IS BASED ON OTHER ANALYST REPORTS**
- **F - THE DATA IS BASED ON TPC RESULTS**
- **G - THE DATA IS BASED ON UNAUDITED SITES (E.G., INTERVIEWS)**

1. MANY OPEN SYSTEMS DATABASES ARE IN PRODUCTION WITH A TERABYTE (OR MORE) OF DATA

APPORTIONMENT OF REPORTED SIZES

INDEXES

DATA

ADMIN

MIRROR AND FREE SPACE

D, A-C,E-F

FOOTNOTE
Ratio: Required Disk to Raw Data

- **Oracle*: 3.61, 6.43, 5.94, 6.59, 5.12**
- **DB2/6000*: 3.77**
- **Teradata*: 8.80, 2.93, 3.28**
- **Non-stop SQL*: 2.86**
- **Informix: 2.5**
- **Sybase: 2.5**

***Data from S. Brobst, *Taming the Data Giants*, DBPD -- computed from published TPC numbers. All other data based on audited sites and vendor recommended computations.**

FOOTNOTE
State of the Art *Single* Database Size

- **Oracle: 1.054 TB total disk space, but...**
  - ABOUT 700 GB DATA OR 300-350 GB RAW DATA
  - 2.4 TB REPUTED, BUT NOT VERIFIABLE (2/15/98)
- **Sybase: 511 GB total disk space**
  - APPROXIMATELY 300 GB OF RAW DATA
  - 3.2 TB TPC, 1.4 TB USER, BUT NOT IN PRODUCTION (10/15/97)
- **Informix: 500 GB data reported**
- **Teradata: 870 GB data reported**
- **DB2/6000: 250 GB estimated (1.13 TB on MPP only)**
- **DB2/MVS: 700 GB estimated**

CORRECTIONS ARE INVITED (MUST BE VERIFIABLE)!

2. DBMSs CAN BE PRODUCED AND CONSUMED AS COMMODITIES

Success factors increasingly obscure

- **DIFFICULT TO IDENTIFY**
- **VERY COMPLEX**
- **EASE OF USE HIDES COMPLEXITY**
- ***VLDB SITES MOST OFTEN FAIL DUE TO M ISMATCH BETWEEN REQUIREMENTS, FUNCTIONALITY, AND USE!***

Implementations differ in important ways

- **BACKUP**
- **DEADLOCK DETECTION**
- **TRANSACTION ISOLATION**

Customers use DBMS specific "workarounds"

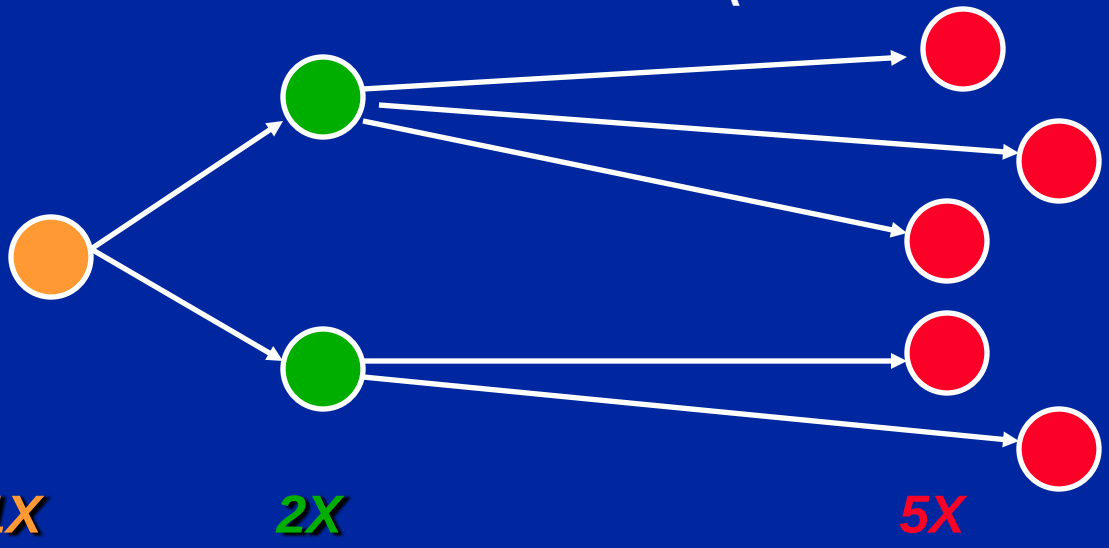
- ***MANY UNSOLVED DBMS SCALABILITY PROBLEMS!***

B, A-G

3. PHYSICAL TRANSACTION RATES MEASURE WORKLOADS FOR SCALABILITY

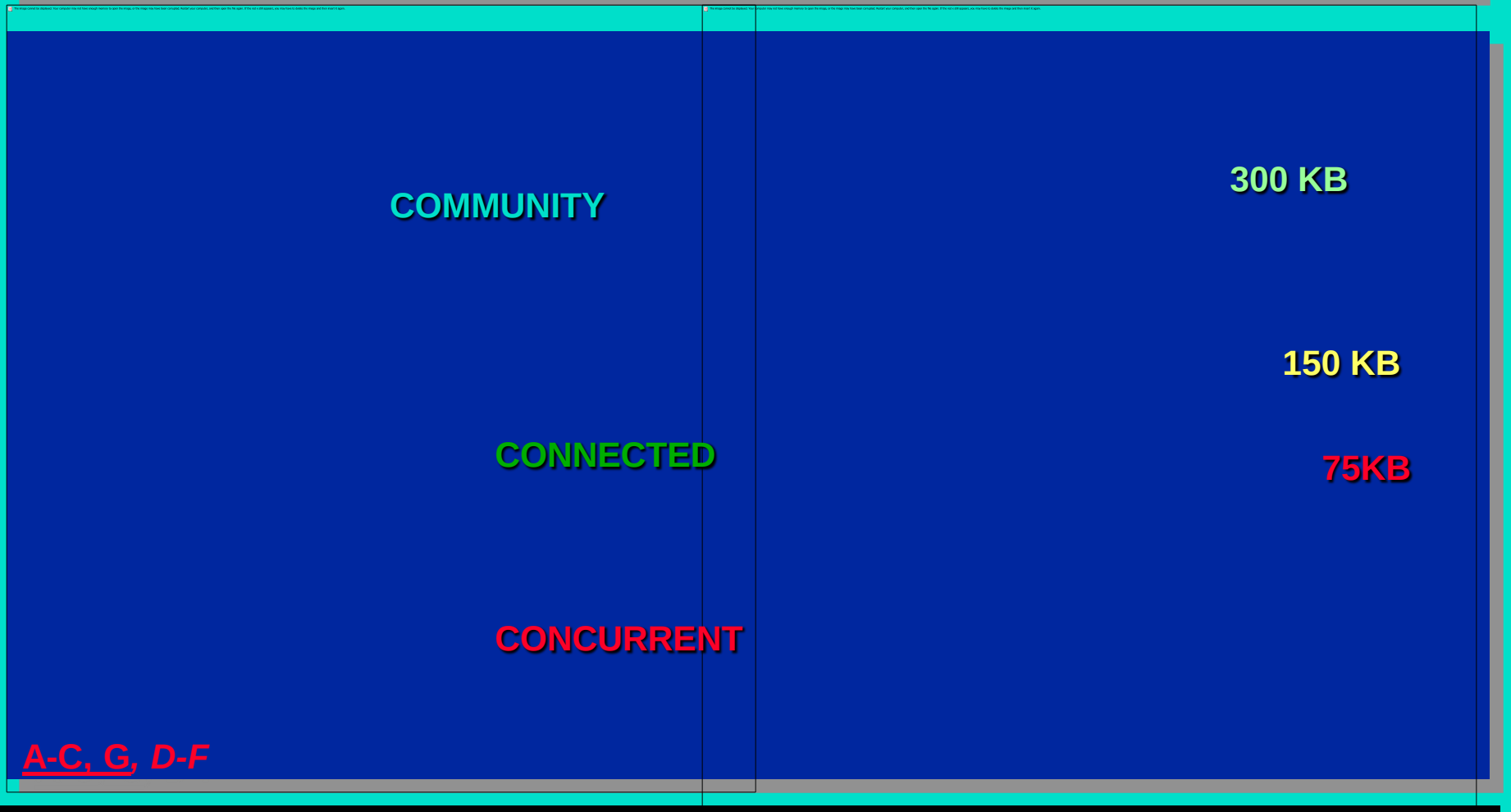
ONLY *BUSINESS* TRANSACTIONS (UNIT OF AUDIT) ARE IMPLEMENTATION INDEPENDENT

- *VERSUS LOGICAL TRANSACTIONS (UNIT OF CONSISTENCY)*
- *VERSUS PHYSICAL TRANSACTIONS (UNIT OF RECOVERY)*



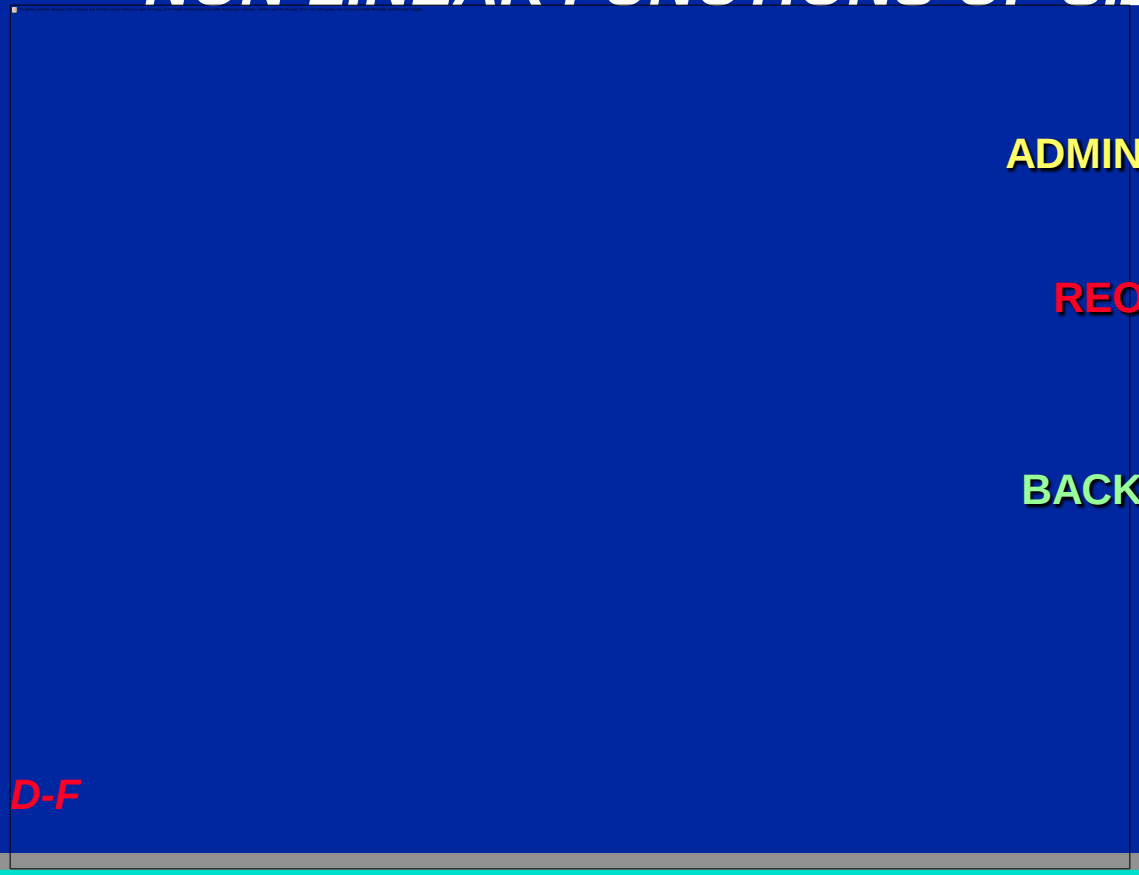
A-C, G, D-F

4. NUMBERS OF USERS SUPPORTED IS A MEASURE OF SCALABILITY



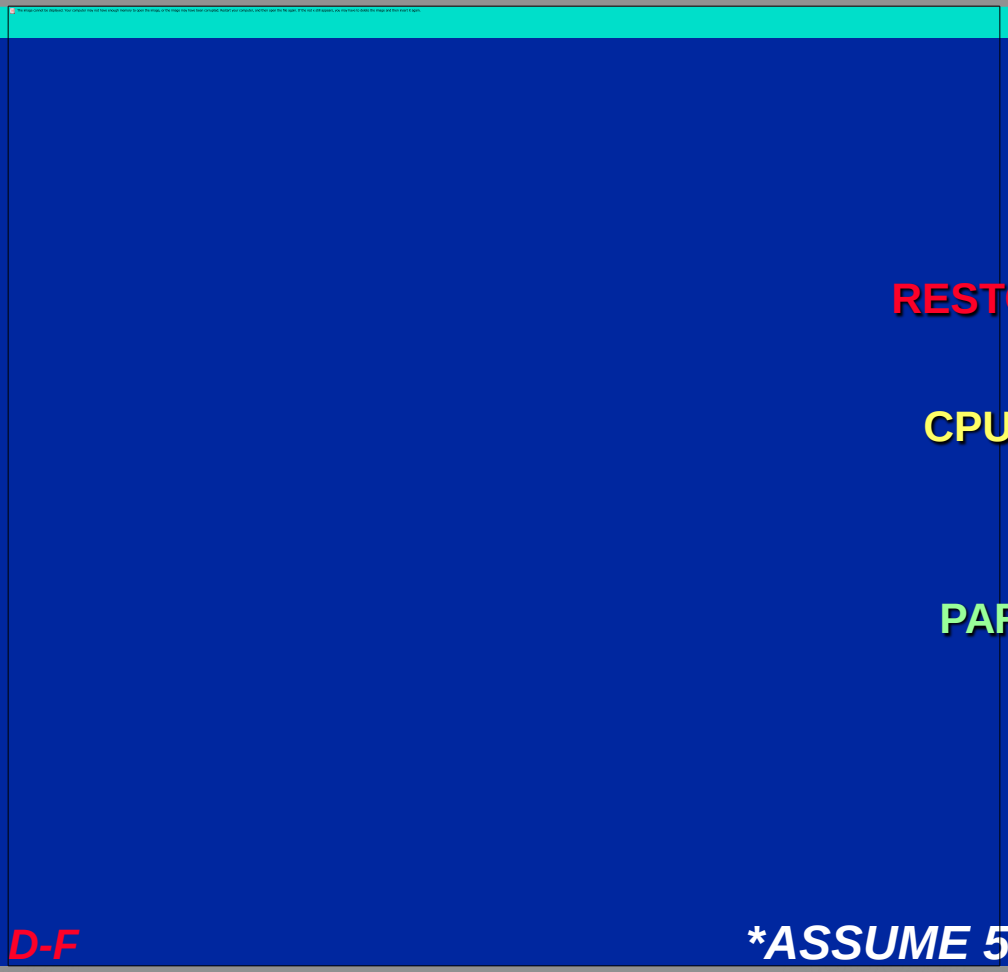
5. SPEED OF ADMINISTRATIVE OPERATIONS DETERMINES ADMINISTRATIVE SCALABILITY

NON-LINEAR FUNCTIONS OF SIZE



A, C, G, B, D-F

6. PARTIAL DATABASE OPERATIONS PROVIDE ADMINISTRATIVE SCALABILITY



RESTORE COMPLEXITY

CPU UTILIZATION

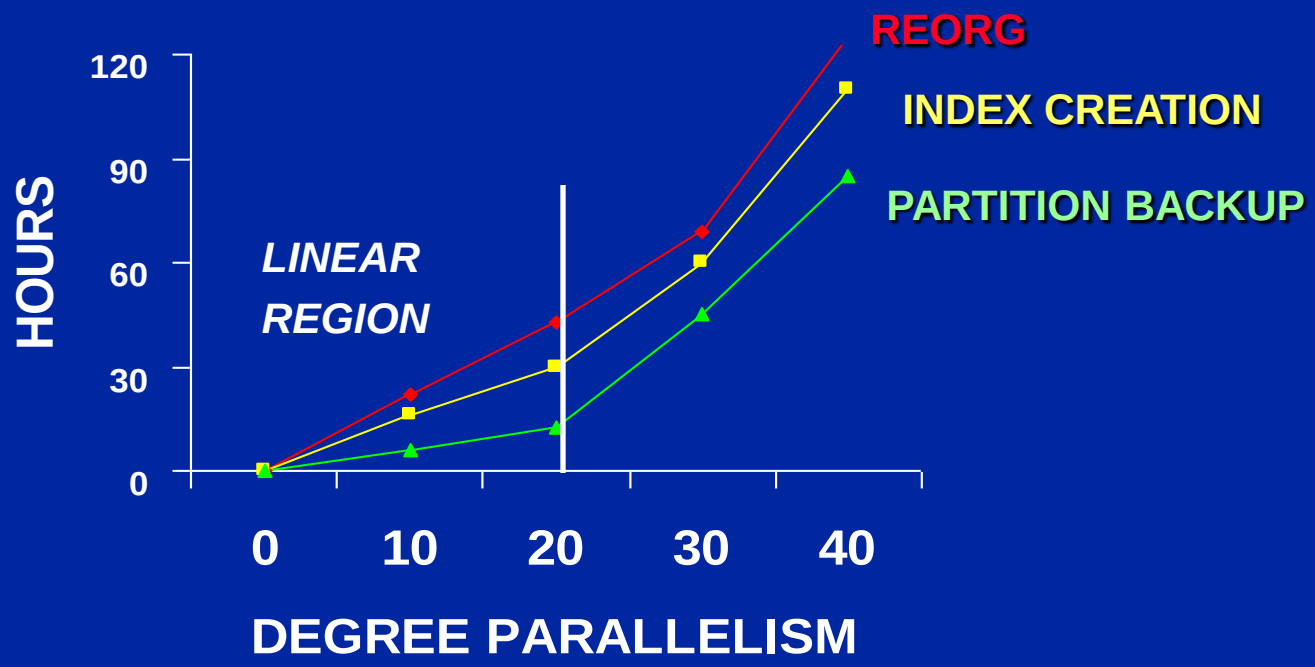
PARTITION BACKUP

A, C, G, B, D-F

****ASSUME 50 GB PARTITIONS***

7. PARALLELISM IS NECESSARY FOR SCALABILITY

CPUs LIMIT LINEARITY



A, C, G, B, D-F

**ASSUME 20 CPUs*

8. STORAGE ADDRESSABILITY IS AN INDICATION OF DBMS SCALABILITY AND VALUE

SUCH LIMITS ARE RARELY TESTED BY THE VENDOR

- TOO COSTLY ($1 MILLION PER TERABYTE)
  - » LOWER COST / GB DOESN'T HELP: HIGHER DENSITIES = HIGHER I/O RATE = LOWER MTBF PER DRIVE
- TOO HARD TO BUILD
  - » EVER FAKE A TERABYTE OF DATA?
- TOO MANY PERMUTATIONS TO TEST

OTHER LIMITS ENCOUNTERED FIRST

OPEN FILES, # SEMAPHORES, # SOCKETS, …

A, C, G, B, D-F

9. HOW THE SPACE IS USED DOESN'T MATTER, AS LONG AS THE DBMS CAN MANAGE IT

2 BILLION ROW TABLE WITH ONE INDEX

(tran_id, tran_date, tran_amount)

DATA:

42 BYTES NATIVE

55 VS. 46 DB FORMAT

BEFORE MIRRORING (INDEX, ADMIN):

526,223,576,699 VS. 255,668,840,309

AFTER MIRRORING:

1,052,447,153,398* VS. *511,337,680,618

D, A-C, E-F

10. DATA AND INDEX SPACE SUPPORT PROVES THE ABILITY OF A DBMS TO SUPPORT LARGE DATABASES

Other space requirements are important too:

- **TEMPORARY SPACE (SORT / REORG), RECOVERY OR LOG SPACE, REDUNDANCY FOR PERFORMANCE, REDUNDANCY FOR AVAILABILITY**

Even if space is supported, non-linear operational issues often dominate. These include:

- **DESIGNING / CONTROLLING TRANSACTION ISOLATION**
- **RO TRANSACTION MANAGEMENT OVERHEAD (IF READ-CONSISTENCY IS REQUIRED)**
- **INCREASING DEADLOCK PROBABILITIES - AVOID, DETECT, AND RESOLVE**
- **ALLOCATION ERRORS AND RECOVERY *(NASTY!)***
- **SPACE MANAGEMENT / ORGANIZATIONAL COMPLEXITY**

A-C, G, E, F

11. DATABASE PARTITIONING CIRCUMVENTS PRODUCT SCALABILITY LIMITATIONS

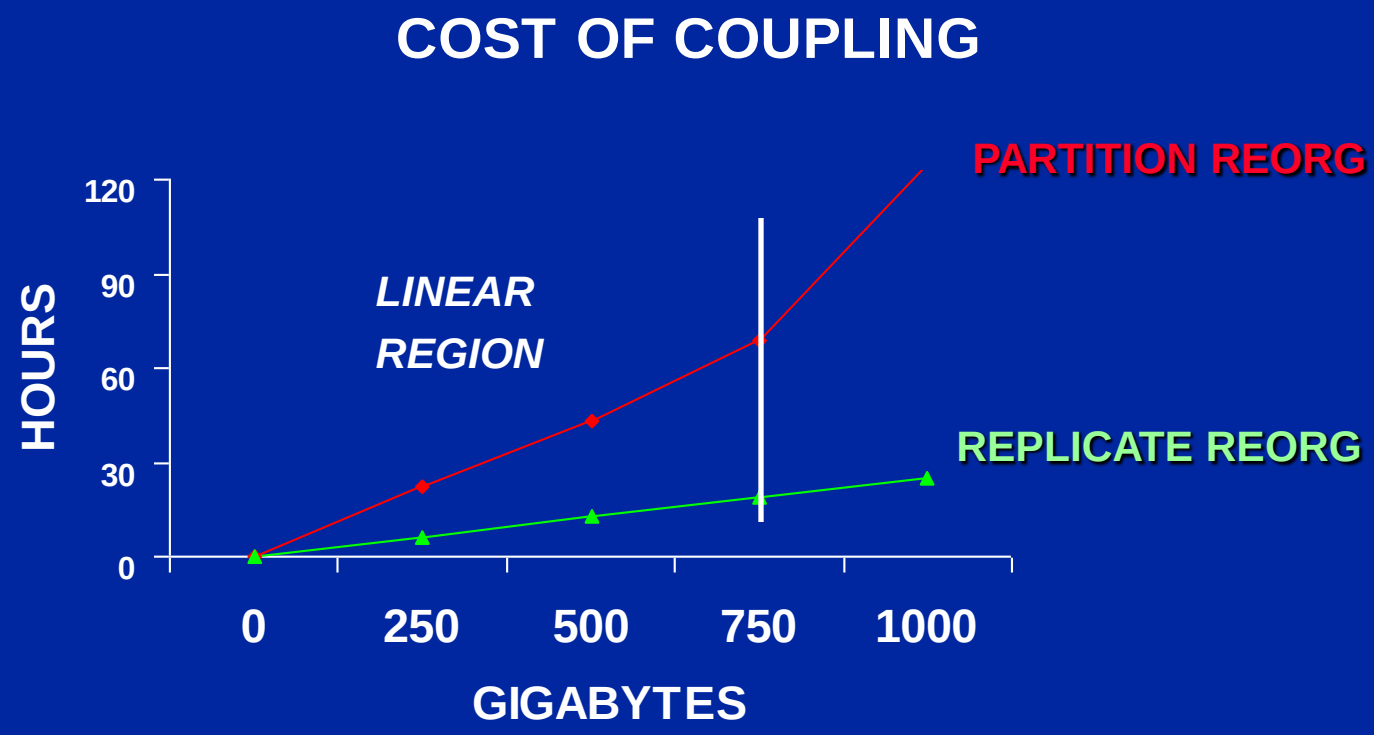
Most large databases are partitioned, but NOT for scalability!

Non-scalability reasons:

- **PLATFORM LIMITATIONS PRECLUDE DBMS SCALABILITY (E.G., 2 GB FILE LIMITATIONS).**
- **SO ADDITIONAL PARTITIONED SYSTEMS CAN BE ADDED ONLINE**
  - » **A BUSINESS AVAILABILITY REQUIREMENT**
- **BUSINESS IMPACT MINIMIZED ON FAILURES (RELIABILTY)**
- **INDIVIDUAL PARTITIONS CORRESPOND TO DISTINCT ASPECTS OF BUSINESS PROCESSING**
- **INDIVIDUAL PARTITIONS CORRESPOND TO DISTINCT PROJECTS**
- **EXISTING STOVEPIPE APPLICATIONS DICTATE THAT PARTITIONS CORRESPOND TO BUSINESS AND POLITICAL DIVISIONS**

A, C, G, B, E-F

12. REPLICATES ARE MORE DIFFICULT TO REORGANIZE THAN TABLE PARTITIONS



A, C, G, *B*

13. MULTIPLE DATABASES / SERVERS ARE WORKAROUNDS FOR DBMS DEFICIENCIES

Multiple databases and/or servers are methods to partition a database and maintain cohesiveness

- **MOST OFTEN BECAUSE IT FITS THE BUSINESS MODEL**
- **A SINGLE, INTEGRATED DATABASE WOULD NOT MEET BUSINESS REQUIREMENTS (EVEN IF THE DBMS COULD SUPPORT IT.)**

A, G, B

14. ASYNCHRONOUS REPLICATION IS USED TO COUNTERBALANCE SCALABILITY LIMITATIONS

Provides cohesiveness among database (versus table) partitions

One technique of many

Best used where tight integration is undesirable or impossible

Sites attempting to use *for scalability* quickly discovered it does not work!

Most often used to improve availability, not scalability

A, C, G, B, E

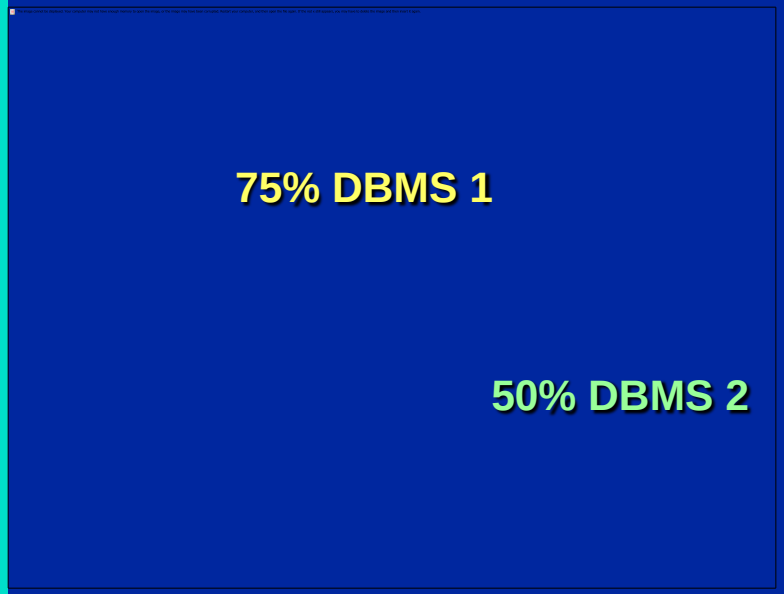
15. DBMS CLUSTERING IS AN IMPORTANT SCALABILITY SOLUTION

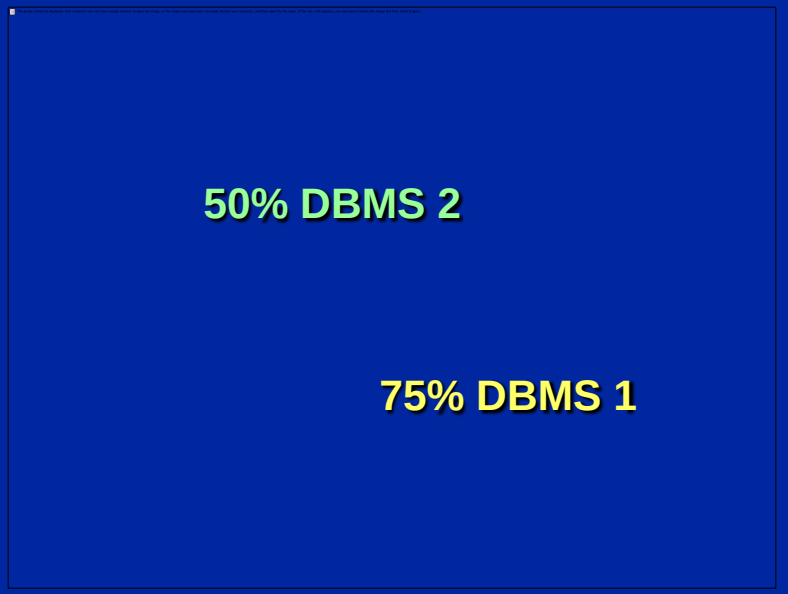
- **Clustering primarily provides, and is used for, high availability**
- **Designers must exercise great care to obtain even moderate scaleup or speedup from cross-node cluster resources**
  - **CLUSTER "SCALEUP" TYPICALLY < 60% OF SMP SYSTEMS!**
- **Designed more like a federation of loosely coupled physical databases**
- **Costs include design time, additional administration, possibly coding, and lock or cache coherence management**

A-C, F, G, D

16. THE MORE SCALABLE THE SYSTEM, THE MORE EFFICIENT AND COST EFFECTIVE THE PRODUCT

WHAT DOES PERCENT SCALABILITY MEAN?

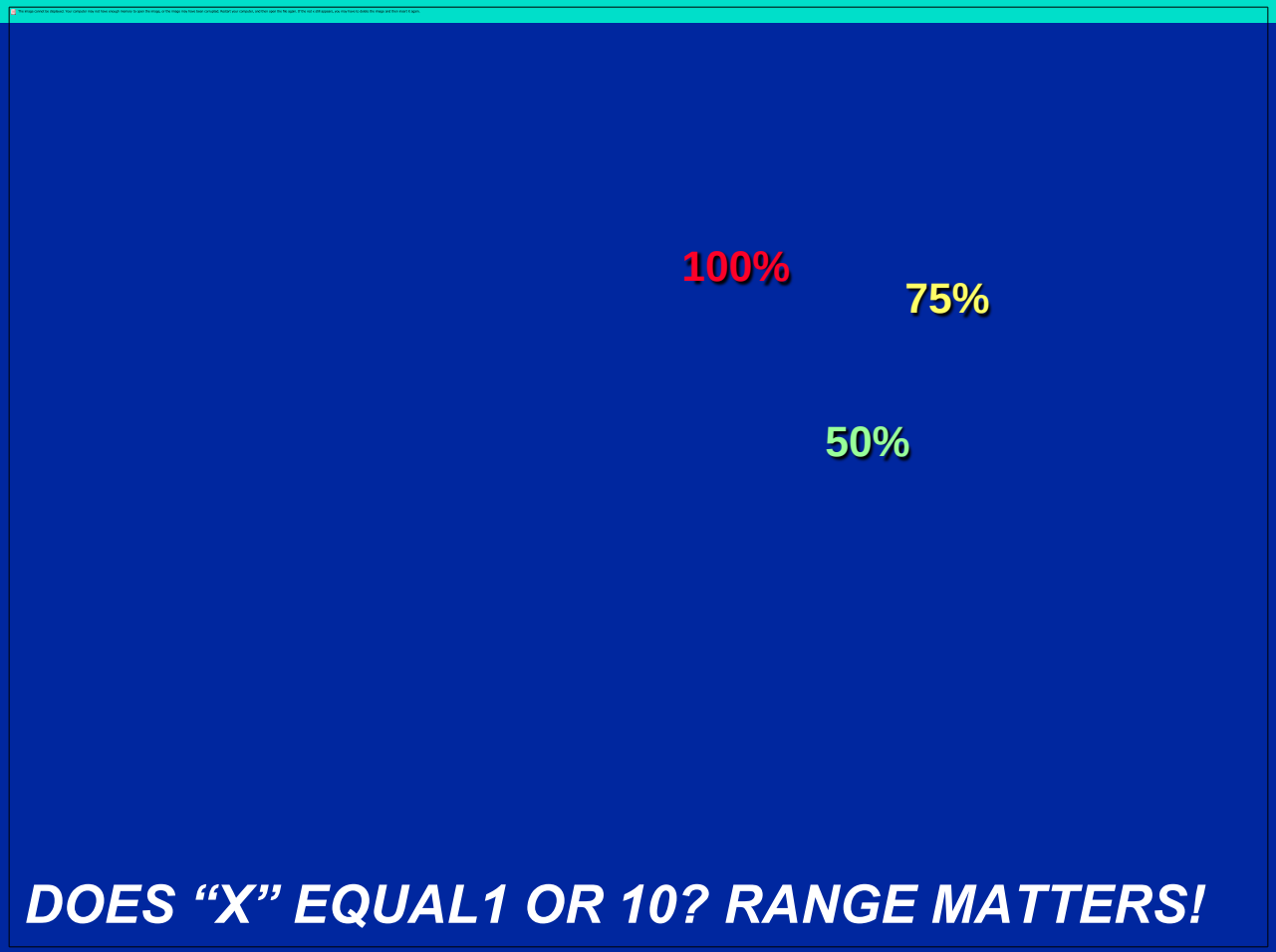




UNLABELED GRAPH **X = 1**

C, B, F

17. A DBMS EITHER IS OR IS NOT SCALABLE

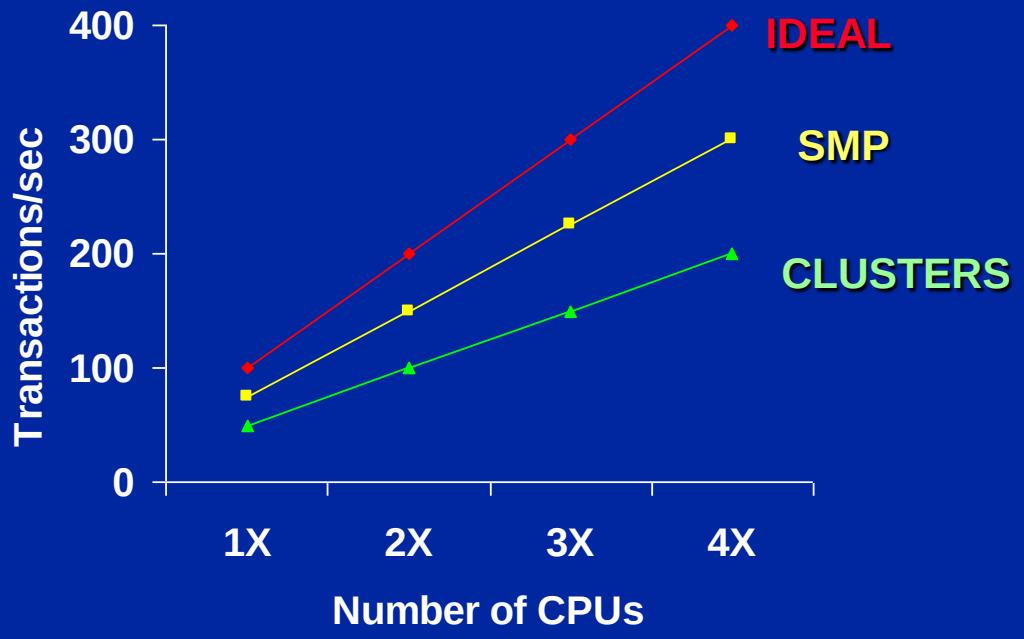


C, G, F

18. SCALEUP OR SPEEDUP CAN BE PROVEN BY EXAMPLE

DBMS SCALEUP AND SPEEDUP ARE:

- PLATFORM AND APPLICATION SPECIFIC
- STRONGLY AFFECTED BY TRANSACTION AND DB DESIGN



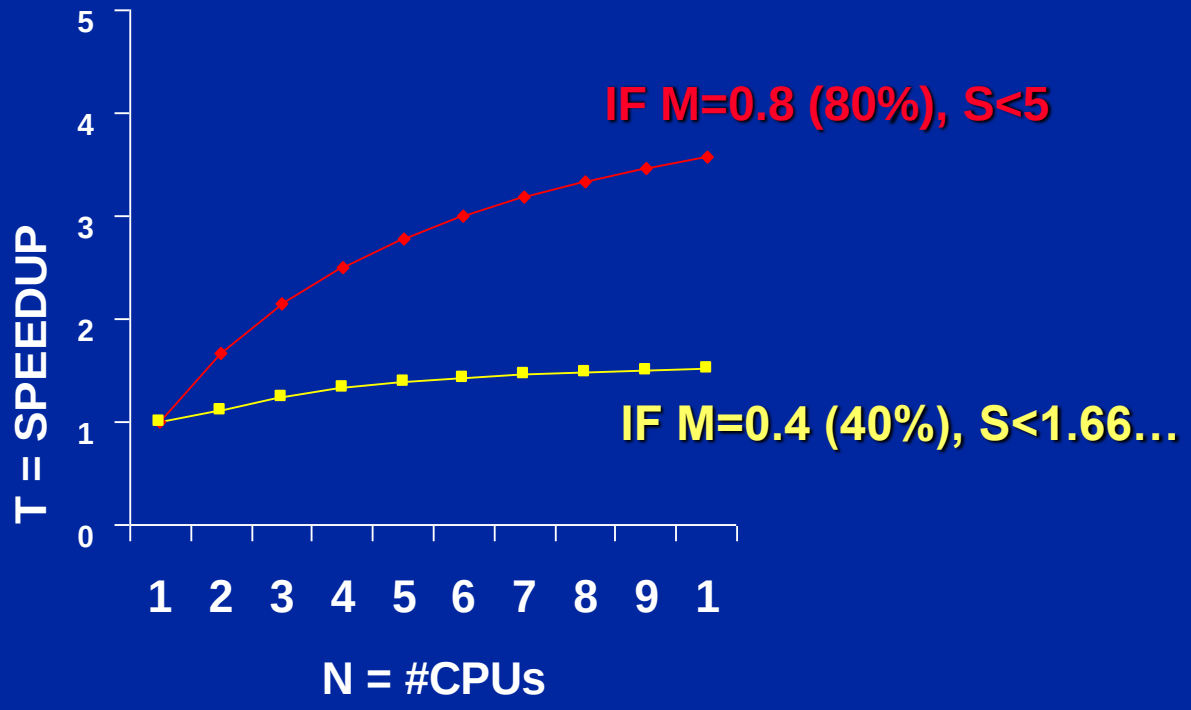
A-C, G, F

Transaction rate versus CPUs

19. GOOD PROCESSOR SCALABILITY CAN PROVIDE ARBITRARY SPEEDUP

PROCESSOR SPEEDUP (S) FOLLOWS AMDAHL'S LAW:

$$S = 1 / ((1 - M) + (M / N))$$



M = RATIO OF TIME TASKS RUN PARALLEL

C, B, E-F

20. OBJECT SUPPORT WILL HELP THE PLIGHT OF VLDB!

The Reality Will Be *Decreased:*

- ***AVAILABLE SPACE***

  DUE TO **HIGHER SPACE OVERHEAD**

- ***TRANSACTION RATES***

  DUE TO **SLOWER ACCESS TIMES & POORER OPTIMIZATION**

- ***CONCURRENCY***

  DUE TO **MORE COMPLEX LOCK MANAGEMENT**

- ***AVAILABILITY***

  DUE TO **LONGER BACKUP AND RESTORE TIMES**

- ***PORTABILITY*** **AND** ***RE-USE***

  DUE TO **NON-STANDARD ACCESS**

A, C, D, G, B, E

APPENDIX A
SOME SCALABILITY GUIDELINES

- ***NEVER PLAN BASED ON SMALLER SCALE SYSTEMS***
  - **EXPECT SERIOUS CHANGES AT 100 GB, 600 GB, 1 TB, AND ABOVE**
  - **SMALL APPLICATION INEFFICIENCES BECOME ENORMOUS**
- ***MAKE CERTAIN YOU UNDERSTAND WHAT IS REAL***
  - **ANECDOTAL, MARKETING, AND "TECHNICAL" SPECS ARE EASILY MISUNDERSTOOD**
- ***BUILD ON TESTED CONFIGURATIONS***
- ***PLAN FOR NON-LINEARITY***
  - **EXPECT N-SQUARED TIME AND SPACE COST BEHAVIOR**
- ***PLAN FOR 2.5X THE STORAGE (5X FOR HIGH AVAILABILITY)***
  - **REMEMBER BACKUP, RECOVERY TIME ISSUES**

APPENDIX B
QUESTIONS FOR YOUR VENDOR

- ***WHAT IS THE LARGEST AMOUNT OF DATA YOU'VE***
  - **BACKED UP, RESTORED, INDEXED WITHOUT ERRORS**
  - **LARGEST INDEX BUILT WITHOUT ERRORS**
- ***WHAT IS THE COMPLEXITY WITH SIZE OF...***
  - **BACKUP AND RESTORE (EACH)**
  - **INTEGRITY CHECKING**
  - **HARD ERROR RECOVERY**
- ***IDENTIFY YOUR THREE LARGEST PRODUCTION SITES***
  - **CONFIRM REPORTED AMOUNT OF *DATA***
  - **CONFIRM REPORTED AMOUNT OF READ/WRITE ACTIVITY**
  - **CONFIRM REPORTED NUMBER OF CONCURRENT TRANSACTIONS**
  - **CONFIRM REPORTED ADMINISTRATIVE COMPLEXITY**

Disclaimer

The information and opinions presented in this report are exclusively those of Alternative Technologies, except where explicitly quoted and referenced. Although all opinions and information are reviewed for technical accuracy, the products discussed have not been subjected to formal tests and it is impossible to verify every statement made by sources. No guarantees or warrantees of correctness are made, either express or implied.. For information about this or other reports, or other products and services, including consulting and educational seminars, contact Alternative Technologies directly by telephone, mail, or via our Web site:

Alternative Technologies
13150 Highway 9, Suite 123
Boulder Creek, CA 95006
Telephone: 408/338-4621 FAX: 408/338-3113
Internet: mcgoveran@AlternativeTech.com
www.AlternativeTech.com

BIOGRAPHY

David McGoveran is a well-known relational database consultant and president of Alternative Technologies (Boulder Creek, CA), specialists in solving difficult relational applications problems since 1981. He publishes The Database Product Evaluation Report Series; authored (with Chris Date) A Guide to SYBASE and SQL Server; and is completing Advanced Client /Server: Design Concepts, Techniques, and Principles. Portions of this presentation are based on his workshops and seminars.