ENTERPRISE-WIDE SCALABILITY
THE MYTHS

The Enterprise Architecture Conference
Orlando
December 2, 1997
10:00 A.M.

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

PLEASE FILL OUT YOUR EVALUATIONS...

Thank you!

SOME HOUSEKEEPING

Presentation Changes (if any...)

- COPY AVAILABLE FROM WEBSITE

  www.AlternativeTech.com

  Login name is “DCI” and password is “EAORL97”

  Scalability Report (Executive Overview) available too!
- OR LEAVE YOUR BUSINESS CARD

Scalability Myths

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

THIS IS AN ON-GOING STUDY

Sources

- MARKET REQUIREMENTS, ANALYST REPORTS, CASE STUDIES, USER SURVEYS, 20 YEARS OF CLIENTS

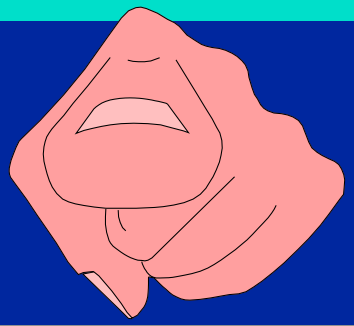
Case study sites

- MANY PLATFORMS AND ARCHITECTURES (SMP, CLUSTER, MPP)
- PUSH SOME OR ALL DBMS PRODUCT LIMITS
- ORACLE AND SYBASE CASE DETAILED STUDIES TO-DATE
- PRELIMINARY STUDIES OF INFORMIX, DB2, AND OTHERS
- ALL VENDORS WILL HOPEFULLY PARTICIPATE

Purpose

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

WE WANT YOU



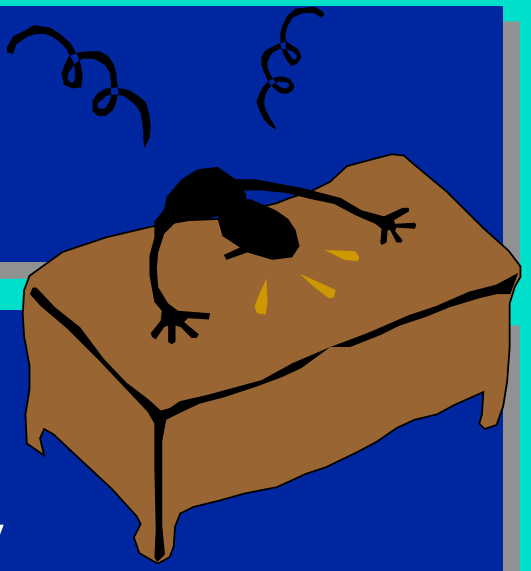
- **Interested in being a participant in the DBMS Scalability Project?**
  - **LARGE DATABASE?**
  - **HIGH TRANSACTION RATES?**
  - **LARGE USER POPULATION?**
  - **DOES YOUR APPLICATION HAVE *ANY* OF THE ABOVE?**
  - **WANT A FREE 3-DAY AUDIT WITH RECOMMENDATIONS?**
- **Contact Alternative Technologies for more information**
  - **LEAVE YOUR BUSINESS CARD, JUST WRITE "PARTICIPANT" ON THE BACK**
  - **SEND E-MAIL TO: mcgoveran@AlternativeTech.com**
  - **TELEPHONE 408/338-4621**

MARKET REQUIREMENTS

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 BUSINESS TRANSACTIONS**

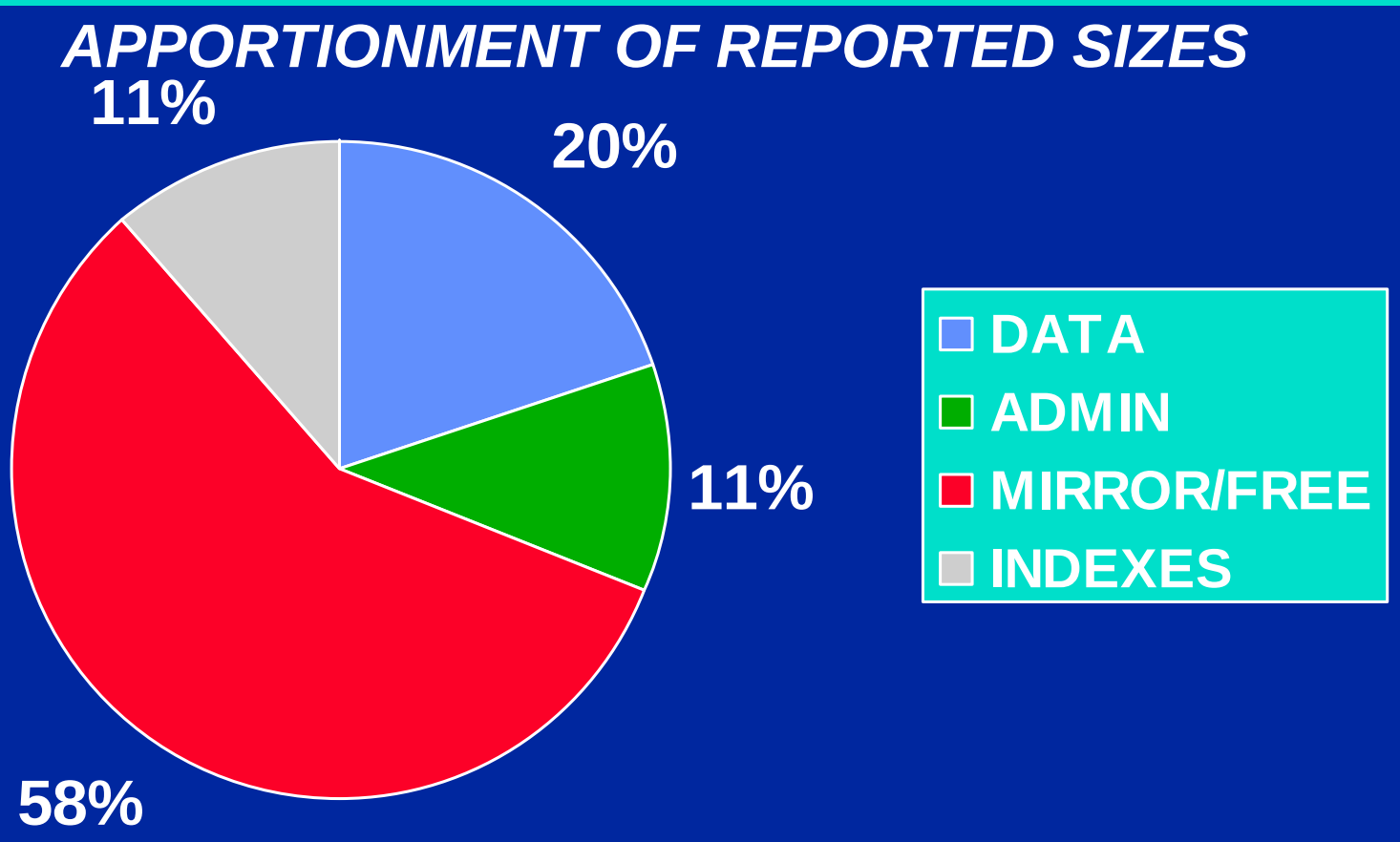
GOTCHA!



- 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

THE TOP TWENTY MYTHS

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



FOOTNOTE
Raw Data to Total Disk Ratio

- **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 Giants*, DBPD**

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 (10/15/97)
- **Sybase: 511 GB total disk space**
  - APPROXIMATELY 300 GB OF RAW DATA
  - 1+ TB VERIFIED, 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**

2. DBMSs CAN BE PRODUCED AND CONSUMED AS COMMODITIES

Success factors increasingly obscure

- **DIFFICULT TO IDENTIFY**
- **VERY COMPLEX**
- **EASE OF USE HIDES COMPLEXITY**

Implementations differ in important ways

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

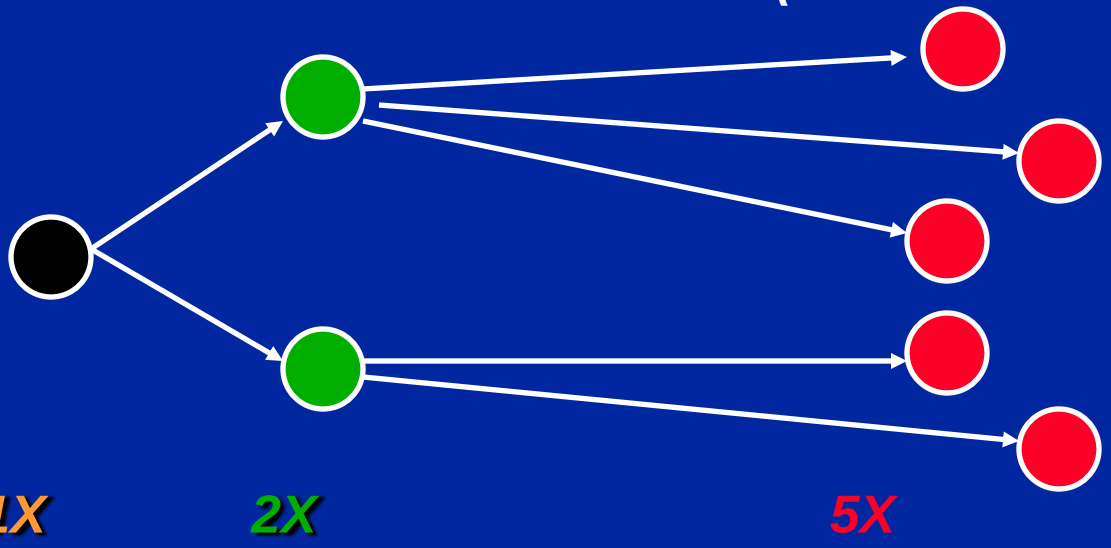
Customers use DBMS specific "workarounds"

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

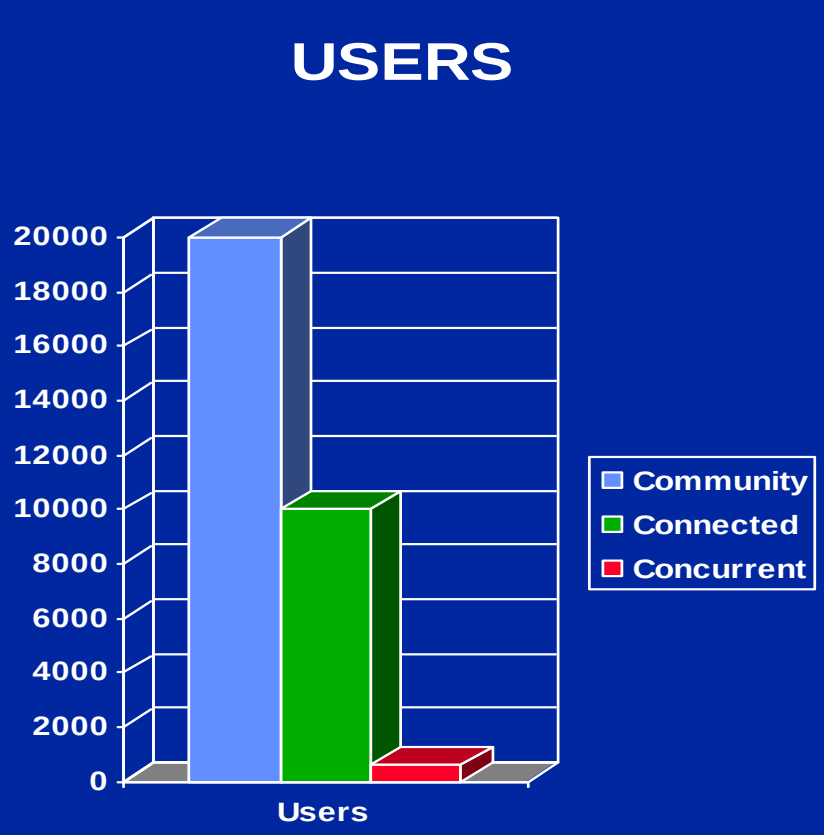
3. PHYSICAL TRANSACTION RATES MEASURE WORKLOADS FOR SCALABILITY

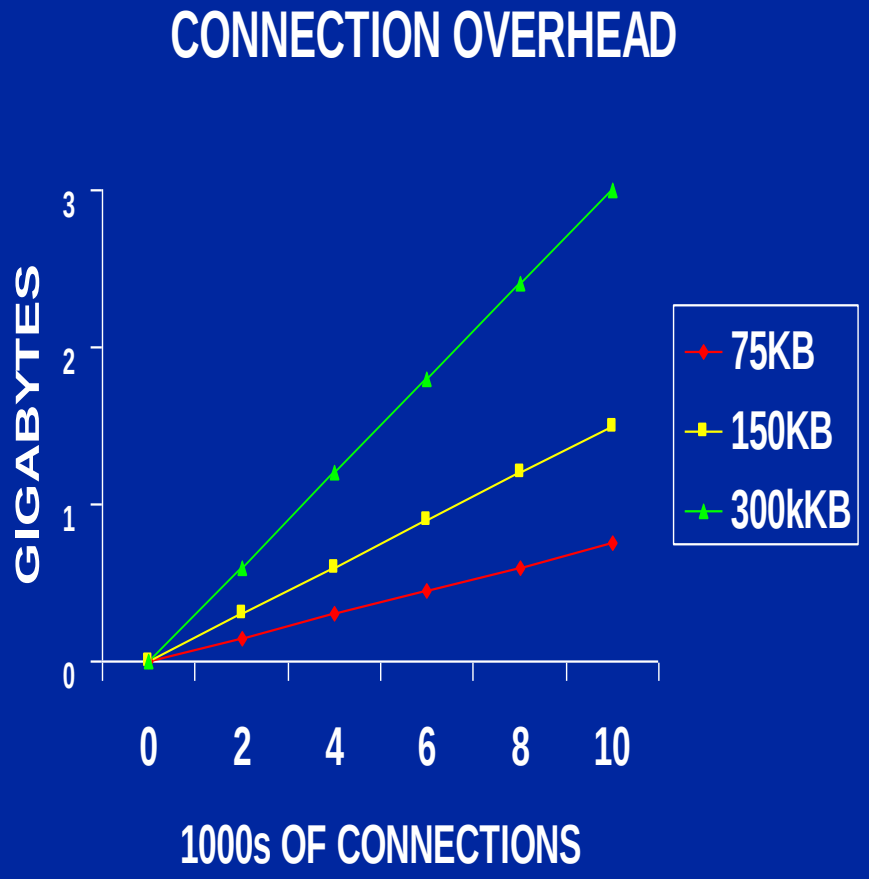
ONLY *BUSINESS* TRANSACTIONS (AUDIT) ARE IMPLEMENTATION INDEPENDENT

- ***VERSUS LOGICAL TRANSACTIONS (CONSISTENCY)***
- ***VERSUS PHYSICAL TRANSACTIONS (RECOVERABLE)***



4. NUMBERS OF USERS SUPPORTED IS A MEASURE OF SCALABILITY

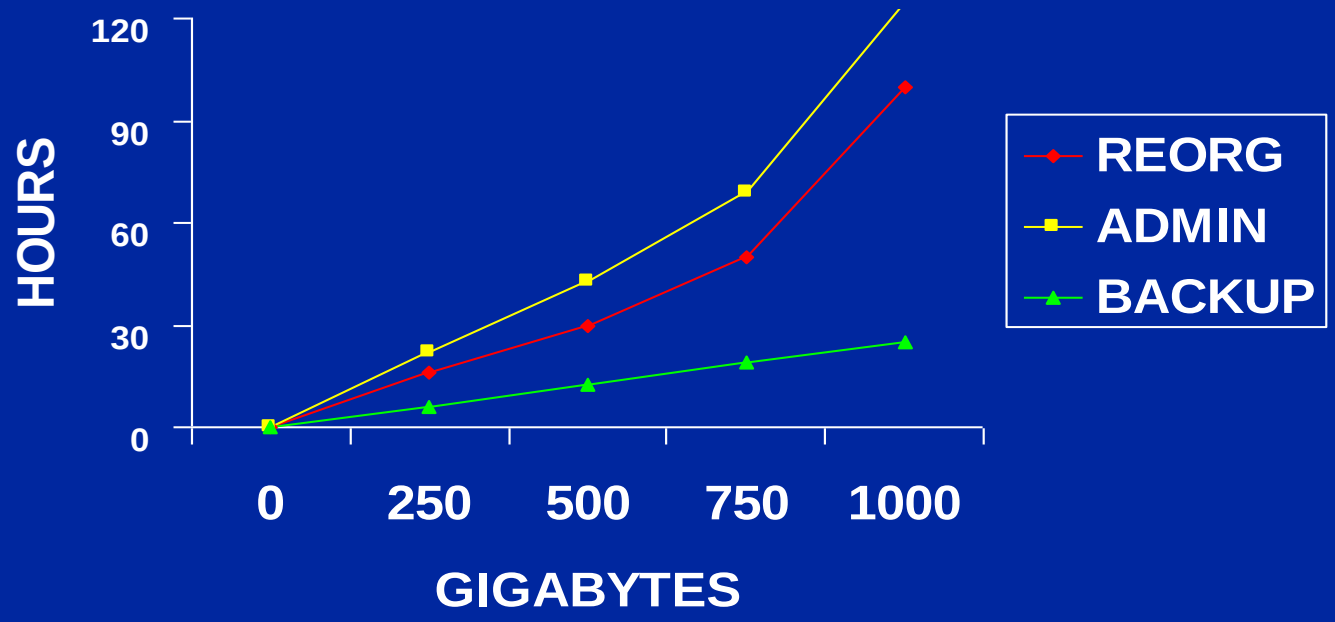




5. SPEED OF ADMINISTRATIVE OPERATIONS DETERMINES ADMINISTRATIVE SCALABILITY

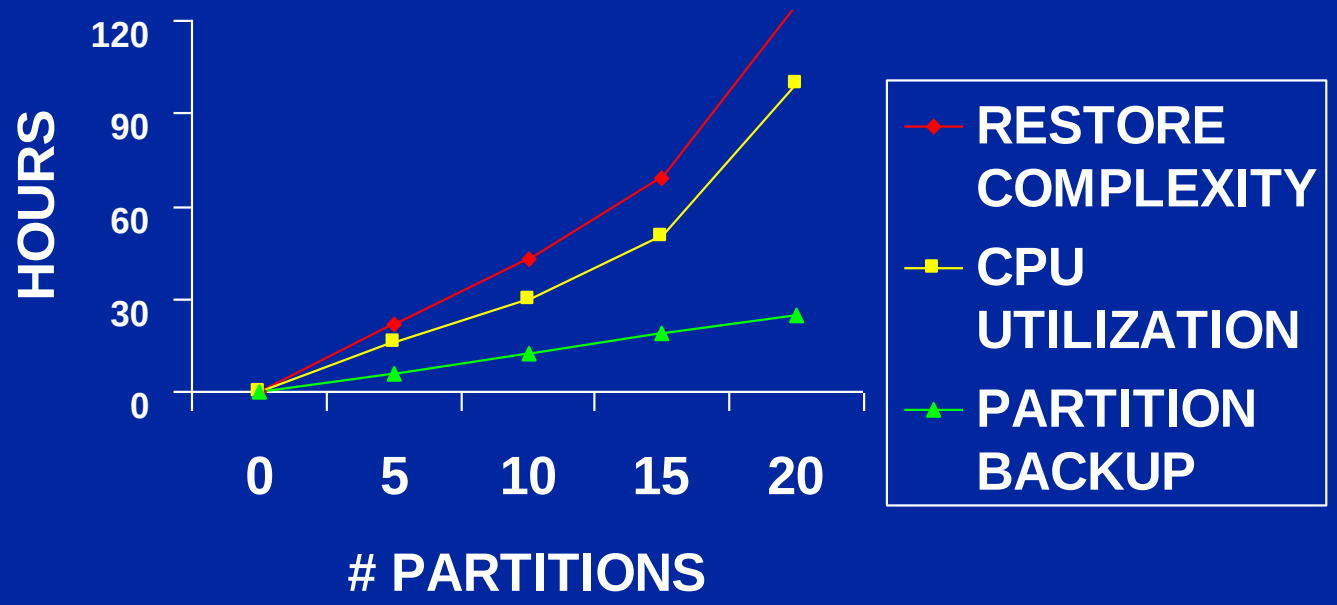
NON-LINEAR FUNCTIONS OF SIZE

TIME COST VS. DB SIZE



6. PARTIAL DATABASE OPERATIONS PROVIDE ADMINISTRATIVE SCALABILITY

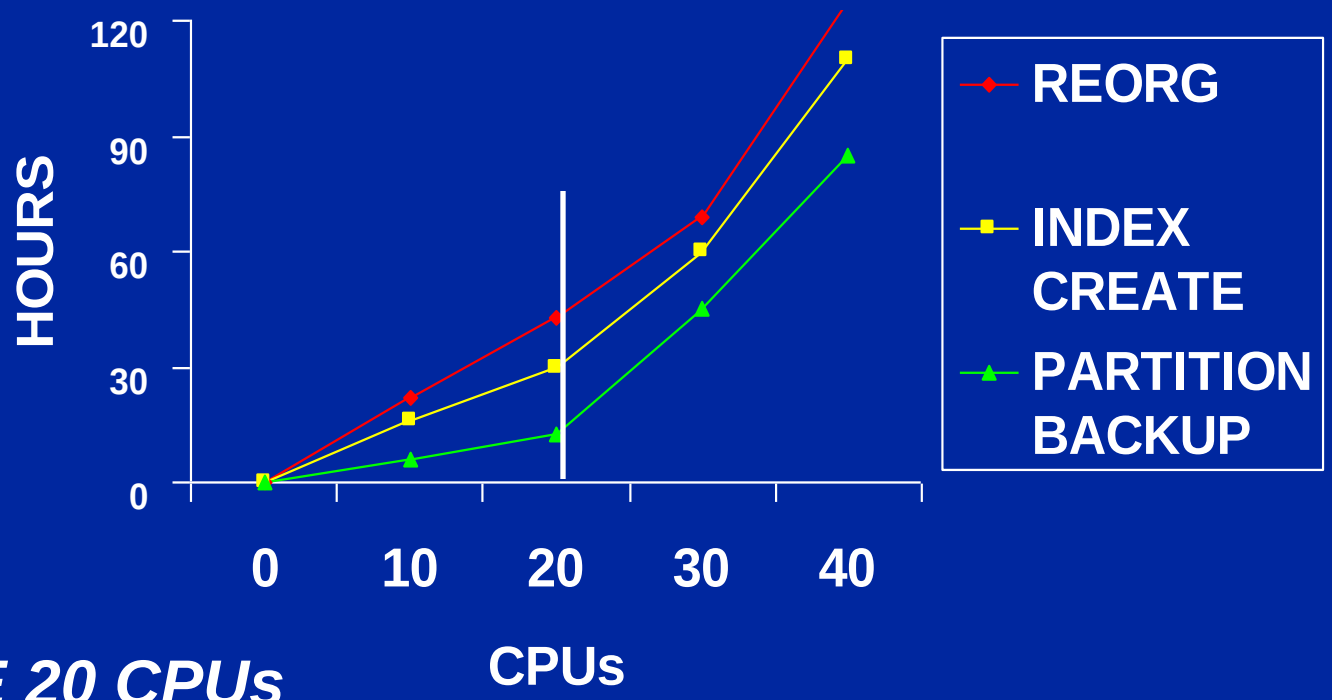
TIME COMPLEXITY VS. PARTITIONS



ASSUME 50 GB PARTITIONS

7. PARALLELISM IS NECESSARY FOR SCALABILITY

CPUs LIMIT LINEARITY



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 = HIGHER 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, ...

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***

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**

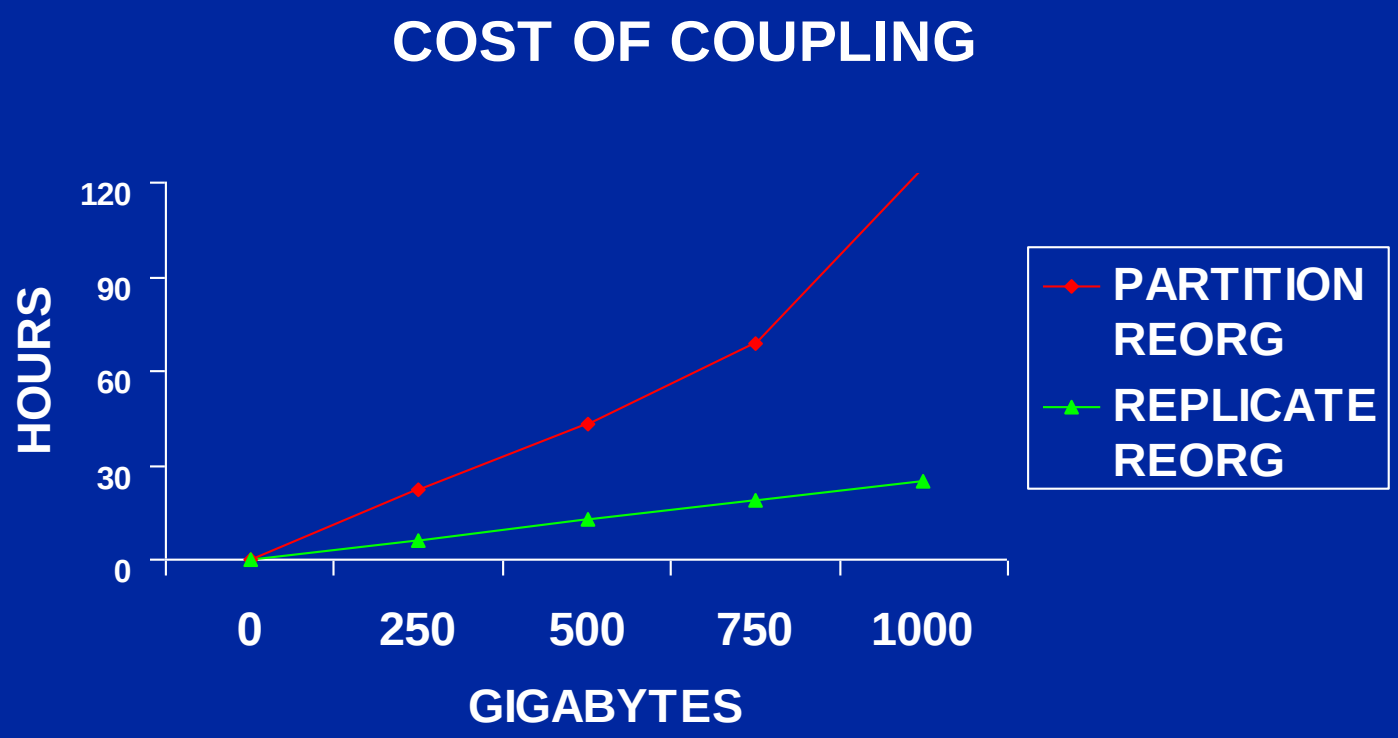
11. DATABASE PARTITIONING CIRCUMVENTS PRODUCT SCALABILITY LIMITATIONS

Most large databases are partitioned

Non-scalability reasons:

- **PLATFORM LIMITATIONS PRECLUDE DBMS SCALABILITY (E.G., 2 GB FILE LIMITATIONS).**
- **ADDITIONAL PARTITIONED SYSTEMS ADDED ONLINE**
- **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**

12. REPLICATES ARE MORE DIFFICULT TO REORGANIZE THAN TABLE PARTITIONS



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.)**

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

15. DBMS CLUSTERING IS AN IMPORTANT SCALABILITY SOLUTION

Clustering primarily provides, and is used for, high availability

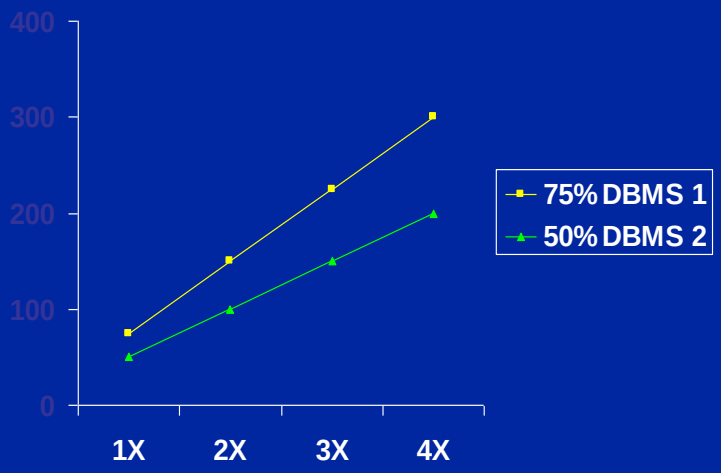
Designers must exercise great care to obtain even moderate scaleup or speedup from cross-node cluster resources

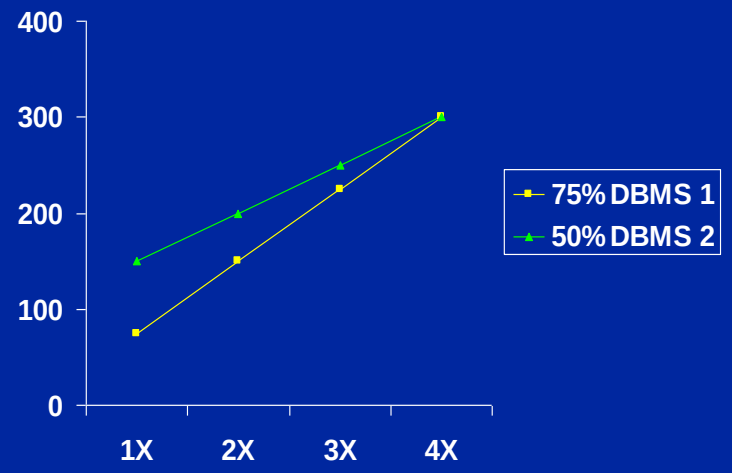
Designed more like a federation of loosely coupled physical databases

Costs include design time, additional administration, possibly coding, and lock or cache coherence management

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

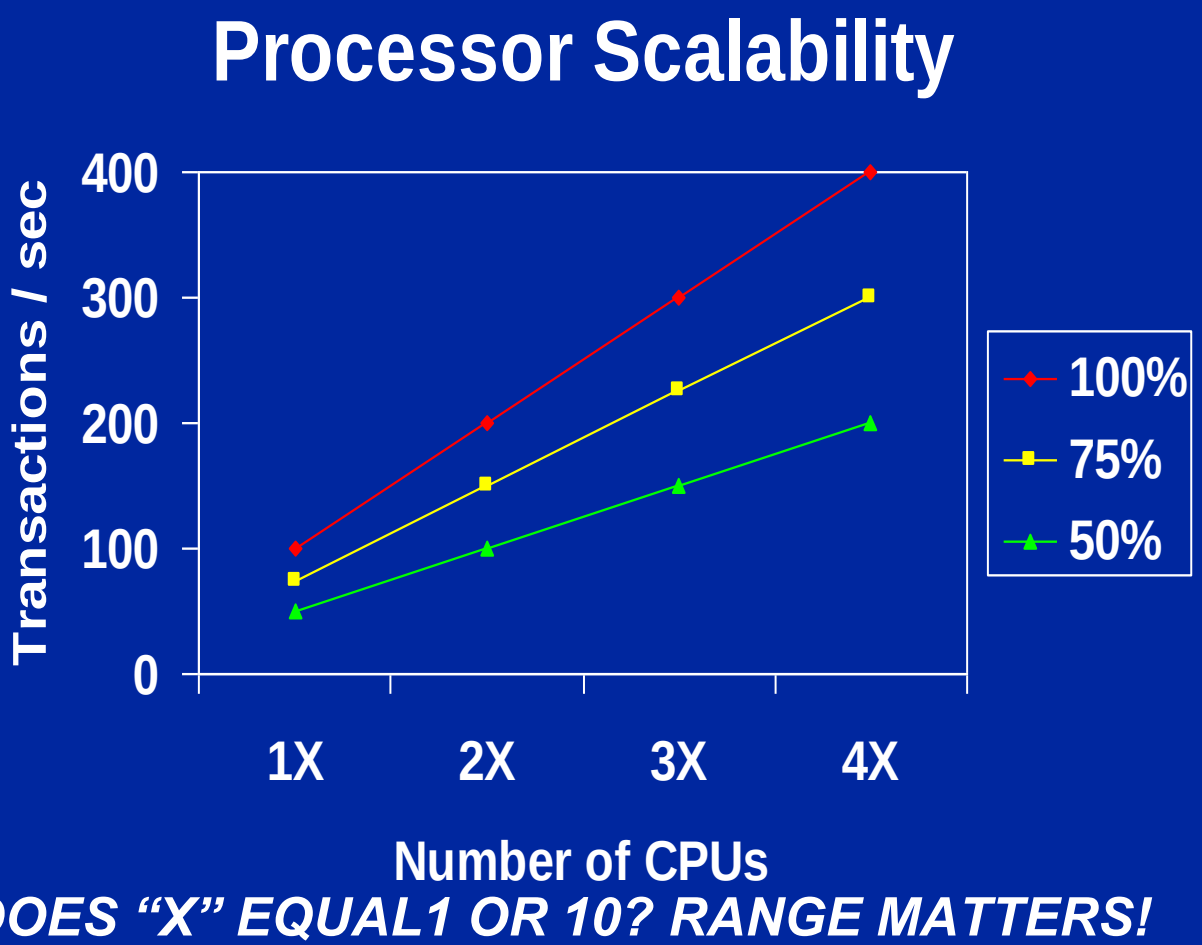
WHAT DOES PERCENT SCALABILITY MEAN?





UNLABELED GRAPH **X = 1**

17. A DBMS EITHER IS OR IS NOT SCALABLE

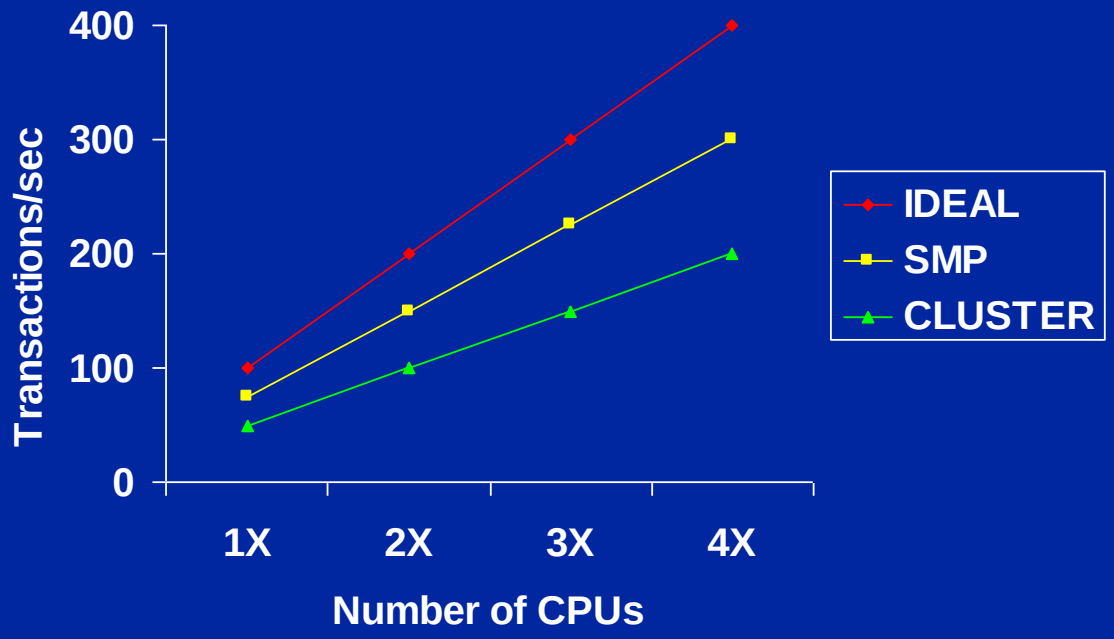


DOES "X" EQUAL1 OR 10? RANGE MATTERS!

18. SCALEUP OR SPEEDUP CAN BE PROVEN BY EXAMPLE

DBMS SCALEUP AND SPEEDUP ARE:

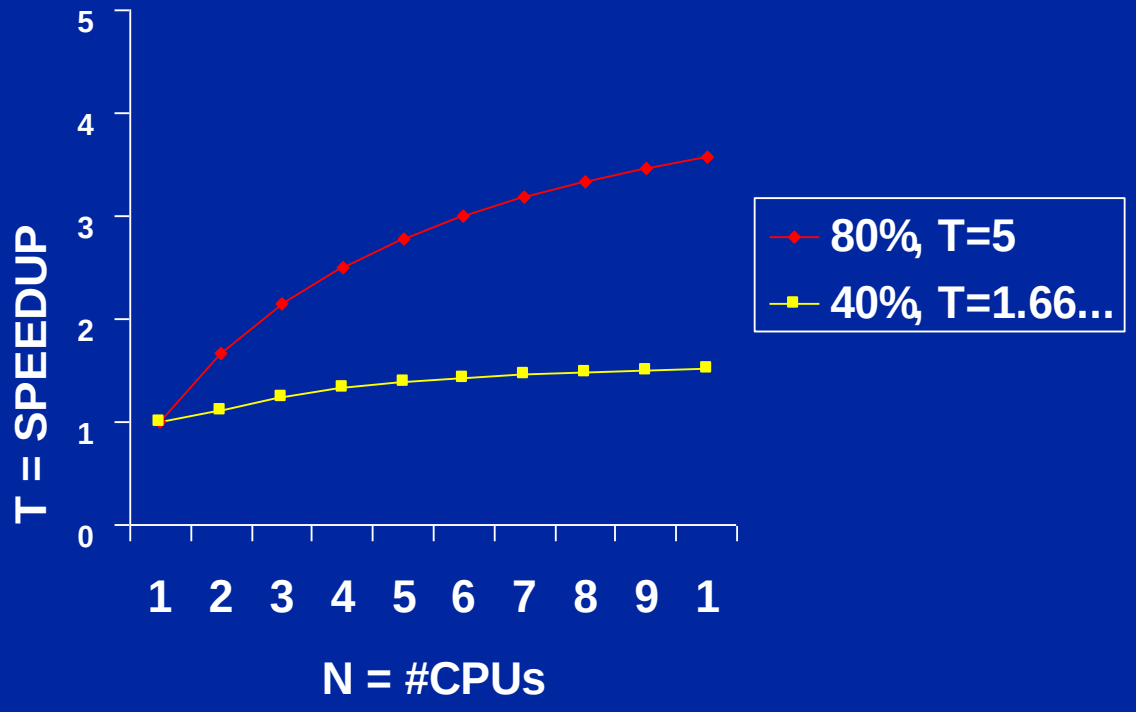
- PLATFORM AND APPLICATION SPECIFIC
- STRONGLY AFFECTED BY TRANSACTION AND DB DESIGN



Transaction rate versus CPUs

19. GOOD PROCESSOR SCALABILITY CAN PROVIDE ARBITRARY SPEEDUP

PROCESSOR SPEEDUP FOLLOWS AMDAHL'S LAW:
T = 1 / ((1 – M) + (M / N))



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**

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 AT SMALL SCALE 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 5X THE STORAGE
  - BACKUP, RECOVERY TIME

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.

PLEASE FILL OUT YOUR EVALUATIONS...

Thank you!